Acoustic-based Assistive Technology Tools for Dysarthria Management

Tolulope Bamidele Ijitona

Department of Electronic and Electrical Engineering

University of Strathclyde

Glasgow, United Kingdom

This dissertation is submitted for the degree of

Doctor of Philosophy

2019

To Oluwafemi.

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Tolulope Ijitona

2019

Acknowledgements

The research in this thesis has been carried out at the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom.

I would like to appreciate my supervisors Doctor Hong Yue, Professor Anja Lowit and Professor John Soraghan for their excellent and invaluable support, encouragement, guidance and contribution towards the development and successful completion of the research work presented in this thesis. Under their exceptional supervision, I learnt about all I need to be a well-rounded researcher including being analytical, creative, resourceful, as well as skills for idea generation and technical presentation.

I would like to thank my colleagues at the Centre for Signal and Image Processing and the Speech and Language Therapy Research groups for their peer reviews, support and collaboration. I also appreciate my friends at the Nigerian Society at Strathclyde for their support and encouragement during my research. Moreover, I would like to appreciate all my family and friends who stood by me, cheered me up, prayed for me and encouraged me throughout these years.

This research would have been impossible without the research grants from the University of Strathclyde, Klik2Learn and the Carnegie Trust Fund.

Finally, I thank God for His grace, sustenance and protection throughout my research and write-up.

Abstract

The research presented in this thesis addresses the concepts of application of important digital signal processing algorithms in the detection and treatment of dysarthria, a neurological motor speech disorder. The novel algorithms presented in this thesis include a silence, unvoiced and voiced segmentation technique for dysarthric speech based on linear prediction error variance (LPEV), an automatic diadochokinetic (DDK) analysis and segmentation scheme for dysarthric speech, the application of speech processing algorithms in the extraction of prosodic, voice quality, pronunciation and wavelet features for the detection and severity classification in dysarthric speech and the modification of dysarthric speech features using speech enhancement techniques to improve the intelligibility of dysarthric speech in a stress production exercise for the treatment of dysarthria.

In particular, an improved silence, unvoiced and voiced segmentation technique for dysarthric speech is proposed. This method is an enhanced technique that makes use of a two-layer segmentation approach which combines the short-time-energy (STE) and LPEV to distinctly differentiate between the silence and voiced segments despite the reduced/inconsistent intensity, pauses, voice breaks and slow speech rate experienced in dysarthric speech. Including the LPEV into the segmentation process has proved to be advantageous in eliminating segmentation errors due to the similarity observed between the STE profiles of the silence and voiced segments in dysarthric speech. The experimental results have shown that this segmentation method is also effective and efficient in reducing the effects of artefacts introduced in dysarthric speech.

A novel automatic DDK analysis scheme is proposed in this research to extract individual DDK syllables and analyse them for consistency. This method is based on a speaker-specific moving average threshold (rather than a fixed threshold) which addresses the varying intensities in the DDK sounds produced by speakers with dysarthria. This method also addresses the challenge of intra-syllable breaks introduced in dysarthric DDK syllables using a minimum distance merging approach. In addition, the algorithm analyses the segmented

DDK syllables by calculating the individual DDK rates and their variance in order to measure the DDK syllable production consistency. The high accuracy of the proposed method is tested and verified using both dysarthric and healthy controlled databases.

Three novel schemes for automatic detection and severity classification of the dysarthric speech are also proposed in this research. One extracts an extended speech feature called centroid formant (which is a representation of energy concentration in the frequency spectrum) and classifies these centroid formants using neural network classifiers for the detection of dysarthria. The centroid formant-based detection scheme also forms the backbone for the development of the second and more robust detection scheme which combines centroid formants with prosodic, voice quality, pronunciation and wavelet features for more efficient classification. A third scheme is developed specifically for the classification of dysarthria into three severity levels using the same features as in the second scheme. The efficiencies of these detection and severity classification schemes are evaluated by calculating the accuracy, sensitivity and specificity of the classifiers.

The effects of modification of prosodic cues used in stress production on the ability of listeners to correctly identify the position of the stressed word in sentences are also investigated in this research. This investigation is focused on the three prosodic cues used by healthy controlled speakers in stress production; namely intensity, duration and fundamental frequency. These three features are modified acoustically and presented to untrained listeners in an aim to evaluate the effects of the individual and combined modifications on the listeners' perception. The findings of this investigation will help clinicians, including speech and language therapists, make an informed decision on the prosodic feature to focus on during stress production exercises for the management of dysarthria.

Finally, the dysarthria management schemes proposed in this research are developed into user-interactive tools in MATLAB from which speaker-specific information and reports can be generated and downloaded for progress monitoring and further analytical purposes.

Tolulope Ijitona
University of Strathclyde, 2019

Table of Contents

List of Figures

List of Tables

Abbreviations

AAF	Altered Auditory Feedback
AC	Alternating Current
ACF	Autocorrelation Function
AD	Ataxic Dysarthria
ALS	Amyotrophic Lateral Sclerosis
AMDF	Average Magnitude Difference Function
AMP	Utterance Requiring Amplification
AMR	Alternating Motion Rate
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
CATRIS	Computerized Assessment and Treatment of Rate, Intonation and Stress
CFDA	Computerised Frenchay's Dysarthria Assessment
CL	Clipping Level
CNS	Central Nervous System
CPAP	Continuous Positive Airway Pressure
CSD	Cepstral Separation Difference
CSRs	Continuous Speech Recognition Systems
DAF	Delayed Auditory Feedback
dB	Decibel
DC	Direct Current
DDK	Diadochokinetic
DEB	Dysarthria Examination Battery
DME	Direct Magnitude Estimation
DNN	Deep Neural Networks
DTP	Dysarthria Treatment Programme
DWT	Discrete Wavelet Transform
F0	Fundamental Frequency
FDA	Frenchay's Dysarthria Assessment
FSF	Frequency Shifted Feedback
FT	Fourier Transform

GFCC	Gammatone Frequency Cepstral Coefficients
GOF	Goodness of Fit
HC	Healthy Control
HNR	Harmonic to Noise Ratio
IFT	Inverse Fourier Transform
IPC	Utterances with Inappropriate Pitch Contours
KNN	k-Nearest Neighbours
LDA	Linear Discriminant Analysis
LPC	Linear Prediction Coding
LPEV	Linear Prediction Estimation Error Variance
LSVT	Lee Silverman Voice Treatment
MFB	Mel-Frequency Banks
MFCC	Mel Frequency Cepstral Coefficients
MPT	Maximum Performance Task
NCCF	Normalised Cross-Correlation Function
PD	Parkinson's Disease
PVOC	Phase Vocoder
PwD	People with Dysarthria
RAPT	Robust Algorithm for Pitch Tracking
RDP	Robertson Dysarthria Profile
RMS	Root Means Square
SLT	Speech and Language Therapists
SNR	Speech to Noise Ratio
STE	Short Time Energy
STFT	Short-Time Fourier Transform
SUV	Silence-Unvoiced-Voiced
SVM	Support Vector Machines
SVOA	Subject-Verb-Object-Adverbial
UBM	Universal Background Model
VAD	Voiced Activity Detection
VU	Voiced-Unvoiced
ZCR	Zero-Crossing Rate

Chapter 1

1 Introduction

In recent years, the management of speech disorders using acoustic and instrumental methods has gained increasing research interests due to the need to improve how these speech disorders are managed by offering less subjective approaches that are based on advanced speech processing principles [1]. Dysarthria is a neurological motor speech disorder that affects the production of sounds due to the weakness of the muscles and nerves involved [2]. This includes impairment in the movement of the lips, larynx, vocal cords, tongue and/or nasal air passage [3]. The effects of dysarthria are seen in the speed, variation (in loudness, pitch and duration), consistency, and rhythm/movement accuracy in speech production [2, 4, 5].

Generally, dysarthria often results from damage to either or both the upper or/and lower motor neurons [2]. In some cases, dysarthria can be accompanied by apraxia; a state where information from the brain to the mouth is disrupted resulting in the production of wrong sounds and movements [3]. Dysarthria can be also be accompanied by aphasia; language disorder due to neurological damage [4]. The causes of these various speech and language disorders (dysarthria, apraxia and aphasia) are often different [2].

More specifically, dysarthria can be caused by various neurological conditions [6]. In order to give a structured way of describing the different types of its occurrence, dysarthria has been divided into six main categories namely; Spastic Dysarthria, Ataxic Dysarthria, Hypokinetic Dysarthria, Hyperkinetic Dysarthria, Flaccid Dysarthria and Mixed Dysarthria [7]. Dysarthria subtypes are classified by examining the five primary speech subsystems namely; phonation, resonance, prosody, respiration and articulation [4]. These dysarthria subtypes will be discussed in Section 2.2 of this thesis.

Dysarthria often results in a decrease in speech intelligibility (when compared with the speech from healthy controlled speakers) [8]. Common causes of dysarthria include stroke [9], Amyotrophic Lateral Sclerosis (ALS) [10], Parkinson's disease [11-14], multiple sclerosis [15], degenerative diseases [16], brain injury [17], tumours [2], etc. The inability to produce speech with high intelligibility is often triggered by various control and articulatory factors [18]. The articulatory factors consist of the speech production organs – which are also called the articulators- such as tongue, lips, larynx, nasal cavity and jaws, whereas, the neuro-muscular mechanisms control the movement of these articulators [19]. These disorders affect the control of speech articulators but do not necessarily impair language production and comprehension [20].

1.1 Motivation

Dysarthria is one of the most common communication disorders representing over 40% of neurological disorders referred to speech pathologists at Mayo clinic yearly [2]. In the UK, dysarthria is also one of the commonly referred disorders to speech therapists [21]. Unfortunately, the research attention given to the assessment, management and treatment of dysarthria does not follow this trend.

The current dysarthria assessment methods used by therapists involves both the physical examination of the speech production system (lips, tongues, larynx and nasal cavity) and speech assessment [1]. The scoring of the speech assessment is largely based on the perception of the clinicians rather than objective quantitative measures of the relevant speech features [22]. There is, therefore, a need to acoustically measure these features in order to characterise the speech impairment and classify the severity objectively. The acoustic measurement will give quantifiable scores with respect to the various aspects of the speech assessment. The scores can then be used to monitor the progress of the speakers before, during and after therapy sessions. This classification will also aid the development of a progressive treatment tool that can be used on computers and hand-held devices which can be monitored remotely by clinicians, including Speech and Language Therapists (SLTs).

Dysarthria can also have long-term psycho-social impacts on the patients by affecting their ability to communicate intelligibly thereby impacting their interpersonal relationships, family and career [23]. Early and accurate detection, as well as management of this disorder, is therefore of uttermost importance. This can be achieved by automating the detection process, independent of human subjectivity and perception.

Another research gap identified is the focus of existing dysarthria classification techniques. Most related recent research works [14, 24, 25] are based on the assessment of Parkinson's disease; which is a progressive speech disorder with little attention paid on other non-progressive speech disorders [21]. There is, therefore, a need to research advanced speech processing techniques that can be used in the management and treatment of the various types of dysarthria.

1.2 Research Aim & Objectives

One of the main objectives of this study is to explore the application of digital signal processing (DSP) principles in the management (assessment and treatment) of dysarthria by developing automatic techniques for the extraction of speech features from dysarthric speech signals. These acoustic-based techniques will offer objective methods that are independent of human perception and expertise which can be prone to error due to variability in individual perception [2]. This research will also focus on the application of machine learning techniques in the detection and classification of dysarthric speech.

Another objective of this research is to develop an algorithm for the measure of the ability of dysarthric speakers to produce short syllables in a fast-repetitive manner (also called, diadochokinetic skill) since dysarthria is characterised by reduced speech rate. This will help the clinicians in identifying and assessing the deterioration in dysarthric speakers' control and coordination over time.

The treatment of stress production deficiencies in dysarthria requires a detailed understanding of the speech features associated with stress production and their impact on speakers' intelligibility. One of the objectives of this study will be to examine the effects of the modification of individual and multiple speech features on dysarthric speakers' intelligibility during stress production activity. This will

enable clinicians to make informed decisions during the management of stress production deficits in dysarthria.

In summary, the aim of this research is to develop assisted technology tools for the management - assessment, treatment and monitoring - of dysarthria in order to achieve the following objectives:

1. To apply DSP principles in the extraction of acoustic features from dysarthria speech
2. To reduce the analysis errors by developing an improved silence-unvoiced-voiced segmentation technique for dysarthric speech
3. To develop novel algorithms for the automatic detection and severity classification of dysarthric speech
4. To assess and analyse dysarthric speakers' diadochokinetic (DDK) ability
5. To develop a tool to assist clinicians in making informed-decision in the management of stress production deficits in dysarthria.

1.3 Contributions

There are six main novel contributions in this research which includes:

A. Development of an automatic algorithm for the segmentation of dysarthric speech into silence, unvoiced and voiced parts using short-time energy (STE), zero-crossing rate (ZCR) and linear prediction error variance (LPEV) in a two-layer segmentation technique (Chapter 4).
B. Novel automatic DDK analysis technique for the assessment of dysarthric speech using speaker-specific moving average syllable segmentation threshold and minimum distance merging (Chapter 5).
C. Automatic detection of dysarthric speech using extended speech features (prosodic, pronunciation, voice quality and wavelets features) and machine learning classification techniques (Chapter 6, Section 6.3 and 6.4).
D. Novel automatic classification of dysarthric speech into three severity levels (mild, moderate and severe) using extended speech features and machine learning classification techniques (Chapter 6, Section 6.5).
E. Analysis of stress marking deficits and effects of prosodic features (intensity, fundamental frequency and duration) modifications on the

speakers' intelligibility for the clinical management of dysarthria (Chapter 7).

F. Development of assistive technology human-machine interfaces for the management of dysarthria using DSP and machine learning techniques (Chapter 8).

1.4 Publications

The academic outputs of this research include:

- Ijitona, T.B., Soraghan, J.J., Lowit, A., Di-Caterina, G. and Yue, H., December, 2017. Effects of acoustic features modifications on the perception of dysarthric speech—Preliminary study (Pitch, intensity and duration modifications). In *3rd IET International Conference on Intelligent Signal Processing (ISP 2017)* (pp. 1-6). (Conference paper contribution - Chapter 7 of this thesis)
- Ijitona, T.B., Soraghan, J.J., Lowit, A., Di-Caterina, G. and Yue, H., December, 2017. Automatic detection of speech disorder in dysarthria using extended speech feature extraction and neural networks classification. In *3rd IET International Conference on Intelligent Signal Processing (ISP 2017)* (pp. 1-6). (Conference paper contribution - Chapter 6, Section 6.3 of this thesis)
- Lowit, A., Ijitona, T., Kuschmann, A., Corson, S. and Soraghan, J., May, 2018. What does it take to stress a word? Digital manipulation of stress markers in ataxic dysarthria. *International Journal of Language & Communication Disorders, 53*(4), pp.875-887. (Journal paper contribution - Chapter 7 of this thesis)
- Lowit, A., Ijitona, T., & Soraghan, J. (2017). The effects of F0, intensity and durational manipulations of the perception of stress in dysarthric speech. 106, July 2017. Presented at *7th International Conference on Speech Motor Control, Groningen, Netherlands*. (Conference poster contribution - Chapter 7, Sections 7.5 and 7.6 of this thesis)
- Ijitona, T., Lowit, A. and Soraghan, J., September, 2017. The effects of acoustic modification of fundamental frequency, intensity and duration on

the perception of stress in dysarthria speech. In the *Royal College of Speech and Language Therapists Conference 2017*. (Conference poster contribution - Chapter 7, Sections 7.5 and 7.6 of this thesis)

- Ijitona, T.B., Soraghan, J.J., Lowit, A., and Yue, H., October, 2020. Automatic diadochokinetic analysis of dysarthric speech using speaker-specific thresholding and minimum-duration merging. *The 21st Conference of the International Speech Communication Association (INTERSPEECH 2020).* Submitted (Conference paper contribution – Chapter 5)
- Ijitona, T.B., Yue, H., Soraghan, J., and Lowit, A., February, 2020. Improved Silence-Unvoiced-Voiced (SUV) Segmentation of Dysarthric Speech Signals using Linear Prediction Error Variance. *The 5th International Conference on Computer and Communication Systems (ICCCS 2020).* Accepted (Conference paper contribution – Chapter 4)
- Ijitona, T.B., Yue, H., Soraghan, J.J., and Lowit, A., August, 2020, Automatic detection and severity classification of dysarthria using prosodic, pronunciation, wavelets and voice quality features. *The 28th European Signal Processing Conference (EUSIPCO 2020).* Submitted (Conference paper contribution – Chapter 6, Sections 6.4 and 6.5)

1.5 Outline of the Thesis

This thesis consists of nine (9) chapters. The first chapter is an introductory chapter which includes the motivation for this study, research aims & objectives, novel contributions and academic outputs (publications) of this research. The next two chapters (Chapters 2 and 3) consists of the review of past and current literature related to this research. Chapter 2 covers the review of techniques used in the detection and treatment of dysarthria including perceptual and acoustic detection techniques, as well as, treatment strategies proposed by researchers over the years. Chapter 3 consists of the review of feature extraction techniques for the analysis of dysarthric speech, as well as, machine learning classification techniques for dysarthria management. The novel contributions of this research are presented in Chapters 4 to 8. The first novel contribution, which is the automatic Silence-Unvoiced-Voiced segmentation of Dysarthric speech using Short-Time Energy, Zero Crossing Rate and Linear Prediction Error Variance, is presented in Chapter

4. In Chapter 5, a novel technique for the analysis of diadochokinetic syllables in dysarthria is presented. Automatic detection of dysarthric speech and severity classification techniques are presented in Chapter 6 while the analysis of stress production deficits in dysarthria speech is presented in Chapter 7, to assist clinicians in making informed decisions when using stress production tasks in the management of dysarthria. Five dysarthria management tools, developed in MATLAB, are presented in Chapter 8 which includes Speech Examination Tool (SETool), Automatic DDK Analysis Tool (DDKTool), Automatic Dysarthria Detection Tool (DyDECTOOL), Dysarthria Severity Classification Tool (DySECTOOL), and Stress Marking Task Tool (SMAT). The sixth tool called the Dysarthria Assessment and Treatment Tool (DySATTOOL) gives users access to the other five tools. The last chapter of this thesis, Chapter 9, gives a summary of the contributions in this research work as well as the future works relevant to this study.

Chapter 2

2 Review of the Techniques used in Dysarthria Detection and Treatment

2.1 Introduction

In this chapter, a review of how technologies are used in the detection and treatment of dysarthria is discussed. In order to understand the roles played by these technologies in its management, dysarthria will be described while highlighting its causes, characteristics, symptoms, and effects on the lifestyle of people living with this disorder. Current clinical techniques used in the management of dysarthria will be an area of focus, providing a review of how the clinicians diagnose, assess and treat dysarthria, while the current gaps in the use of technologies will be identified. In addition, different pathological speech management techniques proposed over the years will be reviewed, while identifying the limitations posed by these techniques based on the availability of resources and underlying clinical factors such as user-ability, reliability and effectiveness. Various measures used in classifying dysarthria into different severity levels will also be discussed with a focus on how relevant the measures are and their degree of subjectivity. More specifically, the latest advances in assistive technologies and their impacts on dysarthria management will be discussed in this chapter.

2.2 Causes, Characteristics and Symptoms of Dysarthria

Occurrences of dysarthria are generally categorised into six subtypes depending on their prosodic, articulatory, resonance, respiration and phonation characteristics [10]. Four of the six types of dysarthria (spastic, hyperkinetic, hypokinetic and ataxic dysarthria) result from damages to the upper motor neurons [26]. Pyramidal tract damage results in spastic dysarthria [2]; whereas, damages to extrapyramidal tract result in hypokinetic or hyperkinetic dysarthria [4]. Ataxic dysarthria, on the other hand, is caused by lesions in the cerebellum [27]. The type of dysarthria resulting from lower motor neurons damage is called the flaccid dysarthria [10].

This includes damages to the cranial nerves - with motor components. There are some occurrences where both the upper and lower neurons are damaged [4, 10]. These occurrences are called mixed dysarthria [3]. An example of mixed dysarthria is Amyotrophic Lateral Sclerosis which consists of the combination of Flaccid Dysarthria and Spastic Dysarthria [15]. The neuromuscular characteristics of various types of dysarthria are presented in Table 2-1.

Table 2-1. Neuromuscular Characteristics of Various Types of Dysarthria

Characteristics	**Spastic Dysarthria**	**Hypokinetic Dysarthria**	**Hyperkinetic Dysarthria**	**Ataxic Dysarthria**	**Flaccid Dysarthria**
Rhythm (individual/repetition)	Regular	Regular	Irregular	-/Irregular	Regular/Normal or Slow
Rate (individual /repetition)	Slow/slow	Slow/Fast	Slow/Slow	Slow/Slow	Reduced/Normal or Slow
Range (individual/repetition)	Reduced/ Reduced	Reduced/ Very Reduced	Reduced to Excessive	Excessive to Normal	-/Reduced
Tone	Reduced	Excessive	Excessive	Reduced	Reduced
Direction	Normal	Normal	Inaccurate	Inaccurate	Normal
Force	Excessive	Reduced	Reduced to Excessive	Normal to Excessive	Weak

The prognosis and characteristics of these various subtypes of dysarthria differ in terms of their effects on the primary speech subsystems namely; resonance, phonation, respiration, prosody and articulation [2]. The ability to articulate consonants is affected by all categories of dysarthria which often results in a slurred speech [28]. Difficulty in the articulation of vowels may also occur in cases with high severity [2]. The level of intelligibility also varies across the different types of dysarthria [4]. These variations in the primary speech subsystems and etiologies are discussed in Section 2.2.1 through to Section 2.2.6.

2.2.1 Spastic Dysarthria

Spastic dysarthria results in difficulty in coordination and increase in the muscle tone. This type of dysarthria usually affects the larynx, lips, throat, cheeks and/or

velum [19]. According to the data from Mayo Clinic Speech Pathology, spastic dysarthria accounts for 7.3% of all dysarthria [2]. Research has also shown that stroke in the brainstem is one of the major causes of this type of dysarthria [5]. In spastic dysarthria, the pitch is often low [4] and in some cases, breaks in pitch occur [5]. Consonant articulations are often imprecise and the speech prosody shows an abnormal pattern (stress, intonation and duration) [5]. There is typically a mild occurrence of hypernasality without nasal emission [3, 4]. Spastic dysarthria is also characterised by distorted vowels, slow speech rate, mono-loudness, reduced stress, mono-pitch and inability to produce long phrases [2]. These features will be useful in characterising this dysarthria subtype for research purposes. The etiological distribution of spastic dysarthria is shown in Figure 2-1 indicating that the causes of spastic dysarthria cannot be streamlined to a specific disease or situation.

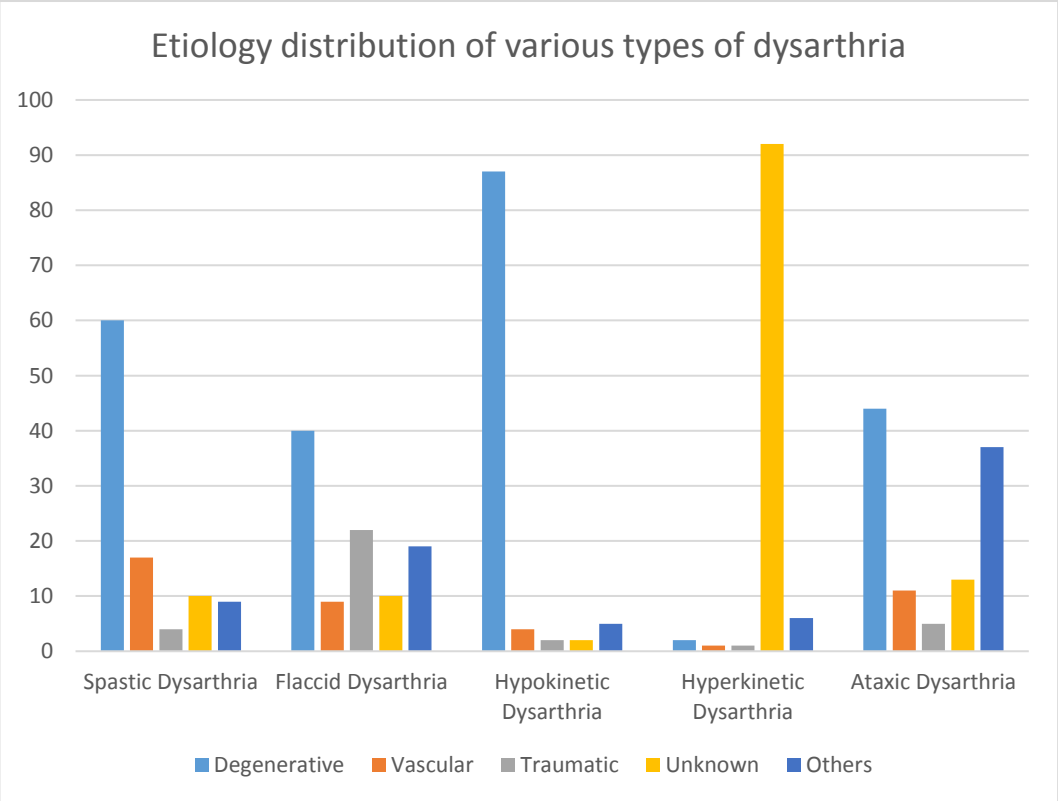


Figure 2-1. Etiologies of Various Types of Dysarthria [2]

2.2.2 Flaccid Dysarthria

Flaccid dysarthria, on the other hand, causes weakness and inevitably affects speech production. Depending on the position (adducted or abducted) of the vocal fold paralysed, the voice can be harsh, breathy and characterised by low loudness [2].

Due to the vocal fold paralysis, there is low variability in pitch and loudness resulting in mono-pitch and mono-loudness [4]. Other notable symptoms of flaccid dysarthria in speech are nasal emission when speaking and reduced loudness over time [3]. Research [2] has shown that unlike spastic dysarthria, 40% of flaccid dysarthria is caused by degenerative diseases. Only 22% of reported flaccid dysarthria is caused by surgical and non-surgical trauma whereas 19% are equally caused by vascular and unknown diseases as illustrated in Figure 2-1. The other occurrences are from infections, tumours and anatomic malformation [2].

2.2.3 Hypokinetic Dysarthria

Hypokinetic dysarthria is one of the most researched types of dysarthria because of its association with Parkinson's disease [29, 30]. Hypokinetic dysarthria is mainly caused by injury in the subtantia nigra [4]. It affects all speech subsystems – respiration, phonation, prosody, articulation and resonance [2]. Intelligibility is greatly affected in hypokinetic dysarthria and patients experience decreased movement and hoarseness in voice quality [2]. Hypokinetic dysarthria is also characterised by hypernasality, mono-pitch and mono-loudness [4]. Repetitions of syllables, freezing movement, rigidity together with reduced force and decreased range in movement can also occur [2]. According to the data from Mayo clinic on the occurrence [2], 87% of reported hypokinetic dysarthria are degenerative, 4% from vascular-related diseases, 3% from traumatic situation and infections. The other occurrences are from multiple and undetermined causes. Research in [2, 13] has shown that intelligibility of speech is largely affected by hypokinetic dysarthria; the loudness of speech is reduced, the pitch is more varied and the rate of speech is increased. These prosodic features can be used to distinguish hypokinetic dysarthria from healthy controlled speech.

2.2.4 Hyperkinetic Dysarthria

Hyperkinetic dysarthria, on the other hand, is associated with a lesion in the basal ganglia [2] which affects the control circuits of speech production which can also be manifested in all the five speech subsystems; articulation, prosody, phonation, respiration and resonance [31]. This type of dysarthria causes involuntary movement characterised as abnormal, unpredictable, fast or slow and sometimes

irregular in pattern [3]. It is characterised by strained, harsh or strangled voice quality [4] which is similar to that experienced in spastic dysarthria. A common resonance symptom is a hypernasality in speech production [4]. The exact distribution of the causes of hyperkinetic remains unknown [2]. This may be due to the fact that this type of dysarthria can be caused by various processes associated with damages to the control circuitry [2]. This further emphasises the need for early detection and classification.

2.2.5 Ataxic Dysarthria

This type of dysarthria affects the main speech subsystems but is more prominent in prosody and speech articulation. It is characterised by less frequent hypernasality, harsh voice quality, high variability in loudness patterns resulting in speech explosiveness. In terms of prosody, equal stress is placed on all words and syllables resulting in incorrect and excessive stress. The speech is slurred due to coordination and control deficiencies or breakdown [2]. The prosody of the speech is also characterised by an inaccurate duration, intensity [32], intonation and variability in articulation [4].

2.2.6 Mixed Dysarthria

According to the research from Mayo Clinic, the combination of any two or more types of dysarthria called mixed dysarthria, is the most frequent type of dysarthria accounting for about 30% of reported cases of dysarthria and also the most common type of motor speech disorders [2]. Examples of mixed dysarthria include amyotrophic lateral sclerosis, Friedreich's ataxia, central pontine myelinolysis, Wilson's disease, multiple system atrophy, hypoxic encephalopathy, multiple sclerosis etc. According to research statistics [2], 78% of etiologies of dysarthria are degenerative, 7% from vascular diseases and demyelinating diseases account for 3%. Other occurrences are from traumatic, neoplastic, toxic-metabolic, multiple and other causes [2].

Now that the various subtypes of dysarthria have been discussed in terms of their causes and characteristics, it is important to note that the causes of most of these dysarthria subtypes are unknown which makes it clinically challenging to detect the disorder at the very early onset. This challenge can be tackled by developing an

acoustic-based technique that is able to detect mild dysarthria before the symptoms become more severe, with the advantage of being less prone to errors due to external factors such as human interference. In addition, some of the speech subsystems affected by the various types of dysarthria (prosody, articulation and pronunciation) can be examined by analysing the speech samples using signal processing techniques.

Over the years, researchers, especially from clinical-related backgrounds, have reviewed ways of characterising dysarthria using perceptual methods and by physical examination, which introduces subjectivity and varying results based on the physician experience, training and/or exposure [16, 33, 34]. The question now is: "Are these (subjective) methods fair?". Also, "Should they (the methods) be conclusive especially when high variability in results can be expected?", "Are there alternative ways of examining dysarthria more objectively?" and "Can techniques that are strictly based on speech processing with little or no physical manipulation/intrusion be developed?" These questions form the basis of this research while highlighting the need to analyse the performance of new "objective techniques" and validate their applications in dysarthria management.

In the next sections, the existing and current methods used in the assessment and treatment of dysarthria will be critically reviewed which will include the review of various severity levels of dysarthria (presented in Section 2.3) and the review of perceptual and acoustic-based techniques used in dysarthria assessment and treatment (presented in Sections 2.4 and 2.5). In Section 2.6, a review of the strategies used in developing treatment tools for dysarthria will be presented which will be accompanied by a review of various treatment tools developed over the years for dysarthria treatment and the gaps identified in the implementation/application of these tools, cutting across both speech therapy and engineering/speech processing fields. After these critical reviews, the focus of this research will be discussed in line with the identified gaps which will form the basis of the review of the speech features presented in Chapter 3 and the justification for the contributions presented in Chapters 4 to 8 of this thesis.

2.3 Review of Severity Measures in Dysarthria

Apart from describing dysarthria based on the six subtypes as presented in Section 2.2, the occurrence of dysarthria can also be described by their severity levels. Severity in speech disorder simply is the measure of the extent to which the speech disorder has affected the patient's voice quality, the ability of the patients to express themselves verbally and how understandable the patient's speech is to the listeners [35]. In essence, speech disorder severity gives a measure of how far apart the speech is from what is regarded as "healthy" speech [35]. In this section, the severity levels in dysarthria will be reviewed, as well as, the measures used in quantifying the severity levels in practice and in current studies.

2.3.1 Intelligibility as a Measure of Severity in Dysarthria

The question then is: "How is the severity in dysarthria measured or quantified?" Over the years, there has been an increased focus on intelligibility as a measure for assessing how far apart the dysarthric speech is from "healthy" speech [28, 36-38]. Intelligibility, in the real sense, is described as the degree to which the listeners understand the speech with respect to the phonetic realisation of the speech [39]. This should not be confused with comprehensibility which also includes semantic, syntactic and pragmatic characteristics of the speech [38, 39].

A review of the literature shows that the intelligibility score is often used to describe dysarthria since it is a function of speech deficiencies [39]. Over a nine-year period, 70% of published articles on dysarthria severity made use of intelligibility as the primary severity measure [40]. The intelligibility scores are generated by asking experienced judges to score the utterances based on an agreed perceptual rating scale [38]. There are various perceptual rating techniques used for scoring speech intelligibility in dysarthria which can be grouped into two categories; relative, with a reference point and absolute, without a reference point. The choice of the rating scale is often determined by the intended applications and availability, or lack, of a potential reference sample. The effects of using either of these two categories of rating scales and their limitations are reviewed in Sections 2.4.1 and 2.4.2. Please note that the terms "Relative" and "Absolute" were introduced by the author of this

thesis as a way of grouping the various intelligibility rating scales based on whether, or not, a reference point is required.

A. Absolute Intelligibility Measurement Techniques

In these intelligibility measurement techniques, the listeners are required to listen to the utterances and write down (transcribe) what they heard. Usually, the speakers would be given single-words, sentences, passage or paragraph to read and their utterances recorded before asking the listeners to transcribe the utterances. The listener's transcriptions are then compared with the original passage/paragraph given to the speakers. The number of words correctly transcribed is then divided by the total number of words in the speech/utterance. This results in a percentage representation of the intelligibility score (0-100%) [41].

The absolute techniques which require the listeners to transcribe single-words are called the single-word intelligibility measures [40]. The use of single-word intelligibility measure has many advantages, one of which is that single-words are less demanding than sentences or paragraphs [40]. This also helps in reducing the potential for fatigue due to long reading sessions. In previous studies, semantically predictable words are often recommended to reduce the semantic distance between single words and words used in natural communication [40].

Phoneme intelligibility measure is another absolute intelligibility measurement technique which involves the examination of perceptual errors in consonants, vowels and diphthongs pronunciations [40]. A typical example of this is "The Phoneme Intelligibility Test" which was developed by Yorkston et al. [42]. Single-words are also used in phoneme intelligibility test but these words have target phonemes examined. The phoneme intelligibility test allows the phoneme error profile of speakers to be generated in a systematic way [40].

The sentence intelligibility measure is another absolute technique for measuring speech intelligibility [40]. Utterances comprising of sentences are given to listeners to transcribe and score based on the percentage of correctly transcribed words with respect to the total number of words in the sentences.

A review of studies on intelligibility measurements has shown that the single-word intelligibility scores can be different from sentences intelligibility scores based on

the severity of the speakers [40, 43]. For instance, a study by Yorkston et al. in [43] shows that speakers with highly severe dysarthria often show higher single-word intelligibility scores when compared with their sentence intelligibility scores. This is not the case with speakers with mild dysarthria who produce higher sentence intelligibility scores when compared with their single-word intelligibility scores. The accuracy of any proposed tool for the classification of the severity levels based on intelligibility needs to be high enough to compensate for the variations due to the choice of intelligibility task.

B. Relative Intelligibility Measurement Techniques

In relative intelligibility measurement techniques, a speech sample is chosen as the reference point and other speech samples are rated with respect to the reference point. Usually, the speakers would be given single words, sentences, passage or paragraph to read and their utterances recorded. After this, the listeners will be asked to judge how intelligible the speech is. A review of the literature shows that researchers make use of varying scales when performing relative intelligibility measurement. For example, in a study in [44], the authors made use of a 5-point measurement scale varying from 0, representing unintelligible, to 4, representing perfectly intelligible. This 5-point scale introduces subjectively to the intelligibility measure because not all the listeners can score the speech samples in a consistent manner. In another study by Hazan and Markham [45], a 2-point subjective intelligibility scale was used; where listeners rated the presented speech samples as "good" or "poor", which can often lead to erroneous results due to limited options. Other point scale measures have been used in assessing the intelligibility in dysarthria (2-point, 5-point, 4-point, 7-point and even 10-point scales [44, 45] [46]). These point scale measurement methods will be referred to as N-point scales. Generally, N-point scales in intelligibility measures are subjective, relative and are functions of the listener's exposure, training and experience which may lead to high variation in the results [45]. There is, therefore, a need to explore other ways to either compensate for these errors (for example, by training the listeners) or generate more objective measurement techniques that are less prone to errors.

However, some intelligibility measurement methods are based on the comparison of the speakers' utterances with a reference stimulus, regarded as a good reference

point of midrange intelligibility [47]. This technique is called the direct magnitude estimation (DME) [47, 48]. The reference stimulus can either be chosen by the listeners or by the examiners (researchers or clinicians) setting up the experiment. In a study by Joan et al [49], listeners were given the opportunity to choose the first utterance they listened to as the reference stimulus after which they compared the other utterances with the chosen stimulus. The first utterance is assigned 100 and the other utterances are rated with reference to the first utterance. If the intelligibility of the next utterance is perceived better than the first utterance, the listener will give it a score greater than 100 (200 if the listener thinks the intelligibility is twice better than the first utterance). If the listener perceives that the intelligibility of the next utterance is half as good as the first utterance, the listener gives it a score of 50. This way, relative scores are generated with respect to the first utterance [49]. This can often result in varied results across listeners (especially when the utterances are randomised).

A review of the literature shows that the choice of reference stimulus can affect the intelligibility measures [47, 49] and this is why some researchers have made use of a reference stimulus deemed to have midrange intelligibility [47]. This will ensure that all the listeners involved in the assessment make use of a single reference stimulus thereby reducing the variability of the measures and make the results more comparable. The DME techniques can also be applied with or without modulus. With modulus means that the reference stimulus is assigned a specific score (for example, 10 or 100) and without modulus (also known as free modulus) means that the listeners are allowed to give their own preferred score to the reference stimulus [47, 49].

The N-point scaling and the DME methods are based on the listeners' perception which can be prone to errors especially when the number of listeners is limited. To overcome this disadvantage, the number of listeners should be increased so that an average can be taken and reduce the effect of any spurious score from a listener. The accuracies of the absolute intelligibility measures are more consistent because they are less subjective than relative intelligibility measures. However, if the assessment is carried out by only one listener, an N-point scaling technique will be more appropriate. Apart from the choice of the intelligibility measurement

technique, there are, however, other factors that affect the intelligibility measures which will be discussed in Section 2.3.2.

2.3.2 Factors Affecting Intelligibility Measures

The intelligibility score is influenced by a number of factors which include the type of utterance (single words, repetition or sentences), how the speech is presented to the listeners (live speech or recorded speech), nature of listeners (native or non-native speakers) and how transcription is carried out (sentence completion or single word transcription) [39]. Environmental factors affecting intelligibility measurement have been examined by Dykstra et al [40] and Yorkston et al [50] and one of the main factors identified is the set-up where the speech recordings were taken. Factors such as lights, noise, closeness to recording equipment and the number of persons present can influence the ability of the speakers to speak naturally which inadvertently contributes to the speaker's perceived intelligibility [40]. Another environmental factor affecting the intelligibility measure is the support received by the speakers from their friends, family members, peers and sometimes, by the clinicians [40]. When assessing intelligibility in dysarthria it is important to consider these factors and minimise the impact, if possible.

However, some of the factors that affect speech intelligibility are inherent in speech attributes. There are four dimensions of speech intelligibility as suggested in [38] including articulation, nasality, voice quality and prosody. Out of these four dimensions, articulation accounts for the strongest correlation with the perceived intelligibility (0.82) followed by prosody (0.55), voice quality (0.46) and nasality (0.32). Three of these dimensions (articulation, prosody and voice quality) will be explored when developing the automatic algorithm for the detection of dysarthric speech in Chapter 6 of this thesis.

2.4 Review of Perceptual Techniques used in Dysarthria Assessment

Some of the current research in the assessment of dysarthric speech includes perceptual analysis [51], acoustic measurement [7, 13, 52] and intelligibility assessment [28, 53]. There are three main techniques traditionally used in the

perceptual assessment of dysarthria which include Frenchay's Dysarthria Assessment (FDA) [19], Robertson Dysarthria Profile [2] and Dysarthria Examination Battery (DEB) [21].

2.4.1 Frenchay's Dysarthria Assessment

After the term Dysarthria was first used to describe deviations in speech and pronunciation due to neurological disorders in 1976, it became important to develop methods for assessment of this disorder and distinguish it from other speech-related disorders. At this period, a descriptive method of assessment was conventionally used which is subjective, unreliable with low sensitivity [33]. Different adjectives were used to describe the speech produced by patients with dysarthria [10, 54]and this brought up the need to develop a standard method for assessment of dysarthria. The Frenchay's dysarthria assessment (FDA) was developed in 1980 as a more reliable assessment of dysarthria compared to the conventional descriptive method. The researchers at the Frenchay Hospital in Bristol (FDA was named after this hospital) felt that the dysarthria assessment should be applicable to therapy, sensitive to change, short and easy to use, require little training and have results that are easy to interpret [33].

FDA was based on the review of how speech therapists assess and analyse the patient's behaviours focusing on relevant features of their speech production [33]. To adequately describe these behaviours, different activities were designed under eight sub-headings; which includes reflex, respiratory, lips, jaw, palate, laryngeal, tongue, and intelligibility scores resulting in 29 tasks in total. Under each sub-heading, specific activities were rated and the average is taken to give a score between a (least severe) to e (most severe) relying on how skilled they are. In fact, it has been shown in [33] that the inter-scorer reliability of FDA increases with therapist's training. Since then different versions of the FDA [29, 55] have been developed but most of them have not been able to completely solve the problem of subjectivity in the assessment.

2.4.2 The Robertson Dysarthria Profile (RDP)

The RDP was developed by SLTs in 1982 to assess motor speech disorders [56] with the aims to provide:

- Profile of patient's speech abilities and disabilities
- Information to aid in dysarthric classification
- A basis for dysarthria therapy and management

Although the first two aims were met in the first publication [56], a program was developed 5 years later to meet the third aim [57]. In this profile, the speech samples of the patients are scored using a 5-grade system from *Normal* to *Good*, to *Fair*, to *Poor* and, finally, to *None* leaving it prone to inter-therapist variation based on their experiences and training levels. In RDP, the patients are assessed under eight sub-headings which include respiration, phonation, facial musculature, diadochokinesis, oral reflexes, articulation, intelligibility and prosody [30]. These parameters were weighted differently using the maximum scores [58] as illustrated in Table 2-2.

Table 2-2. The Robertson Dysarthria Profile Scoring

Parameter	Number of Tasks	Maximum Score
Respiration	5	20
Phonation	12	48
Facial musculature	20	80
Diadochokinesis	11	44
Reflexes	7	28
Articulation	5	20
Intelligibility	6	24
Prosody	5	20
Total	**71**	**284**

Unlike in FDA, where 29 tasks are carried out, 71 tasks are carried out in RDP across the eight different parameters each carrying a score of 4 [58]. Also, it is important to note that the highest number of tasks is carried out under facial musculature which involves physical examination [30].

2.4.3 Dysarthria Examination Battery

Dysarthria Examination Battery (DEB) was first introduced by Drummond in 1993. DEB examines the patients in all the five speech production subsystems. In DEB, five tasks are carried out under the respiration sub-heading involving resting breathing, vital capacity, maximum performance tasks (MPT), s:z ratio and word per inhalation. Fundamental frequency, intensity, intensity range, maximum loudness, maximum pitch and speech quality are also measured under phonation. DEB also contains an additional task of oral sensitivity test during the tactile stimulation. The major difference between DEB and other assessment tools discussed above is the introduction of acoustic measurements to the assessment. Only a few pieces of literature were found on DEB. This could be due to the fact that some of the measured parameters are similar to those in RDP and FDA. Some of the measures in DEB are also subject to the therapist's perception which can vary based on experience and training. All the tasks carried out in the three assessment techniques discussed above are presented in Appendix A.

2.4.4 Computerised Frenchay's Dysarthria Assessment

The CFDA is a more recent tool for dysarthria diagnosis. It was proposed in 2015 by James Carmichael [19]. It is based on digital signal processing methods to quantitatively assess speech signals in a quest to diagnose the traits of dysarthria. CFDA procedure consists of two main parts; respiration test and phonation test [19]. The respiration test involves asking the user to take a deep breath using the mouth. After this, the user is asked to exhale the taken breath slowly in a way that exhalation can be heard. On the other hand, in the phonation test, the user is asked to say the phoneme "AH" and prolong/sustain it as long as they can. The two tests are then graded from the highest score A to the least score E. While this technique works well for phonation and respiration measurement (when compared with traditional assessment method), the pitch detection and excessive airflow (hypernasality) have not been taken into consideration [19]. The limitations identified in the FDA are also present in CFDA and the problem of the subjectivity of users still exists.

The perceptual analysis is very useful in the initial clinical diagnosis and assessment of dysarthria. However, due to the fact that human beings have different perception capabilities and styles, perception scores can vary considerably. This can be seen in the global statistical analysis of various dysarthric speech as recorded by DeMino and Dynamics in 2011 [26] and the scoring of the Darley, Aronson and Brown speech samples [10]. Now that the state-of-the-art perceptual techniques used in dysarthria assessment have been reviewed, it is important to also review published studies on the use of acoustic features in dysarthria assessment and classification which were presented in Section 2.5.

2.5 Review of Acoustic-based Techniques used in Dysarthria Assessment

In the past two decades, the focus of the research community in dysarthria assessment has shifted from using traditional perceptual techniques to developing acoustic techniques that are based on speech processing technologies [12, 59]. This shift is mainly attributed to the need to objectively assess dysarthria using signal processing and machine learning techniques applied in dysarthria speech recognition [60-62], loudness detectors [63], pitch trackers [64], speech rate detection [52], articulatory feature extraction [59], and respiration and phonation analysis [19]. Speech processing techniques such as linear prediction coding, Mel-frequency cepstral coefficients, perceptual linear prediction and machine learning techniques have also been used in some of these acoustic assessments [9, 19, 52]. One of the major difficulties in acoustic dysarthria assessment is the development of techniques that can easily detect subtle changes in dysarthric speech features as most of the existing speech processing methods work well for healthy speech but do not give good and consistent results when analysing dysarthric speech [62]. Another gap identified is the fact that some of the machine learning techniques used (for example DNN) require a large amount of training data; which are not readily available for dysarthric speech.

2.5.1 Choice of Speech Features

There are several aspects that can be used to acoustically assess dysarthria and research [65] has shown that a single method may not be appropriate for every case

of speech disorder, thereby necessitating the need to choose appropriate acoustic features for specific assessment application and/or speech disorder. Researchers found a slow syllable rate with high variability in dysarthric speakers [10, 65] and unlike in dysphonia, where only the instantaneous acoustic features of the speech are distorted, duration-related features and speech dynamics are also distorted in dysarthria [65]. Another study in [42] shows that the articulation movements of dysarthric speech are also affected due to poor speech coordination and weakness.

In other studies [65-67], researchers have noted that spectral features such as fundamental frequency, energy, jitter, shimmer, and harmonics-to-noise ratio are good indicators of pathological speech. Some other studies [68, 69] suggested that cepstral features (such as Mel Frequency Cepstral Coefficients) are also good indicators of dysarthric speech. The list of relevant features keeps growing over the years even though there are few overlaps on the chosen features by researchers. A review of these speech features and their relevance to dysarthria assessment will be presented in Chapter 3 of this thesis.

2.5.2 Dataset/Corpus Used

Apart from the intended application, another factor that determines the choice of speech features is the available dataset. A review of the literature shows that the dataset used by researchers in developing dysarthria assessment techniques were recorded by experienced personnel, mostly clinicians, within a controlled environment [65]. Even though recording and collection of speech data within a controlled environment might be more realistic for research purpose, it sometimes does not provide the true presentation of the speaker's day-to-day natural communication. Another challenge is that the available data might have been collected for a different purpose (for example, for speech recognition) other than for automatic assessment [65].

Lack of availability of data is one of the limitations researchers face in the development of automatic assessment techniques. Another common challenge in terms of the dataset is that there is no specific standard for how data for automatic dysarthria assessment should be collected [65]. Likewise, the distribution of the available dataset varies considerably across various age groups, gender and sometimes, demography. This makes cross-validation of assessment techniques

across the various dataset challenging [65]. In addition, age and gender mismatch between the dataset from dysarthric speakers and that of controlled speakers often pose challenges when analysing and validating the automatic assessment techniques [2, 65].

In the development of an efficient dysarthria assessment tool, it is important to take the challenges introduced due to the dataset into consideration. This can be done by using age and gender-matched dataset and performing appropriate pre-processing on the speech signals before features are extracted.

2.5.3 Acoustic-based Dysarthria Assessment Techniques Over the Years

A summary of automatic dysarthria assessment techniques proposed by researchers in the past decade is presented in Table 2-3 which includes the performance of the various techniques. One of the notable studies on automatic dysarthria assessment was presented by Jangwon Kim et al [70] where sentence-level speech features, extracted under three subsystems: prosody subsystem, voice quality subsystem, and pronunciation subsystem, were used for the classification problem. Their choice of speech features was based on the evidence that dysarthric speakers often find it difficult to produce certain speech sounds which alter their prosody, pronunciation and of course, voice quality [2]. Under the prosody subsystem, pitch and durational related features were extracted whereas harmonics to noise ratio, jitter and shimmer features were extracted under voice quality subsystem and syllable duration, pause to syllable ratio, vowel duration and MFCC-based features were extracted under the pronunciation subsystem [70]. Also, the performance of three classification techniques, support vector machines (SVM), k-nearest neighbours (KNN) and Linear Discriminant Analysis (LDA) were compared. Using the TORGO database, a maximum accuracy of 71.3% was realized for prosody subsystem (using SVM classification), 71.7% for pronunciation subsystem (using SVM classification), 68.9% for voice quality subsystem (using LDA classification) and 72.0% for a combination of the three subsystems (using LDA classification).

Table 2-3. Summary of Automatic Dysarthria Assessment Techniques Proposed by Researchers in Published Studies in the Past Decade

S/N	Study	Year	Corpus used	Classifier	No of Classes	Features	Accuracy	Focus
1.	Automatic intelligibility classification of sentence-level pathological speech [70]	2015	TORGO and NKI CCRT databases	SVMs, KNNs	2	Pitch, Duration, HNR, Jitter, Shimmer, MFCCs, Formants and Pause to syllable ratio	72.0%	Pronunciation, voice quality, and prosody
2.	Automatic detection of Parkinson's disease in running speech spoken in three different languages [71]	2016	100 Spanish, 176 German, and 36 Czech speakers	SVMs	2	MFCCs, Formants, F0, Energy, Duration, Pauses, Bark band energies	97%, 94.3%, 85% respectively	Prosody, pronunciation
3.	Automatic estimation of Parkinson's disease severity from diverse speech tasks [72]	2015	Parkinson's and Eating Condition Sub challenges	SVMs with fifth-order kernel	2	MFCCs, GFCCs, GBF, MFBs, F0, Energy, HNR, Jitter, Shimmer, segmental features	76.2%	Pronunciation, voice quality, and prosody
4.	Perceptually Enhanced Single Frequency Filtering FNR Dysarthric Speech Detection and Intelligibility Assessment [8]	2019	UASPEECH (16 dysarthric 13 health speakers)	GMM-UBM	4	PLP, MFCCs, Multitaper MFCC, CQCC, PE-SFCC	93.64%	Articulation
5.	Automatic Evaluation of Articulatory Disorders in Parkinson's Disease [59]	2014	46 native speakers of Czech	SVMs	2	Formants, DDK rate, pace and fluctuation, VSQ, VOT	83.3%	Articulation
6.	Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge [73]	2016	Nemours and Torgo databases	GMMs, SVMs and hybrid	4	MFCCs, external, middle and internal ear model features	97.2%	Phonation and Auditory model
7.	Classification of speech intelligibility in Parkinson's disease [46]	2014	240 running speech samples from 60 PD and 20 HCs	SVMs	2 and 3	ZCR, F0, STE, MFCCs, Cepstral Separation Difference (CSD)	92% (2 classes) 85% (3 classes)	Phonation, articulation and prosody

S/N	Study	Year	Corpus used	Classifier	No of Classes	Features	Accuracy	Focus
8.	Assessment of dysarthric speech through rhythm metrics [74]	2013	11 males with dysarthria and one non-dysarthric adult	Gaussian Bayes	2	rPVI, nPVI, vocalic intervals and duration of voiced & non-voiced intervals	62.98% (Dysarthric) 90.3 % (HC)	Rhythm metrics
9.	Fully automated assessment of the severity of Parkinson's disease from speech [75]	2015	168 PD speakers	SVM, Ridge and Lasso Regression	2	Pitch, Jitter, Shimmer, Loudness, HNR, MFCCs	Mean Absolute error: 5.5	Loudness, Voice quality & articulation
10	Automated Intelligibility Assessment of Pathological Speech Using Phonological Features [76]	2009	211 speakers (51 HC)	Linear regression	2	Phonemic, Phonological, Context-Dependent Phonological features	RMSE 3.9	Phonemic and Phonological measures
11.	Detection of Amyotrophic Lateral Sclerosis (ALS) via Acoustic Analysis [77]	2018	67 ALS and 56 healthy speakers	SVM	2	MFCC, Spectral variance, Spectral Entropy	83% female and 79% male	Statistical feature selection
12	Intelligibility Classification of Pathological Speech Using Fusion of Multiple Subsystems [78]	2012	NKI CCRT Speech database	Bayesian	2	Phoneme probability feature, Pitch, Duration, MFCCs, Jitter, Shimmer, Formants, HNR, pause	79.9% (76.8% on test set)	Phoneme, Prosody and Intonation
13	Automatic Assessment of Dysarthria Severity Level Using Audio Descriptors [79]	2017	UA speech and TORGO database	Artificial Neural Network	4	Energy, STFT magnitude & power, ERBFFT & Gammatone, Multi-taper Harmonic & Power	96.44% (UA) and 98.7% (TORGO)	Audio Descriptors
14	Automatic Detection of Speech Disorder in Dysarthria using Extended Speech Feature Extraction &Neural Networks Classification[80]	2017	10 Ataxic Dysarthric and 10 HC speakers	Neural Networks	2	Centroid Formant	75.6%	Articulation

A few other studies have extended Kim's technique to include more features and more languages other than English. One of those studies was presented in [71] in 2016 where the authors proposed a method for automatic detection of Parkinson's disease in three European languages; Czech (36 speakers), German (176 speakers) and Spanish (100 speakers). For enhanced speech feature extraction, the authors initially segmented the speech samples into silence, unvoiced and voiced segments [71]. This was followed by extracting the MFCC features from both the unvoiced and voiced segments. Prosodic features (fundamental frequency, energy, duration and pauses), as well as noise and formant measures, were extracted from the voiced segments. Finally, the energy of the unvoiced segment was distributed into twenty-five filter banks. The study made use of SVM classification technique due to its success in similar related studies [70, 81]. The speech samples were classified into two groups namely; dysarthric speech (Parkinson's disease -PD) or healthy control speech (HC). Accuracies of 97%, 94.3% and 85.0% were realized for Spanish, German and Czech group respectively. The technique proposed in this study is not fully automated. This can lead to errors and delays during the clinical application of the technique. There is also a need to validate the performance of the proposed technique using sentences and passage reading dataset.

The choice of features for the classification feature vector is a function of the type of speech task in the available dataset. For instance, a study in [72] in 2015 showed that the effective features needed for a specific classification task vary depending on the type of speaking task. In their study, the authors proposed a fully automated Parkinson's disease severity classification technique where frame-level and utterance-level features were extracted from the speech samples. The extracted frame-level features included MFCCs, Mel-Frequency Banks (MFBs), spectral features, Gabor features, spectro-temporal features, Gammatone Frequency Cepstral Coefficients (GFCCs) and prosodic features (fundamental frequency, energy (RMS)), and voice quality features (HNR, jitter and shimmer). Whereas, the utterance-level features were extracted using functional and i-vector extraction methodologies (using the Universal Background Model (UBM)). Combining these features resulted in an accuracy of 76.2% on the training dataset and 74.6% on the test dataset using SVM classification technique with fifth-order kernel [72]. Although this study suggested that feature fusion system using unsupervised

learning has the tendency to increase the accuracy, the proposed techniques suffer from high complexity which may cause a delay in real-time clinical applications. Also, there is a need to improve the performance (accuracy) of this technique.

Some researchers have focused on single speech subsystems (the effects of the five speech subsystems on different types of dysarthria are discussed in Section 2.2) in tackling the automatic assessment problem. In [59] an automatic articulatory disorder assessment for Parkinson's disease was presented although phonation and prosodic cues in the speech signals were not considered in this technique. Thirteen articulatory features (including voice quality, coordination of Supralaryngeal and Laryngeal activity, occlusion weakening, the precision of consonant articulation, speech timing, and tongue movement features) were extracted from the speech samples [59]. SVM was used in the classification process. An accuracy of about 87.1% was reached using this technique. Because the technique was based on diadochokinetic tasks (repeating /pa/-/ta/-/ka/ syllables at a fast rate), the results of these techniques are not directly comparable to other techniques proposed in [70-72]. Performance validation is needed for this articulation-focused technique.

One of the widely published automatic assessment techniques is the phonetic score histogram [65]. This method was motivated by a study by Green and Carmichael [55] where a new feature called the goodness of fit (GOF) to Hidden Markov Models (HMMs) was investigated. The GOF feature was used to develop an automatic intelligibility metric system for single words [55]. The authors in [65] however used the HMMs to align recorded speech samples with their phonetic transcriptions. The resulting parameter (phoneme log-likelihood) is then normalised by the duration of the speech and used to determine if the speech sample was dysarthric or not. It is expected that the distribution of the normalised phoneme log-likelihood follows a specific pattern for healthy speakers with peaks falling in the same histogram bin. Their study showed that this is different for dysarthric speakers for whom the location of the peaks are moved to lower histogram bins as the severity of the disorder increases [65]. This technique is based on the statistical distribution of the phoneme log-likelihood and not on the phonetic, prosody and articulatory features in speech. Also, to be able to use this method, a phonetic transcription of the speech sample is needed which makes the process semi-automatic in nature.

Although most of the studies on automatic dysarthria assessment are based on binary classification (presence or absence of dysarthria), few studies are also focused on the ability to automatically classify dysarthria into various severity levels. Most of these studies have made use of a 4-point severity level classification with "0" indicating the absence of dysarthria, "1" indicating mild severity, "2" indicating moderate severity and "3" indicating high severity. For instance, in a recent study by Kadi et al [73], a 4-point dysarthria severity was carried out on Nemours and Torgo dataset. Auditory knowledge was used to simulate the models for the external, middle and inner sections of the ear. The features from the auditory models were then combined with MFCC to form a feature vector. Gaussian Mixture Models (GMMs), SVM and a hybrid of the SVM/GMM classifier were used for the classification. The technique produced an accuracy of 93.2% [73] as against an accuracy of 85% realised in an earlier study by Khan et al in [46]. There is a continued interest in developing techniques with better performance using speech features only and this is what this research work intends to achieve.

The list of literature keeps growing over the years as both clinicians and researchers are interested in automatic technologies that make the assessment of dysarthria easy and accessible as presented in Table 2-3. In summary, it is noteworthy that most of the classification techniques developed in the past five years are focused on the application of the SVM classifier or its variant. This is mainly due to its proven performance in automatic speaker and emotion recognition in pathological speech [82-86]. A review of the literature indicated that the performance of the SVM classifier is highly consistent even with the increasing number of the dataset used [77, 84]. The application of SVM classifier in speech processing application will be further reviewed in Section 3.7 of this thesis. Also, the performance of various classification techniques, including the SVM, will be compared in Chapter 6 of this thesis.

2.6 Review of Strategies Used in the Treatment of Dysarthria

Over the years, a majority of treatment strategies employed in the treatment of dysarthria focused on the speech production subsystems using either behavioural, instrumental or prosthetic techniques [18, 41, 87-92]. Behavioural techniques use traditional approaches in teaching patients new compensatory skills in speech rate,

intensity and repetitions whereas instrumental techniques involve the application of modern technology in measuring and giving feedbacks to patients on specific speech production skills and the prosthetic techniques use technologies that alter the physical properties of the patient's speech production subsystem. Review of these techniques, in different speech production subsystems, are described below.

2.6.1 Speech Rate

In literature, manipulation of speech rate was used as a treatment technique for improving the intelligibility of dysarthric speakers [18, 41, 87-90]. This dysarthria treatment technique was first presented by Beukelman and Yorkston [93] where alphabet boards were used to slow down the speech rate which helps in improving patients' speech intelligibility [21]. The effects of speech rate dysarthria on the treatment of Parkinson's disease (hypokinetic dysarthria) was also discussed in [88] where both dysarthric and control speaker groups showed considerable improvement in the proportion of pauses located at syntactic boundaries. However, the effect on this syntactic improvement on speech intelligibility was not examined.

Moreover, in a later study by Dagenais et al [94], speech samples from four dysarthric and two control speakers were manipulated (normal speech rate, 30% slower speech rate and 30% faster speech rate) and presented to listeners. The listeners rated the utterances in terms of intelligibility (number of correct words) and acceptability (perception of utterances). While the acceptability increased with faster speech rate for intelligible speakers and a slower rate for less intelligible speakers, the intelligibility remained unchanged for most speakers across the different speech rates [94].

Additionally, there is a varied level of success with studies that compare the effect of reduction in speech rate. Research [50] suggested that slowing down the speech rate does not significantly affect how natural or intelligible the patient's speech was. Pilon's study [95] on the other hand showed that speech rate manipulation had different effects on the patient's speech depending on their severity as speech rate reduction resulted in increased intelligibility for severe dysarthric speakers and decreased intelligibility for mild dysarthric speakers [95]. The direct impact of varied speech rate on dysarthria treatment remains debatable.

2.6.2 Resonance

Some studies have also considered how resonance can be used to manage dysarthria. One of the methods involves the use of continuous positive airway pressure (CPAP) to exercise the soft palate during the production of speech [21]. This produces resistance to the velopharyngeal muscles [92]. Another technique involves the use of a palatal lift which consists of a lift part (along the surface of the soft palate) and a retention part (covering the hard palate). This is fastened to the teeth by prosthodontist using wires. Although this treatment method has proven to be effective, it remains intrusive and not easily accessible [21].

2.6.3 Oro-motor

Exercises involving oro-motor have also been used in the treatment of dysarthria [21]. One of the oro-motor treatment techniques involves the use of active exercises focused on strength training leading to an increase in the tension that can be sustained in a muscle over a period of time and the speed of produced tension. Passive exercises are also used which involves deep massage, belly tapping and stretching. Studies [21, 91] suggests that strength training can help in producing required force and speed for improved intelligibility. However, according to a study in [91], there is a need to continue these exercises until a state of fatigue is reached.

2.6.4 Articulation

A comparison of speech, articulation and alternating motion rates in [96] suggested that articulation treatments showed a promising impact on the treatment of dysarthria. When combined with oro-facial muscle movement, a study [34] showed that articulation exercises produced a considerable improvement in intelligibility. This treatment which involved eight patients with dysarthria post-stroke lasted for 10 weeks involving 45 minutes of weekly oro-facial movement and articulation exercises with each exercise repeated four to five times, three times daily. RDP was used to assess the speaker's progress before and after the therapy. The study showed improved intelligibility by an average of 9.9% across speakers. The effects of these articulation-based treatment methods, however, need further investigation using a larger sample size.

2.6.5 Prosody

Prosody techniques use a combination of various speech elements: pitch, intensity, intonation and stress [21] which can either be individually manipulated [97, 98] or combined [21, 67] in management of dysarthria. The Lee Silverman Voice Treatment (LSVT) exercises focus on increasing the intensity (loudness) of the utterances in an aim to increase the speech intelligibility [97]. Some researchers [22, 63, 98, 99] have reviewed the LSVT programme and found improvements in speech rate, phonation, articulation and intelligibility.

Although LSVT gives increased loudness and improved vowel space area, the direct effect of this treatment on intelligibility has not been established [63]. Also, the perceptual based articulatory ratings after LSVT treatment has not shown substantial improvement [63]. Another shortcoming of the LSVT is sensitivity to subtle articulatory differences especially in consonant production [63] [98]. People with difficulty with other speech subsystems, (prosody, resonance, respiration and phonation) may find the use of LSVT alone insufficient for treatment and improved intelligibility.

Another study [100] investigated the effects of combining the use of LSVT with physiotherapy of the wall of the upper chest. The initial study lasted 4 weeks with an hour of therapy per week resulting in an improvement in the carryover from sustained vowels compared to using LSVT alone. After which respiration exercises were added for another 10 weeks resulting in improvement in reading ability, lung measures, speech sustainability and intelligibility. The use of prosody in dysarthria treatment has proven to be more effective when combined with other therapy techniques.

Over the years, clinicians, have combined traditional and modern methods in the management of dysarthria and there is continued research of the most effective method(s) based on the type of dysarthria and severity of the disorder as described in Section 2.3. A review of current trends in dysarthria treatment is presented in Section 2.7 with a focus on identifying existing gaps in the treatment techniques.

2.7 Review of Current Techniques used in Dysarthria Treatment

The review in this section will be focused on the aims of treatment, the treatment structure, and the state-of-the-art treatment techniques. Specifically, treatment techniques proposed by researchers [19, 22, 63, 97, 99] over the last few years will be critically reviewed.

2.7.1 Aims of Treatment

The majority of current treatment techniques are aimed at improving the speech intelligibility of dysarthric speakers [21]. Increased loudness has been used by a few researchers to achieve this aim [22, 41, 97, 98]. In an aim to improve the speech intelligibility, some researchers [41, 90] have also considered the use of speech rate. Further evidence on the use of speech rate in treating dysarthria has been discussed in Section 2.6.1. Other researchers have focused on improving a specific speech aspect; for example, fundamental frequency [21], respiration [100], duration [41], and intonation [90].

An aspect that has gained research attention in dysarthria treatment is the quantification of the measure of therapy needed (or targets to be met) to improve the speech intelligibility. For example, in stress marking exercise for dysarthria treatment, research [32] has shown some deficiencies in loudness, pitch and duration in dysarthric speech and recommended therapy include working on these deficiencies. But it is unclear how to manage the deficiencies using quantitative scores. In this research (in Chapter 7), this will be addressed and an evidence-based treatment tool for dysarthria using the stress marking exercise will also be proposed.

In conclusion, the ultimate aim for dysarthria treatment remains “to improve speech intelligibility and communication effectiveness” As discussed in Section 2.7, most researchers focus on using speech rate, resonance, oro-motor, articulation and/or prosody tasks to achieve improved intelligibility. In Section 2.7.2, the structure of some of the clinically accessible and/or published treatment tools will be reviewed.

2.7.2 Treatment Structure

The treatment recommended by the therapists to a particular individual is determined by the type of dysarthria, nature of the symptoms and, the severity of

the dysarthria which will often require a combination of multiple treatment techniques [21]. There are identified cases where the combination of a selection of treatment techniques and communication strategies were recommended for various individuals [21]. Although the combination of treatment techniques is often used in therapy, there is limited evidence to its effectiveness in therapy.

One key factor that contributes to the choice of treatment structure is the availability of the resources needed for the possible treatment techniques [21]. For example, practical tools needed for a specific treatment technique might be expensive or not accessible. Another factor is the willingness of the patients. Some dysarthria patients might not be open to the use of treatment techniques that will require physical intrusion or can cause any inconvenience (like wearing palatal lift [21]). Therefore, research attention is gradually shifting towards the development of speech-based treatment technologies that require little or no physical intrusion. A behavioural experiment will also be carried out in this research (Chapter 7) to examine the effect of the combination of prosodic cues in the treatment of dysarthria in stress marking exercise.

2.7.3 Other Treatment Techniques

Apart from the LSVT, some researchers have come up with alternative treatment techniques that showed promising results for dysarthria management. Many of these published techniques, however, still require clinical validation.

A. Dysarthria Treatment Programme

One of the published dysarthria treatment programmes named "The Dysarthria Treatment Programme (DTP)" was designed to involve 53 activities in 18 different tasks, nine (9) of which are speech-based [101]. The guidelines and task-specific stimulus were provided as part of the programme. Due to the large volume of tasks and activities available to choose from, support on the choice of task, the duration of tasks and complexity of intervention was also provided. Although the performance of this programme has not been extensively reviewed by researchers, the effects of DTP has been examined in two patients [21, 101]. The DTP treatment was administered in 7 sessions within a period of 3 weeks. One patient showed improved intelligibility after the treatment but the condition of the second patient

got worse (not necessarily because of the therapy) after the 3 weeks [101]. This programme has not been validated and lacks therapy effectiveness evidence.

B. Altered Auditory Feedback (AAF)

AAF, which was initially developed for people with stuttered speech, was proposed in 2010 [99] as another method of controlling the speech rate of people with dysarthria (PwD). The AAF involves having the speaker wear an audio device through which the speaker hears the altered version of their speech, therefore, making the speaker slow down when speaking. Research [99] suggested that the use of AAF can have a positive impact on the improvement of intelligibility of people with Parkinson's disease although this method has not been tested on other types of dysarthria. It is also important to point out that based on the results of this research there is no correlation between the severity and treatment progress [99]. Other variants of AAF that might be useful for dysarthria treatment include delayed auditory feedback (DAF) where the speech heard by the speaker is delayed by some time (50 to 200ms) and frequency-shifted feedback (FSF) which involves sending frequency shifted (distorted pitch) version of the speech. These methods, however, require testing on other dysarthric speakers.

C. Computerised Assessment and Treatment of Rate, Intonation, and Stress

CATRIS was developed in a quest to assess other dysarthric speech features other than loudness, which was the focus in LSVT [97]. There are two major aims of CATRIS; development of an assessment tool for stress, intonation and rate of speech and the development of a computerised speech therapy tool. Based on related research [102], it has been discovered that intensive speech rate (reduced rate) and intonation treatment (contrast in the final intonation pattern of questions and statements) can improve the intelligibility of speech in neurological diseases such as Parkinson's disease. This method has also not been tested extensively on different types of dysarthria and analysis of other respiration and phonation features in dysarthric speech is lacking.

D. Music Therapy for Dysarthria Treatment

One interesting study on dysarthria treatment uses music therapy treatment protocol based on research in biomedical theories which showed that musical stimulations can help in improving neurological conditions [103]. The music treatment protocol involved the use of preparation, respiratory, oro-motor, rhythmic, and melodic exercises [103]. The study also suggested that the use of vocal intonation tasks and therapeutic singing can help to enhance the prosodic cues in patient's speech [103]. This suggested protocol has not been extensively clinically validated although there is theoretical evidence of its potential effectiveness in dysarthria treatment. However, the use of music-protocols in the treatment of dysarthria is not the focus of this research but an extensive review of the various music-protocols for the treatment of dysarthria and their limitation can be seen in [104].

E. "Be Clear" Therapy

Another notable dysarthria treatment technique recently published in [105] is called "Be Clear", an intensive dysarthria treatment programme focused on improving speech intelligibility designed originally for adults with non-progressive dysarthria. This dysarthria treatment protocol comprises of two stages; the pre-practice stage and an intensive practice stage based on four categories of tasks including functional phrases, service request, functional speech tasks and homework tasks [105]. The protocol relied heavily on the repetition of tasks. The performance of this treatment protocol was examined using 8 participants with non-progressive dysarthria. The results showed an improvement in word intelligibility and sentence intelligibility by 3.2% and 8.6% respectively. There is, however, a need to cross-validate these results using more dysarthric speakers across different severity levels.

F. Combined Therapy Approaches

A review of the literature showed an interesting trend in combining two or more therapies to improve the speaker's intelligibility during dysarthria treatment [100, 106]. Combined therapy approaches are not only common in speech-based therapies but also in speech and non-speech therapy combinations. Tamplin in [106] combined the use of vocal exercises with the therapeutic singing of familiar songs. This method helped to improve the naturalness of the speech and reduce the duration of pauses between words. Solomon et al also proposed, in their study [100],

a combination of intensity and respiration based treatment which involved the combination of LSVT therapy [97] and breathing exercises with physical therapy. This produced improvements in intelligibility and sound pressure level, although several speech measures were not sustained after the treatment [100]. Unfortunately, most of these studies were carried out on single or very few patients and require clinical cross-validation.

As discussed in Sections 2.5 and 2.6, this research will be focused on the automatic assessment of dysarthria using speech features. The feature extraction and classification techniques of these speech features will be reviewed in Chapter 3.

2.8 Summary

In this chapter, the term dysarthria has been introduced as a neurological motor speech disorder that is grouped into six categories based on the characteristics and causes. A review of current clinical techniques used in the assessment and treatment of dysarthria has also been presented which includes both perceptual techniques and acoustic techniques. An overview of severity levels as a function of the speakers' intelligibility scores have been presented and studies involving the use of non-speech/speech-based features have been reviewed with respect to the various strategies used in dysarthria treatment. Finally, existing and current techniques used in the treatment of the dysarthria have been reviewed while identifying the research gaps and limitation of these techniques.

In the next chapter, speech processing technologies for the assessment and treatment of dysarthria will be reviewed. The different methods used in the extraction of the relevant speech features identified in this chapter (Chapter 2) will be reviewed and their limitations discussed in Chapter 3. This review will form the basis for the chosen research methodology and contributions presented in Chapters 4-8 of this thesis.

Chapter 3

3 Feature Extraction and Classification Techniques in Dysarthria Management

3.1 Introduction

Extraction of features from speech signals is a fundamental requirement for most speech processing, recognition and identification applications which includes mathematical modelling, time-domain analysis, spectral analysis and/or cepstral analysis of the speech signals. The features to be extracted from speech signals are determined by the intended applications. In disordered speech processing applications, the goal of the feature extraction scheme is to describe each speech signal using reliable representations such that dissimilar utterances can be differentiated. In this chapter, feature extraction techniques for dysarthric speech processing will be reviewed and a comparison of the standard techniques used in speech feature extraction will be carried out. A review of state-of-the-art techniques used in the silence-unvoiced-voiced segmentation will also be presented, as well as, a review of machine learning schemes used in various speech disorder detection applications.

3.2 Pre-processing of Speech Signals

In signal processing, pre-processing is often carried out in order to enhance the performance of the feature extraction algorithms [107, 108]. Amplitude normalisation, noise reduction, pre-emphasis filtering, and direct current (DC) component removal are some of the commonly used pre-processing techniques in speech processing applications which will be introduced in this section.

3.2.1 DC Component Removal

Recorded audio signals often contain a constant component with a non-zero mean [109]. This could be due to the DC bias of the equipment used for recording and storing the audio signals which creates a DC offset that carries no useful information [109]. The DC offset can affect the signal energy calculation if not removed. The effect of

DC offset is reduced by subtracting, from the audio signal, the mean amplitude of the audio signal given by

$$s'[n] = s[n] - \mu_s \quad (1),$$

where s[n] is the original audio signal of length N, s' is the signal after removing the DC component and μ_s is the mean amplitude of the audio signal defined by

$$\mu_s = \frac{1}{N}\sum_{n=0}^{N-1} s[n] \quad (2).$$

The effect of DC component removal is illustrated in Figure 3-1.

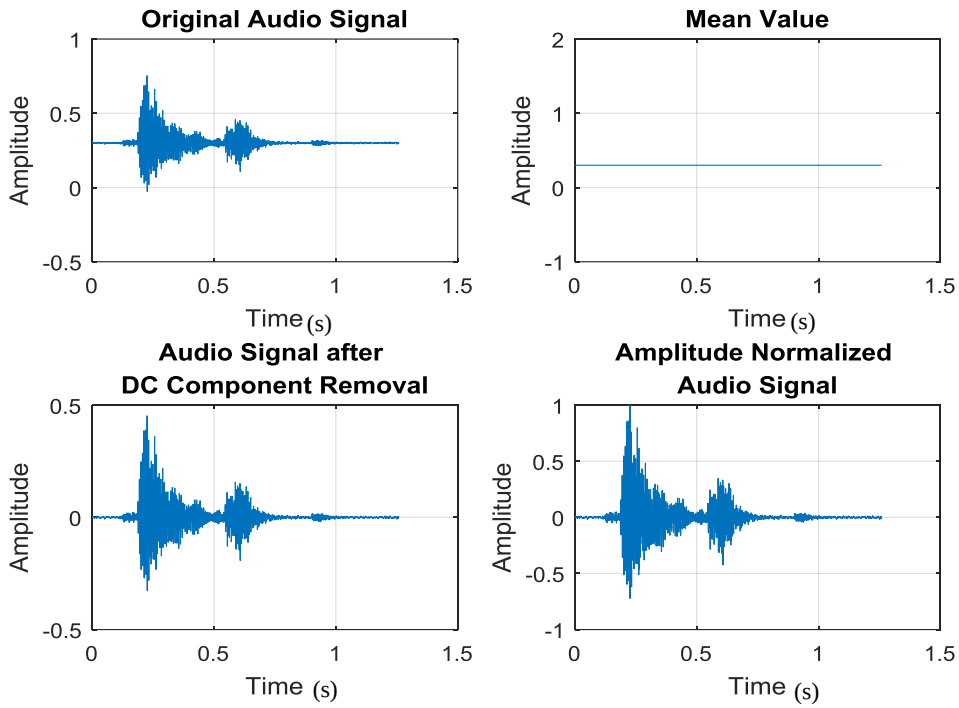


Figure 3-1. Original Audio Signal (top-left), Mean Value (top-right), Audio Signal after DC Component Removal (bottom-left) and after Amplitude Normalization (bottom-right)

3.2.2 Amplitude Normalization

The maximum amplitudes of recorded audio signals are a function of the standard environmental conditions, the size and shape of the room, the distance of the recording object from the mouth of the speakers, the type of equipment used for recording and other external factors [108]. Recorded audio signals can, therefore, have varying amplitudes based on these factors. One way to standardise the amplitudes of the audio

signals is by amplitude normalisation achieved by dividing their values by the absolute maximum amplitude in each signal as illustrated in (3) below.

$$s_N[n] = \frac{s'[n]}{\max(|s'[n]|)} \quad for\ n = 0,1,2,3,\dots,N-1 \tag{3}$$

where $s_N[n]$ is the normalized audio signal. This point-by-point division of the signal by its absolute maximum amplitude thereby constraining the dynamic range of the signal between -1.0 and +1.0. This will eliminate the effect of varying energy range. Another state-of-the-art normalization method is based on dividing each instantaneous signal value by the variance of the audio signal [109]. However, this method does not constrain the dynamic range of the normalized signal [109]. The effect of amplitude normalization in an audio signal is illustrated in the bottom-right plot in Figure 3-1.

3.2.3 Noise Reduction

Sometimes, the audio recordings may be corrupted by noise [110], due to how the audio recordings were taken, where the recordings were taken and the type of equipment used. It is therefore important to remove or reduce the effects of these noise components from audio signals before analysing the signals in order to:

- Improve the perceptual quality of the distorted speech
- Improve objective intelligibility and Speech to Noise Ratio (SNR) of the signal
- Enhance the robustness of feature extraction and speech processing applications

Research has shown that there are different filters used in removing noise from audio signals [111]. One of the commonly used time domain-based noise filters in speech processing is the Wiener filter [112, 113]. The Wiener filter, developed in the 1940s by Norbert Wiener, was one of the first applications of stochastic signal models in optimizing filters based on prior knowledge of the signal [114]. It is assumed that the signals can be modelled using stochastic processes that are stationary with known power spectral density. To reduce the computational complexity, the FIR linear discrete time filter is used. The initial (silent/non-speech) part of the signal is modelled to get the noise and the filter is applied to the speech part of the signal. The Wiener filter improves the SNR and the Mean Square Error (MSE) [113], with a trade-off in

speech distortion [111]. Figure 3-2 illustrates the effects of the Wiener filter on an audio signal corrupted by noise. The right plots show a cleaner waveform and spectrum. The noise reduction/removal technique is also useful in audio signal enhancement [115].

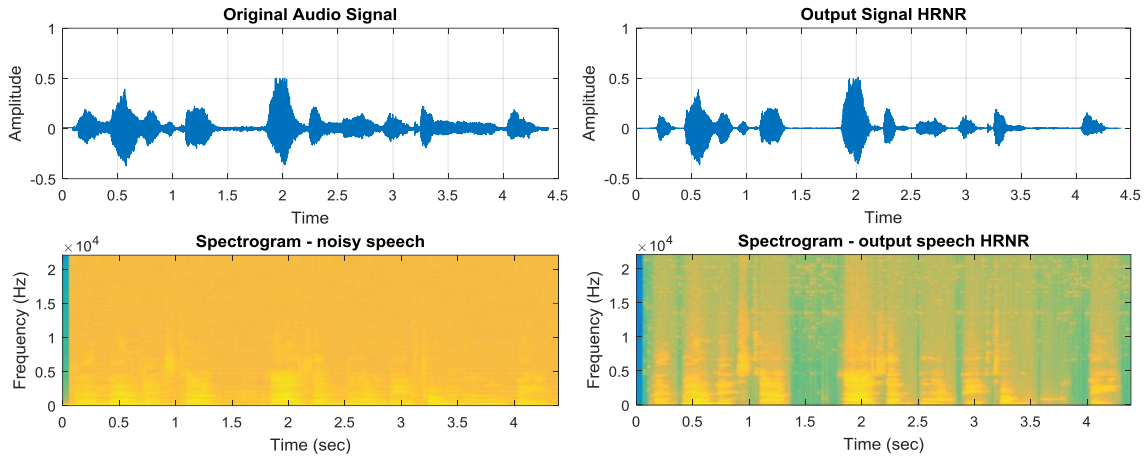


Figure 3-2. Corrupted Audio Signal before and after noise reduction using Wiener filter

3.2.4 Pre-Emphasis Filtering

In speech processing applications, the pre-emphasis filter is useful in flattening of the dynamic range of the audio signal's power spectrum [109]. Pre-emphasis filtering is also applied to compensate for the effect of the suppressed high-frequency components during speech production [116]. This involves increasing the amplitude of high-frequency bands and reducing the amplitude of low-frequency bands thereby flattening the spectral tilt. The most often used pre-emphasis filter is given by

$$H_p(z) = 1 - az^{-1} \qquad (4)$$
$$0.9 \leq a \leq 1,$$

where a is the pre-emphasis filter co-efficient that can be tuned to adjust the filtering effect. This filter is a high pass FIR filter whose magnitude response is shown in Figure 3-3 when $a = 0.97$.

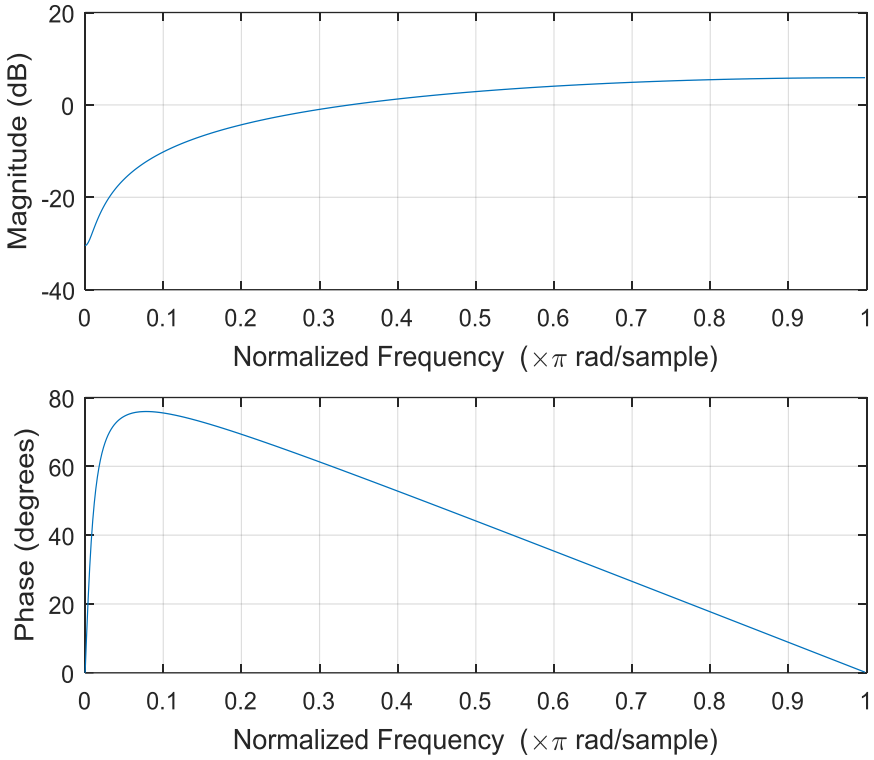


Figure 3-3. Magnitude and Phase Response of a Pre-emphasis Filter; a=0.97

3.2.5 Resampling

Audio recordings taken at different times and using different devices can be sampled at different frequencies. There is, however, a need to have all the audio signals used for the same application sampled at the same frequency, which can affect the perceptual quality of the signal. To adequately sample a speech signal, without aliasing, the sampling frequency should be more than double of the maximum frequency of the signal based on the Nyquist sampling criterion [117] given by

$$f_s > Nq = 2f_{max} \quad (5),$$

where f_s is the sampling frequency, Nq is the Nyquist frequency and f_{max} is the maximum frequency of the signal. Figure 3-4 shows the periodogram of four audio signals sampled at 44,100Hz. This one-sided magnitude response shows frequency components between the ranges of 0 Hz to 22,050 Hz. Human speech has frequency components in the audio frequency range between 20 Hz and 20 kHz in the frequency spectrum [111]. This means that the Nyquist frequency when considering the whole frequency range is 40 kHz. However, there is a need to consider the range of frequencies where the majority of the speech information resides. This can be done by

analysing different audio signals in the frequency domain. As illustrated in Figure 3-4, it can be seen that most of the frequency components (or information) in the audio signals are located between the 0-8000Hz range. Therefore, a sampling frequency of 16 kHz has been adopted in this research.

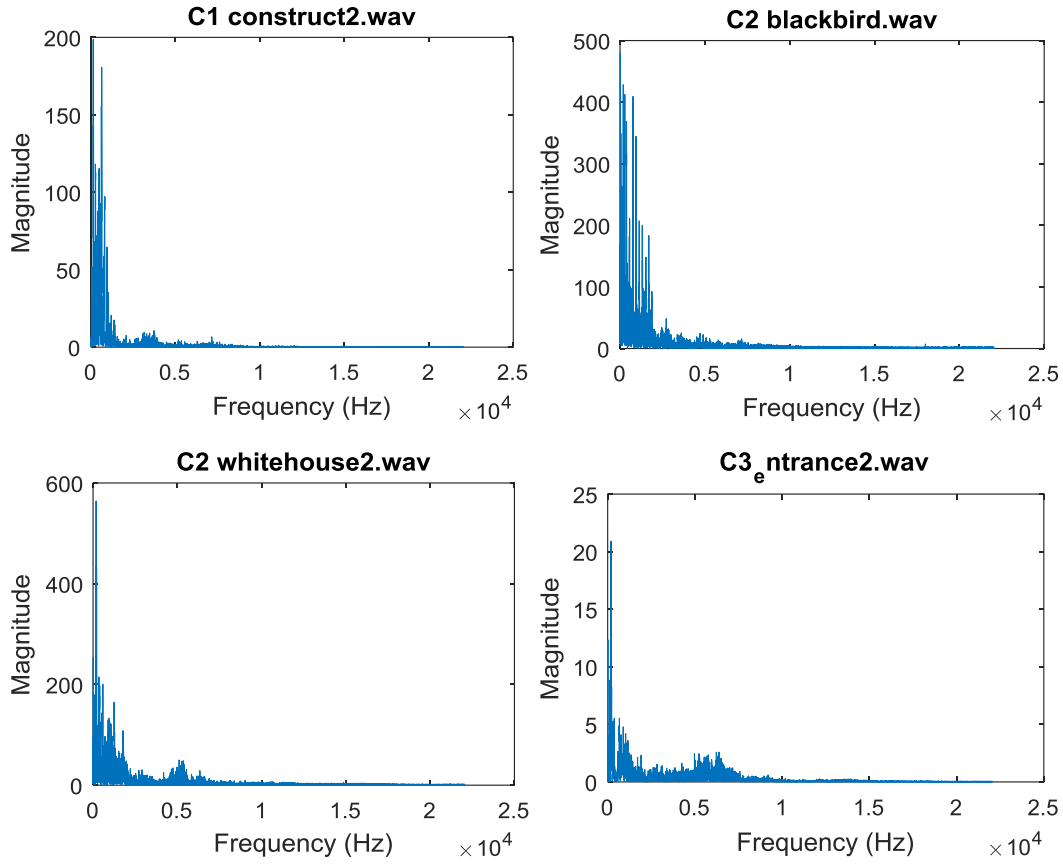


Figure 3-4. Periodogram of 4 Audio Signals Sampled at 44,100 Hz

3.2.6 Frame Blocking and Windowing in Speech Processing

Although waveforms can be used to describe some phonetic information in a speech signal, it is not useful in illustrating time-varying properties of the signal [108], which include phonetic, perceptual, voice quality and frequency characteristics of the speech signal [118]. Moreover, speech signals are non-linear and have varying frequency characteristics. It is, therefore, useful to analyse the speech signal in small time segments. Each frame can be analysed independently as a linear signal. The process of breaking down the speech signal into smaller overlapping segments of speech is called framing. The choice of frame size is a function of the intended application. Smaller frame sizes give better time resolution but poor frequency resolution [111]. This is

called Wideband analysis. Narrowband analysis, on the other hand, uses bigger frame size resulting in better frequency resolution but poor time resolution.

For speech processing applications, there is a need to consider the range of the fundamental frequencies of human speech. Humans speak at varying fundamental frequencies depending on their age, gender and voice quality [111]. The fundamental frequency of a typical adult female is between 165 and 255 Hz and that of a typical adult male is between 85 and 155 Hz [119]. The minimum fundamental frequency across children and adults is 50Hz. Consequently, at least 20ms frame size (that is, 1/50Hz) is needed for a good frequency resolution. A 20ms segment of a speech signal sampled at 16 kHz is equivalent to 320 samples. These 20ms frames are applied every 5ms (75% overlap). This reduces the effect of discontinuity in the analysis between consecutive frames. The frame duration, frame periods and frame feature vectors are illustrated in Figure 3-5. Speech features are extracted from each speech frame as illustrated.

Windowing is also an important process in speech processing. It involves the application of window filter to each speech segment (after framing) before processing. Windows are used to reduce the effect of spectral leakage and scallop losses in the audio signal to be processed [118]. Windowing also helps to reduce the effect of discontinuities in consecutive frames by tapping the edges of each frame [118]. The most commonly used window in speech processing is the Hamming window [111].

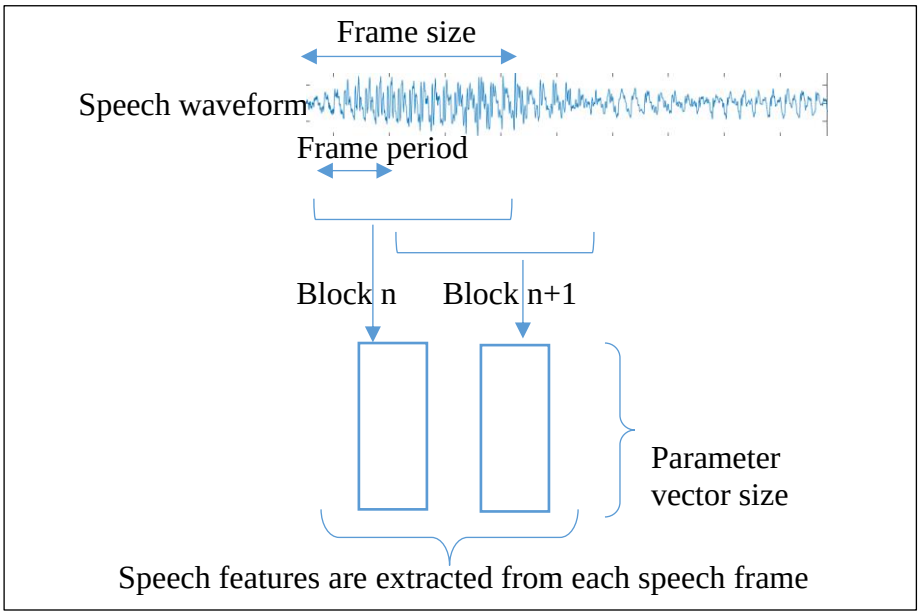


Figure 3-5. Framing in Speech Processing

3.3 Time-domain Feature Extraction

Time-domain features used in speech processing include duration [120], intensity [22], short-time energy [121], zero-crossing rate [121] and speech rate [122]. Their applications in dysarthric speech assessment and classification are also presented in this section.

3.3.1 Short-Time Energy

Short-time energy (STE) is the measure of the total energy of the speech signal in a short-time speech segment [25]. Speech signals are time-varying and nonstationary in nature and thus the energy associated with speech signals is also time-varying [123]. It is therefore important to know how the energy of a speech signal varies from a short-time segment to another. This results in STE measurements. The STE in a speech segment is described as the summation of the square of the signal amplitude given by

$$E_m = \sum_{m=0}^{F-1} x^2[m] \quad (6),$$

where F is the length of the frame and $x[m]$ is a short-time speech segment of $s_N[n]$.

In speech processing applications, the STE has been used in estimating the loudness or intensity of a speech segment [25]. High STE in speech segments indicates high loudness and vice versa. Likewise, STE has been applied in the segmentation of speech signals into silence, voiced and unvoiced segments [124, 125]. Silent parts of the speech signals will normally have the least STE when compared with voiced and unvoiced segments. Another application of STE is the detection of the start and endpoints of an utterance. In a study by Enqing et al [123], STE was used to detect voice activity in speech signals, although the study also shows that STE is very prone to noise in the environment. Combining the STE with adaptive noise reduction techniques, however, increases the robustness of the analysis [123].

Moreover, STE is useful in other speech processing techniques such as speech recognition [124], speech recovery [126], speaker recognition [127] and blind speech separation [128]. Most of these applications have been targeted toward healthy speech since disordered speech signals have very high variability in STE which can introduce

artefacts to the analysis as shown in Figure 3-6. There is a need to further investigate these variabilities for dysarthria detection application.

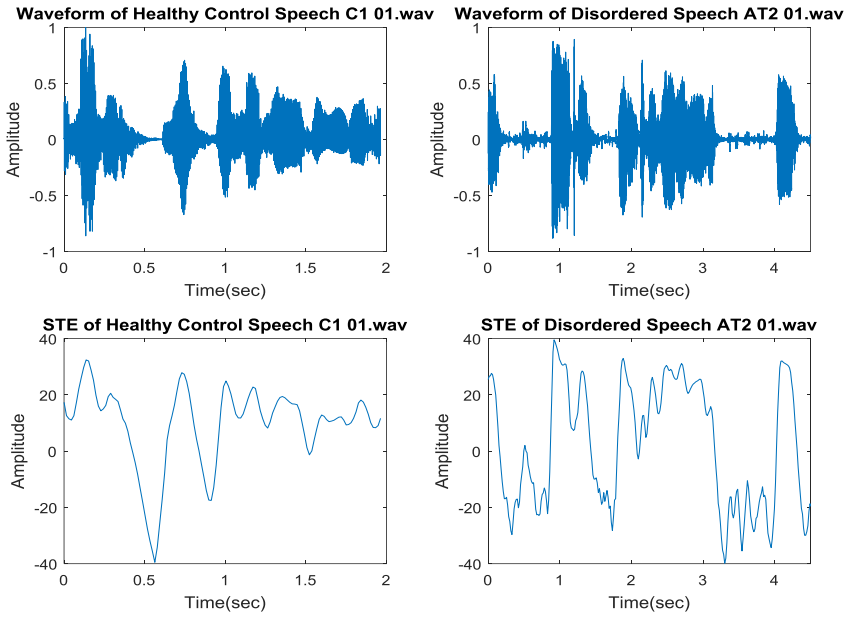


Figure 3-6. Waveform and Short Time Energy of Two Audio Signals from Healthy Controlled (left) and Disordered (right) Speakers

The differences between the STE of healthy control speech and dysarthric (AD) speech are graphically illustrated in Figure 3-6. The top plots are the waveform of the recorded audio signals when the speakers uttered the sentence "The model wrote her memoirs in Lima." The lower plots show the corresponding STE of the two signals. It is expected that the STE of the audio signals will vary across the sentence, however, the STE of the disordered speech showed higher variability. The dysarthric speaker could not sustain high intensity for a long period of time causing dips and breaks between syllables and extended pauses between words. These observations make the STE a useful feature in differentiating dysarthric speech from healthy speech.

3.3.2 Zero-Crossing Rate

Zero crossing rate (ZCR) is another important time-domain feature used in speech processing applications. The ZCR is the measure of the number of times the speech signal crosses the zero amplitude line [25]. The ZCR is estimated across all frames and the ZCR for the nth frame of an audio sample is given as:

$$ZCR(n) = \sum_{m=2}^{M} |sgn(x_l[m]) - sgn(x_l[m-1])| \quad (7)$$

$$l = 1, 2, 3, 4, \dots, F$$

$$sgn(x_{n,m}) = 1 \ \ when\ x_l[m] \geq 0$$

$$= 0 \ \ when\ x_l[m] < 0$$

where F is the number of frames in the speech sample, M is the length of each frame and $x_l[m]$ is the lth short-time speech segment of $s_N[n]$. The ZCR can be measured as the number of zero crossings per frame or the number of zero crossings per second, derived by dividing the number of zero crossings in a frame by the length of the frame in seconds.

In speech processing applications, the ZCR is often used in the classification of a speech segment into silence, unvoiced and voiced segments [25, 129]. The model of speech suggests that there a relationship between zero crossing rate and energy distribution across frequencies; high frequency components will have high number of zero crossings and low frequency components will have low number of zero crossings [130]. Silent segments are usually characterised by negligible ZCR. Unvoiced speech segments are characterised by high ZCR because most of the energy of the unvoiced segments are found in high frequencies (by extension high ZCR) whereas voiced speech segments are characterised by low ZCR because most of the energy of the voiced segments are found in low frequencies (by extension low ZCR). Over the years, the ZCR segmentation threshold has been a major topic for discussion among researchers. Most researchers have proposed the use of a fixed segmentation threshold [129, 131] which often results in misclassifications between voiced and unvoiced segments. The study by Jalil et al [25] also shows that the ZCR varies from speaker to speaker and is a function of their gender. Female speakers tend to have higher ZCR than male speakers because they (female speakers) have higher fundamental frequency than male speakers. Consequently, a fixed gender-independent ZCR threshold, in itself, is not sufficient to adequately classify a speech segment into silence, unvoiced and voiced segments. The use of a speaker-dependent segmentation threshold will be explored in this research.

Apart from the segmentation of speech into silence, unvoiced and voiced segments, the ZCR can also be used in differentiating between healthy control speech and

dysarthric speech. Due to the poor voice quality and high variability (in intensity and fundamental frequency) experienced in dysarthric speech samples, it is expected that their ZCR will be higher than those observed in healthy control speakers. An example is seen in Figure 3-7 where the range of the zero-crossing of a healthy control speech signal (bottom-left) was between 30 and 120 per frame whereas that of disordered speech (bottom-right) was between 40 and 720 per frame (with a frame size of 50ms). The differences observed in the ZCRs of healthy and dysarthric speech are due to increase in the energy concentration at high frequencies in dysarthric speech which also affects the voice quality. These differences show that ZCR can be useful in disordered speech classification.

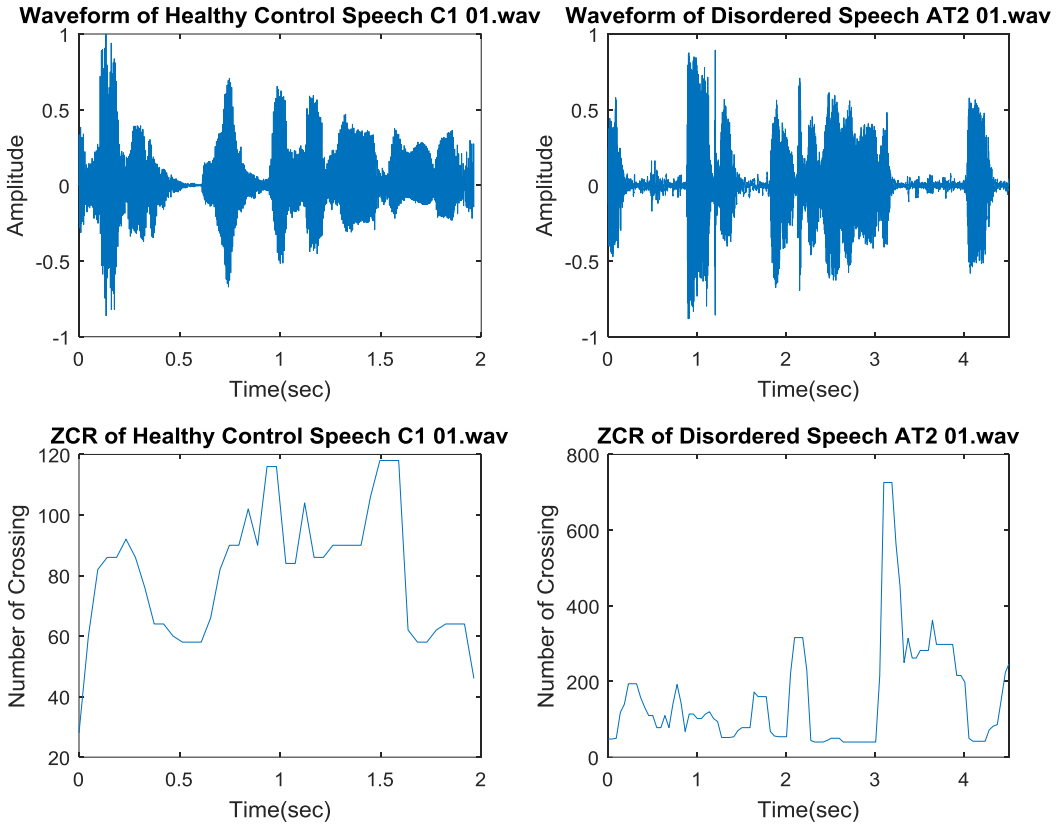


Figure 3-7. Waveform and Zero Crossing Rate of Two Audio Signals from Healthy Controlled (left) and Disordered (right) Speakers

3.3.3 Duration-related Features

The duration of a speech signal is a measure of how long the speech signal is in seconds. In speech processing applications, the duration of an utterance can often be used to describe the prosodic and stress characteristics of the utterance [132]. Long

duration is often used to show an emphasis on a particular segment/word in an utterance with respect to the other parts of the utterance [32].

Pauses are also duration-related features that are often used to separate words or sentences, to indicate breaks in utterances and sometimes to mark stress [133]. Pauses are measured in seconds or milliseconds as the period of silence between speech segments. The application of pauses in word/syllable segmentation, speech breaks quantification and stress marking will be explored in this research.

In addition, the speech rate can be measured as the number of syllables produced by a speaker per second [18]. Speech rate should not be confused with word rate which is the number of words produced per unit second [134]. The word rate is partially a function of the sentence composition whereas the speech rate is a function of how fast or slow a speaker speaks. As discussed in Section 2.6.1, one of the strategies researchers have used in the treatment of dysarthria is targeted towards improved speech rate. The application of speech rate will be explored in the development of a novel DDK analysis tool presented in Chapter 5.

Extended features derived from the time-domain features described above are used in various speech processing applications discussed in Section 3.5.

3.4 Spectral and Cepstral Features Extraction

While the time domain representation shows how a signal varies with time, the frequency domain representation shows the spectrum of the power distribution of the signal over a range of frequencies. The frequency-domain representation, called the spectrum contains information about the frequency distribution and phase shift required to reconstruct the original signal in the time domain [135]. A signal can be converted from the time domain to the frequency domain (and vice versa) using a pair of equations called transforms [111, 135]. A good example of the time-frequency transform is the Fourier transform. The Fourier transform (FT) convert a time domain signal into frequency domain representation of the signal [111, 136]. Inverse Fourier transform (IFT), on the other hand, converts the frequency-domain function to the time-domain representation of the signal. The cepstrum is, however, derived when the IFT of the logarithm of the Fourier transform of a signal is taken [111]. (Cepstrum is derived from the word “spectrum” by reversing spec to form ceps) Just as in the case

of the frequency spectrum, there exists power cepstrum, real cepstrum, complex cepstrum, and phase cepstrum [111].

In this section, various spectral and cepstral features used in speech processing applications, especially in the analysis of disordered speech will be discussed. Their applications in dysarthria detection and classification will also be reviewed.

3.4.1 Fundamental Frequency

The fundamental frequency (F0) of a speech signal is defined as the lowest spectral component which corresponds to the natural frequency at which the vocal cords vibrate (open and closes) during speech production [137]. The F0 of a signal can be measured by estimating the separation in pulses in the time domain called the pitch period [111]. The F0 in the voiced segments of a speech signal depends on two factors; the variation in the length of the vocal cords and how the aerodynamic factors adjust to suit the vibration in the vocal cords [111]. When the vocal cords are short and thick, voiced sounds with low F0 are produced whereas when the vocal cords are long and thin, voiced sounds with high F0 are produced [111]. However, during speech production, the glottis vibrates resulting in less periodic signal [138]. These glottal vibrations result in variation in amplitudes, speech rate and waveform shape [139]. Due to these variations, speech signals are not perfectly periodic which makes the F0 estimation a challenging problem [138], although, variations in F0 are useful in prosody and lexical differentiation in tonal languages [138].

The range of F0 for an individual is also a function of their gender or age [111]. As mentioned in Section 3.2.6, the F0 for a typical male adult ranges from 85Hz to 180Hz whereas that of typical female ranges between 165Hz and 255Hz. Children have the highest range of F0, between 250Hz and 500Hz [111].

Over the years, researchers have used different techniques in tackling the problem of F0 estimation in speech signals which include the autocorrelation-based pitch detection method [139, 140], cepstral pitch detection algorithm [139], average magnitude difference function [109, 139], robust algorithm for pitch tracking [141] and YIN fundamental frequency estimator [138]. In this section, the various F0 estimation techniques will be reviewed while highlighting their limitations.

A. Autocorrelation-based Pitch Extraction Algorithm

The autocorrelation method is one of the frequently used pitch detection techniques. It was first proposed by Sondhi in a study where he presented three methods for pitch extraction, two of which were based on the autocorrelation function [142]. These methods include pitch extraction by minimum phase compensation, pitch extraction by autocorrelation of spectrum flattened speech and pitch extraction by centre clipping and autocorrelation.

In general, the autocorrelation function (ACF) of a periodic signal is also periodic where the period of the ACF is defined by the period of the signal [111, 140]. To calculate the pitch period in a short-time speech segment, the ACF of the segment is derived and the period of the ACF is estimated. The F0 is then estimated by taking the inverse of the pitch period. The autocorrelation function of a signal is defined by

$$ACF(\tau) = r_l(\tau) = \sum_{i=1}^{F} x(i)x(i+\tau) \quad (8),$$

where $ACF(\tau)$ is the autocorrelation function of lag τ of windowed signal $x(i)$ where F is the frame (window) size.

The major sdrawback of the autocorrelation-based pitch extraction algorithm is the presence of false peaks due to harmonics. To overcome this limitation, researchers [137, 140] have introduced the use of a low pass filter before the application of the autocorrelation function to the speech segment. A low pass filter will remove high-frequency components from the speech signal which will, in turn, reduce the effects of high-frequency formants. The use of low pass filter does not, however, remove the low-frequency formants [140] (especially the first formant).

A wide variety of methods have been proposed to remove/reduce the effects of the low-frequency formants in the extraction of the F0 in a speech signal. These include spectral flattening techniques such as centre clipping [142], peak and centre clipping [140], inverse filtering by linear prediction, and spectral flattening by linear prediction [140, 143]. These methods, however, do not totally eliminate the problem of false peaks introduced by the formants.

The autocorrelation method has its own challenges due to the interference of low-frequency formants. Another pitch extraction technique was developed based on the cepstral analysis. This is discussed in the subsection below.

B. Cepstral-based Pitch Extraction Algorithm

The theory of cepstral analysis is based on the fact that the Fourier transform of a speech signal can result in peaks with regular spacing which are the spectral harmonics of the signal [139]. Taking the logarithm of the spectrum reduces and re-scales the amplitudes of these peaks. This also results in a waveform that is periodic in nature whose period is related to the pitch period of the original signal [139].

Unlike the autocorrelation-based technique, the cepstrum method does not make use of a low pass filter [109]. Audio signals are first divided into frames of 20ms and 5ms overlaps as discussed in Section 3.2.6. A Hamming window is then applied to the framed section. Cepstrum analysis is performed on the windowed signal after which the peak value is detected between 50Hz and 500Hz range. This peak value is used to determine the pitch period along the time axis. It is important to point out that the cepstral analysis was designed based on the assumption that the signal to be analysed has regularly spaced spectral harmonics [139]. Any deviation from this assumption results in errors in the pitch extraction algorithm [139].

C. Average Magnitude Difference Function (AMDF) Pitch Extraction Algorithm

The AMDF is also widely used for extraction of pitch in speech recognition and speech processing applications. This method is a variation of the autocorrelation method discussed above. AMDF gives a better pitch measurement resolution than the conventional autocorrelation method [109]. The AMDF-based pitch extraction algorithm is faster than the autocorrelation method in terms of computation speed and cost [141]. Similar to the autocorrelation method, the audio signal is first passed through a low-pass filter of cut-off frequency 900Hz after which a Hamming window is applied. The average magnitude difference function is calculated as

$$AMDF(\tau) = \frac{1}{N}\sum_{i=1}^{F}|x(i) - x(i-\tau)| \quad (9),$$

where x(i) is the windowed speech signal and F is the window length. The pitch period is estimated by computing the minimum value of the AMDF between the pitch period

range of 20ms (50Hz) and 2ms (500Hz). Although there is an improved measurement resolution, the AMDF pitch extraction method also suffers from the false peak problems introduced by the spectral harmonics as experienced in the autocorrelation-based method [141]. To overcome these challenges, a cross-correlation based method called the Robust Algorithm for Pitch Tracking (RAPT) was proposed by Talkin [141].

D. Robust Algorithm for Pitch Tracking (RAPT)

The RAPT was designed to be robust to noise and speaker group (male or female) while maintaining the pitch tracking accuracy [141]. Pitch-tracking technique is based on the normalised cross-correlation function (NCCF). After the peak values of the NCCF are estimated, dynamic programming is carried out to select the best F0 candidates in each frame [109]. The selection is based on both the local and the global (contextual) evidence. This reduces the drastic jumps and discontinuities in F0 estimation. A very similar pitch extraction method is called the Integrated Pitch Tracker based on the NCCF of the linear prediction residue of the signal rather than the NCCF of the signal [139].

E. YIN Fundamental Frequency Estimator

The YIN fundamental frequency estimator was developed by two researchers in the early 21st century [138]. It was named after the yin-yang philosophy of balance in an attempt to strike a relationship between autocorrelation and cancellation in the estimator. The YIN estimator was developed to solve the problem of sub-harmonic peaks introduced by the autocorrelation technique. The YIN estimator is based on a function that minimizes the difference between a discrete signal and its delayed replica. This function is called the difference function which is represented by (10).

$$d_n(\tau) = \sum_{i=1}^{F} (x(i) - x(i-\tau))^2 \tag{10}$$

where $d_n(\tau)$ is the difference function of lag τ of windowed signal x(i) where W is the window size. The YIN estimator also makes use of the cumulative mean function to de-accentuate high-period dips in (10). This helps to reduce the errors due to subharmonic peaks. The YIN estimator also incorporates the use of parabolic interpolation to further reduce the effect of the estimation error.

The performance of the five F0/ pitch extraction methods discussed above is examined and illustrated in Figure 3-8 and Figure 3-9. In Figure 3-8, two speakers, one healthy control speaker and one dysarthric speaker were given a sentence to read out. The texts in the sentence are "The model wrote her memoirs in Lima". The utterances from the two speakers were recorded and analysed. Four of the five F0 estimation techniques outperformed the RAPT technique. The estimation errors recorded in the RAPT method were more pronounced as the pitch tracker performed poorly within individual words and across words. The profiles from the autocorrelation method, cepstral, AMDF and YIN methods show that the peak pitch was clearly marked as well as the variation across the sentence. In terms of consistency, the cepstral and the AMDF methods gave a good tracking consistency for healthy control speech. Most pitch tracking techniques perform well when analysing healthy speech, however, that was not the case when analysing disordered speech. For example, the cepstral method showed a good performance in tracking the F0 in the healthy control speech but gave a poor performance in tracking the F0 in disordered speech resulting in pitch doubling. The autocorrelation, AMDF and YIN methods performed relatively better in tracking the F0 in the disordered speech.

As illustrated in Figure 3-9, these techniques were also used to estimate the F0 in a single word (construct) from a healthy speaker and a dysarthric speaker. The F0 of the healthy control speaker is shown on the left and that of the dysarthric speaker is shown on the right. As discussed earlier, the RAPT technique performed poorly both in healthy control speech and dysarthric speech samples. For, the first part of the word ("con") in the healthy speech sample, the other four techniques performed well. However, for the middle part of the word (*"-st-"*), the cepstral method resulted in false peaks. Also, in the last part of the word (*"-ruct"*), the autocorrelation method resulted in pitch-halving, the Cepstral method resulted in pitch doubling and the AMDF method performs poorly.

Moreover, using the RAPT method in the estimation of F0 in dysarthric speech sample resulted in poor performance with a lot of errors. The Cepstral method resulted in false peaks for consonant sounds in the words. The autocorrelation method also resulted in a few false peaks but the AMDF and the YIN methods gave good performance with very few estimation errors. The analysis of these results shows that the AMDF and the Yin techniques gave good and comparable results with low estimation errors for both

healthy control and dysarthric speech samples. It is important to point out that out of the five techniques and across the different speaker groups, the YIN pitch estimation technique outperformed the other techniques for both single-word and sentence samples and will be used for F0/pitch extraction in this research.

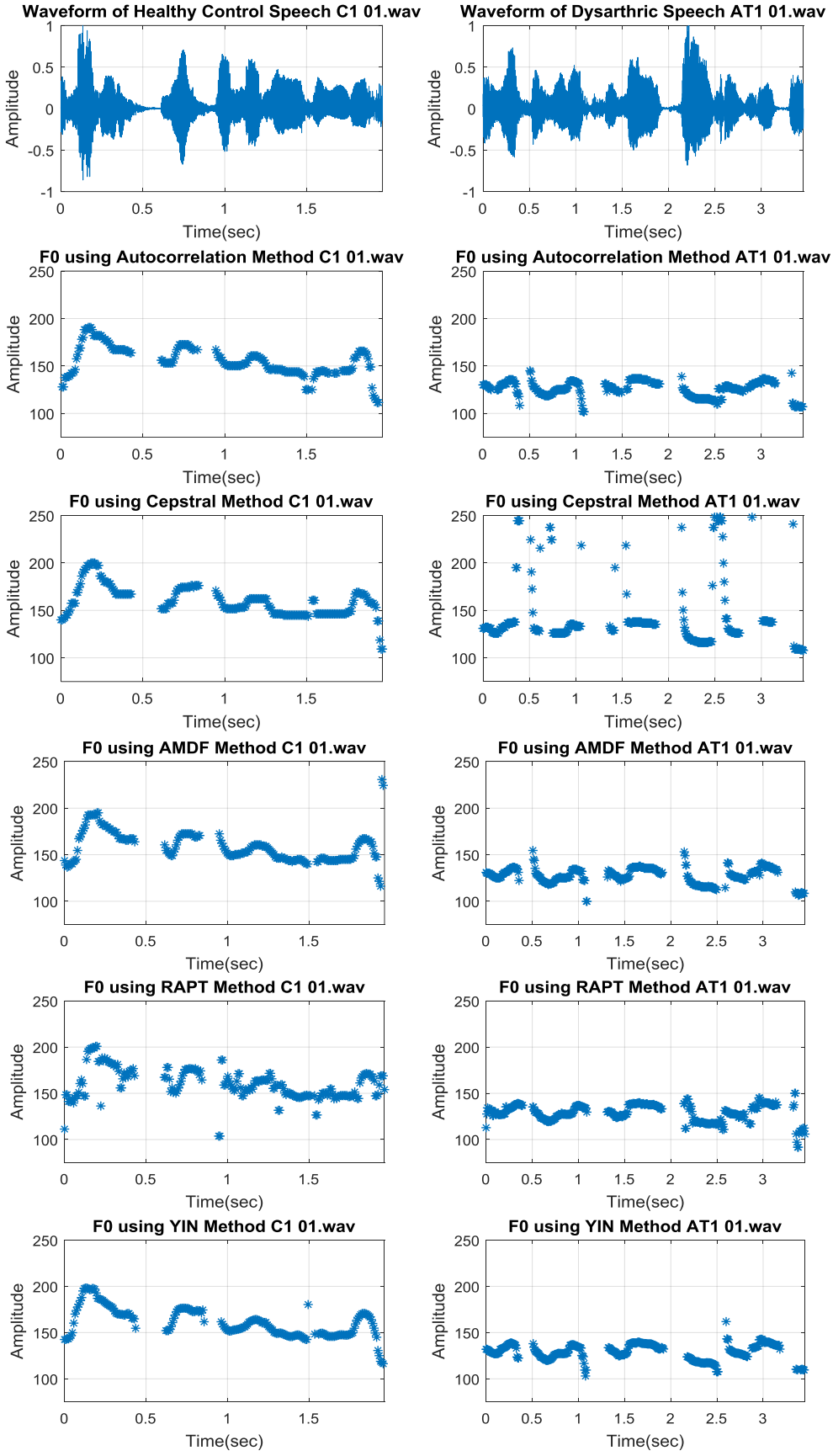


Figure 3-8. Comparison of five pitch detection techniques and their performance in **sentences** produced by healthy (left) and dysarthric (right) speakers

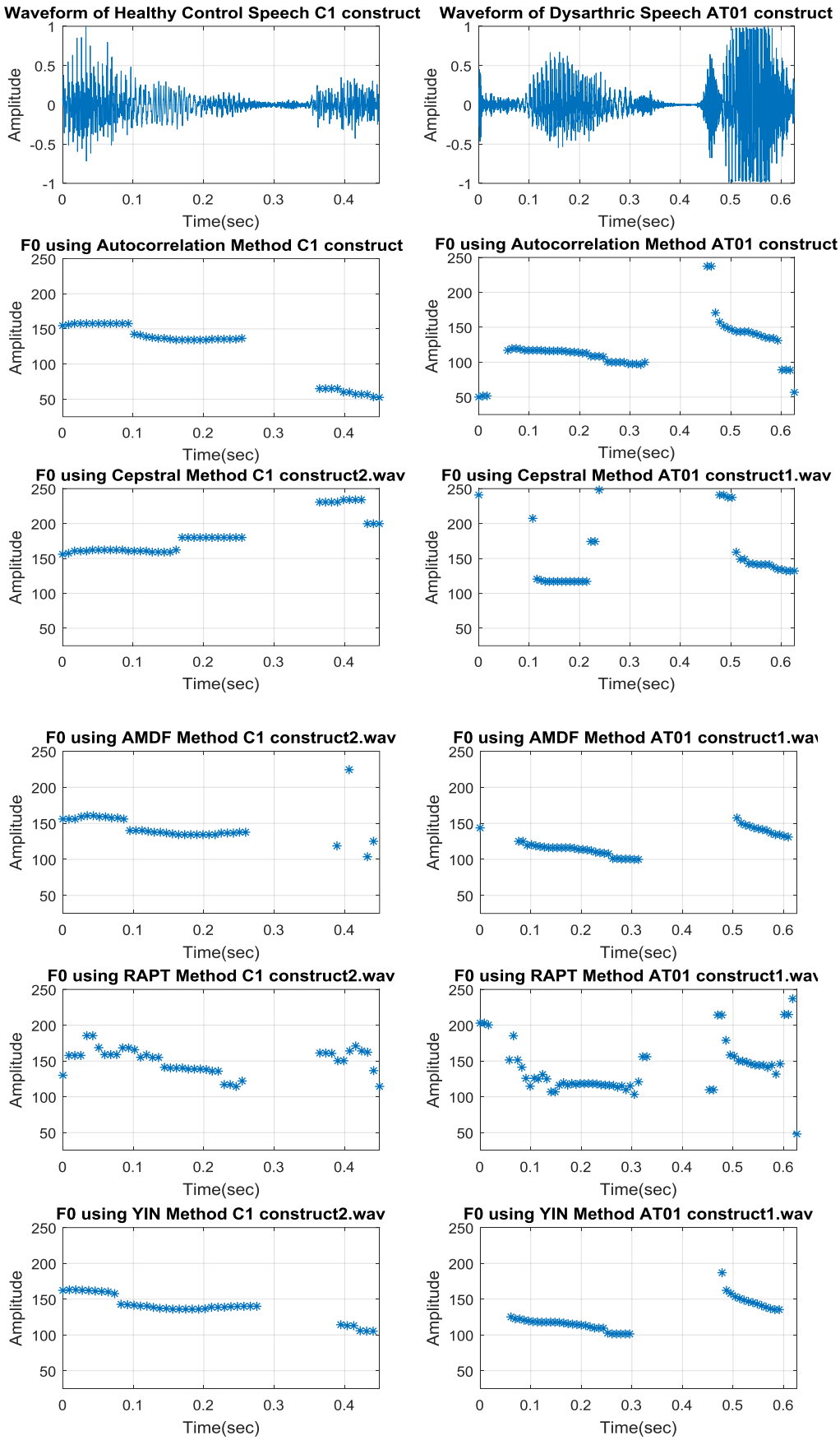


Figure 3-9. Comparison of five pitch detection techniques and their performance in **words** produced by healthy control (left) and dysarthric (right) speakers

3.4.2 Linear Prediction Coefficients

Linear Prediction Coding, also known as LPC, is a spectral analysis technique used for encoding a signal in a way that the current value of a signal is taken as a linear function of the previous values [144]. The coefficients generated after LPC are called the linear prediction coefficients. The LPC analysis assumes that the human vocal tract can be modelled as a tube with varying diameter. This results in a mathematical model which gives an approximation of the human vocal tract response [145]. This linear prediction error (P_e) is given by

$$P_e = E\{e^2[n]\} = E\left\{\left(x[n] - \sum_{k=1}^{N} a_k x[n-k]\right)^2\right\} \quad (11),$$

where a_k is the kth LPC coefficient, N is the order of the linear prediction and $x[n]$ is the speech signal. The speech sample x[n] is represented as a weighted linear sum of N previous samples; given that N is the order of the LPC estimation. This results in a prediction system where the next sample is predicted by the sum of N preceding samples. The resulting coefficients of the LPC are used in estimating the formants; the frequency characteristics of a speech signal over time. Formants are frequencies within the speech spectrum where acoustic energy are concentrated [146]. For example, a speech signal sampled at 8kHz and encoded at 8 bit per sample will have a bit rate of 64kbits/sec. However, performing a linear prediction will reduce the rate to 24kbits/sec.

Furthermore, research [144, 147] has shown that even though the bit rate is reduced during linear prediction coding, the estimated speech signal remains audible and comprehensible. Due to these attributes, the LPC is useful in speaker identification and also in speech coders with low or medium bit rate [148]. The LPC also offers a robust and reliable way of estimating the main frequency components of speech signals (formants) [148]. Nevertheless, the LPC analysis gives a poor performance in the detection of emotion due to reduced speech quality [144]. There is usually a trade-off between emotion prediction optimisation and speech quality.

In addition, the LPC analysis will be useful in considering the frequency characteristics of the dysarthria speech. The accuracy of the LPC-based technique in formants estimation is high [9] compared to other feature extraction techniques [149]. The LPC is robust to noise unlike the MFCC feature extraction technique [145]. Also, the

formants estimation is used as a tool for measuring the intelligibility and pronunciation features in spoken language [62] which will be explored in this research.

3.4.3 Formants

Formants are resonance in the frequency spectrum of a speech signal [98]. There are two main methods of extracting formants from a speech signal. The first method involves LPC root solving and the second uses an adaptive bandpass filter to estimate the formants [98]. The latter is robust to speaker type and non-stationary background noise, however, requires high computational cost [109].

The LPC formant extraction is based on the energy distribution of the signal in the frequency domain also called the power spectral density. The formants positions are chosen in such a way that they match this distribution of energy. These formants are frequencies with bandwidths of less than 400Hz. Therefore, frequency bands with a high concentration of energy and bandwidths less than 400Hz are located as the formants of the speech signal. Using LPC analysis, the order of the linear prediction is a function of the sampling frequency of the speech signal given by the rule of thumb given by

$$P = 2 + \frac{Fs}{1000} \quad (12),$$

where P is the order of the LPC and Fs is the sampling frequency [111]. This is because if the number of poles does not match the number of resonances present in the signal, the model spectrum will lead to errors as poles will be placed in-between the actual formants [111, 150, 151]. The estimated LPC coefficients are converted from rectangular form to polar form and the phases of the coefficients with bandwidths less than 400Hz and positive phase are extracted as the bands of resonance of the spectrum. The extracted positive phases are called formants. In Figure 3-10, the formants of two audio signals ae extracted, one from a healthy control speaker and the other from a disordered speaker using the LPC-based formant extraction algorithm. The LPC-based formants extraction algorithm performed well in being able to track the formants in both speaker groups. The formants alone might not be sufficient to classify dysarthric speech but the combination of these formants with other features will be of interest in this research.

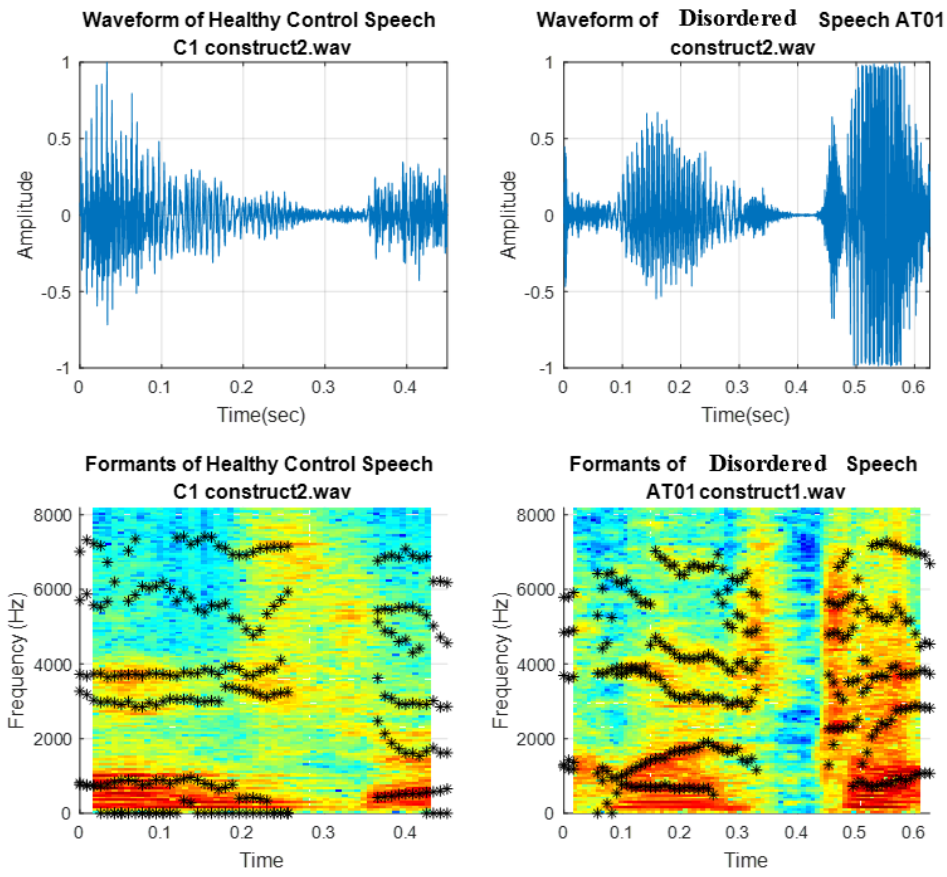


Figure 3-10. Formants extracted from Two Audio Signals from Healthy Controlled (left) and Disordered (right) Speakers

3.4.4 Mel Frequency Cepstral Coefficients (MFCC)

MFCC is one of the extensively used feature extraction techniques in speech processing. This feature extraction technique was first proposed by Davis and Mermelstein in [152] and since then many variations of the original algorithm have been developed [153]. The MFCCs are used to estimate the power spectrum of a speech signal using overlapping triangular frequency bins with equal height. Construction of M-length Mel filter bank will result in M overlapping triangular filters with equal height. The *ith* triangular filter can be represented as

$$H_i(k) = \begin{cases} 0 & for\ k < f_{b_{i-1}} \\ \dfrac{(k - f_{b_{i-1}})}{(f_{b_i} - f_{b_{i-1}})} & for\ f_{b_{i-1}} \le k \le f_{b_i} \\ \dfrac{(f_{b_{i+1}} - k)}{(f_{b_{i+1}} - f_{b_i})} & for\ f_{b_i} \le k \le f_{b_{i+1}} \\ 0 & for\ k > f_{b_{i+1}} \end{cases}, \quad i = 1,2,3,\dots,M \quad (13),$$

where b_i is the ith band, f_{b_i} is the *ith* band frequency, M is the number of filters and k is the discrete Fourier transform index.

Although MFCC has been widely used in speech processing applications, there are certain limitations to its computation and usage. Research has shown that the MFCC computations are not robust to noise especially additive noise [148]. Another major limitation of the MFCC is the fact that the performance of the algorithm is dependent on the number of filters used and also on the range of the filter banks [154]. When the number of the filters is reduced below an optimum point, the overall performance of the MFCC reduces. When the number of filters is increased above an upper optimum point, the overall performance of MFCC is negatively affected [115]. These lower and upper optimum numbers of filters are dependent on the specific application of the feature extraction technique.

Moreover, change in the Mel filter shape also affects the overall performance of the system. In a filter shape analysis experiment, research [155] showed that using a critical masking curve gives a better performance in speaker verification than the conventional triangular filter shape. The architectural complexity of MFCC algorithm is also high compared to other feature extraction techniques [108], resulting in high computational time and cost.

The MFCC algorithm has been modified in various ways to overcome some of the limitations discussed above. These modifications include cubic compression in MFCC to increase accuracy [156], a combination of principal component analysis with MFCC [157] and the fusion MFCC technique [158]. For the purpose of this research, a modified MFCC technique (with liftering) will be used to extract some articulatory and phonation speech features in conjunction with other feature extraction techniques and machine learning techniques discussed in Section 3.7 for dysarthric speech classification.

3.5 Extended Feature Extraction

Speech features derived from time-domain and spectral features are called extended features since they are derived from other speech features. Techniques used in estimating these features are discussed below with emphasis on their applications in disordered speech analysis.

3.5.1 Jitter

Jitter, also known as fundamental frequency perturbation, is a pitch-based speech feature which measures the variation in fundamental frequency in speech signals. It is calculated by estimating the short-term pitch period perturbation [109] which is given by

$$Jitter = \frac{1}{N-1}\sum_{i}^{N-1}|T_{0i} - T_{0i+1}| \quad (14),$$

where T_{0i} is the pitch period of the ith frame and N is the number of frames. As the fundamental frequency varies from frame to frame, the jitter value increases. This extended feature is very useful in quantifying pitch variability in speech.

Over the past decade, researchers have used the jitter in the measurement of speech intelligibility [159], stress and emotion classification [160], and speech encoding [111]. Another interesting application of jitter measurements is in the detection of pathological disorders in speech signals [161, 162]. For example, Silva et al [162] proposed a method for the detection of pathological voices using a jitter estimation algorithm. In this research, the use of jitter in the detection of dysarthria and its classification into various severity levels will be explored.

3.5.2 Shimmer

Shimmer, on the other hand, is the measure of variation in peak amplitudes in a speech signal between consecutive frames. Shimmer also known as the short-time amplitude perturbation is measured from cycle to cycle and it shows the transient change in energy in speech signals [109]. The shimmer of an audio signal is given by

$$Shimmer = \frac{1}{N-1}\sum_{i-1}^{N-1}|20\log(A_i/\ A_{i+1})| \quad (15)$$

measured in decibels, where A_i is the amplitude of the ith frame and N is the number of frames. As the amplitude varies from frame to frame, the shimmer value increases. This extended feature is also useful in quantifying amplitude variability in speech. Just as the fundamental frequency perturbation (jitter), shimmer has been used in various speech processing application such as in stress and emotion classification [160], speech emotion recognition [163]and pathological speech classification [161].

3.5.3 Harmonic to Noise Ratio

Harmonic to noise ratio (HNR) gives the measurement of the relative amount of noise in a signal [164]. During phonation, noise can be added to the speech signal due to the turbulent airflow through the glottis. Partially closed vocal cords can allow the passage of excessive airflow resulting in turbulence. Another source of additive noise in speech signals is due to the aperiodic vibrations of the vocal cords [164]. The HNR shows the ratio of the periodic (harmonics) to the aperiodic (noisy) components of the speech signal which is measured in decibels (dB). The value of HNR for healthy adults ranges from 11-13 dB whereas that of isolated vowels can be as low as 7.4 dB [164]. The HNR of a speech signal is defined by (16).

$$HNR = 10\ log_{10}\left(\frac{\text{Signal Energy} - \text{Noise Energy}}{\text{Noise Energy}}\right) \quad (16)$$

Since the autocorrelation of the signal at lag zero ($R_{xx}[0]$) is equal to the total signal energy and the harmonic energy can be represented by the autocorrelation at the peak period ($R_{xx}[T_0]$), the HNR can also be defined as:

$$HNR = \ 10\ log_{10}\left(\frac{R_{xx}[T_0]}{R_{xx}[0] - R_{xx}[T_0]}\right) \quad (17)$$

Furthermore, the HNR is related to the quality of the speaker's voice. Research [67, 164] has shown that the HNR can be used to determine the perceptual impression and define the physiological aspects of the voice. The HNR is, therefore, a significant feature in determining the voice quality and measuring the deviation of a speech signal from the expected quality range.

Ferrand in his study on the effects of vocal ageing on HNR showed that the HNR changes (decreases) as the speakers become older due to voice instability [164]. Another factor that impacts the variation seen in HNR is the effect of pathological influence on voice quality [67, 164]. People with pathological speech tend to have lower voice quality leading to an increase in the additive noise in the speech [164]. This invariably results in decreased HNR. HNR measures have also proven to be more sensitive to differences in voice quality than jitter measures [164]. The effects of combining the HNR and Jitter measures in the detection and classification of speech disorders have not been widely researched and will be investigated in this study.

3.5.4 Wavelets

The wavelet transform is a very useful tool in analysing non-periodic signals as well as noisy or transient signals [165] mainly because the wavelet transform analysis can examine both the time and frequency characteristics of a signal simultaneously [166, 167]. The history of the wavelet transform is dated as far back as 1909 when the mathematician Alfred Harr proposed the Harr wavelet, however, the concept of wavelet was first introduced in 1981 by Jean Morlet who is a geophysicist [168]. Since then, there have been various methods for the wavelet transform. The wavelet transform makes use of wavelets which are short finite length waveforms whose mean amplitude is zero [165]. The Discrete Wavelet Transform (DWT) is based on subband decomposition of signals which is also similar to subband coding in speech signal coding applications [168, 169].

In DWT, the signal to be analysed is decomposed by passing the signal through digital filters with specified cut-off frequencies [166]. Wavelets are generated by the process of iteration of filtering and rescaling of the signal through a pair of low pass and high pass filters as shown in

Figure 3-11 (Mallat tree) [170]. The outputs of both the high pass and low pass filters are decimated by two and that of the low pass filter is further decomposed using another layer of low pass and high pass filter [170].

Over the years, the wavelet transform has been applied in diverse fields such as condition monitoring [171], climate forecasting [172], financial analysis [173, 174] and biomedical applications [175, 176]. This has led to an increasing number of studies

and research on the use of wavelet transform and its modifications for signal analysis, manipulation and investigation.

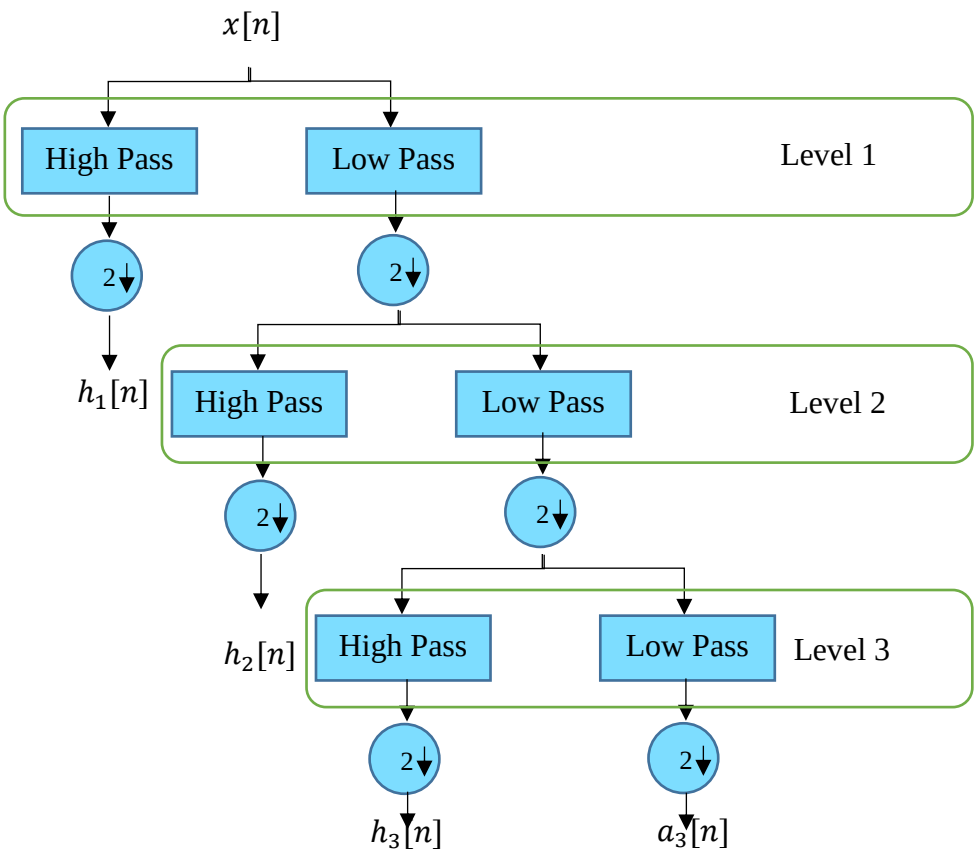


Figure 3-11. Mallat-tree Diagram showing Different Levels of DWT

Illustrated in Figure 3-12 are results of four-level wavelet analysis carried out on ataxic dysarthric and healthy control speech samples. The percentage of residual energy after the wavelet decomposition at every level in dysarthric speech samples is higher than that of healthy control speech samples. This difference, however, reduces as the level increases [166].

The wavelet transform has been used in various speech processing applications including music and speech separation [177], emotion recognition [178], automatic detection of swallowing difficulty [179] and pathological voice detection [180, 181] with promising outcomes even though its application in automatic speech disorder detection and severity classification has not been examined. This research aims to fill this gap and explore the use of wavelet transform in severity classification in dysarthria.

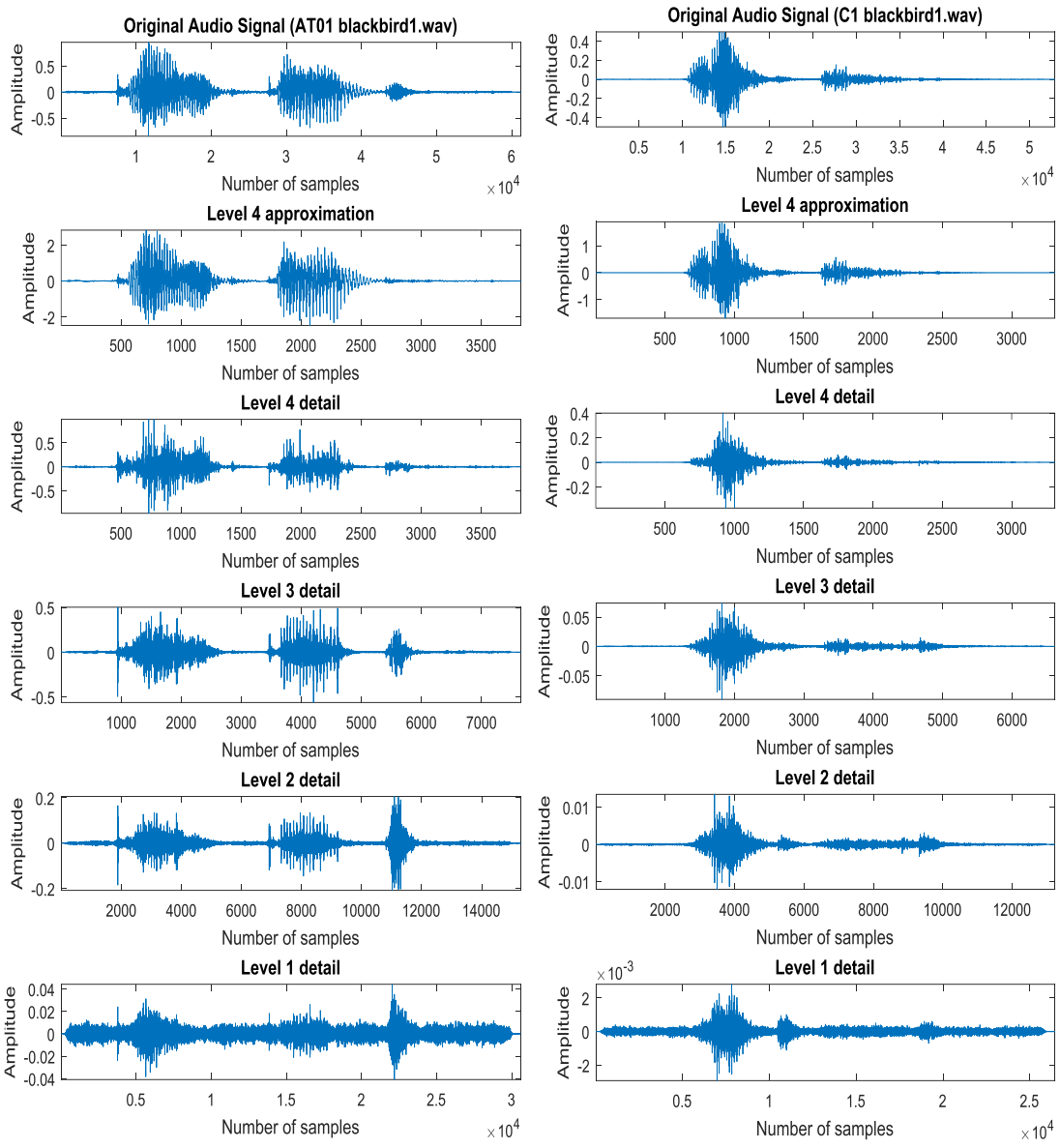


Figure 3-12. Comparison of Four-Level Wavelet Analysis of Dysarthria (left) and Healthy Controlled (right) Speech Signals

3.6 Review of Silence-Unvoicing-Voicing Segmentation Techniques

The classification of human speech into three distinct sections: silence, unvoiced and voiced (SUV) is significant in many speech processing applications which are based on whether or not the vocal cords vibrate during the production of the speech segment. The silence segment is a period within the human speech where no sound is produced which could occur at the start of statements, during pauses in between words/syllables and at the end of statements. Speech segments produced when the vibration of the vocal cords is aperiodic are called unvoiced segments. Whereas voiced segments are

produced when the vocal cords vibrate in a periodic manner. SUV segmentation is one of the prerequisites in feature extraction algorithms such as pitch detection, formant extraction, word separation, and syllable segmentation.

Being a three-class problem, the SUV segmentation is a more challenging classification problem than voiced activity detection (VAD) and voiced-unvoiced (VU) classifications consisting of two classes only. Research [182] has shown that SUV segmentation can be performed by combining both VAD and V/U segmentations. This will require prior knowledge of the noise statistics of the speech signal [182], thereby making the classification problem dependent on the accuracy of the noise statistics. The SUV segmentation is, therefore, often treated as a unique problem.

In existing and current research works, a number of SUV classification methods have been developed involving the use of unsupervised learning [182], ZCR [125] [131], pattern recognition algorithms [183], cumulants [184], autocorrelation algorithms [25], spectral parameters [185] and combinations of two or more of these methods [25, 186]. There is, however, a need to develop a method that reduces the architectural cost and also improves robustness at low signal energy (intensity) experienced in dysarthric speech. Moreover, finding fine boundaries between voiced and unvoiced segments is a major limitation of most existing methods [25].

A few researchers have also focused on the use of STE and ZCR in the SUV segmentation of speech [124, 125]. This involves setting up a segmentation rule for both STE and ZCR based on pre-defined thresholds. Voiced sounds are classified as sounds with a relatively low number of zero-crossing and high STE, unvoiced sounds have a high number of zero-crossing and low STE and silence segments have low ZCR and low STE. Using fixed ZCR and STE thresholds do not give good performance across various speech samples [25], especially in varying intensity profiles experienced in dysarthric speech.

3.7 Review of Machine Learning Techniques for Dysarthric Speech Classification

Over the years, researchers have used various classification techniques ranging from statistical/mathematical methods [26, 64, 187], to perceptual methods [6, 188], to machine learning methods [67, 68, 189] for the classification of dysarthria into

different severity levels. But in recent years, machine learning techniques have gained more research attention due to their performance [190-192]. In this section, different machine learning techniques will be reviewed with respect to their application in dysarthric speech classification.

3.7.1 Neural Networks

Artificial neural networks (ANN) are models that are biologically-inspired by how information is processed in the neural systems [193] which allows the computer to learn patterns and observations from large data. ANN consists of a number of neurons that are interconnected and working together to achieve a particular goal which could be to recognise patterns [189], solve a problem [194], identify subjects [195], classify data [79] or detect an abnormal situation [69]. Due to its non-linear mapping function, an ANN has the potential to effectively learn data with non-linear models such as speech [196]. Researchers have achieved significant results using ANN with just one layer of hidden neurons to recognize patterns and predict outputs from non-linear input data [189, 197]. In recent years, however, multi hidden layers have been more used due to the increased complexity of the dataset.

ANN has been used in many speech processing applications for different purposes such as speech recognition [189, 196, 198], speaker identification [195, 199], emotion detection [200] dysarthria automatic detection [79] and speech classification [6, 189]. Of all these applications, little research attention has been given to the application of ANN in speech classification especially in the area of dysarthria severity classification [189], whereas a lot of attention has been given to using ANN techniques in speech recognition within the research community [60, 201]. There is, therefore, a need to explore the possibility of applying different variants of ANN in the dysarthria detection and severity classification problem.

3.7.2 Support Vector Machines

The support vector machine (SVM) was introduced by researchers in [202] as a supervised learning technique for mapping input and output data using a classification or regression function [203]. In recent years, the SVM classification technique has been applied in various disciplines because of its high accuracy, flexibility and ability to handle complex data [203]. SVM classifiers belong to the kernel methods family.

Kernel methods are learning methods whose dependence on the training data is only by dot-products [204]. The kernel functions convert the input data to a feature space with high dimension [203]. This allows the classifier to perform dot-products in the high dimension and feature vector space with complex structure. The SVM classifier has the ability to produce non-linear boundaries and work on data set with no fixed dimension in the vector space [204] which makes the SVM useful in speech-related applications. In speech processing, SVMs are mainly used for regression analysis, class separation (also known as classification) and detection of irregularity [205].

The simplest form of SVM is a binary classifier consisting of two distinct classes that are separated by a hyperplane. New observations are mapped as points in the SVM model space and then classified to either side based on their position with respect to the hyperplane [205]. The SVM classifier is set to find the optimal hyperplane that can best differentiate the classes by minimizing the classification error. The data points that are closest to this optimal hyperplane are called the support vectors [205]. The optimal hyperplane maximizes the margin between the support vectors on either side of the plane [204].

SVMs are grouped into different types based on the type of classifiers used. Broadly speaking, SVM classifiers can either be linear, (with linear classification boundaries) or nonlinear (with nonlinear classification boundaries). The linear classifiers provide simplicity in training whereas the nonlinear classifiers provide a better training accuracy especially in a linearly inseparable dataset [204]. Quadratic, cubic, fine Gaussian, medium Gaussian, coarse Gaussian are common types of nonlinear SVM. Nonlinear SVMs are differentiated by their hyperparameters such as the degree of the polynomial kernel (for example cubic, quadratic), the width of the Gaussian kernel (gamma) and the soft margin constant (C) [204].

The SVMs have proved to be very useful in the classification of complex dataset especially when nonlinear kernels are used. However, the major limitation of the SVM classifiers is that they are binary classifiers in nature. Researchers have however used the pair-wise classification technique to overcome this limitation in multiclass classification problems [206]. In this research, the performance of the various types of SVM classifiers will be examined in automatic dysarthric speech detection and severity classification.

3.7.3 k-Nearest Neighbours

The k-nearest neighbours (kNN) is a classification technique that involves identification of the k elements that are the closest neighbours in the training dataset to the test element [205]. The test element is classified to the dominant class within its k closest neighbours. The kNN classifier is also called the memory-based classifier because the labels of the training elements are needed at run-time [207].

In kNN classification, the first step includes labelling of the training dataset in the various classes. The test data elements are then located in the dataset space and their k-nearest neighbours identified. The distance between the elements can either be estimated using the Euclidean distance function or a multidimensional distance function. The voting for the most suitable class can be done either by a majority vote or a weighted vote (to achieve a well-balanced vote) [207].

Applications of the kNN classification technique in speech processing include speech emotion detection [206, 208], speech recognition [209], and repetition detection in stuttering [210]. Although the kNN classification technique is simple to implement, flexible and can naturally handle multi-class problems, it requires storage memory for the pre-labelled elements and a large data search [207]. Another limitation is the need to describe a distance function that fits the classification problem. In other speech-related applications, the question of the most suitable size of k neighbours also arises. The performance of the kNN classifier in comparison to other types of classifiers will be of interest in this research.

3.7.4 Deep Learning

Over the years, the type and nature of the neural network architecture have evolved from the simple models to a variety of highly complex and sophisticated models. Researchers have made use of a variant of single-layer, multi-layer, self-organizing, self-recurring, time delay, and even adaptive neural network models to address various nonlinear, multidimensional classification problems [205]. There exists a continuous need, within the research community, to improve these models to achieve higher intelligence levels that closely match human intelligence. This quest for stronger and more robust artificial intelligence has opened a research path for deep learning within and beyond the machine learning research community.

Deep learning algorithms are often motivated by a need for higher classification efficiency, increased data complexity, availability of more computer processing power and recent advances in the cognitive neuroscience field [205]. Deep neural networks, spiking neural networks, hierarchical temporal memory and cortical algorithms are some of the examples of the deep learning techniques [205], most of which are biologically inspired and are quite promising in speech-related applications [198, 211, 212]. Although deep learning algorithms are relevant in supervised, unsupervised and reinforcement learning, there are a few limitations in their implementation [213]. One major set-back of deep learning is the requirement for a very large amount of training data [214] which are not readily available in dysarthric speech analysis. Due to increased complexity and cost, deep learning is often not applied when the performance of other classifiers is deemed good enough.

3.8 Summary

In this chapter, various feature extraction techniques relevant to this research have been reviewed. The review also involves the techniques used in pre-processing of speech signals to enhance the performance of the feature extraction algorithms. Techniques used in the extraction of time-domain features, as well as spectral and cepstral features, are reviewed and their performance compared both for healthy control and dysarthria speakers. In addition to this, extended features derived from time-domain, spectral and cepstral features are also reviewed and their applications to disordered speech discussed. Furthermore, the major gaps in the segmentation of disordered speech into silence, unvoiced and voiced segments are identified and the review of current machine learning techniques relevant to the automatic detection and severity classification in the dysarthric speech is also presented. This chapter provides the foundation for the contributions of this research work presented in Chapters 4 to 8 of this thesis.

As discussed in this chapter, traditional techniques for the SUV segmentation of dysarthric speech are prone to errors which motivated the development of a novel SUV segmentation technique presented in the next chapter that gives a better segmentation performance both in healthy controlled and dysarthric speech. One of the ways of assessing dysarthric speech is to measure the ability of the speakers to produce repetitive sounds at a fast rate using DDK syllables. In Chapter 5 of this thesis, an

automatic DDK analysis technique which is based on moving average segmentation and duration-based merging is proposed. Furthermore, machine-learning techniques are proposed in Chapter 6 for the automatic detection and severity classification of dysarthric speech using speech features that characterise dysarthric speech which includes: jitter, shimmer, HNR, centroid formants, MFCCs, wavelets and prosodic features. In addition, an assessment of the impact of manipulation of three prosodic features (intensity, duration and pitch) on the listeners' ability to correctly identify the location of the stressed word in dysarthric sentences will be presented in Chapter 7, to enable clinicians to make informed decisions when administering prosody-based therapies. The techniques proposed in this thesis are developed into interactive dysarthria management tools in MATLAB which are presented in Chapter 8 of this thesis.

Chapter 4

4 Novel Silence Unvoiced Voiced (SUV) Segmentation in Dysarthric Speech

4.1 Introduction

In this chapter, a novel algorithm for the segmentation of dysarthric speech into silence, unvoiced and voiced (SUV) segments will be described. The proposed algorithm will be based on the combination of short-time energy (STE), zero-crossing rate (ZCR) and linear prediction error variance (LPEV). Extending the previous work in this field, the proposed method will address the difficulties in distinguishing between voiced and unvoiced segments in dysarthric speech. More precisely, the error variance of the linear prediction coefficients will be used to design a three-fold decision matrix that can accommodate the high variability in loudness experienced in dysarthric speech. In addition, a moving average threshold approach will be proposed in order to provide an "as-fit" segmentation technique that is fully automated and that will be able to handle highly severe dysarthric speech with varying loudness and ZCRs. The ability of the proposed fully-automated algorithm will be validated using real speech samples from healthy speakers, and speakers with ataxic dysarthria. Furthermore, the performance of the algorithm in real-time segmentation and its application extraction of speech features will be presented in this chapter.

4.2 SUV Segmentation Algorithm for Dysarthric Speech

The proposed SUV segmentation is divided into five stages namely; pre-processing, ZCR estimation, STE estimation, linear prediction and decision stage. The block diagram of the proposed SUV segmentation algorithm is illustrated in Figure 4-1. The audio sample to be segmented is first pre-processed after which the segmentation features are extracted. The features are extracted in short-time frames (the importance of short-time analysis in speech processed are discussed in Section 3.2.6). The size of the frames used in this algorithm is 31.25 milliseconds (that is, 1/32 second or 512 samples at a sampling rate of 16 kHz as discussed in Section 3.2.5). The feature

extraction step is carried out by estimating the number of zero crossings in each speech frame, calculating the STE and then performing a linear prediction to estimate the prediction error variance of each frame. The last stage of the process involves the classification of each speech frame to one of the three SUV classes.

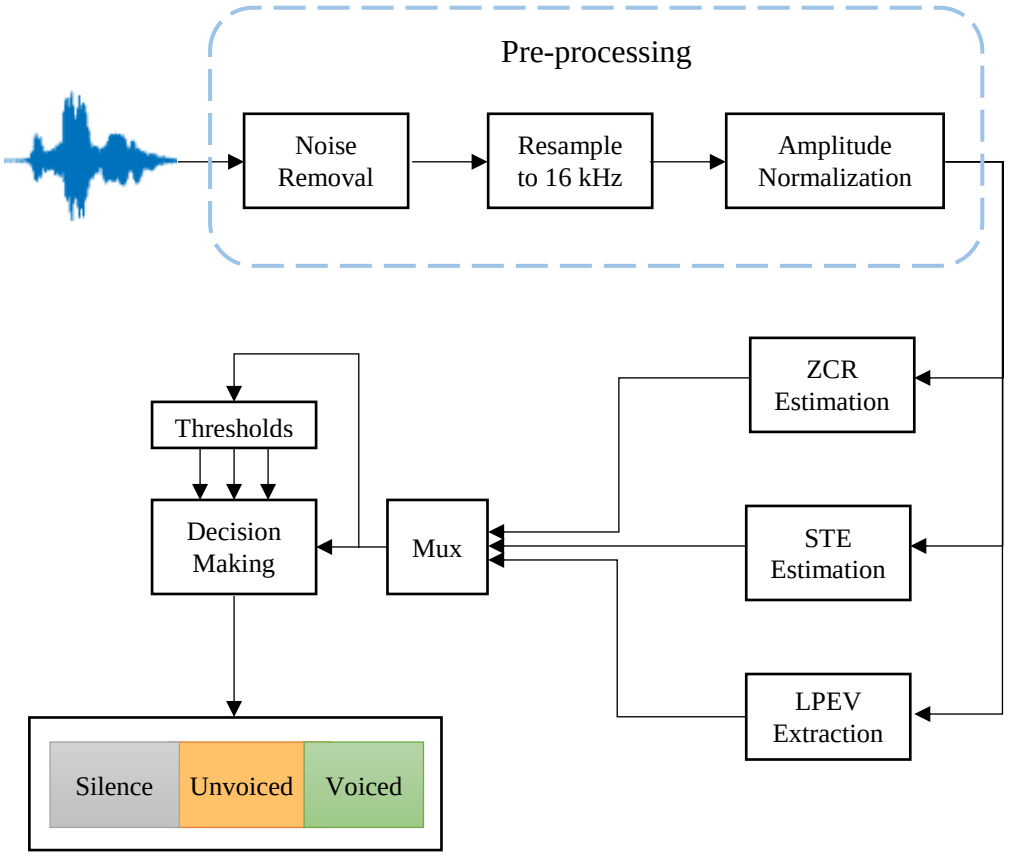


Figure 4-1. The Block Diagram of the Proposed SUV Segmentation Technique

4.2.1 Pre-processing

There are three steps involved in the pre-processing of the audio signal in readiness for feature extraction. These include noise removal, resampling and amplitude normalization. Additive noise in a speech signal can often interfere in the feature extraction process. As discussed in Section 3.2.3, noise removal/reduction helps in improving the perceptual quality of distorted speech.

The Wiener filter is used to remove the undesired additive noise due to the recording equipment or background noise. The FIR Wiener filter was chosen because of its lower computation cost since the SUV segmentation is to be carried out in real-time [113].

After denoising, the audio signals are resampled to 16 kHz. This is to ensure that all the signals are sampled at the same rate. The choice of this sampling rate is as a result of the range of human speech frequencies and the Nyquist criterion (as discussed in Section 3.2.5). Sampling all the audio signals at the same rate also helps in equalizing the frame size which will normalize the ZCR to be estimated in Section 4.2.2.

Afterwards, the amplitudes of the resampled audio signals are normalised. Due to the variations in speaker volume and microphone distance, the amplitude of the audio samples is normalised such that the signal lies between -1.0 and +1.0 (without changing the sign of the signal values as discussed in Section 3.2.2). Amplitude normalisation is achieved by dividing the signal by its maximum absolute value. The resulting denoised resampled and normalised audio signals are divided into overlapping frames of 512 samples each with 75% overlap between consecutive frames. Using overlapping frames is targeted towards improvements in the segmentation process.

4.2.2 Zero-Crossing Rate Estimation

The first step in the feature extraction stage is the estimation of the ZCR across all frames. The ZCR for unvoiced speech segment is expected to be higher than that of the voiced segment [129]. This is because unvoiced frames are less periodic with varying frequencies, leading to a high rate of change in amplitude and signal value sign. For silent frames of the speech, the ZCR is expected to be approximately zero, provided the speech signal is free of background noise [129]. However, the number of zero crossings within a speech segment is affected by the quality of the recording [124]. A speech sample recorded using different devices will give variable ZCRs across devices based on their impulse responses. To reduce the effect of this variability in ZCRs, the use of a signal-specific threshold that is a function of the ZCRs across all the frames of the audio sample is proposed. The threshold of the ZCR segmentation is therefore based on the quality of individual speech sample rather than a fixed value. The proposed ZCR threshold is given by (18). As the ZCR varies across devices, the threshold also varies to match the range of variation in recording quality.

$$ZCR_{Thres} = k \cdot [\max(ZCR) - \min(ZCR)] \quad (18)$$

where k is the voicing threshold factor. For this study, by inspection, k is taken to be 0.3. This choice was made after analysing 700 single words from both healthy controlled and dysarthric speakers. Differences in ZCR values are used to section the speech signal into silence, unvoiced and voiced classes.

4.2.3 Short-Time Energy Estimation

The STE is also estimated across all frames and measured to decibels. Since the amplitudes of the signals have been normalised in the pre-processing stage, the STEs can be used to distinguish the silence frames from the other parts of the signal in a healthy speech. The speech signals from healthy control speakers are expected to have very low STE in silent frames and the voiced frames are expected to have the highest STE. However, this is not what is observed in dysarthric speech. The loudness of the dysarthric speech varies considerably, therefore, affecting the STE values.

High variability in loudness in dysarthric speech samples leads to a mismatch in STE for silence, unvoiced and voiced frames. This implies that the voiced or unvoiced frames in dysarthric speech can often have low STE due to reduced loudness, especially at the end of the utterance. In addition, there can be bursts of loudness within the utterances which can affect the performance of the STE-based SUV segmentation proposed in [124, 125]. The STE in itself is, therefore, not sufficient to adequately segment the dysarthric speech into SUV segments.

4.2.4 Linear Prediction Error Variance

The third feature extracted in this proposed SUV segmentation algorithm is the linear prediction estimation error variance (LPEV). The LPEV is derived from the linear prediction of the speech signal. Linear predictive coding (LPC) is used in speech processing to model audio signals for feature extraction [147], speech recognition [134] and speech synthesis [215]. This is achieved by predicting the next signal value based on the last P values where P is the linear prediction order. The P previous values are weighted by the LPC coefficients and added to give the next signal value. The LPC gives a very close approximation to the original signal [134]. The residual signal after the estimated signal being removed from the original signal is called the linear

prediction estimation error (LPE). The variance of the LPE is measured for each frame, which is called the LPEV.

The LPEV is useful in SUV segmentation because LPC shows different performance for different parts of the speech sample. In the voiced frames, most of the signal energy is extracted in the LPC coefficients as formants and the residual signal contains very low energy in comparison with the original signal energy. However, for unvoiced frames, the residual signal energy is usually higher than the linearly predicted signal energy because most of the signal energy is decomposed in the residual signal. Despite the high variability in dysarthric speech, the linear prediction algorithm is still able to accurately predict the next signal value given the previous P values. The order of the linear prediction algorithm (that is, the P-value) used in this proposed method is given by (12). This is based on the rule of thumb for formant estimation [80, 216]. The linear prediction error, on the other hand, is given by (11) as discussed in Section 3.4.2.

The analysis of the LPEV has shown that, in dysarthric speech signals, the variance of the LPE for unvoiced frames is lower than that of voiced frames for the same speaker. The lowest LPEV is recorded in silent frames. The residual signal in an unvoiced frame does not vary significantly since it consists mainly of the original signal energy. For voiced signal, on the other hand, there is a high variance in residual signal energy due to the change in formant energy across different sounds. The threshold value is taken as the median LPEV. This gives a basis for segmenting speech samples into silence, voiced and unvoiced segments.

4.2.5 Segmentation Decision Criteria

Combining the ZCR, STE and LPEV criteria, the audio samples are segmented into the three classes; Silence, Unvoiced, and Voiced at the decision stage. The flowchart of the proposed algorithm is illustrated in Figure 4-2. In the case of clean speech samples, silence segments have zero ZCR and very low (approximately 0) LPEV while unvoiced segments have high ZCR and low LPEV, and voiced segments have low ZCR and high LPEV. However, in realistic conditions, audio samples are not usually noiseless due to environmental conditions and recording quality. To mitigate these challenges, more refined segmentation criteria are used.

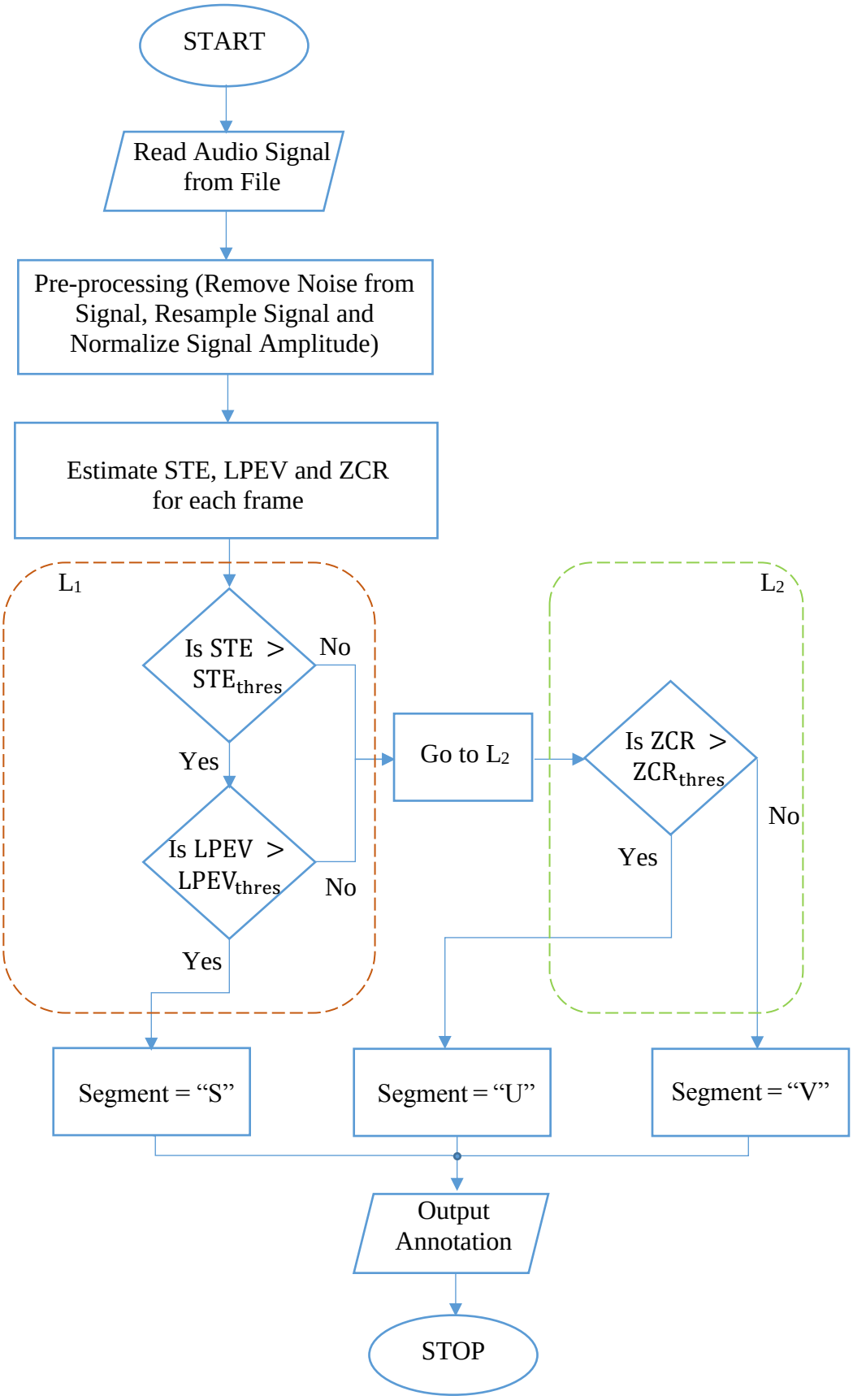


Figure 4-2. Flow Chart of the Proposed Algorithm

As illustrated in Figure 4-2, a three-fold criterion was used for the segmentation process. The thresholds used in this segmentation process are speaker-dependent and are a function of the range of extracted features (ZCR and STE). The use of speaker-specific thresholds ensures that the process works automatically without any interference from the users. Also, fixed thresholds are known to result in low accuracy in dysarthric speech since their intensity profiles are not as observed in healthy speech (due to high variability and bursts of loudness).

The threshold for the ZCR measures is given in (18) and the threshold for the LPEV measures is given by the median of the LPEV values in the whole signal and the threshold of the STE is fixed at 0 dB. To keep the threshold within an acceptable range, the ZCR threshold is, however, kept within the range of 100 and 200 zero crossings per frame and the LPEV threshold is kept at a minimum of -100 dB. These thresholds are used to determine if the measured values are "high" or "low". For example, if the measured ZCR for a particular frame is less than the ZCR threshold (ZCR_{thres}), the ZCR for the frame is marked as "low" and vice versa. Apart from "low" and "high", those instances, where the measured values are approximately equal to zero, are also noted. Table 4-1 gives a summary of the decision matrix.

Table 4-1. Three-fold SUV Segmentation Criteria

Layers	**STE**	**LPEV**	**ZCR**	**Dysarthric Speech**
Layer L_1	Low	Low	Approximately 0	Silence
	Low	Low	Low	Silence
	Low	Low	High	Silence
Layer L_2	Low	High	Low	Voiced
	High	Low	Low	Voiced*
	High	High	Low	Voiced
	Low	High	High	Unvoiced
	High	Low	High	Unvoiced*
	High	High	High	Unvoiced

*occurs only during bursts of loudness

Using the decision matrix, the decision-making stage involves two layers (L_1 and L_2). The first layer, L_1, is where the silence frames are separated from the rest of the speech signal. The separation is based on the LPEV and the STE thresholds. When the LPEV and STE values for a particular frame are lesser than the thresholds the frame is assigned 'S' (silence). The silent frames are characterised by low LPEV and low STE. If one or both of these two parameters is/are "high", the ZCR is needed to decide the class in the second decision layer, L_2. After the silence frames have been separated in L_1, the remaining components of the signal are taken through the second layer, L_2, where the voiced and unvoiced segments are separated based on the ZCR threshold. If the ZCR is higher than the threshold, the segment is classified 'U' (unvoiced), otherwise, the segment is classified 'V' (voiced). This two-layer process ensures that the boundaries between the 3 classes are well defined.

4.3 Experimental Results

The proposed method was tested on **385 audio samples** recorded in an echo-free environment. The audio samples are from 20 speakers, 10 of which are ataxic dysarthric speakers and 10 age and gender-matched healthy control speakers. Each speaker produced 20 single word speech. Each group consisted of 5 males and 5 females. The recorded audio samples were prescreened and 15 were removed due to the quality of the recording. This corpus was taken from the dataset reported by [27]. The audio samples were manually labelled using Praat (using the established Praat voicing labels) whereas MATLAB was used for speech analysis. The results of the experiment were compared with that of the state-of-the-art segmentation technique proposed in [129] and [131].

Figure 4-3 shows the results of one of the audio samples where a speaker said the word '*differ'*. Grey, Red and Blue colours were used to represent the 'S', 'U' and 'V' segments respectively. The word is divided into two syllables (di-ffer). The proposed method accurately removes the silence segments and automatically separates the unvoiced segments from the voiced segments. The first syllable contains voiced sounds whereas the second syllable contains both unvoiced and voiced sounds. Due to reduced loudness at the start of the second syllable, the STE of the sound 'ff' was lower than the STE threshold. Without considering the LPEV, the unvoiced sound ('ff')

would have been classified as silence. Figure 4-4 shows the results of the SUV segmentation based on the method proposed in [129] and [131].

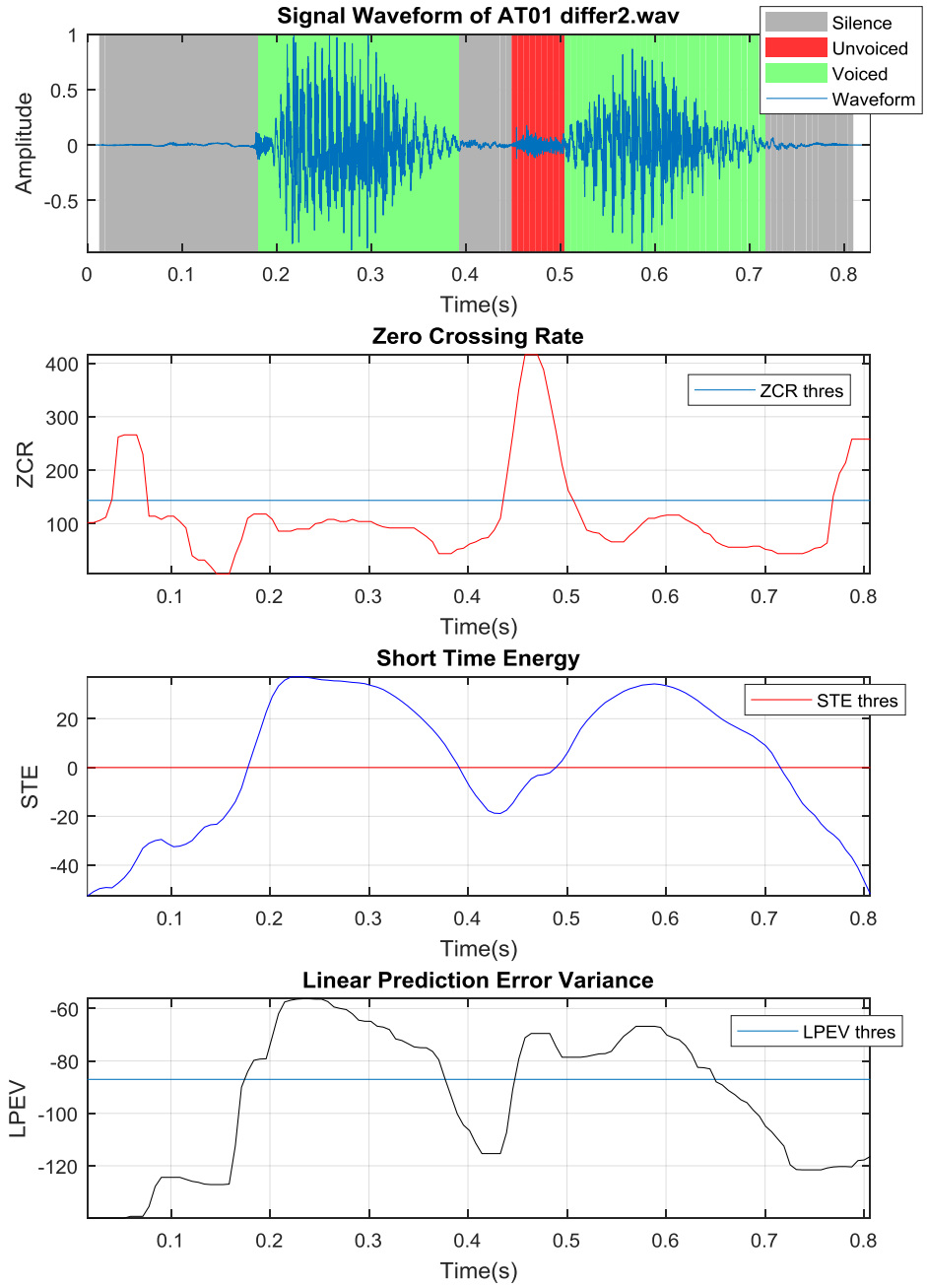


Figure 4-3. SUV Segmentation of the Word "Differ" using the proposed method

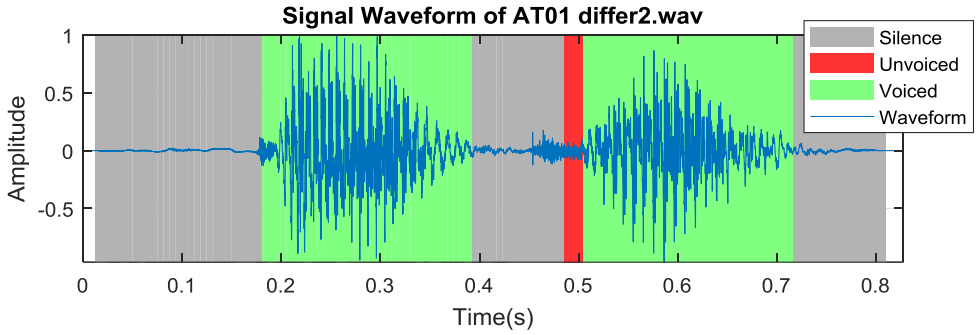


Figure 4-4. SUV Segmentation of the word "Differ" using ZCR+STE method [129]

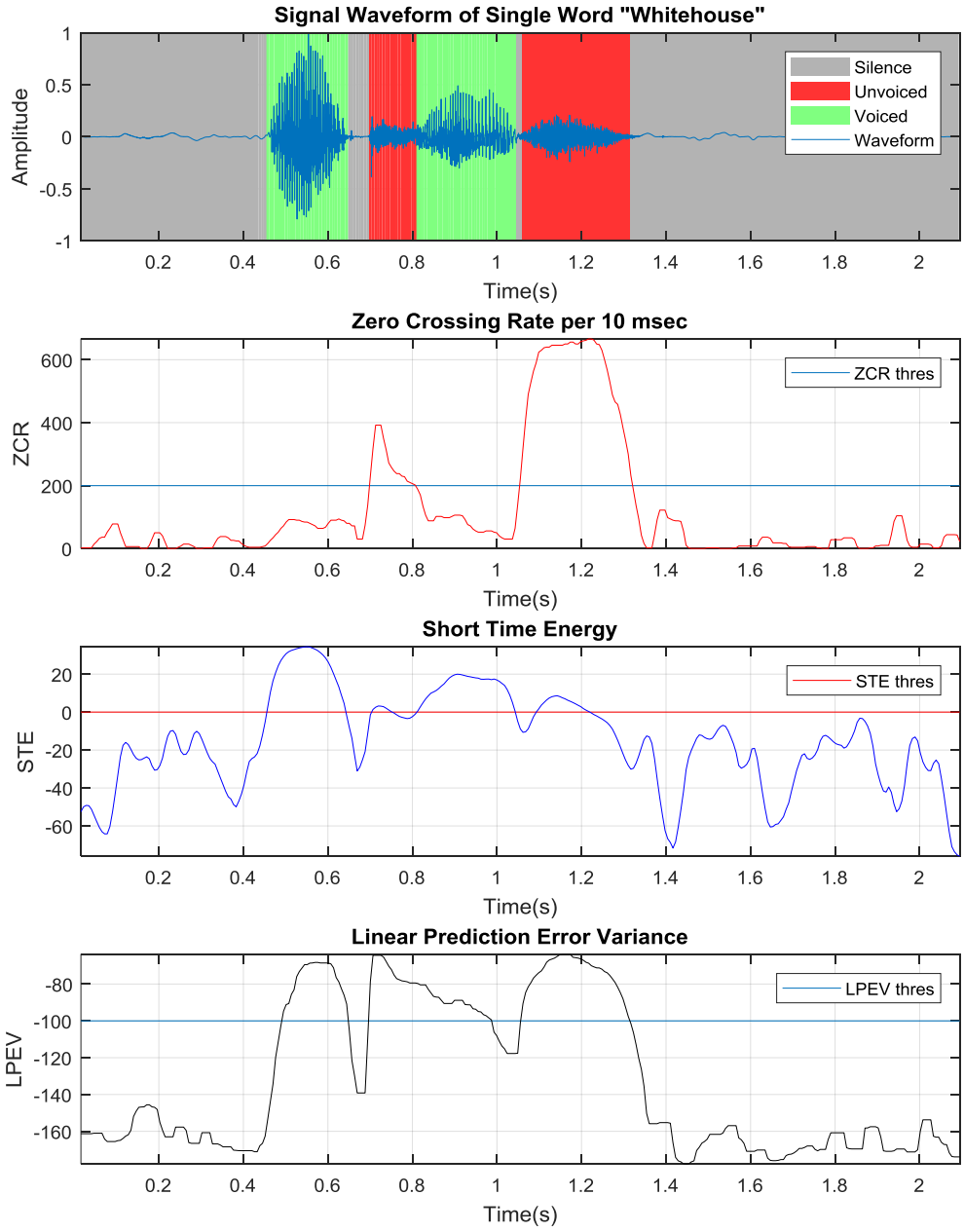


Figure 4-5. SUV Segmentation of "Whitehouse" using the proposed method

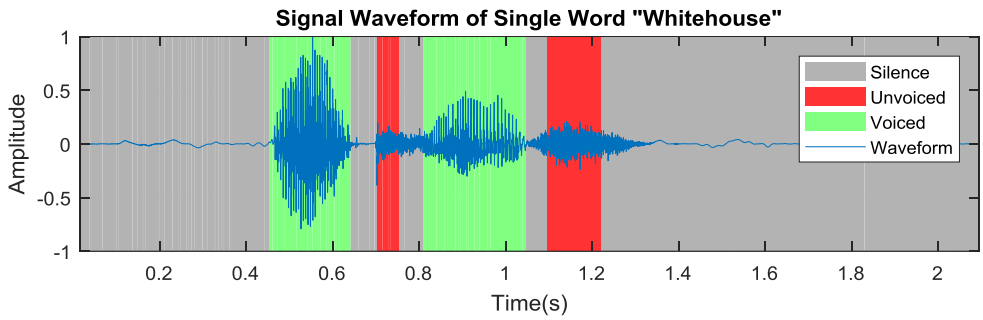


Figure 4-6. SUV Segmentation of the Word "Whitehouse" using ZCR+STE method in [129]

Figure 4-5 shows the result of the proposed method in another single word "Whitehouse". The proposed method correctly identified all the voiced, unvoiced and silence segments despite the high variability in intensity. The ZCR+STE method in [129], on the other hand, resulted in multiple errors as shown in Figure 4-6. It can be seen in Figure 4-6 that modification of the STE threshold (to a lower value) will introduce more errors into the segmentation results, thereby, reaffirming that STE alone is not sufficient for segmenting the dysarthric speech due to high variability in intensity.

The experimental results were compared with the manually labelled annotation for the 385 audio samples and the errors in the proposed segmentation method were measured. An SUV segmentation described in [129, 131] using STE and ZCR was also implemented and compared with the experimental results from the proposed three-fold method. The results of this comparison are presented in Figure 4-3, Figure 4-4, Figure 4-5, and Figure 4-6. The average accuracy of traditional STE+ZCR was 94.8% as illustrated in Table 4-2. With the same dataset, the average accuracy of the proposed method across the 3 classes was 98.9% and the average percentage error rate was 1.1%.

Table 4-2. Performance of the proposed algorithm on dysarthric data set comprising of 385 audio signals

Segment	**Accuracy**	
	STE+ZCR Method	**Proposed Method**
Silence	96.2%	99.4%
Unvoiced	89.3%	98.4%
Voiced	99.0%	99.0%
Average	94.8%	98.9%

One major advantage of the proposed method over the STE + ZCR method in [129] is its high sensitivity at low energy. An example is illustrated in Figure 4-3 and Figure 4-4. This is because periodic and quasi-periodic speech components are detected in linear prediction and the LPC method is independent of the signal energy. In terms of consistency, since the LPEV is very sensitive even at low signal energy, there is a reduction in classification gaps due to fluctuations in signal energy, therefore improving the detection of the inter-class boundaries.

Furthermore, voice onset is well defined in the proposed method (represented by a very steep rise in LPEV). Consequently, this method is very useful in voice activity detection applications. Using STE in voice activity detection does not give well-defined boundaries between silence and the first voiced/unvoiced segment in the speech samples. When compared with the manually labelled annotation, the proposed method produced consistently accurate results for both male and female speakers.

4.4 Summary

In this chapter, a novel algorithm for the segmentation of dysarthric speech into silence, unvoiced and voiced segments has been presented. This method uses a three-fold segmentation decision matrix based on the ZCR, STE and LPEV of the signals. The algorithm reduces the segmentation error due to reduced loudness with high variability experienced in dysarthric speech. The LPEV ensures that the unvoiced and voiced segments with low intensity are accentuated while suppressing the silence segments. In this algorithm, speaker-specific thresholds are used for the segmentation problem thereby reducing the errors due to fixed segmentation thresholds. Experimental results show that the proposed algorithm leads to accurate segmentation of dysarthric speech with fine boundaries between the three classes despite the high variation in intensity in the signals. This algorithm can also be applied in the extraction of voice-based and spectral features such as fundamental frequency, formants, voiced activity detection, syllable duration, speech rate, etc. This algorithm can, therefore, be extended to other speech processing applications.

Chapter 5

5 Novel Automatic DDK Analysis for Assessment of Dysarthria

5.1 Introduction

A novel technique for the automatic analysis of diadochokinetic (DDK) samples from dysarthric speakers is presented in this chapter. The proposed algorithm will be based on the automatic segmentation of DDK syllables and estimation of the DDK rate as well as the peak intensity of the DDK syllables. Improving on previous techniques for DDK syllable segmentation, the proposed technique will make use of a moving average threshold to reduce the effect of intensity bursts and high variability in intensity due to articulatory breakdown. In addition, a minimum duration merging method will be proposed in order to reduce over-segmentation due to intra-syllable pauses between the consonant and the vowel sounds. The performance of the proposed algorithm will be validated using 284 DDK samples from 71 speakers. These include speakers with ataxic dysarthria, Parkinson's disease, young speakers and healthy control speakers. Furthermore, the reliability of the proposed technique will be validated by increasing and decreasing the threshold by 2 dB and measuring the performance of the technique at these threshold values. The applicability of the proposed technique in supporting clinicians during the assessment of dysarthria will also be presented in this chapter.

5.2 Diadochokinetic Skill in Speech

Diadochokinetic (DDK) skill is the ability of speakers to rapidly repeat alternating movements in speech [217]. DDK tasks often involve repetition of syllables at a very fast rate. The most common examples of syllables used in DDK tasks are /pʌ/ /tʌ/ and /kʌ/. DDK rate also called the alternating motion rate (AMR), is the measure of the number of syllables repetitions within a period of time (typically 1 second). The average DDK rate in healthy adults ranges from 5 to 7 repetitions per

second, however, this varies with age [217]. DDK tasks are routinely used by speech and language therapists (SLTs) to assess speech difficulties both in children and in adults. Analysing and interpreting the results can be challenging as the clinicians will have to manually transcribe and annotate the recorded speech signals to measure the peak loudness, the duration of the repetitions, syllable repetition rate and the variability of the repetitions [217].

Recently, research interests in quantitative analysis of DDK tasks using instrumental methods have increased [217]. These instrumental methods offer time-savings, improved reliability, the ability for post-treatment analysis, progress tracking, and more information about the speech samples. In this study, a fully automated instrumental DDK analysis algorithm is proposed which will record the DDK samples from the patients (using a microphone connected to a computer), analyse the recorded signals, extract the peak loudness and the DDK rate.

5.3 Participants

The details of the participants involved in this study are shown in Table 5-1. The dataset consisted of 71 individuals; 23 of which have been diagnosed with Parkinson's disease, 8 participants diagnosed with ataxic dysarthria, 13 young participants and 27 healthy control participants. Each of these participants carried out 4 DDK tasks on fast repetitions of /pʌ/, /tʌ/, /kʌ/, and /pʌtʌkʌ/. The data is collected in noise and echo-free environment. All data used in this study are recorded using the same equipment set-up. The recorded audio samples are labelled according to the assigned speaker number shown in Table 5-1

Recorded speech samples for each task are stored in separate folders with the speaker number. The data is stored in a secured cloud server which is accessible for research purposes. For this study, the recorded audio samples were manually labelled in Praat and separated to individual DDK syllables. The automatic segmentation algorithm was, however, developed in MATLAB using the methodology proposed in Section 5.4.

Table 5-1. Participants for the Automatic DDK Analysis Study

Parkinson's Disease Speaker Group		Ataxic Dysarthria Speaker Group		Control Group	Speaker
Speaker	Gender	Speaker	Gender	Speaker	Gender
PD01	M	AD01	M	C01	F
PD02	F	AD02	F	C02	M
PD03	M	AD03	F	C03	F
PD04	M	AD04	M	C04	F
PD05	F	AD05	F	C05	M
PD06	M	AD06	M	C06	M
PD07	M	AD07	M	C07	M
PD08	M	AD08	M	C08	M
PD09	M	Young Speaker Group		C09	F
PD10	M			C10	F
PD11	M	Speaker	Gender	C11	M
PD12	M	Y01	F	C12	M
PD13	M	Y02	F	C13	M
PD14	F	Y03	M	C14	M
PD15	M	Y04	F	C15	M
PD16	M	Y05	M	C16	M
PD17	M	Y06	F	C17	M
PD18	M	Y07	F	C18	F
PD19	M	Y08	F	C19	M
PD20	F	Y09	M	C10	F
PD21	M	Y10	M	C21	M
PD22	M	Y11	M	C22	M
PD23	F	Y12	M	C23	F
		Y13	F	C24	M
				C25	M
				C26	M
				C27	M

5.4 Proposed Methodology

The proposed automatic DDK analysis algorithm is designed to tackle three of the main issues faced by instrumental DDK analysis methods as discussed in [217]. These include errors due to reduced or increased intensity, over-segmentation due to inter-syllable pauses and under-segmentation due to the choice of threshold. The aim of the proposed algorithm is to automatically extract the DDK rate and peak loudness of DDK syllables from DDK audio samples by the following:

- Accurately segmenting the DDK syllables using intensity thresholding
- Reducing the errors due to inter-syllable pauses
- Reducing the errors due to dips between consonant and vowel segments
- Setting a changing segmentation threshold to account for erratic intensity or reduced peak intensity caused by an articulatory breakdown
- Eliminating the skewed DDK rates at the start and the end of each recording

The block diagram of the proposed algorithm is illustrated in Figure 5-1. The methodology includes pre-processing, syllable segmentation, peak intensity extraction, syllable duration measurement and the estimation of the DDK rate and its covariance.

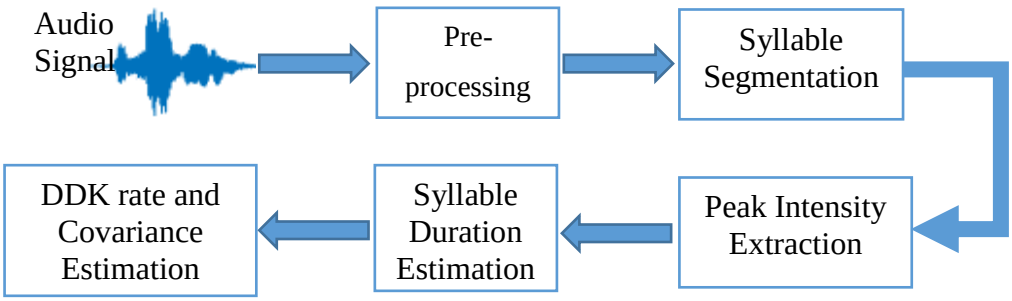


Figure 5-1. Block Diagram of the Automatic DDK Analysis Tool

5.4.1 Pre-processing

The recorded audio signals from the DDK tasks are first pre-processed to remove additive noise and resample the signal. The noise due to the recording equipment and recording environment are removed using a Wiener filter. After which, the audio signals are sampled at 16 kHz to ensure a uniformed sampling rate across all audio samples (based on the Nyquist sampling criterion as described in 3.2.5).

5.4.2 DDK Syllable Segmentation

Estimating the peak intensity and the syllable duration requires that the utterances are segmented into the individual DDK syllables as produced by the speakers. Therefore, the next stage after the pre-processing involves the segmentation of the audio signals into the individual syllables. This is achieved by detecting the boundary point between consecutive syllables (that is, the point where a syllable ends and the next one begins). The boundary points are the low turning points in the intensity profile which are determined by the difference function of the intensity profile. At turning points (maximum and minimum turning points), it is expected that the first-order derivative (the first-order difference function for the discrete-time system) of the intensity function will tend to zero. At the minimum turning point, the second-order derivative of the intensity function is expected to be greater than zero.

However, to locate the troughs (minimum turning points) in the intensity profile, there is a need for a segmentation threshold below which the minimum turning points can be searched for (this will reduce the search region thereby reducing the computational time). The choice of segmentation threshold was based on two factors. The first factor being that the range of the intensity profile varies from speaker to speaker as well as by age and gender. The second factor is that dysarthria speech samples are often characterised by the highly varied intensity which can be accompanied by an articulatory breakdown. There is, therefore, a need to apply a segmentation threshold that is unique to an individual's intensity range but varying in nature to cope with varying intensity.

The first type of threshold considered is the mean intensity threshold. Using the mean intensity of the speaker's intensity profile ensures that the threshold is associated with a particular speaker which is more appropriate than using a fixed threshold for all speakers. Even though this helps in addressing the factor of the speaker's intensity range, it does not consider cases where the speaker's intensity varies considerably within the same utterance as shown in Figure 5-2. Although in this example, some of the syllables are correctly segmented, a good number of the syllables are omitted because their corresponding intensity troughs are located

above the segmentation threshold. This results in under-segmentation of the DDK audio sample as illustrated in Figure 5-2.

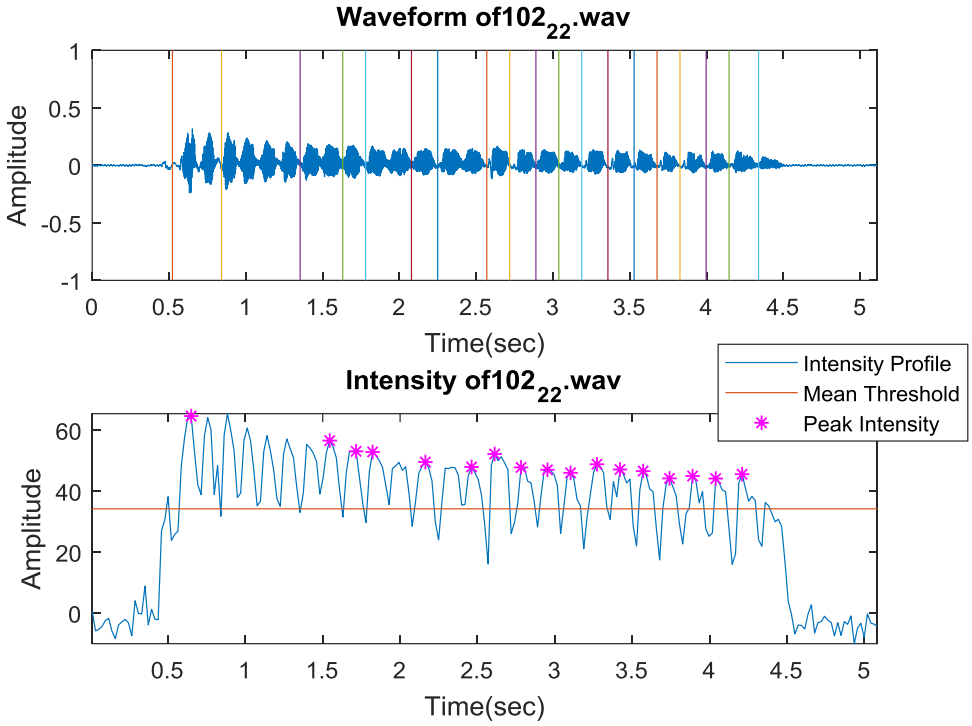


Figure 5-2. Waveform of a DDK audio signal (top) and the intensity profile (bottom) showing the identified peak intensity using speaker-specific mean intensity segmentation

To address this segmentation challenge, a moving average threshold approach is introduced with an averaging window of approximately half a second. This window size is chosen because previous research [217] shows that the expected range of DDK syllable duration is between 100 milliseconds and 200 milliseconds. The moving average threshold is unique to the speaker, not fixed and a function of the moving range of the intensity profile of the signal as well as the variability of the speaker's intensity. The moving average threshold of the ith frame of the audio signal is given by (19).

$$Threshold_{moving}(i) = \frac{1}{N}\sum_{k=1}^{N} INT_i(k) = \frac{1}{N}\sum_{k=1}^{N} |x_i[k]|^2 \qquad (19)$$

where N is the length of the frame, INT_i is the intensity profile of the ith frame, which is the square of the amplitude of the audio signal $x_i[k]$.

The effects of using the moving average threshold are illustrated in Figure 5-3. All of the omitted syllables in Figure 5-2 were accurately segmented in Figure 5-3 using the moving average threshold as shown by the yellow line on the intensity profile. (NB: Figure 5-2 and Figure 5-3 show the same DDK audio sample segmented using the mean threshold approach and moving average threshold approach respectively). The use of moving average threshold helps in reducing the errors likely to be introduced due to reduced peak intensity caused by articulatory breakdown. In addition, using a fixed threshold value can result in too high threshold which can result in the omission of syllables with reduced peak intensity or too low threshold which can lead to syllables merging.

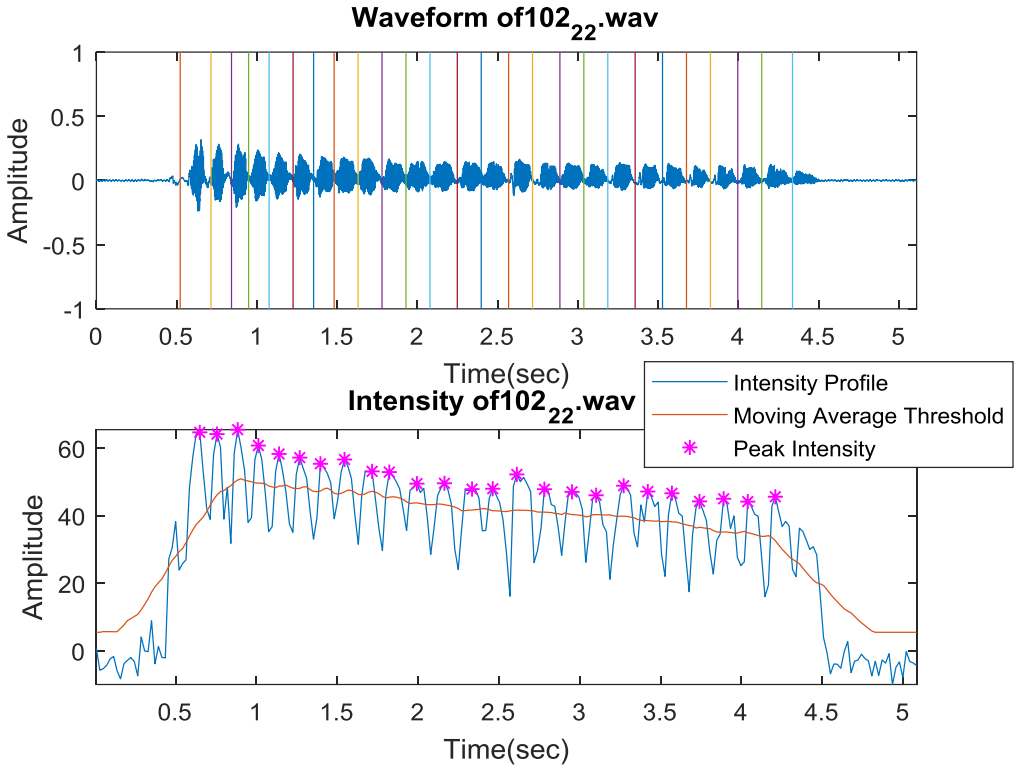


Figure 5-3. Effects of Moving Average Threshold on DDK Segmentation

Moreover, another challenge faced by current instrumental DDK segmentation techniques is over-segmentation of syllables. DDK syllables are often over-segmented when multiple troughs are detected within the same syllable. In Figure 5-4, one of the DDK syllables has been over-segmented due to an intensity dip between the consonant and vowel sounds as shown in black. This leads to an increase in the number of syllables and reduces the measured mean DDK rate for the speaker.

To address the issue of over-segmentation, the proposed algorithm makes use of a minimum syllable duration criteria to reduce the over-segmentation errors due to inter-syllable pauses. An analysis of multiple speech samples shows that the maximum number of syllables achievable by a very fast speaker is 9 syllables in a second. This means that the minimum syllable duration is 0.11 second (or 111 milliseconds) and that any syllable with duration less than 100 milliseconds has a high probability of being incomplete and most likely a pseudo-syllable (that is, a segment of a syllable). Two consecutive short pseudo syllables are therefore merged to form one syllable. This is termed minimum syllable duration merging.

Furthermore, this approach caters for scenarios where the syllable has been wrongly segmented into more than 2 pseudo syllables by merging neighbouring pseudo syllables based on the minimum syllable duration criteria. Apart from reducing the errors due to inter-syllable pauses, this minimum duration-based merging also helps in reducing (or eliminating) the errors introduced by dips between the consonant and vowel segments in the DDK syllables; thereby addressing the challenges of over-segmentation.

An illustration of the advantages of minimum duration-based merging is shown in Figure 5-5. Although, one of the /tʌ/ syllables has been over-segmented in Figure 5-4 due to the inter-syllable dip between the consonant and the vowel sounds, using the minimum duration syllable merging approach discussed above, the pseudo syllables are detected and merged as shown in Figure 5-5, in which all the DDK syllables are correctly segmented and the error in the DDK rate estimation eliminated. Apart from inter syllable dips (between constant and vowel sounds), this minimum duration merging method can also help in reducing over-segmentation due to pauses within each syllable; which is common in dysarthric speakers.

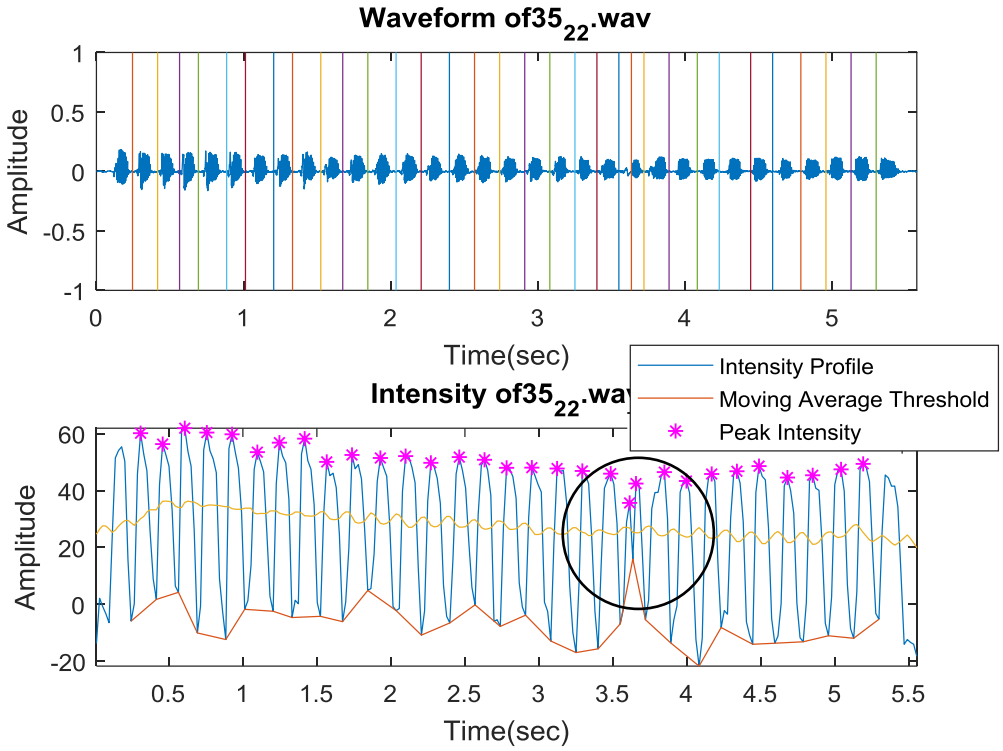


Figure 5-4. Automatic Syllable Segmentation of a DDK Audio Signal illustrating Over-segmentation due to inter-syllable dip.

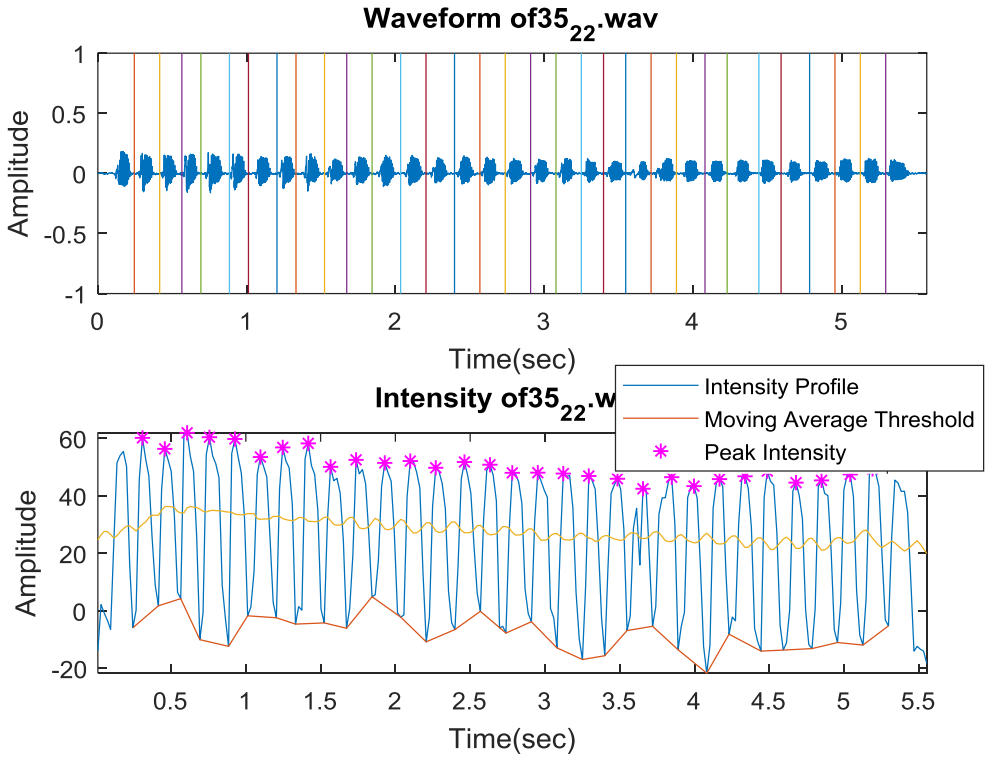


Figure 5-5. Corrected Syllable Segmentation using the Minimum Duration Pseudo-Syllable Merging Approach

5.4.3 Feature Extraction

After syllable segmentation, the peak intensities and the duration of the individual syllables are extracted from the segmented signal. The summary of the features extracted are shown in Table 5-2. The peak intensities are estimated as the STE of the syllables given by (6) as discussed in Section 3.3.1. Whereas the syllable duration is estimated by calculating the difference between two consecutive syllable boundaries. The final stage of the proposed automatic DDK analysis tool involves the estimation of the coefficient of variation of the peak intensity and duration of the individual syllable. The coefficient of variation is defined as the standard deviation of a variable (that is, DDK rate or Peak Loudness) divided by the mean of the variable. The peak intensities are measured in decibels, the DDK rates measured in syllables per second and the coefficients of variation measured as percentages.

Table 5-2. Summary of Extracted Features for Proposed Automatic DDK Analysis Algorithm

S/N	Variable	Description	Unit
1	*DDKrate*	Number of syllable repetitions per second	/s
2	*DDKpi*	Peak Intensity of DDK syllables	dB
3	*DDKavrate*	Average DDK rate	/s
4	*DDKavpi*	Average Peak Intensity	dB
5	*DDKavd*	Average DDK syllable duration	s
6	*rDDKrate*	Range of DDK rate [min, max]	/s
7	*rDDKpi*	Range of Peak Intensity [min, max]	dB
8	*stDDKrate*	Standard Deviation of DDK rate	/s
9	*stDDKpi*	Standard Deviation of Peak Intensity	dB
10	*cDDKrate*	Coefficient of Variation of DDK rate	%
11	*cDDKpi*	Coefficient of Variation of Peak Intensity	%

5.5 Experimental Results

In this section, the performance of the proposed DDK analysis algorithm is examined experimentally. All the audio samples were manually labelled by marking the start and the endpoint of each DDK syllable and the individual syllable duration were measured. These manually labelled results were then compared with the results from the automatic segmentation algorithm. An example of the manually labelled DDK audio sample is illustrated in Figure 5-6. It is important to note that the first and the last syllables were removed from the analysis to reduce the errors introduced due to the skewed DDK rates at the start and the end of each recording. As shown in Figure 5-6, the syllable boundaries are marked by the blue lines. In this figure, 23 syllables were extracted (after the removal of the first and the last syllables). The DDK rate for each syllable was measured as the reciprocal of the syllable duration. These calculated DDK rates were then compared with the DDK rates estimated using the proposed algorithm.

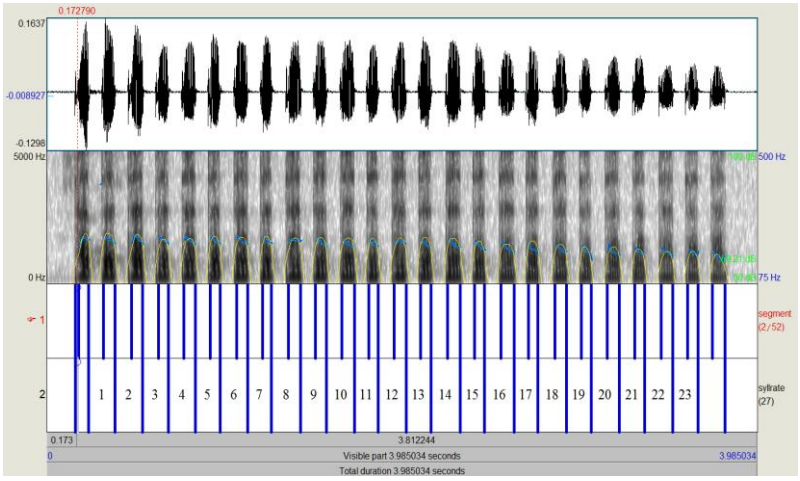


Figure 5-6. Manually Labelled DDK Audio Sample for Speaker C20 Performing the Fast Repetition of /pʌ/ Task

The difference between the average DDK rates estimated automatically and those calculated manually was measured. Audio samples where the difference was more than 1 syllable per second were noted. Table 5-3 shows that all the DDK audio samples from control speakers were accurately segmented and their corresponding mean DDK rates correctly match the manually measured DDK rates resulting in a mean DDK rate accuracy of 100%. Automatic segmentation and analysis of the DDK audio signals from Parkinson's disease, ataxic dysarthria and young speaker

groups resulted in mean DDK rate accuracies of 98.91%, 96.88% and 88.46% respectively. As shown in Table 5-3, the automatic DDK analysis algorithm gave a good performance with an average accuracy of 97.18% across the 4 speaker groups.

Table 5-3. Performance of the Automatic DDK Analysis Algorithm on Mean DDK Rate Estimation

Speaker Group	Size	Correct	Incorrect	Total	%Accuracy
PD	23	91	1	92	98.91
AD	8	31	1	32	96.88
Control	27	108	0	108	100.00
Young	13	46	6	52	88.46
All	**71**	**276**	**8**	**284**	**97.18**

To verify the reliability of the proposed algorithm, the segmentation thresholds for the same set of DDK audio samples were varied. The threshold for each audio sample was first estimated as the moving average of the peak intensities. Thereafter, two other thresholds that are 2 dB higher or 2 dB lower than the moving average threshold were used. This choice of ±2 dB was made because research [217] has shown that control speakers often vary their peak intensities by about 2 dB during DDK tasks. It is expected that the automatic DDK segmentation algorithm should still perform reliably if the threshold is increased by 2dB or decreased by 2dB.

The reliability of the algorithm was measured by comparing the mean DDKavrate, DDKavpi, DDKavd, stDDKrate, and stDDKpi for these two thresholds. The outcomes of this reliability test are shown in Table 5-4. Comparing the mean and the standard deviation values of the variables at +2 dB and -2 dB threshold show that varying the threshold by ±2 dB has little or no effect on the performance of the proposed algorithm. The resultant mean of the parameters and their corresponding standard deviation for the two thresholds are largely comparable. In addition, the mean and standard deviation of the absolute difference between the results from the two thresholds were calculated as shown in Table 5-4. These values are very low compared to the actual values in columns 2 and 3 of this table. Very low mean and standard deviation of the absolute difference between the two measurements recorded for the five variables shows that the results from the two threshold values are highly comparable.

Table 5-4 Outcomes of DDK Analysis Variables at Thresholds of +2 dB and -2 dB of the estimated moving average threshold

Variable	Mean ± STD		
	+2 dB	-2 dB	Absolute Difference
DDKavrate	5.59 ± 1.05	5.59 ± 1.04	0.09 ± 0.19
DDKavpi	51.98 ± 14.58	51.69 ± 14.63	0.45 ± 1.44
DDKavd	0.20 ± 0.07	0.20 ± 0.05	0.006 ± 0.038
stDDKrate	1.08 ± 0.44	1.08 ± 0.44	0.07 ± 0.12
stDDKpi	6.17 ± 2.71	6.62 ± 3.10	0.52 ± 1.08

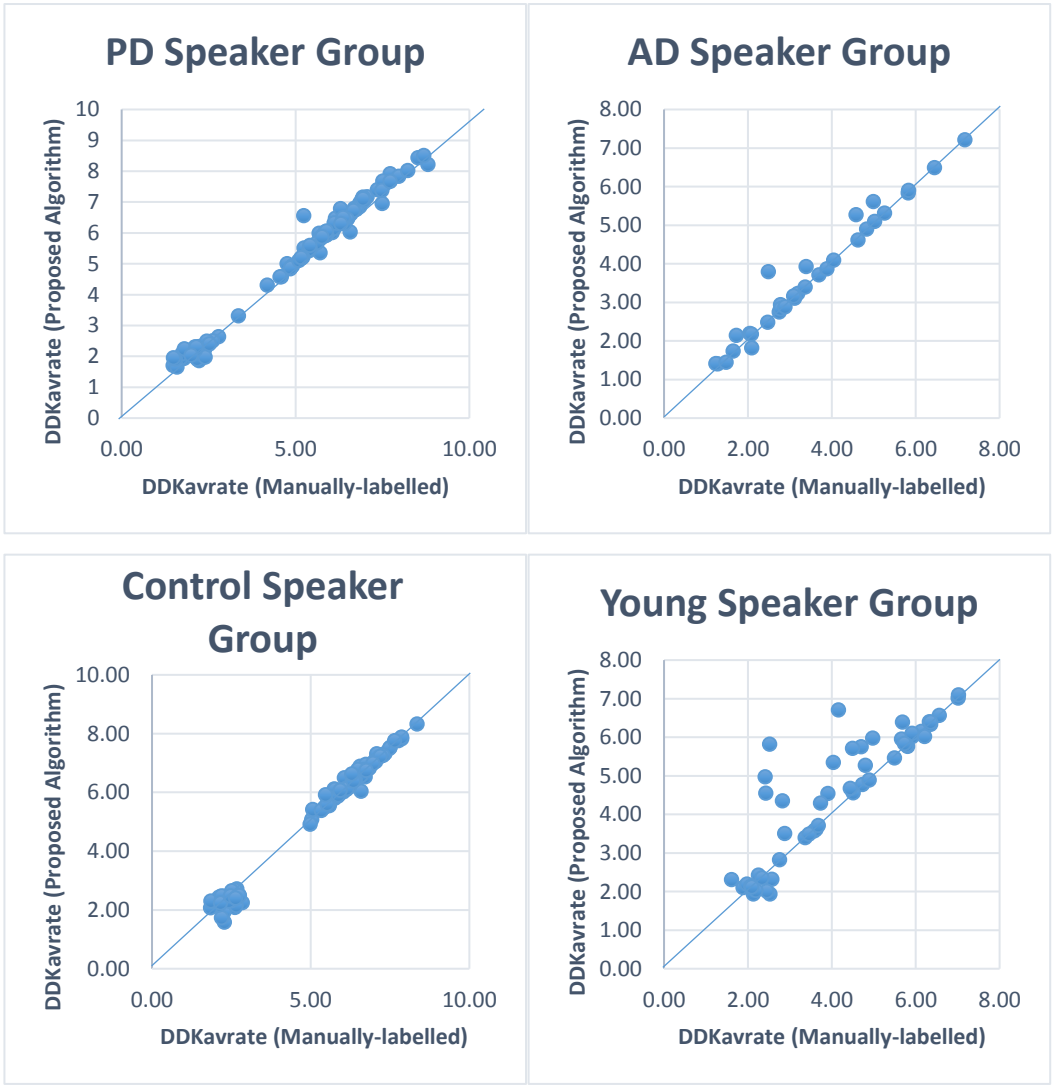


Figure 5-7. Scatter Plots Validating the Performance of the Proposed Automatic Algorithm across the Four Speaker Groups

To further examine the proposed algorithm, the resultant DDKavrate and DDKavd from the proposed algorithm were compared with those estimated from the manual-

labelling. The comparisons were carried using scatter plots for individual speaker groups as shown in Figure 5-7.

The results in the proposed automatic algorithm and the manually-labelled measurements are more consistent in the control speaker group, PD speaker group and AD speaker group with correlation coefficients of 1.00, 0.99 and 0.98 respectively. The correlation coefficient in the Young speaker group was, however, 0.89 due to varying DDK syllable distance in young speaker group. The DDK syllable distance in the young speaker group varied from 100 to 410 milliseconds which differs from the expected range of 100 to 200 milliseconds [217].

5.6 Discussion

This study has addressed the major limitations in the segmenting DDK syllables such as reduced peak intensity due to articulatory breakdown, over-segmentation due to inter syllable pauses, the dips between the consonant and vowel segments and under-segmentation due to the inappropriate choice of the threshold. In this study, two speaker-specific threshold approaches have been examined. Although the mean threshold approach works well for DDK samples from control speakers, its performance is reduced in dysarthric speakers. The use of a speaker-specific moving average threshold approach improved the performance of the algorithm, especially when analysing disordered speech. The moving average threshold ensures that the threshold value is adjusted as the intensities vary within the audio signal. An averaging window of about half a second is used to ensure that the averaging is carried out across multiple syllables.

Average accuracy of 97.8% has been achieved across the four different speaker groups. The accuracy was estimated with a minimum difference of 1 syllable per second. One of the factors that contributed to high accuracy is the ability to accurately detect the boundaries between consecutive syllables thereby making the segmentation process less prone to errors. The use of speaker-specific moving average threshold also contributed to the ability to detect the boundaries between two consecutive syllables.

Moreover, the reliability test carried out using two threshold values 2 dB greater than or less than the estimated threshold shows satisfactory reliability across five

measured parameters as shown in Table 5-4. With very low mean absolute difference between the two measurements, the high performance was maintained. This reliability test also reveals that the algorithm will still give a satisfactory performance when the average intensity is varied by 2 dB (as experienced in audio samples from control speakers).

Furthermore, the algorithm reduces errors due to inter syllable pauses and dips between the consonant and the vowel sounds. Dysarthric speakers often add inter syllable pauses during DDK syllable production leading to over-segmentation; thereby resulting in the extraction of pseudo-syllables. The proposed algorithm uses the minimum duration merging method to merge multiple consecutive pseudo-syllables to form syllables; thereby reducing over-segmentation.

The proposed automatic DDK analysis technique will be useful to clinicians in the assessment of oro-motor skills of dysarthria speakers. Apart from assessing the ability to produce repetitive movements, the proposed technique can potentially be used in other speech processing applications such as syllable segmentation and speech recognition in dysarthric speech.

5.7 Summary

In this chapter, a novel algorithm for the automatic segmentation and analysis of DDK audio samples has been presented. This automatic algorithm is designed to support clinicians in assessing and analysing DDK audio samples from dysarthric speakers. This algorithm uses a speaker-specific moving average threshold approach to segment DDK audio signals into syllables. It also uses minimum duration criteria to merge multiple pseudo-syllables previously over-segmented. This technique helps in reducing errors due to high variability in intensity, errors due to articulatory breakdowns experienced by dysarthric speakers, errors due to inter-syllable pauses and under-segmentation errors due to inappropriate thresholding. Experimental results show that the moving average thresholding coupled with minimum duration merging leads to higher segmentation accuracy when compared with manually labelled syllables. The reliability of this automatic DDK analysis algorithm has been tested and verified using a database comprising of 71 speakers.

Chapter 6

6 Novel Automatic Detection and Severity Classification of Dysarthric Speech

6.1 Introduction

In this chapter, novel algorithms for automatic detection and classification of dysarthria into various severity levels will be presented. The first novel algorithm presented will be designed to analyse and detect ataxic dysarthria in speech using an extended speech feature referred to as Centroid Formants combined with neural networks classification technique. Then a more robust dysarthria detection algorithm will be presented. This robust detection algorithm is based on a feature vector consisting of 29 speech features. These features are selected and extracted based on the prosodic, phonetic and vocal quality characteristics of the dysarthric speech. The performance of this fully automated detection algorithm will be examined using various machine learning techniques. Finally, a novel automatic classification technique for classifying dysarthric speech in to three severity levels using various machine learning techniques will be presented. These techniques were validated using a dataset from ataxic dysarthria speakers and gender and age-matched healthy control speakers.

6.2 Corpus

The dataset used for this study consists of 1400 audio samples from 20 speakers, 10 of which are ataxic dysarthric (AD) speakers and 10 gender-matched healthy control speakers. Each group consists of 5 males and 5 females. These audio samples consist of single words and sentences from the two speaker groups. Each speaker produced 20 single words and 50 sentences. This corpus was taken from the dataset reported by [23]. The ataxic dysarthric speakers have no cognitive deficiency neither do they have any visual and hearing impairment. The severity of the ataxic dysarthric speakers varied from mild to severe cases as illustrated in

Table 6-1. In addition, all of them were monolingual speakers of Standard Southern British English or Standard Scottish English.

Table 6-1. Details of AD participants involved in the study

Participant	Age	Gender	Etiology	Intelligibility Score (%)
AT_01	46	M	CA	74
AT_02	60	F	CA	67
AT_03	28	M	FA	6
AT_04	52	F	CA	25
AT_05	28	F	FA	9
AT_06	65	F	SCA6	58
AT_07	72	M	CA	19
AT_08	51	M	CA	44
AT_09	56	M	SCA8	82
AT_10	57	F	FA	80

CA: Cerebellar ataxia of undefined type, FA: Friedreich's ataxia and SCA: spinocerebellar ataxia

The intelligibility scores for the ataxic dysarthric speakers varied from 6 to 82 as shown in Table 6-1. These intelligibility scores were estimated from the average scores from five trained listeners during a passage reading task as presented in [23]. The etiologies of these participants are either cerebellar ataxia (50%), Friedreich's ataxia (30%) or spinocerebellar ataxia (20%). The available dataset from this corpus consists of 6 different types of audio samples namely; limericks, oro-motors, reading, sentences, single words and story retell, among them only sentences and single words are used in this study since they are the most relevant to the features to be extracted. This dataset was used for training, testing and validation of the three novel techniques presented in this chapter.

6.3 Automatic Detection of Ataxic Dysarthria using Extended Feature

6.3.1 Methodology

The block diagram of the proposed algorithm is illustrated in Figure 6-1. The first stage is pre-processing followed by feature extraction. After extraction of the speech features, a two-class classification is carried out using neural networks classification methods.

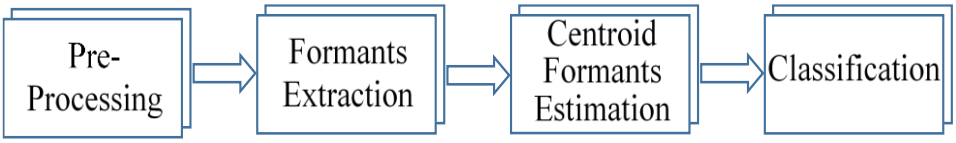


Figure 6-1: Block Diagram of the Proposed Algorithm

6.3.1.1 Pre-processing

Audio signals are not stationary in nature and thus it is essential to analyse these signals in short time intervals by dividing the audio signals into uniform short interval frames. The resulting amplitude normalised audio signals sampled at 16 kHz are divided into overlapping frames of 256 samples each with 80% overlap between consecutive frames. The overlapping is used to improve the segmentation process.

6.3.1.2 Formants Extraction

The formant extraction algorithm, in this proposed technique, is based on the Linear Prediction Coding (LPC) analysis, which gives a smoothed approximation of the power spectrum of the original signal [13]. The formant extraction process is described in Section 3.4.3 and the order he linear prediction function is given by (12). Figure 6-2 and Figure 6-3 show the formants for the word defer extracted from ataxic dysarthric speech and health speech respectively. Although the two speakers pronounced the same word “defer”, their formants differ and the energy concentration in the frequency spectrum also differs. The formants and formants energy for various dysarthric speakers were compared with healthy speaker and the differences observed motivated the extraction of an extended feature called Centroid Formant.

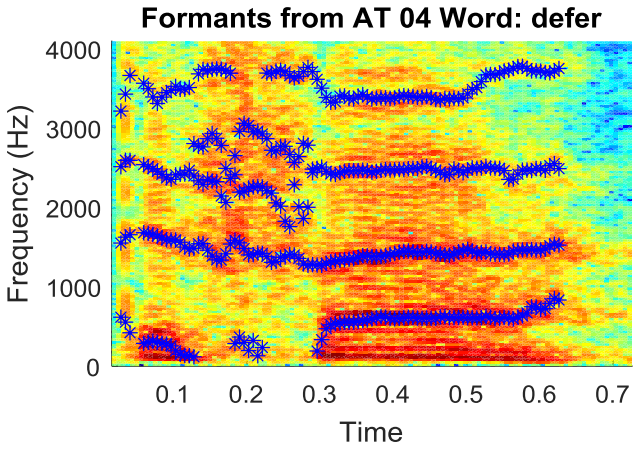


Figure 6-2. Formants extracted from Ataxic Dysarthric speech

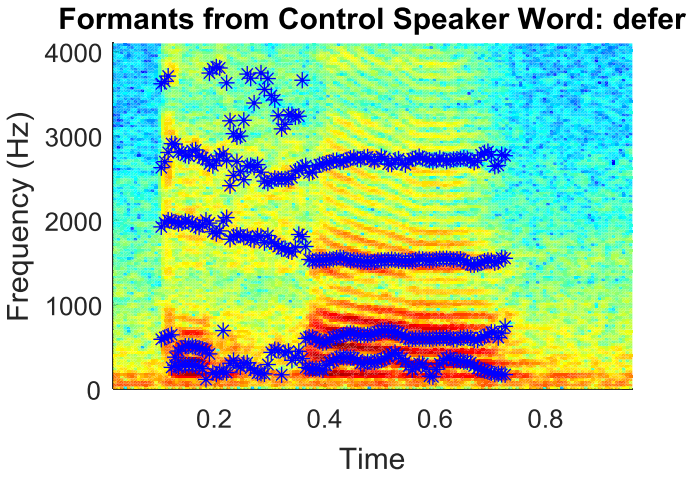


Figure 6-3. Formants extracted from Healthy Control speech

6.3.1.3 Centroid Formants Extraction

Centroid formants are the weighted averages of the formants in each frame in the short-time frequency spectrum. The formants are weighted by their corresponding energy, thereby resulting in a measure of where the power in the frequency spectrum of an audio signal is centralised. For instance, if the majority of the power in the spectrum resides in high-frequency components, then the centroid formant will lie in the high-frequency range. Figure 6-4 and Figure 6-5 illustrate the centroid formants of the audio files shown in Figure 6-2 and Figure 6-3 respectively for an ataxic speaker and a healthy speaker. Given that F1n, F2n, F3n, and F4n are the four formants of the nth frame of an audio signal and the corresponding formants energy are E1n, E2n, E3n and E4n respectively. The centroid formant of the nth frame is given by CFn as in (20).

$$CF_n = \frac{E_{1,n}F_{1,n} + E_{2,n}F_{2,n} + E_{3,n}F_{3,n} + E_{4,n}F_{4,n}}{4} \quad (20)$$

The centroid formant can be used to measure the rate of change of the formants and the intonation pattern of the audio signal. This is because as the formants change from low frequency to high frequency, the centroid formants also change in the same pattern. The weighting of the individual formants also ensures that frequency components with highest power contribution are given the highest weight. Therefore, the effects of picking weak peaks as formants will be reduced.

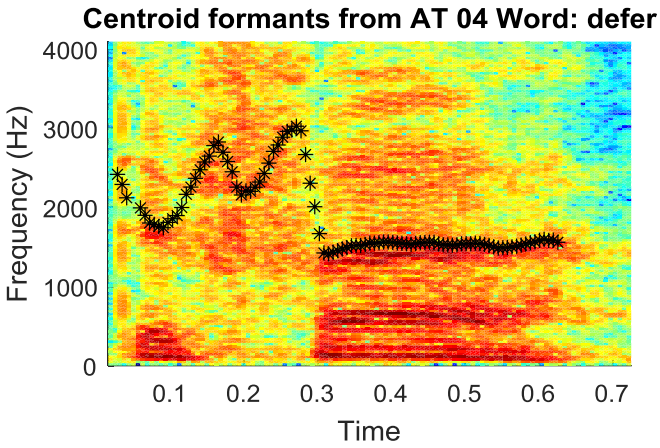


Figure 6-4. Centroid formants for AT speech

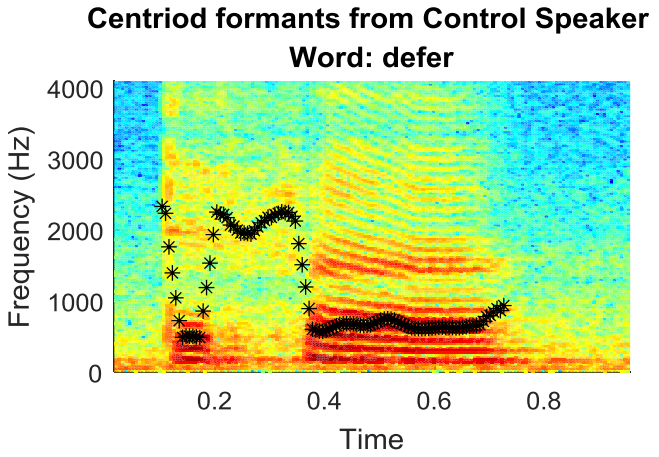


Figure 6-5. Centroid formants for healthy speech

In addition, there is a close relationship between pitch and centroid formants profiles for healthy control speech. If the energy contribution of each formant remains the same within a speech segment, the centroid formant will give a pattern similar to the fundamental frequency. However, this similarity with pitch profile does not apply for audio signals with rapidly changing intonation patterns, which is

the case in dysarthric speech. The centroid formant is very sensitive to the rapid changes in pitch and intonation which means that the high pitch variability in dysarthric speech can be effectively tracked using centroid formants, thereby making centroid formants a suitable feature to be explored for dysarthria detection.

6.3.1.4 Classification

Using the extracted centroid formants, the audio samples are classified into healthy control and dysarthric classes. One of the commonly used machine learning methods is ANN. This classification technique is robust and it combines pattern recognition with acoustic-phonetic methods [21]. In this ANN learning technique, knowledge of the acoustic and phonetic characteristics of the speech is used to generate rules for classifiers [22]. The classification was carried out using different settings of hidden layer neurons by varying the number of neurons (J) and finding the optimum J with the highest accuracy. The neural network classifier with one hidden layer and 10 neurons had the highest performance and was used for the classification. The excitations (inputs) are the centroid formants and the observations (outputs) are the binary signals indicating whether or not the corresponding audio sample is dysarthric (0) or healthy (1).

The performance of the chosen classifier was analysed using the confusion matrix which is a tabular representation of the correctness of the outputs of the classifier when compared with the targets (outputs versus targets) as shown in Table 6-2.

Table 6-2. Confusion Matrix Relationship between Output and Target Classes

Confusion Matrix		Target Class: Negative	Target Class: Positive	
Output Class	Negative	True Negative	False Negative	Neg. Prediction $\frac{TP}{TP+FP}$
	Positive	False Positive	True Positive	Precision $\frac{TP}{TP+FP}$
		Specificity $\frac{TN}{TN+FP}$	Sensitivity $\frac{TP}{TP+FN}$	Accuracy $\frac{TP+TN}{N}$

In the confusion matrix, the number of audio signals whose output classes match the expected target classes are indicated in green (true positives, TP, and true negatives, TN) and the number of data set whose output classes do not match the expected target classes are indicated in red (false positives, FP and false negatives, FN). The accuracy of the classifier is calculated as the percentage of the correctly classified data set as indicated in Table 6-3. The other classification performance parameters include specificity, sensitivity and precision as described in Table 6-3.

Table 6-3. Description of Classification Parameters

Parameters	Description	Annotation
Positive	Dysarthric audio samples	1
Negative	Healthy control audio samples	0
Total	Total number of audio samples	N
True Positive	The number of audio samples correctly classified as dysarthric	TP
True Negative	The number of audio samples correctly classified as healthy control	TN
False Positive	The number of audio samples incorrectly classified as dysarthric	FP
False Negative	The number of audio samples incorrectly classified as healthy control	FN
Accuracy	The measure of the ability of the classifier to correctly classify all audio samples	$\frac{TP+TN}{N}$
Specificity	The measure of the ability of the classifier to correctly classify healthy control audio samples	$\frac{TN}{TN+FP}$
Sensitivity	The measure of the ability of the classifier to correctly classify dysarthric audio samples	$\frac{TP}{TP+FN}$
Precision	The measure of the exactness of the classifier in identifying dysarthric audio samples	$\frac{TP}{TP+FP}$

6.3.2 Experimental Results

The classification was carried out using the Neural Network Toolbox in MATLAB software. The audio samples were distributed randomly as follows; 70% of the audio samples were used for training, 15% for testing and 15% for validation. The confusion matrix of the classifier is illustrated in Figure 6-6.

Figure 6-6. Confusion Matrix for the Outputs of The ANN Classifier

Even though a single hidden layer has been used for this classification, the overall accuracy recorded was 75.6% using 10 neurons. The confusion matrix for the trained neural network is illustrated in Figure 6-6. The first two columns of the confusion matrixes indicate the two target classes (0 for healthy speech and 1 for dysarthric speech). Likewise, the first two rows of the confusion matrixes show the two output classes (0 or 1) whereas the third row shows the sensitivity, specificity and accuracy of the network respectively. The third column represents the negative predictive value, precision and accuracy of the network respectively. The training dataset gives an accuracy of 74.3%, the validation dataset gives an accuracy of

80.3%, and whereas the test data set gives an accuracy of 77.0% bringing the total accuracy to 75.6%.

6.3.3 Discussion

The application of an extended speech feature in the classification of dysarthric speech from healthy speech using neural networks has been explored in this section. The extended feature, called centroid formants, proposed in this study resulted in an accuracy of 75.6% with just one hidden layer and 10 neurons. This classification has been carried out across different levels of severity of ataxic dysarthria from mild to highly severe cases. Classification using other artificial intelligence techniques such as Deep Neural Networks (DNN), Support Vector Machine (SVM), LQV and Hidden Markov model, however, needs to be explored. In addition, this study opens up new research opportunities for the application of the centroid formants in speaker identification, speech recognition and emotion detection in disordered speech. The performance of the detection algorithm can be improved by combining the centroid formants features with other spectral and cepstral features; which can also be extended to other classification applications. This will be explored in the next section, where a more robust dysarthria detection algorithm which makes use of centroid formants and other relevant speech features will be presented.

6.4 Novel Robust Automatic Dysarthria Detection Algorithm

Presented in this section is a novel automatic dysarthria detection algorithm that models dysarthric speech using prosodic, voice quality, phonetic and wavelet features. These features are selected based on the characteristic differences between dysarthric and healthy control speech samples. This algorithm consists of four stages namely; pre-processing, acoustic analysis, feature vector design, and classification. The pre-processing stage involves techniques used in enhancing the speech samples in preparation for the acoustic analysis stage. The speech signals are segmented into silence, unvoiced and voiced parts in the pre-processing stage. The second stage involves modelling of the speech signals by extracting features that describe the prosody, phonation, voice quality, and articulation characteristics of the speech signals. The feature vector design stage involves the combination of these extracted features to form a feature vector that is suitable for the classification

of the speech samples into dysarthric and healthy control classes. Finally, in the classification stage, multiple classification techniques are used, and their performances are compared using four performance parameters namely; accuracy, sensitivity, precision, and specificity.

6.4.1 Pre-processing

The speech pre-processing stage comprises of resampling of the audio signal, amplitude normalisation, pre-speech and post-speech silence removal and SUV segmentation. All the audio signals are resampled at 16 kHz ensuring that they are not under-sampled with respect to Nyquist criteria as discussed in Section 3.2.5. After resampling, the amplitudes of the audio signal are normalised to keep the values between +1.0 and -1.0. Subsequently, the silence segments before and after the speech (that is, pre-speech silence and post-speech silence) are removed using a cut-off threshold of at least 10% of the maximum signal amplitude. This is followed by segmentation of the audio signal into silence, unvoiced and voiced parts. The SUV segmentation is carried out using the novel three-fold (STE, LPEV and ZCR) segmentation approach presented in Chapter 4.

6.4.2 Acoustic Analysis

The acoustic analysis stage aims at extracting important and useful information from the speech signals that can adequately distinguish dysarthric speech from healthy control speech whilst discarding the less relevant information. This acoustic analysis process makes use of multiple features in an aim to consolidate any speech information that could be missing in a single feature. This makes the system more robust and reduces the classification error. As discussed in Chapters 2 and 3, there are features that differentiate disordered speech signals from healthy control speech signals. These features are grouped under prosody, voice quality, pronunciation, and wavelet analysis subheadings. It is important to note that at the start of this study, over 50 features were identified but the relevance of each feature was tested and the features with the highest classification accuracies were selected for this study. The choice of selected features is also based on the characteristics of the dysarthric speech as well as the speech subsystem to be modelled.

6.4.2.1 Prosody Analysis

Prosody analysis involves the extraction of duration, fundamental frequency (F0) and intensity-based features from speech signals. It, however, requires the speech signals to first be segmented into silence, unvoiced and voiced parts as the prosodic features are extracted from the voiced segments of the speech signals. As discussed in Section 2.2, the characteristics of the prosodic features extracted from the dysarthric speech are different from those extracted from the controlled speech in terms of value (or amplitude), range and variability. These differences are, therefore, explored in distinguishing between dysarthric speech samples and healthy control speech samples. The features extracted under the fundamental frequency subset include the mean and standard deviation of F0 whereas the features extracted under the intensity subset include the peak amplitude and mean amplitude.

6.4.2.2 Voice Quality Analysis

One of the characteristics of dysarthric speech is hoarse voice quality. A review of literature presented in Section 3.5.3 has shown that the harmonic to noise ratio (HNR) is one of the speech features that describe the voice quality of a speaker. Amplitude and frequency perturbation (shimmer and jitter) have also been used in recent studies to model speakers' voice quality. These features (HNR, shimmer and jitter) have been popularly used to assess the voice quality in speech disorders using sustained vowel sounds. In this study, these features will be extracted from single words and sentences. To model the differences, in voice quality, between dysarthric speakers and healthy control speakers, six (6) voice quality features are extracted from the audio signals which form the voice quality subset in the feature vector: mean jitter, mean shimmer and HNR at cut-off frequencies of 500Hz, 1500Hz, 2500Hz and 3500Hz.

6.4.2.3 Pronunciation Analysis

Control and coordination difficulties often lead to pronunciation errors which also contributes to reduced intelligibility in dysarthria. A review of previous studies presented in Section 3.4.4 shows that Mel frequency cepstral coefficients (MFCCs) can be used to model the pronunciation variations in speech. The Mel-scale, which maps a linear frequency scale to a non-linear frequency scale, is based on the

auditory perception of human which is a function of the pronunciation. In the estimation of the short-time MFCCs, hamming windows that are 16ms (256 samples at a sampling rate of 16 kHz) in length with 75% overlap are used. This frame size gives a good balance between resolution quality and complexity of extracted features. The MFCCs are extracted using twelve triangular frequency banks. The mean MFCCs are estimated for each frequency bank across all frames resulting in 12 features per utterance.

Moreover, another set of features used in modelling pronunciation in speech signals are the centroid formants. As discussed in Section 6.3.1.3, the centroid formants are the weighted average of the first four formants in each frame. As the pronunciation varies, the centroid formants also vary in value. This variation makes centroid formants useful in detecting dysarthria in speech. In modelling the pronunciation characteristics, the peak centroid formants and mean centroid formants are estimated for each utterance bringing the total number of pronunciation features to 14 features per utterance.

6.4.2.4 Wavelet Analysis

The last group of features used in this study are wavelets features. Although wavelet analysis has been applied to various speech processing applications such as emotion detection and speech separation, its application in automatic detection of dysarthria has not yet been explored. As presented in Section 3.5.4, the percentage residual energy of dysarthric speech samples after the wavelet decomposition at every level varies considerably from that of healthy control speech samples. Although these variations occur at every level of decomposition, they are more pronounced at the lower levels. In this study, four-level wavelet analysis is carried out on all the utterances resulting in five wavelet features. These include the total energies in level 4 approximation, level 1 detail, level 2 detail, level 3 detail and level 4 detail signals.

6.4.3 Design of Feature Vector

The feature vector consists of four feature blocks namely; prosody feature block, voice quality feature block, pronunciation feature block, and wavelet feature block as illustrated in Table 6-4. These feature blocks contain 4, 6, 14 and 5 features

respectively. The description of the features in each feature block is also presented in Table 6-4.

Table 6-4. Summary of the Extracted Features for the Automatic Detection of Dysarthria

S/N	Feature Block	Feature	Description
1.	Prosody	*meanF0*	Mean fundamental frequency
		stdF0	The standard deviation of the fundamental frequency
		meanInt	Mean intensity
		peakInt	Peak intensity
2.	Voice Quality	*meanJit*	Mean jitter
		meanShim	Mean shimmer
		HNRs	Harmonic to noise ratio at cut-off frequencies of 500 Hz, 1500 Hz, 2500 Hz and 3500 Hz
3.	Pronunciation	*meanCF*	Mean centroid formant
		peakCF	Peak centroid formant
		MFCCs	Mel-frequency cepstral coefficients using 12 triangular frequency banks
4.	Wavelets	*ED1*	The energy of level 1 detail signal
		ED2	The energy of level 2 detail signal
		ED3	The energy of level 3 detail signal
		ED4	The energy of level 4 detail signal
		EA1	The energy of level 4 approximate signal

Under the prosody feature block, the extracted features per utterance include mean fundamental frequency, the standard deviation of the fundamental frequency, peak intensity and mean intensity. The mean jitter, mean shimmer, and harmonic to noise ratio at cut off frequencies 500 Hz, 1500 Hz, 2500 Hz and 3500 Hz are the extracted features under the voice quality feature block. The features extracted under the pronunciation feature block are mean centroid formant, peak centroid formant and twelve Mel-frequency cepstral coefficients.

Finally, the features extracted under the wavelet feature block include the energy of the level 4 approximate signal as well as energies of level 1, 2, 3, and 4 detail signals. This makes up a total of 29 features in the feature vector as shown in Table 6-4. The prosody and voice quality features are extracted from the voiced segments of the audio signals whereas the pronunciation features are extracted from the unvoiced and voiced segments. The wavelet features, on the other hand, are extracted from the whole signal since the focus is on the total energy of the detail signal at every level and total energy of the approximate signal at level 4 of the wavelet analysis.

6.4.4 Classification and Experimental Results

As discussed extensively in Section 3.7, researchers have made used of various classification techniques in detecting abnormalities in speech. Although only a few of these techniques have been applied to automatic detection of dysarthria, some of them have shown promising results in the classification of pathological disorders with accuracies ranging from 62% using statistical methods in [74] to 97 % using support vector machines in [73]. In this study, the performances of variants of six classification techniques were investigated. These classification techniques include decision tree, discriminant analysis, logistics regression, support vector machines, k-nearest neighbours, and neural networks classifiers. The application of multiple classifiers is achieved using the classification learner tool in MATLAB. This tool is advantageous when comparing the performance of more than two classifiers using a single input/output dataset. In the classification learner tool, multiple classifications are carried out simultaneously and the results are compared using its interactive user interface. The tool comprises of 23 different classifiers which are variants of the decision tree, discriminant analysis, logistics regression, support vector machines, and k-nearest neighbours classifiers. The neural networks classification, on the other hand, is carried out using the dedicated neural network tool in MATLAB. This tool allows the dataset to be classified using networks of a varying number of hidden layers and neurons.

The performance of the proposed automatic dysarthric detection technique was examined in an experiment involving multiple classifiers. The experiment was carried out on 1329 audio samples from both ataxic dysarthria and healthy control

speaker groups. The participants in the dysarthric speaker group are as described in Table 6-1. Out of the initial 1400 recorded audio samples for this study, 71 audio samples were removed due to signal distortion, recording errors or/and incomplete recordings. The 29 features described in Section 6.4.3 are extracted from each of the audio signals resulting in an input data of dimension 1329 x 29. The target data contains binary information of which 1 signifies the dysarthric speech sample and 0 signifies healthy control speech sample. For classification purposes, 70% of the dataset was used for training, 15% for validation (in order to prevent generalization) and the remaining 15% for testing. The classification was carried out in two stages. The first stage involves the use of NN Tool in MATLAB for classification and the second stage involves the use of classification learner tool, also in MATLAB, for the other 23 classifiers.

In analysing the performance of the classifiers, four parameters were used. These parameters are accuracy, specificity, sensitivity and precision described in Table 6-3. The ability of the classifier to accurately identify the class of all audio samples belonging to is termed accuracy and the ability of the classifier to accurately detect dysarthric audio samples is termed specificity. Whereas, sensitivity and precision are the measures of the ability to correctly classify dysarthric audio samples and the exactness of the classifier respectively.

In the first stage of the classification, single-layer neural networks with a varying number of neurons were used in classifying the input data into two classes (dysarthric and healthy control). The number of neurons in the hidden layer was varied from 2 to 20 in steps of 2 to get the optimum number of neurons for the classification problem, as illustrated in Figure 6-7. Optimal performance was reached when the number of neurons in the hidden layer was 10 with an accuracy of 99.4%. The confusion matrix of the trained neural network with 10 hidden neurons is presented in Figure 6-8. This trained neural network also resulted in training dataset accuracy of 99.9%, validation dataset accuracy of 97.4% and the test dataset accuracy of 99.0%.

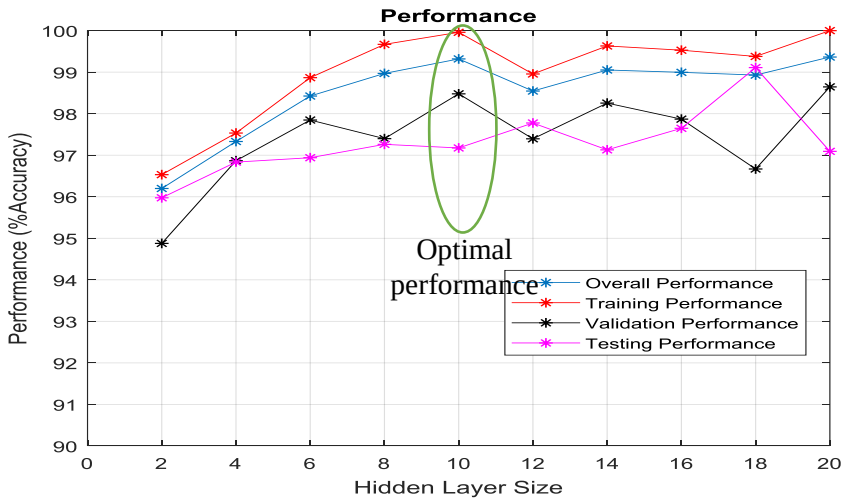


Figure 6-7. Accuracy of the Neural Network Classifiers with One Hidden Layer and Varying Number of Neurons

Figure 6-8. Confusion Matrix of the Trained Single-layer Neural Network with 10 Neurons for the Automatic Detection of Dysarthria

In the second stage with the use of the classification learner, 23 different classifiers were used to classify the audio samples into dysarthric and healthy control classes. The performances of the 23 classifiers were examined by comparing their accuracies as illustrated in Table 6-5.

Table 6-5. Performance of the Various Classification Techniques used

Decision Tree		**Support Vector Machine**		**k-Nearest Neighbours**		**Discriminant Analysis**		**Logistics Regression**	
Classifier	Accuracy	Classifier	Accuracy	Classifier	Accuracy	Classifier	Accuracy	Classifier	Accuracy
Simple Tree	87.1%	Linear SVM	95.4%	Fine kNN	99.0%	Linear Discriminant	79.2%	Logistics Regression	95.0%
Medium Tree	91.7%	Quadratic SVM	98.3%	Medium kNN	97.7%	Quadratic Discriminant	80.1%		
Complex Tree	92.5%	Cubic SVM	98.6%	Coarse kNN	91.5%	Subspace Discriminant	93.4%		
Boosted Tree	96.7%	Fine Gaussian	85.3%	Cosine kNN	97.7%				
Bagged Trees	97.6%	Medium Gaussian	98.5%	Cubic kNN	97.1%				
RUS Boosted Trees	92.9%	Coarse SVM	94.0%	Weighted kNN	98.0%				
				Subspace kNN	72.6%				

The variants of the decision tree classifiers used are the simple tree, medium tree, complex tree, boosted tree, bagged tree, and RUS boosted tree classifiers. The variants of the discriminant analysis classifiers include linear discriminant, quadratic discriminant and subspace discriminant classifiers. The variants of the SVM classifiers include linear, quadratic cubic, fine Gaussian, medium Gaussian and coarse SVM classifiers. Whereas the variants of the kNN classifiers include fine, medium, coarse, cosine, cubic, weighted and subspace kNN classifiers. In addition to these classifiers, the logistic regression-based classifier was also used as shown in Table 6-5.

As presented in Table 6-5, each of the six variants of the decision tree classifiers used gave an accuracy of over 90% except the simple tree classifier. The bagged tree classifier resulted in the highest accuracy of 97.6% in the group. Furthermore, the SVM classifiers resulted in accuracies greater than 90% except for the fine Gaussian classifier with an accuracy of 85.3%. In the SVM group, the medium Gaussian classifier resulted in the highest accuracy of 98.5%. Within the kNN classifier group, the subspace kNN classifier resulted in the least accuracy of 72.6% and the fine kNN classifier resulted in the highest accuracy of 99.0%. Statistical discriminant analysis using linear, quadratic and subspace discriminant analysis resulted in accuracies of 79.2%, 80.1% and 93.4% respectively whereas the logistics regression analysis resulted in an accuracy of 95.0%. Out of these 23 variants of classifiers, the fine kNN classifier resulted in the highest accuracy of 99.0%. Comparing this with the accuracy of the neural network with single hidden layer and 10 neurons, whose confusion matrix is shown in Figure 6-8, shows that the neural network gave the highest total accuracy of 99.4% with only 8 incorrect classifications out of 1339 audio samples.

Furthermore, other performance indices of the single-layer neural network classifier with 10 neurons are presented in Table 6-6. The specificity of the classifier was 100% for training samples, 97.1% for validation samples and 99.0% for test samples. The classifier also resulted in high precision of 100%, 96.8% and 98.9% for training, validation and test audio samples respectively. Although the values of the four performance parameters varied across training, validation and test samples, the value of the assessment parameters for all audio samples was 99.4% each.

Table 6-6. Performance Indices of the Single-Layer Neural Network with 10 Hidden Neurons for the Automatic Detection of Dysarthria

Performance Index	Training Samples	Validation Samples	Test Samples	All Samples
Specificity	100%	97.1%	99.0%	99.4%
Sensitivity	99.8%	97.9%	98.9%	99.4%
Precision	100%	96.8%	98.9%	99.4%
Accuracy	99.9%	97.4%	99.0%	99.4%

6.4.5 Discussion

The automatic detection technique presented in this study is more robust than the one presented in Section 6.3 as it makes use of prosody, pronunciation, voice quality and wavelets features. Unlike in other techniques, this proposed method looks at the features that model the characteristics of dysarthria speech across multiple dimensions. The features were selected based on an initial analysis of the effects of the individual features in the detection of dysarthria in speech. The 29 features extracted in this study were analysed individually to verify their relevance in the automatic detection problem. Although different speech subsystems were considered, the most relevant features were chosen to accurately model the differences between dysarthric and healthy speech samples.

Under the prosody subsystem, the most varying features in dysarthric speech are the fundamental frequency and intensity. Dysarthric speech is characterised by high variability and reduced fundamental frequency, therefore the mean and standard deviation of the fundamental frequency were extracted. In addition, reduced loudness is often experienced in the dysarthric speech which informed the choice of mean and peak intensity features for the classification of dysarthria.

Moreover, the selected voice quality features are based on the amplitude and frequency perturbation as well as the harmonic to noise ratio. An initial study reveals that the harmonic to noise in healthy control audio samples are greater than those in dysarthric audio samples. Furthermore, the mean and peak of the weighted formants called centroid formants, as described in Section 6.3.1.3, are extracted

under the pronunciation subsystem. Based on recent research works in dysarthria detection, MFCCs are also selected under the pronunciation subsystems.

An analysis of wavelet energies reveals that more energy components are decomposed in lower level signal details (level 1 and 2) in dysarthric speech when compared with healthy control speech. The differences observed in the wavelet energies could be due to the high pitch and intensity variations experienced in dysarthric speech. These differences were explored by estimating the total energies in levels 1, 2, 3 and 4 detail signals as well as the level 4 approximate signal.

Moreover, the choice of features has contributed greatly to an increase in accuracy in the proposed technique. Although the same audio samples were used in the method presented in Section 6.3, the accuracy of the automatic detection algorithm has increased from 75.6% to 99.4% in the method proposed in this section. One of the factors that contributed to this increase is the extraction of features from multiple speech subsystems. Another factor is the novel inclusion of wavelet features in the feature vector. Although wavelets analysis has been applied in other speech processing applications, its use in automatic detection of dysarthria is novel and the results are promising which opens up more research opportunities in its application. Recent research has also shown that wavelets can be used to track pitch variations in healthy speakers [218].

The proposed algorithm has consistently shown improved performance across the four measured classification parameters when compared with techniques proposed in recent literature presented in Table 2-3. An accuracy of 99.4% indicates that incorrect classifications of audio samples are 3 in 500 audio samples. This can give clinicians a level of confidence when making decisions. Although the proposed algorithm has only been applied to the detection of ataxic dysarthria, its potential in detection of other types of dysarthria is promising as the selected features are based on the common characteristics of the various types of dysarthria.

Although deep learning has been a major topic of discussion amongst researchers in recent years, its necessity remains contentious. In deep learning, there exists a trade-off in complexity and performance. The necessity of deep learning is often questioned when the performance of the classification technique is already good enough as in the case in this study. Increasing the complexity of the algorithm

whose accuracy is already 99.4% needs to be justified. Another consideration is the size of the available data which is quite limited, in this study, for deep learning. With the availability of more training data, the application of deep learning may be considered in future work.

6.5 Automatic Severity Classification of Dysarthric Speech

6.5.1 Methodology

The application of the automatic detection technique proposed in Section 6.4 was further explored in the classification of the dysarthric speech into different severity levels. For the purpose of this research, the severity levels are described by the scale illustrated in Table 6-7 with respect to absolute intelligibility score. Severity level 0 was assigned to healthy control audio samples. For dysarthric audio samples, three severity levels (1, 2 and 3) were assigned. Severity level 1 was assigned to mild dysarthric with absolute intelligibility score ranging between 70% and 100%. Severity level 2 was assigned to moderate dysarthric audio samples with an absolute intelligibility score range between 40% and 69%. And severity level 3 was assigned to severe dysarthric with an absolute intelligibility score range between 0% and 39%. Using these ranges, three dysarthric speakers (AT_01, AT_09 and AT_10) were classified as mild, three (AT_02, AT_06 and AT_08) were classified as moderate and four (AT_03, AT_04, AT_05 and AT_07) as severe as illustrated in Table 6-8. These ranges make the severity levels well distributed across the dataset available for this study.

Table 6-7. Severity Level based on Intelligibility Score Range

Severity Level	Description	Absolute Intelligibility Score Range
0	Healthy	-
1	Mild Dysarthria	70% - 100%
2	Moderate Dysarthria	40% - 69%
3	Severe Dysarthria	0% - 39%

Table 6-8. Severity Classification of Participants involved in the Study

Participant	Intelligibility Score (%)	Assigned Severity Level
AT_01	74	1
AT_02	67	2
AT_03	6	3
AT_04	25	3
AT_05	9	3
AT_06	58	2
AT_07	19	3
AT_08	44	2
AT_09	82	1
AT_10	80	1

6.5.2 Experimental Results

An experiment was carried out on the 1329 audio samples previously used for the automatic detection of dysarthria. Again, 70% of the audio samples were randomly selected for training, 15% for validation and 15% for testing. For the neural network classification, single-layer classifiers with a different number of neurons were tested to find the optimal number of neurons that will result in improvement in performance as the number of neurons increases. The test was set-up by increasing the number of neurons from 2 to 20 in steps of 2. Optimal performance was reached when the number of neurons was increased to 12. Beyond this point, the performance did not increase significantly. Furthermore, the number of hidden layers was increased to 2 and different combinations of the number of neurons were tested. An analysis of the performance shows that increasing the number of hidden layers does not improve the performance of the classifier. The confusion matrix of the neural network classifier with one hidden layer and 12 neurons is shown in Figure 6-9. The performance parameters of the classifier are presented in Table 6-9. An accuracy of 99.5% was achieved for training samples, that of the validation

samples was 97.0% and the accuracy of test samples was 91.0%. This brings the percentage of accurately classified samples to 97.8%.

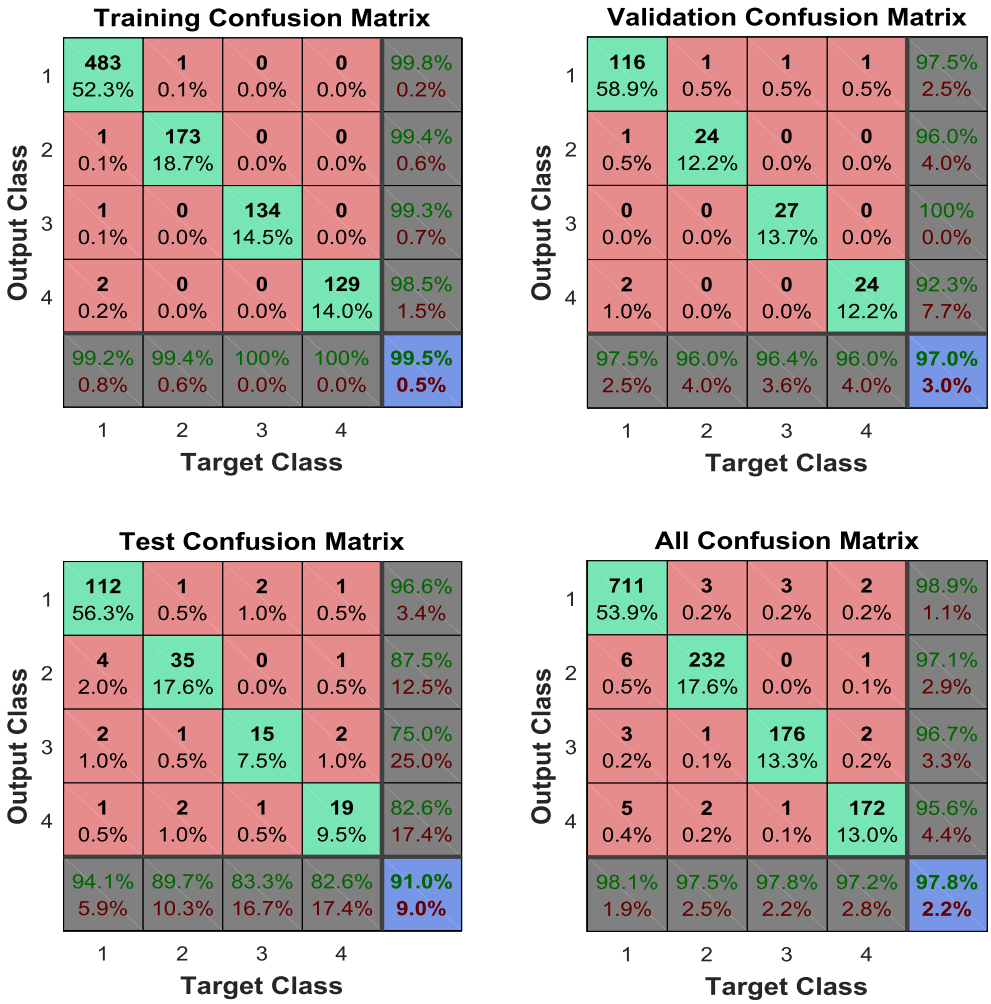


Figure 6-9. Confusion Matrix of the Trained Single-layer Neural Network with 12 Neurons for the Four-class Severity Classification of Dysarthria

Additionally, the ability of the classifier to accurately classify the healthy control audio samples called the specificity was tested. The specificity of the classifier was 99.2% in the training samples, 97.5% in the validation samples and 94.1% in the test samples. The overall specificity was measured to be 98.1% which shows that healthy control speakers are less likely to be classified as dysarthric. The overall sensitivity and precision of the classifier were 97.5% and 96.5% respectively.

Table 6-9. Performance Indices of the Single-Layer Neural Network with 12 Hidden Neurons for Four-class Severity Classification of Dysarthria

Performance Index	Training Samples	Validation Samples	Test Samples	All Samples
Specificity	99.2%	97.5%	94.1%	98.1%
Sensitivity	99.8%	96.1%	86.2%	97.5%
Precision	99.1%	96.1%	83.1%	96.5%
Accuracy	99.5%	97.0%	91.0%	97.8%

6.5.3 Discussion

The same set of features used for the automatic detection of dysarthria were also used for the four-class severity classification, the performance of the classifier showed high consistency across the training, validation and test audio samples. The results of the proposed technique also showed an improvement in performance when compared with other studies on multi-class severity classification techniques presented in [46, 73, 79].

To further investigate the performance of other types of classifiers the *Classification Learner tool* was used. As stated in Section 6.4.4, the classification learner tool consists of 23 classifiers which are variants of decision tree, support vector machines, k-nearest neighbours, discriminant analysis and logistics regression-based classifiers. The results of the classification learner tool are presented in Table 6-10. Out of the 23 classifiers used, 12 classifiers gave an accuracy greater than 90%. The quadratic SVM classifier gave the highest accuracy of 95.6%. Closely followed is the fine kNN classifier with an accuracy of 95.5%. Although these two classifiers resulted in high classification performance, the neural network classifier gave a better performance as illustrated in Figure 6-9 and Table 6-9. The neural network classifier was therefore chosen as the best fit for this severity classification problem.

Table 6-10. Performance of the Various Classifiers used for the 4-class Severity Classification

Decision Tree		**Support Vector Machine**		**k-Nearest Neighbours**		**Discriminant Analysis**		**Logistics Regression**	
Classifier	Accuracy	Classifier	Accuracy	Classifier	Accuracy	Classifier	Accuracy	Classifier	Accuracy
Simple Tree	70.1%	Linear SVM	91.9%	Fine kNN	95.5%	Linear Discriminant	69.3%		
Medium Tree	84.4%	Quadratic SVM	95.6%	Medium kNN	94.4%	Quadratic Discriminant	73.7%		
Complex Tree	85.7%	Cubic SVM	95.3%	Coarse kNN	76.4%	Subspace Discriminant	85.2%		
Boosted Tree	91.6%	Fine Gaussian	58.6%	Cosine kNN	94.7%				
Bagged Trees	94.7%	Medium Gaussian	95.4%	Cubic kNN	93.7%				
RUS Boosted Trees	91.1%	Coarse SVM	85.0%	Weighted kNN	95.0%				
				Subspace kNN	58.4%				

Furthermore, validation audio samples are used to measure the generalisation of the classifier and halt the training when the generalisation stops improving. With a validation accuracy of 97.0%, the neural network classifier performed very well in terms of generalisation. This is also evident during testing as the testing accuracy was 91.0%.

Analysis of the confusion matrix of the neural network classifier illustrated in Figure 6-9 shows that the majority of the incorrect classification was not as a result of inter-class error within the three severity classes but between the healthy control class and the severity classes. If this error can be reduced or eliminated, the performance of the classifier can be improved. One way to address this is by implementing the robust automatic detection algorithm proposed in Section 6.4 first and then perform a three-class severity classification on the audio samples classified as dysarthric. A three-class neural network-based classifier was set-up to investigate this approach.

In the three-class classifier, the dysarthric audio samples were classified into three classes representing mild, moderate and severe dysarthria respectively. The three class problem is a more realistic problem as clinicians will already know if the speakers are dysarthric or not before assess the severity of the speech disorder. Just as in the four-class classifier, the number of neurons in the hidden layer was 12. 70% of the dysarthric audio samples were used for training, 15% for validation and 15% for testing. The confusion matrix of the three-class classifier is shown in Figure 6-10. All the training and validation audio samples were accurately classified into the three severity classes with an accuracy of 100%. Only two audio samples within the test samples (one from the moderate class and the other from the severe class) were misclassified as mild. This brings the accuracy of the test samples to 97.8% and the overall accuracy to 99.7%. This shows that a two-level classification (automatic detection followed by severity classification) can improve the performance of the classifier.

The analysis of the performance of the proposed classification technique showed promising results. The proposed technique can potentially be useful in the classification of other neurological speech disorders. Its application in emotion detection and speech recognition in dysarthria also needs to be explored.

Figure 6-10. Confusion Matrix of the Trained Single-layer Neural Network with 12 Neurons for the Three-class Severity Classification of Dysarthria

6.6 Summary

In this chapter, three novel techniques for automatic dysarthria detection and severity classification have been presented. These techniques involve the use of spectral and cepstral features to distinguish between dysarthric and healthy control speech samples. The first method makes use of an extended speech feature called centroid formants to automatically classify ataxic dysarthria. The second method uses a more robust feature vector which consequently increases the accuracy from 75.6% in the first method to 99.4%. Finally, a neural networks-based severity classification algorithm is presented that classifies audio samples from dysarthria speakers into three severity levels. It results in the classification accuracy of 97.8% in four-class classification (including healthy control class) and 99.7% in three-class classification.

Chapter 7

7 Analysis of Stress Production Deficits in Dysarthric Speech for the Clinical Management of Dysarthria

7.1 Introduction

In this chapter, perceptual analysis of the utterances from ataxic dysarthria speakers during a stress production exercise will be presented. This will include stress marking exercise carried out by 10 ataxic dysarthria speakers together with two listening experiments. Each of the listening experiment will be carried out among 50 untrained listeners. The effects of dysarthria and severity on the ability of the speakers to accurately mark stress will be reviewed while also considering how the position of the target word (word to be stressed) affects their stress marking abilities. An initial study on how dysarthric speakers mark stress with respect to acoustic features (intensity, pitch and duration) will be presented. Also presented in this chapter is an experiment on the effects of acoustic modifications on how listeners perceive stress marking in dysarthric speech. This will give a comparison of acoustic modifications and the perception of the listeners. Finally, clinical recommendations will be presented showing potential target levels for intensity, fundamental frequency and durational amplifications during stress marking therapy exercises in the management of dysarthria. Throughout this chapter, three software will be used; two software for speech processing and one statistical software. These are MATLAB (Version 9.1), Praat (Version 5.4.04) and IBM SPSS Statistics 24.

7.2 Participants

The participants in this investigation include 10 speakers with ataxic dysarthria consisting of 5 males and 5 females as illustrated in Table 7-1. In addition to the ataxic dysarthric (AT) speakers, 10 healthy control (HC) speakers were also recruited. These healthy control (HC) speakers were aged-matched, as well as gender and dialectal background matched, with the AT speakers. These participants are taken from the dysarthric speech data set reported in [133]. The participants

have no cognitive deficiency neither do they have any visual and hearing impairment. Their severity varied from mild to severe cases. In addition, all of them are monolingual native speakers of English.

Their intelligibility scores varied from 18 to 91 as shown in Table 7-1. These intelligibility scores were estimated from the average scores from five trained listeners during the passage reading task described in [21]. The etiologies of these participants are either cerebellar ataxia (50%), Friedreich's ataxia (30%) or spinocerebellar ataxia (20%). Each participant produced 30 sentences using the 10 Subject-Verb-Object-Adverbial (SVOA) structured sentences across three (3) sentence conditions. These sentence conditions are stress on the initial (S), medial (O), and final (A) target words tagged T1, T2, and T3 respectively. For example, T1 implies that the target word to be stressed is in the initial position (subject) of the sentence. The words in the sentences are tagged W1, W2 and W3 representing the subject, object and adverbial respectively. Audio recordings are saved in the format of AT_XX_YY_ZZ where XX is the participant's number (from 01 to 10), YY is the sentence number (from 01 to 10) and ZZ is the stress position (01-initial 02-medial or 03-final position).

Table 7-1. Details of participants involved in the study

Participant	Age	Gender	Etiology	% Intelligibility Score
AT_01	46	M	CA	26
AT_02	60	F	CA	33
AT_03	28	M	FA	94
AT_04	52	F	CA	75
AT_05	28	F	FA	91
AT_06	65	F	SCA6	42
AT_07	72	M	CA	81
AT_08	51	M	CA	56
AT_09	56	M	SCA8	18
AT_10	57	F	FA	20

CA: Cerebellar ataxia of undefined type, FA: Friedreich's ataxia and SCA: spinocerebellar ataxia

The list of the 10 SVOA structured sentences used for this study is given below.

1. The model wrote her memoirs in Lima
2. The gardener grew roses in London
3. The landlord owns dwellings in Reading
4. The lawyer met the model in London
5. The diva made a movie in Venice
6. The minister has a nanny from Norway
7. The widow bought a villa in Ealing
8. The milliner got a memo from Melanie
9. The murderer met his lawyer in Limerick
10. The neighbour plays melodies on her mandolin

In addition, 50 listeners were recruited for the perceptual experiment. These listeners are untrained and do not have any hearing or speech impairment. They are all native speakers of English and mainly university students aged between 18 and 50 years old. Their suitability for the study was tested by engaging them in a practice experiment where each participant is required to attain more than 80% accuracy in a stress identification task in healthy speech. Listeners with less than 80% stress identification accuracy were excluded from the study. In terms of sample size, the total number of audio samples in the first listening experiment is 212 and that of the second experiment is 259 audio samples.

7.3 Initial Study on Stress marking in Healthy Control and Dysarthric Speech

Before carrying out acoustic modifications on the utterances produced by dysarthric speakers, it is important to understand how dysarthric speakers show stress relative to what healthy control speakers will do. This is achieved by carrying out a stress production exercise for both speaker groups. The stress production exercise showed that both speaker groups make use of pitch, intensity and duration to mark stress. That is, speakers emphasized the stressed word by increasing the fundamental frequency, increasing the intensity and/or elongating the duration of the stressed word. An analysis of variance (ANOVA) was also carried out on the data from both speaker groups using IBM SPSS Statistics 24. The results of the ANOVA showed that the three acoustic features (pitch, intensity, and duration) of the target word are significantly independent of the acoustic features of the other words in the sentence.

Furthermore, the comparison of the analysis of how healthy control and dysarthric speakers mark stress, with relation to intensity, pitch, and duration of words, is illustrated in Figure 7-1, Figure 7-2, and Figure 7-3 respectively.

The peak intensities of the words (W1, W2, and W3) were measured for the three sentence conditions (T1, T2, and T3) for the two speaker groups and the average peak intensity was estimated across the 10 healthy control speakers. This was also done for the 10 ataxic dysarthric speakers. The average peak intensities for the healthy control and dysarthric speakers are shown in Figure 7-1. In the first sentence condition (T1), across healthy control speakers, the average peak intensity for W1 was higher than that of W2 and W3. The highest average intensity was recorded for W1 when the sentence condition was T1. Similarly, the highest average intensity was recorded for W2 when the sentence condition was T2. Likewise, the highest average intensity for W3 was observed in sentence condition T3. This trend is the same for dysarthric speakers. This trend was also observed for dysarthric speakers. The major difference being that the average intensities in healthy control speech are larger than those observed in dysarthric speech. It is also important to point out that the average peak intensity for W3 remained unchanged for both T1 and T2 sentence condition whereas is increased in T3 sentence condition. These observations show that both healthy control and dysarthric speaker groups make use of increased intensity to mark stress.

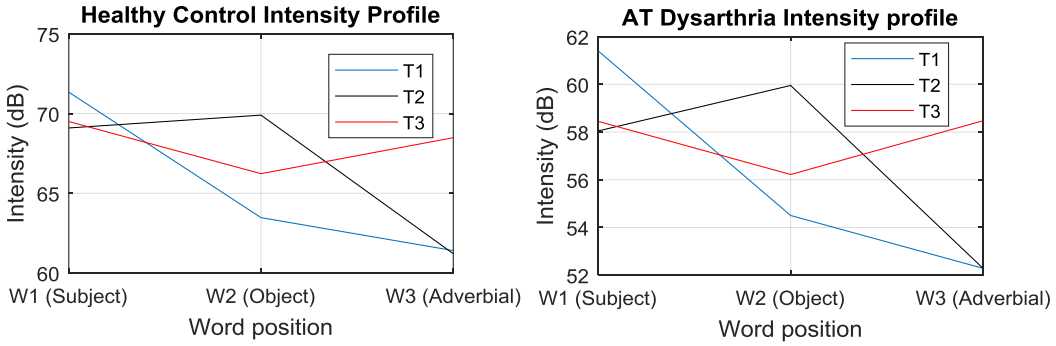


Figure 7-1. Effects of Stress Marking on the Intensity of Words in Healthy Control and Dysarthric Sentences

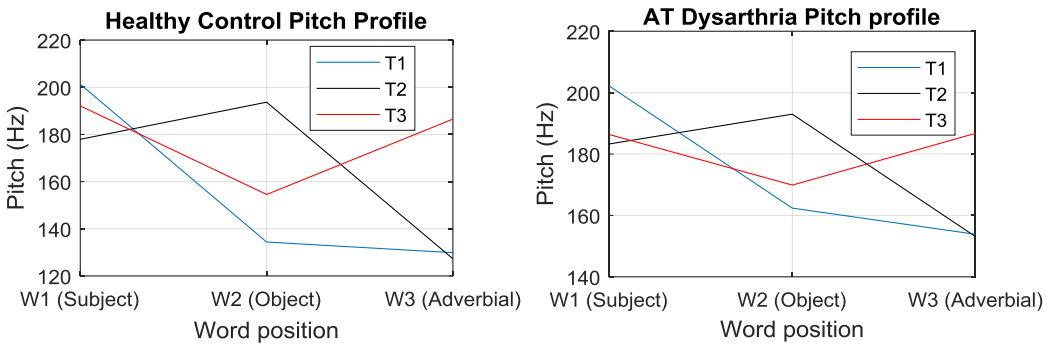


Figure 7-2. Effects of Stress Marking on the Pitch of Words in Healthy Control and Dysarthric Sentences

Moreover, the peak pitch (fundamental frequency, F0) profiles for both healthy control and dysarthric speaker groups were also analysed. These are shown in Figure 7-2. In healthy control speech samples, the highest average F0 for W1 was observed in T1 sentence condition, while the highest average F0 for W2 was observed in T2 sentence condition, and, as expected, the highest average F0 for W3 was observed in T3 sentence condition. These results show that healthy control speakers emphasize (stress) words by increasing the F0 of the word being stressed. Like in the peak intensity profiles (Figure 7-1), the peak pitch profiles for the two speaker groups are quite similar. However, the extents to which the dysarthric speakers reduced or increased the F0 of the words are quite limited. For example, in T1 sentence condition, for healthy control speakers, the difference in the peak intensities of W1 and W2 was about 70Hz whereas the difference observed in dysarthric speakers was only about 40Hz. This implies that dysarthric speakers are unable to fully accentuate the stressed word using F0 as much as healthy control speakers do. In all, both speaker groups use F0 in emphasizing words in a sentence.

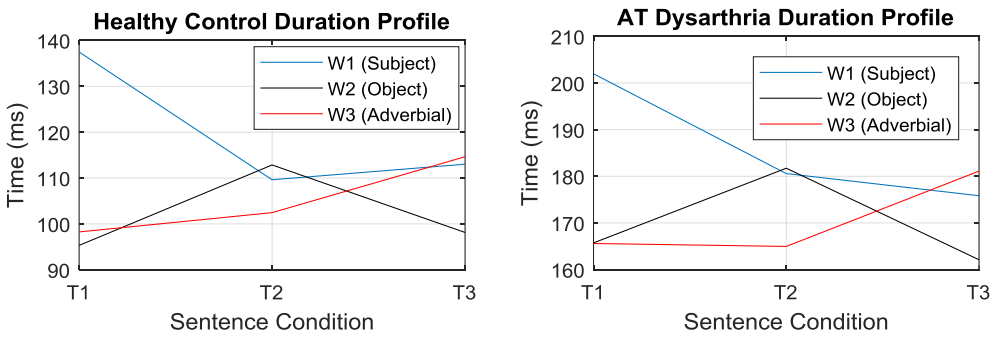


Figure 7-3. Effects of Stress Marking on the Duration of Words in Healthy Control and Dysarthric Sentences

The analysis of the average duration was a bit different from how the peak intensity and peak F0 were analysed. The average duration was analysed across sentence conditions rather than across words. This is because different words are expected to have different durations but the increase in duration (due to stress) can only be observed when comparing the duration of a word in a sentence condition with the duration of the same word in another sentence condition. This is what was illustrated in Figure 7-3. The average duration of W1 (Subject) was longest in T1 sentences, whereas, that of W2 (Object) was longest in T2 sentences, and that of W3 (Adverbial) was longest in T3 sentences. The trends in both speaker groups are quite similar but the average duration of the words spoken by dysarthric speakers are generally longer. This validates the hypothesis that utterances from dysarthric speakers are characterised by slowed speech when compared with healthy speakers.

The analyses of the peak intensity, peak F0 and duration show that both speaker groups use increased intensity, increased F0 and increased duration to mark stress, of words, in sentences. The extent to which the two speaker groups increase these acoustic features is, however, different in the two speaker groups. The healthy control speakers accentuated the three acoustic features more than the dysarthric speakers to mark stress. This explains why listeners are more likely to identify the emphasized word in healthy control speech than in dysarthric speech.

In addition to these studies, a logistics regression on the impact of these prosody parameters on the listeners' ability to correctly mark stress was carried out for healthy control speaker group (in order to understand why listeners found it easier to identify the stressed word in healthy control speaker group) using IBM SPSS Statistics 24. The results are shown in Table 7-2. The significance p-value threshold used for the regression analysis was $p< 0.05$. In T1 utterances, the odds of listeners ability to accurately identify the target word increases as the peak intensity, duration and peak F0 of W1 increases, and the odds increases as the peak intensity of W2, duration of W2 and peak F0 of W3 decreases. In T2 utterances, the odds of listeners ability to accurately identify the target word increases as the peak intensity and peak F0 of W2 increases and the odds increases as the peak intensity of W1, peak F0 of W1, duration of W3 and peak F0 of W3 decreases. In T3, the odds of listeners ability to accurately identify the target word increases as the peak intensity and peak

F0 of the W3 increases and the odds increases as the peak intensity and peak F0 of W2 decreases.

Table 7-2. Logistics Regression of the Effects of Acoustics Features on Listener Accuracy

Variables in the Algorithm									
	T1	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	W1 Peak dB	0.182	0.029	38.910	1	0.000	1.199	1.133	1.269
	W1 Dur	0.009	0.002	20.965	1	0.000	1.009	1.005	1.013
	W1 Peak F0	0.011	0.005	5.592	1	0.018	1.011	1.002	1.020
	W2 Peak dB	-0.101	0.041	6.158	1	0.013	0.904	0.834	0.979
	W2 Dur	-0.008	0.003	9.907	1	0.002	0.992	0.987	0.997
	W2 Peak F0	0.003	0.006	0.175	1	0.676	1.003	0.990	1.015
	W3 Peak dB	-0.049	0.036	1.875	1	0.171	0.952	0.888	1.021
	W3 Dur	0.001	0.002	0.188	1	0.664	1.001	0.996	1.006
	W3 Peak F0	-0.009	0.004	4.905	1	0.027	0.991	0.983	0.999
	Constant	-3.467	1.916	3.274	1	0.070	0.031		
Variables in the Algorithm									
	T2	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	W1 Peak dB	-0.078	0.023	11.989	1	0.001	0.925	0.885	0.967
	W1 Dur	0.002	0.002	0.924	1	0.336	1.002	0.998	1.005
	W1 Peak F0	-0.009	0.004	4.758	1	0.029	0.991	0.982	0.999
	W2 Peak dB	0.271	0.031	74.768	1	0.000	1.312	1.234	1.395
	W2 Dur	0.003	0.002	2.588	1	0.108	1.003	0.999	1.007
	W2 Peak F0	0.011	0.004	8.356	1	0.004	1.011	1.004	1.019
	W3 Peak dB	-0.180	0.032	31.864	1	0.000	0.835	0.784	0.889
	W3 Dur	-0.004	0.002	3.171	1	0.075	0.996	0.992	1.000
	W3 Peak F0	-0.002	0.004	0.268	1	0.605	0.998	0.991	1.005
	Constant	-1.552	1.607	0.933	1	0.334	0.212		
Variables in the Algorithm									
	T3	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	W1 Peak dB	-0.050	0.034	2.191	1	0.139	0.951	0.890	1.016
	W1 Dur	0.002	0.002	1.063	1	0.302	1.002	0.998	1.005
	W1 Peak F0	-0.008	0.005	2.355	1	0.125	0.992	0.981	1.002
	W2 Peak dB	-0.172	0.040	18.691	1	0.000	0.842	0.779	0.910
	W2 Dur	0.001	0.002	0.117	1	0.732	1.001	0.997	1.005
	W2 Peak F0	-0.012	0.006	3.983	1	0.046	0.988	0.977	1.000
	W3 Peak dB	0.289	0.034	73.891	1	0.000	1.336	1.250	1.427
	W3 Dur	0.000	0.002	0.036	1	0.850	1.000	0.996	1.005
	W3 Peak F0	0.023	0.004	40.379	1	0.000	1.023	1.016	1.030
	Constant	-5.067	1.569	10.434	1	0.001	0.006		

a. Variable(s) entered on step 1: W1 Peak dB, W1 Dur, W1 Peak F0, W2 Peak dB, W2 Dur, W2 Peak F0, W3 Peak dB, W3 Dur, W3 Peak F0.

Parameters Description

B: Regression Coefficient for the constant (intercept)	S.E.: Standard error around Coefficient for the constant (B)	Wald: Wald statistic to test the Odds ratio (Exp(B))	df: degrees of freedom for Wald chi-square test	Sig: p-value for Wald chi-square test	C.I.: Confidence Interval

These observations imply that listeners are more likely to correctly identify the stressed word if dysarthric speakers are able to increase the peak intensity, peak F0 and duration of the target word. Increased intensity, F0 and duration, therefore, can be used in therapy to help dysarthric speakers mark stress in sentence thereby increasing their speech intelligibility. But the question is, should these dysarthric speakers increase their intensity, F0 and duration as much as the healthy speakers will? Also, should the therapy be focused on all three features at the same time or some of them? These questions will be answered in the following sections.

7.4 Focus Sentence Selection

A prior study was carried out in [27] on the dataset. This study involved a perception experiment using seven untrained listeners who are native speakers of English and do not have any hearing impairment. Listeners were asked to identify the emphasized (stressed) word in the sentence. Sentences where more than 60% of the listeners could not locate the target word, were identified and selected for this study. These included sentences where no stress has been placed on any of the words and sentences where the AT speakers produced incorrect pitch contours. These identified sentences formed the baseline (focus utterances) for this experiment. These focus utterances were also grouped into two; utterances with appropriate pitch contours but requiring amplification only (AMP) and utterances with inappropriate pitch contours (IPC). For this study, 7 AMP utterances and 8 IPC utterances were selected.

7.5 Stress Marking Features Modifications

Two distinct pitch modifications techniques were implemented based on the category of the focus sentences (AMP or IPC). Pitch incremental modifications were carried out on all the focus sentences while pitch contour modifications were carried out on IPC sentences only.

7.5.1 Pitch Amplification

To establish a reference point for pitch incremental modifications, audio samples from the 10 healthy speakers were examined. These healthy control (HC) speakers

were aged-matched, as well as gender and dialectal background matched with the dysarthric speakers. The HC speakers were given the same SVOA structured sentences and the average increment in the fundamental frequency (F0) of the target word was estimated for the three sentence conditions. As presented in [32], HC speakers mark stress by increasing F0 before a target word and decreasing F0 after the target word. As illustrated in Figure 7-4, the pre-target increments and post-target decrements vary across the sentence conditions but the average pre-target increment and post-target decrement are 14% and 30% respectively.

Consequently, the F0s of the target words were increased to a maximum of 30% (picking the worst case possible) at an incremental rate of 25%, 50%, 75% and 100% (that is, 0.25 of 30% =7.5%, 0.5 of 30% =15%, 0.75 of 30% =22.5% and 1.00 of 30% = 30% respectively). Praat (Version 5.4.04), a speech processing software, was used to modify the pitch incrementally. Figure 7-5 illustrates pitch incremental modifications carried out on AT_08_04_01. The pitch contours are represented by the blue bars in both plots. The F0 of the target word has been increased by 30% while keeping that of other words the same.

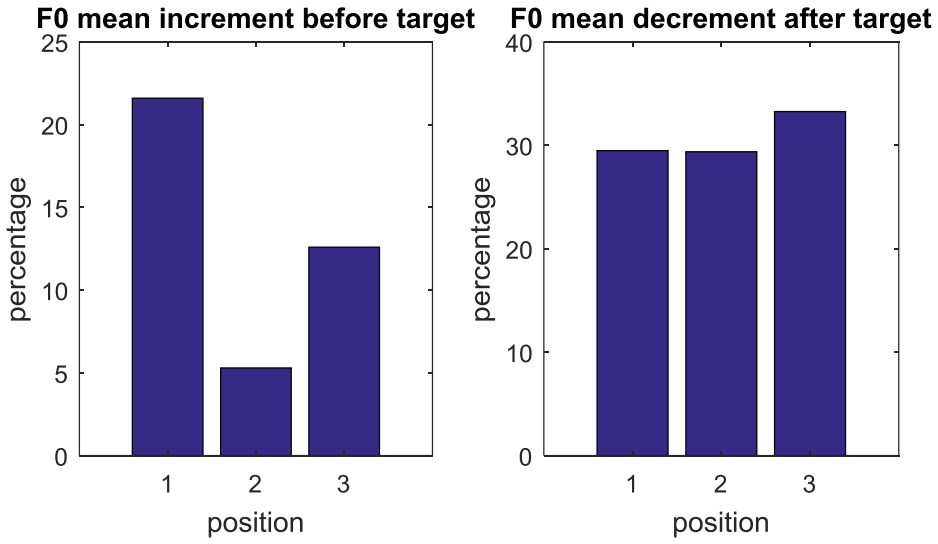


Figure 7-4. Mean F0 change before (left) and after (right) the target

7.5.2 Pitch Contour Modification

Pitch contour modifications, on the other hand, involves the modification of the pitch contours of the IPC sentences. A scenario of IPC is when all the words in the sentence have been stressed equally. This could be due to the fact that the AT

speakers have placed pauses before each word in the sentences. Another scenario is when the stress has been placed on the wrong word or on two words (the target word and another word). The pitch contour modification was implemented using Praat. And the new signal is stored using the synthesis function in Praat. It is important to note that for all pitch contour modifications carried out in this study, the pitch contours of the target words were not modified in any way. Only the pitch contours of other words were modified. This is to ensure that the pitch contour of the target word is preserved for the pitch amplification process.

An example illustrated in Figure 7-6 shows the pitch contour of AT04_06_02 before and after pitch contour modifications. Here, the pitch contour of 'O' word in the SVOA structured sentence 6 from speaker AT_04 has been preserved. Whereas the pitch contours of the other words in the sentence have been modified to correspond with the expected pitch profile for target position 2. Without increasing F0 of the target word, it can be seen that the resulting pitch profile shows the location of the target word. It is important to note that sometimes AT speakers can use the wrong pitch contour (for example, falling F0) within the target word. In this case, a more complex pitch contour modification will be required.

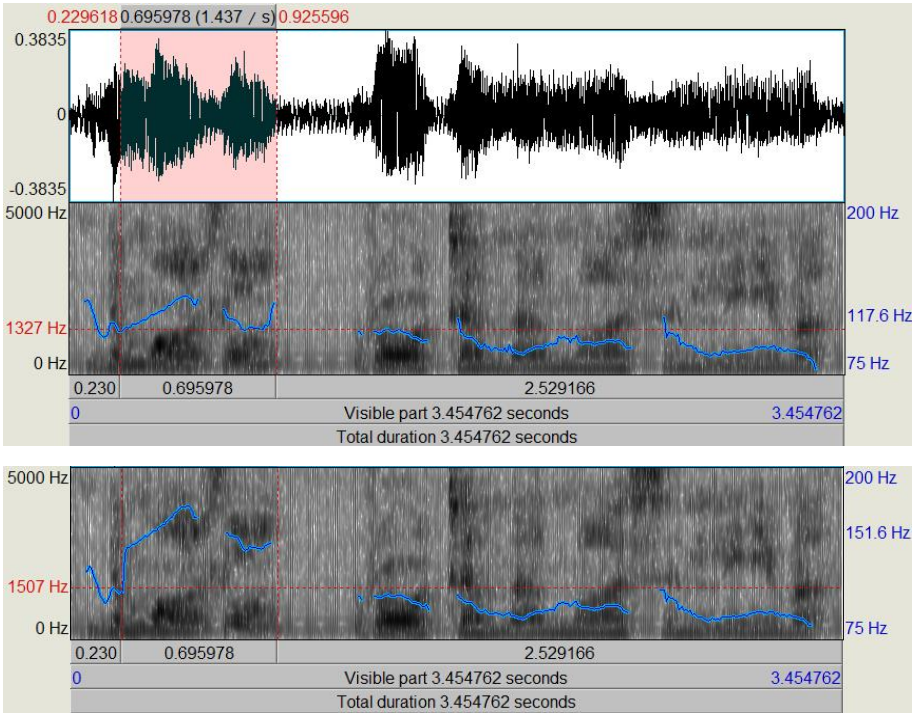


Figure 7-5. AT_08_04_01 speech before (top) and after (bottom) 30 % increment in F0 of the highlighted target word

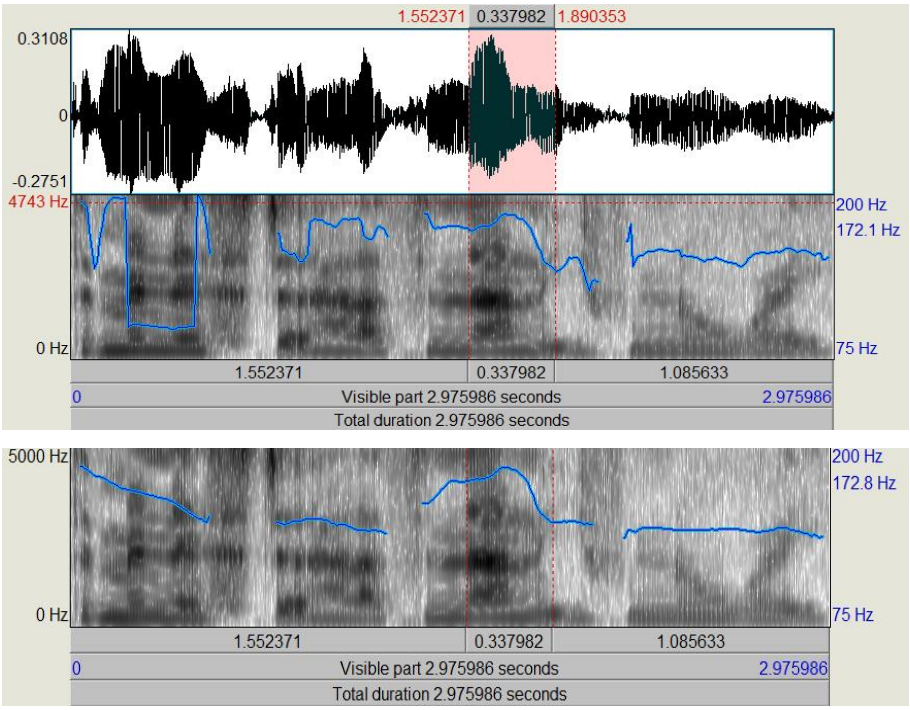


Figure 7-6. AT04_06_02 before (top) and after (bottom) pitch contour modification

7.5.3 Intensity Amplification

Apart from increasing F0, the initial study in Section 7.3 has shown that healthy speakers use increased intensity to mark stress. Healthy speakers increase their intensity just before the target word and decrease the intensity right after the target word. However, for ataxic dysarthric speakers, the variation in intensity is reduced. The relative changes in intensity are dependent on the position of the target word. In this study, the intensities of the target words in the focus sentences (all the 15 focus sentences AMP and IPC inclusive) were modified at 4 incremental rates. The incremental rates used are 25%, 50%, 75% and 100%. The increments are done in MATLAB (Version 9.1). The intensity increment was achieved by multiplying the amplitude of the target word by the incremental factor as given in (21).

$$Int_{new} = Int_{old}(1 + fac) \tag{21}$$

where Int_{new} is the new intensity, Int_{old} is the initial intensity and fac is the incremental factor (fac=0.25, 0.5, .75 or 1).

An example of the intensity modification is presented in Figure 7-7. In this figure, the speaker AT_05 produced sentence 09 with the target word position in the final part of the sentence (03). Looking at both waveforms, for the original and modified signals, it can be seen that the intensity of the target word has been modified by increasing the amplitude of the highlighted segment of the speech signal. Likewise, the intensity profiles in dB before intensity modification shows an occurrence of mono loudness. However, after increasing the amplitude of the target word waveform by 100%, the intensity profile shows an emphasis on the target word.

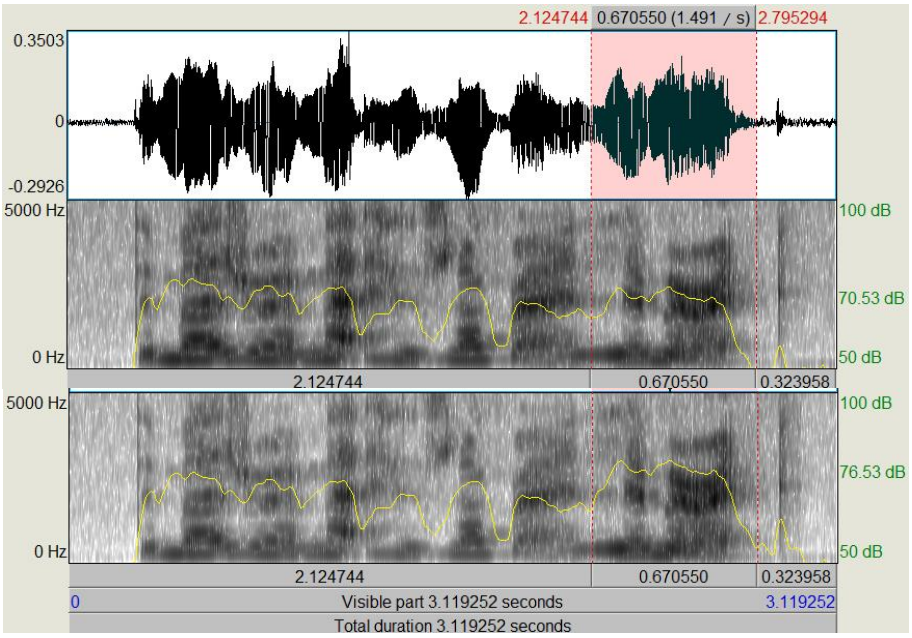


Figure 7-7. AT_05_09_03 before (top) and after (bottom) 100% increment in intensity

7.5.4 Duration Amplification

Modifying the duration of a target word without altering the intensity or the pitch could be challenging. Over the past few decades, researchers have offered techniques for elongating or shortening speech signals based on time-domain [219] or frequency domain analysis [220]. These techniques included resampling, frequency domain interpolation, and phase vocoder.

Resampling techniques involve upsampling or downsampling the speech signal in the time domain. After which the audio signal is saved at the original sampling frequency. This technique is simple and fast to implement. However, the quality of the signal is compromised leading to modifications in pitch and distortion of the

sounds (unnatural sounds). On the other hand, frequency interpolation involves Fourier transform followed by interpolation and inverse Fourier transform. The frequency-domain interpolation alters the signal intensity and pitch quality, therefore, also resulting in unnatural sounds.

For the purpose of this study, the phase vocoder technique was used. The phase vocoder (PVOC) is a well-known audio synthesis technique used for time dilation and pitch scaling. Time dilation or scaling is achieved by modifying the original short-time Fourier transform (STFT) of a signal before performing an inverse short-time Fourier transform (ISTFT) on the modified spectrum [221]. The STFT coefficients are modified by keeping the amplitudes the same and modifying the phase so that there are more or fewer oscillation cycles in each frequency band [222]. Even though the first implementation of PVOC was for a low bit rate speech encoding [221], PVOC has gained high popularity in audio and music processing.

The PVOC is based on the assumption that most audio or music signals consist of resonances of sinusoids and thus the amplitudes and the phases of these sinusoids can be estimated using the STFT function. In the initial application of PVOC, that requires coding and decoding, these amplitudes and phases can be coded (by quantization) and transferred over a channel to a decoder [221]. Over the years, different modifications have been proposed to the original PVOC depending on the required application.

For time-scaling applications, the phase vocoder can be implemented in two ways:

1. Using varying hop factors for the analysis and synthesis stages (hop factor is distance, in samples, between the first samples of consecutive short time frames)
2. Phase interpolation and instantaneous frequency calculations. The phase interpolation guarantees phase coherence; to preserve the correlation between consecutive adjacent frames [223].

A phase interpolation-based PVOC was used to modify the duration of the target words. The audio signal is represented as a summation of sinusoids with time-varying amplitudes and instantaneous phases. The PVOC modifies the STFT of the sinusoidal signal by unwrapping the phases of the STFT coefficients. This is

achieved by using the increment in phase between two successive frames to estimate the instantaneous frequency of close sinusoid in individual channels [223].

An example of the duration modification is presented in Figure 7-8. In this figure, the speaker AT_02 produced sentence 03 with the target word position in the final part of the sentence (03). For the original and modified signals, it can be seen that the duration of the target word has been modified by elongating the highlighted segment of the speech signal by 100%. The duration is now double of the original.

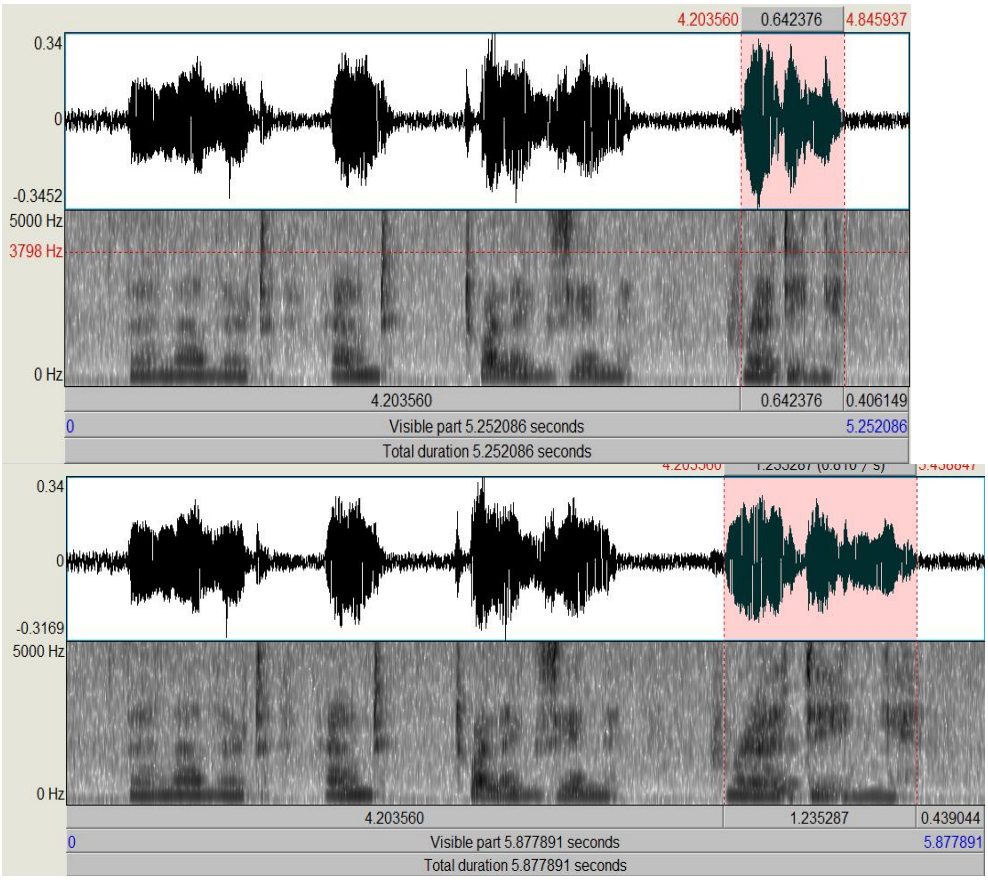


Figure 7-8. AT_02_03_03 before (top) and after (bottom) 100% increment in duration

7.5.5 Addition of Pauses

Another modification carried out on the speech signals is the addition of pauses. Research [27, 32] has shown that speakers also use the addition of pauses to emphasize the target word in sentences. Pauses are added just before the target word and after the target word in two different scenarios.

7.6 Listening Experiments

The listening experiment consisted of two stages. The first experiment involves the modification of the three prosody stress markers, pitch, intensity, and duration individually. The second experiment involves a combination of these features. The second experiment also will involve a combination of the three features and the addition of pauses before the target words. The methodology for the proposed modifications for the individual features (F0, intensity, and duration) has been described in Section 7.5. It is expected that the results of experiment 1 will be a guide to how the modifications in experiment 2 will be combined (for example, whether a 25% increment in F0 is sufficient for stress marking, the subsequent modifications will not go beyond 25% in F0).

7.6.1 Experiment A: Effects of Individual Modifications

This experiment involved individual manipulation of intensity, duration, and fundamental frequency. These features are increased in steps of 25% (that is, 25%, 50%, 75% and 100%). This gives 12 different modifications for each audio sample. For this experiment, 15 audio samples from people with Ataxic dysarthria were used; five audio samples for each sentence condition. The audio samples are also grouped into two; IPC and AMP as described in Section 7.4. The set-up of the modifications involved in the first listening experiment is illustrated in Table 7-3.

For the AMP audio samples, the three acoustic features (intensity, duration, and fundamental frequency) were amplified in increments of 25%. For the IPC audio samples, the acoustic features amplifications were carried out together with the pitch contour modifications as described in Section 7.5.2.

The aims of the first experiment are:

- To investigate the effect of acoustic features amplification on the listeners' ability to correctly identify the target word
- To examine the effect of pitch contour modification on IPC samples
- To determine what level of amplification is enough to give a significant improvement in listener accuracy

- To determine which of the acoustic features has more impact on listener accuracy

In order to validate the listeners' suitability, an initial screening was introduced. The screening process involved a practice experiment. The participants were presented with 10 audio samples (practice sentences) where the target words are well amplified and not wrongly emphasised. Listeners needed to identify at least 8 target words correctly before being allowed to participate in the main experiment (that is, at least 80% accuracy). Out of the 52 participants initially recruited, 2 of them achieved accuracies of less than 80% and therefore were excluded from the main experiment.

Table 7-3. Set-up of Listening Experiment 1

S/N	Sent. Cond.	Modifications	No of Combinations	No of Samples
1	T1, T2, T3	None	-	15
2	T1, T2, T3	Pitch Contour* (IPC utterances)	1	8
3	T1, T2, T3	25% Intensity	1	15
4	T1, T2, T3	50% Intensity	1	15
5	T1, T2, T3	75% Intensity	1	15
6	T1, T2, T3	100% intensity	1	15
7	T1, T2, T3	25% Duration	1	15
8	T1, T2, T3	50% Duration	1	15
9	T1, T2, T3	75% Duration	1	15
10	T1, T2, T3	100% Duration	1	15
11	T1, T2, T3	25% F0	1	15
12	T1, T2, T3	50% F0	1	15
13	T1, T2, T3	75% F0	1	15
14	T1, T2, T3	100% F0	1	15
15	T1, T2, T3	Practice Sentences (for listeners' initial screening)		10

7.6.2 Experiment B: Effects of Combination of Two or More Modifications

In the second experiment, individual amplifications and modifications are combined (that is, 2 or 3 amplifications carried out on a single audio sample). In addition, the effects of combining pitch contour modification and intensity/duration amplification are also examined. The setup of the second experiment is shown in Table 7-4. In terms of sample size, the total number of audio samples in the first listening experiment is 212 and that of the second experiment is 259 audio samples. The same set of listeners were used in both experiments. The summary of the modifications in the two experiments is presented in Table 7-5.

Table 7-4. Set-up of Listening Experiment 2

S/N	Sent. Cond.	Modifications	No of Combinations	No of Samples
1	T1, T2, T3	None	-	15
2	T1, T2, T3	Pitch Contour** and Intensity	4	32
3	T1, T2, T3	Pitch Contour** and Duration	4	32
4	T1, T2, T3	25% Intensity & 25% Duration	1	15
5	T1, T2, T3	25% Intensity & 75% F0	1	15
6	T1, T2, T3	25% Duration & 75% F0	1	15
7	T1, T2, T3	25% Intensity & 50% Duration	1	15
8	T1, T2, T3	25% Intensity & 100% F0	1	15
9	T1, T2, T3	25% Duration & 100% F0	1	15
10	T1, T2, T3	50% Intensity & 25% Duration	1	15
11	T1, T2, T3	50% Intensity & 75% F0	1	15
12	T1, T2, T3	50% Duration & 75% F0	1	15
13	T1, T2, T3	Pauses before target	1	15
14	T1, T2, T3	Pauses after target	1	15
15	T1, T2, T3	25% Duration, 25% Intensity & 75% F0	1	15

The aims of the second experiment are:

- To understand the effects of combining acoustic feature modifications on the ability of the listeners to correctly identify the target word.

- To analyse how corrected pitch contours (for IPC samples) and intensity& durational modifications affect the listeners' accuracy
- To investigate if the effects of the addition of pauses before and after the target have any impact on listeners' accuracy

Table 7-5. Summary of Modifications for the Two Listening Experiments

S/N	Modifications			
1	F0	Intensity	Duration	Pitch Contour
2	F0& intensity	F0& duration	Intensity &duration	F0, intensity & duration
	Pause before target		Pause before target & most salient parameter	

7.7 Experimental Results

7.7.1 Individual Manipulations

The 15 focus sentences (AMP and IPC) were manipulated by increasing the intensity, duration, and fundamental frequency as described in Section 7.5. The listener's accuracy was measured for different sentence conditions (T1, T2, and T3) and utterance groups (AMP and IPC). The effects of duration, intensity and fundamental frequency for AMP utterances are illustrated in Figure 7-9 and that for IPC utterances are illustrated in Figure 7-10.

For AMP utterances, increasing the duration by 25% in T1 utterances, increases the listener accuracy, though an accuracy drop was experienced in T2 and T3 utterances. This could be due to the location of the target word. However, this drop is not significant as it is within the natural variance of the listeners' perception. Increasing the duration further to 50% improves the listener accuracy significantly. A further increase in duration beyond 50% does not give a further significant improvement in listener accuracy. Furthermore, increasing the intensity of the target words in AMP utterances improves the listener accuracy as the increment progresses. However, after a 50% increase in intensity the improvements in listener accuracy became less significant. In addition, increasing F0 gave significant improvements from 25% to 50% to75% and to 100%. The significance of improvements increased across all manipulations.

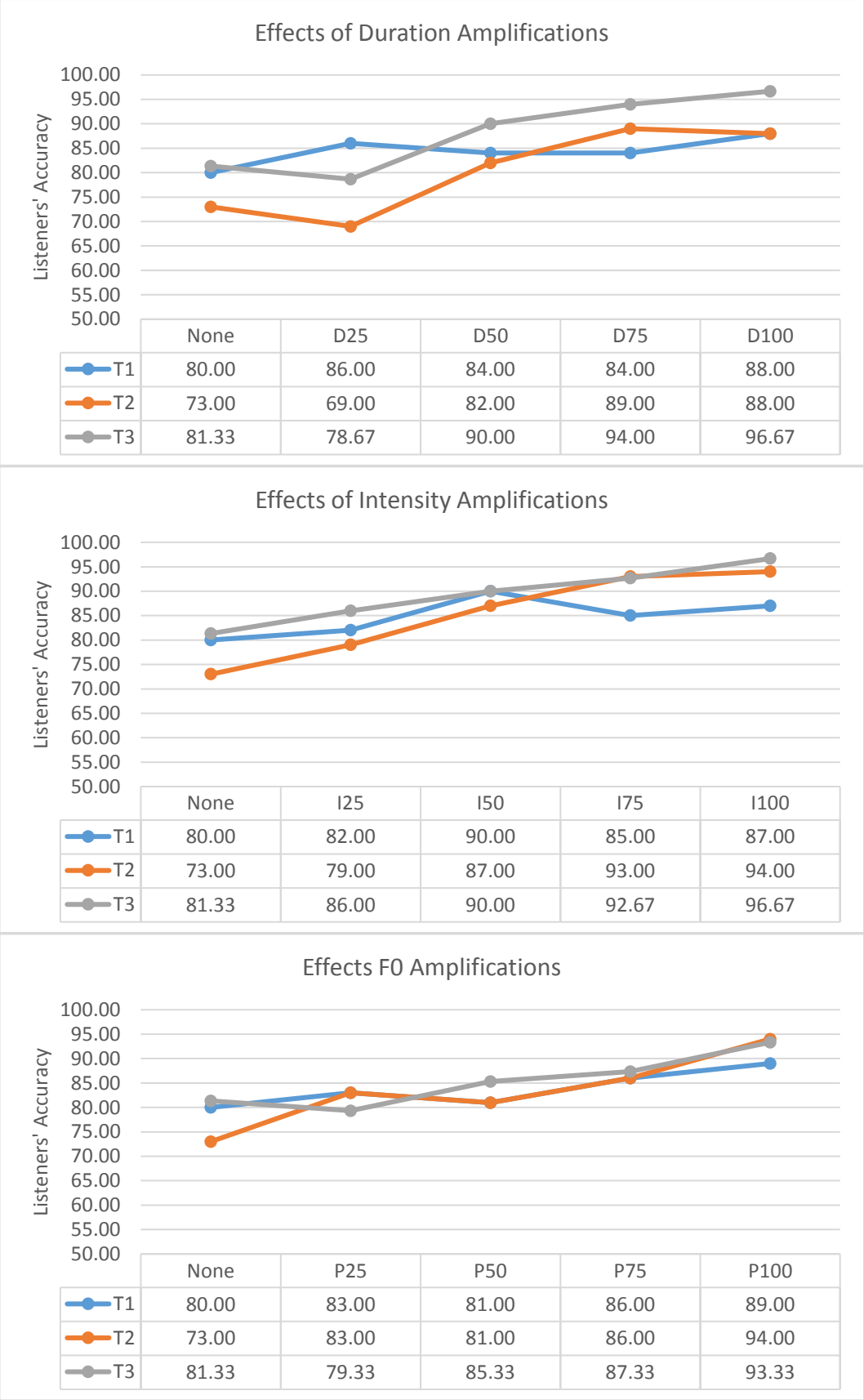
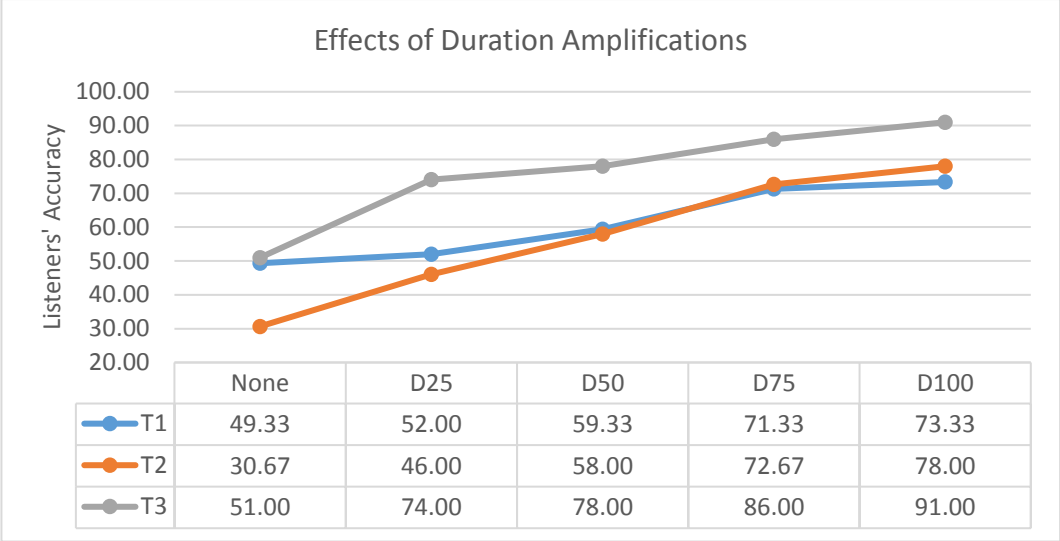
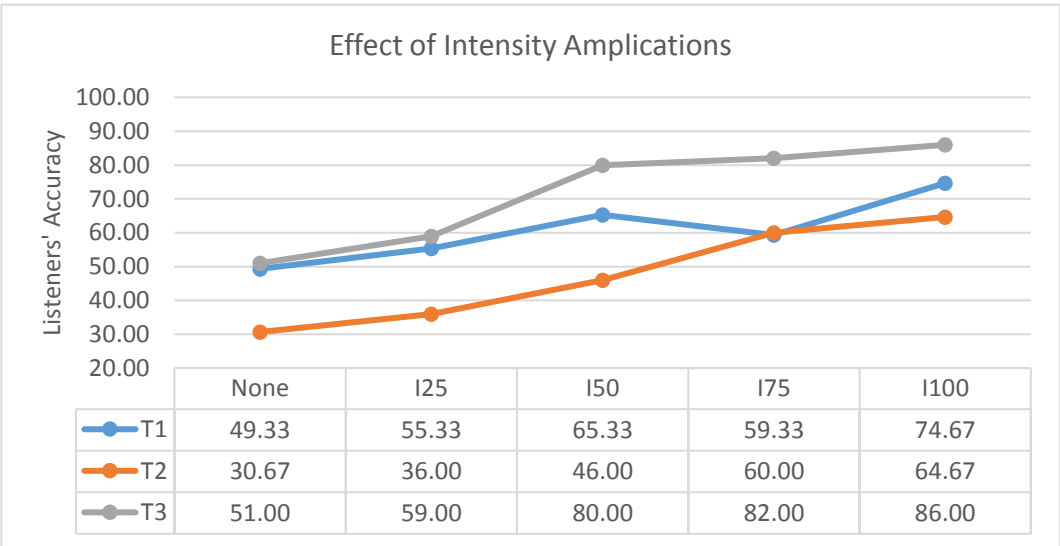


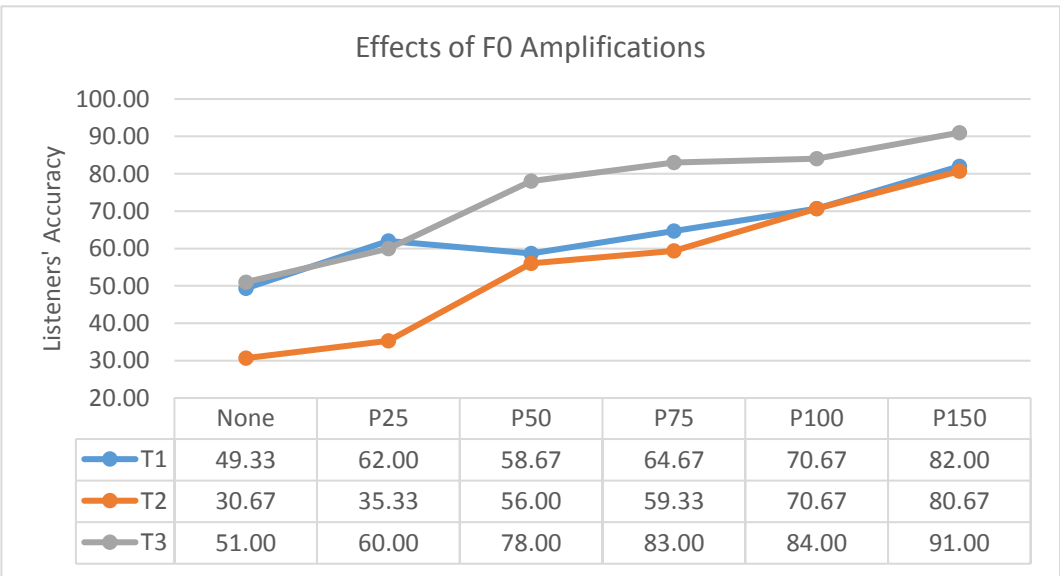
Figure 7-9. Effects of Individual Amplifications on Listener Accuracy in AMP Utterances



	None	D25	D50	D75	D100
T1	49.33	52.00	59.33	71.33	73.33
T2	30.67	46.00	58.00	72.67	78.00
T3	51.00	74.00	78.00	86.00	91.00



	None	I25	I50	I75	I100
T1	49.33	55.33	65.33	59.33	74.67
T2	30.67	36.00	46.00	60.00	64.67
T3	51.00	59.00	80.00	82.00	86.00



	None	P25	P50	P75	P100	P150
T1	49.33	62.00	58.67	64.67	70.67	82.00
T2	30.67	35.33	56.00	59.33	70.67	80.67
T3	51.00	60.00	78.00	83.00	84.00	91.00

Figure 7-10. Effects of Individual F0 Amplifications on Listener Accuracy in IPC Utterances

IPC utterances gave similar significance in listener accuracy improvements (Figure 7-10). The changes in listener accuracy in IPC utterances were relatively higher than those experienced in AMP utterances. For example, 50% increment in intensity improved the listener accuracy by 15% in IPC utterances and the listener accuracy was increased by 10% in AMP utterances. These results imply that 50% increments in duration and intensity are sufficient for improving the listener accuracy significantly. However, a 100% increment in F0 is necessary to significantly improve the listener accuracy. In addition, IPC utterances improved significantly as the acoustic features are increased even though the inappropriate pitch contours were not corrected.

7.7.2 Pitch Contour Modifications in IPC Utterances

Apart from amplification, the effects of pitch contour modifications were also investigated. The pitch contours were modified as described in Section 7.5.2. The results are shown in Figure 7-11. Improvements in listener accuracy were recorded in all three sentence conditions. The highest increment was recorded in T2 utterances and the least increment recorded in T1 utterances. Across the individual amplifications and pitch contour modifications, T1 utterances gave the least improvements in listener accuracy.

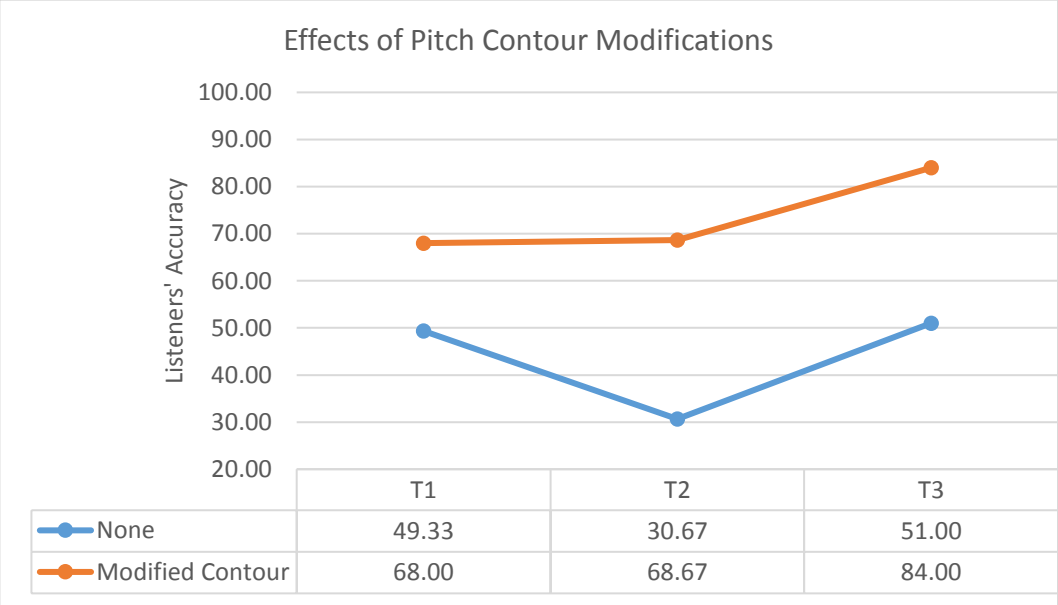


Figure 7-11. Effects of Pitch Contour Modifications on Listener Accuracy in IPC Utterances

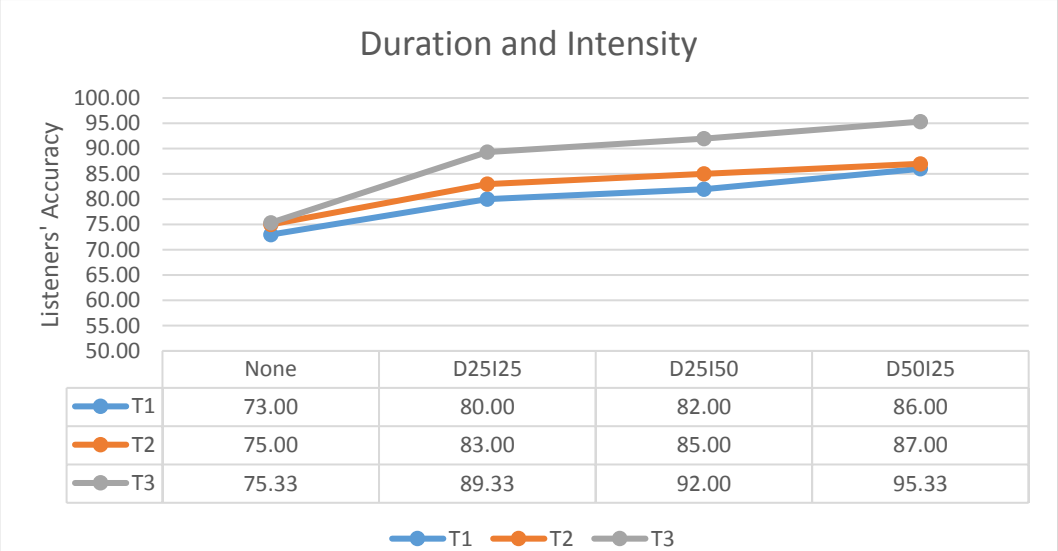
7.7.3 Combination of Manipulations

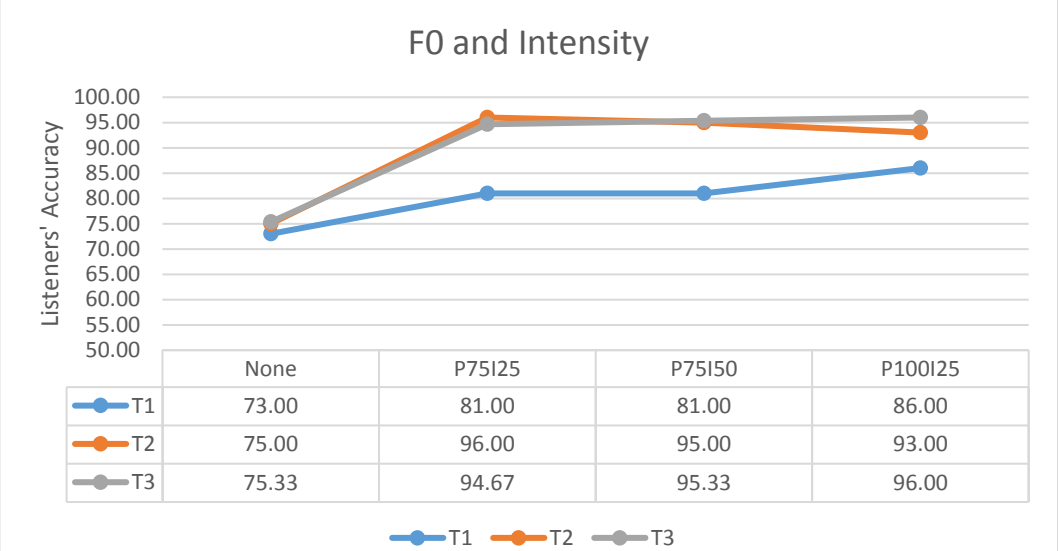
Furthermore, the effects of combining the acoustic features amplifications were investigated. These combinations included duration and intensity, F0 and intensity, F0 and duration and the combination of the three acoustic features. The number of possible combinations was determined by the results from Section 7.6. These combinations are as illustrated in Table 7-5.

As shown in Figure 7-12, combining intensity and duration gave a considerable improvement in listener accuracy. A 25% increment in duration and 25% increment in intensity gave a significant improvement in listener accuracy. Keeping the duration constant at 25% and increasing the intensity further to 50% resulted in higher accuracy for the three sentence conditions. On the other hand, keeping the intensity increment at 25% and increasing the duration to 50% resulted in even greater listener accuracy (compared to D25I0). This shows that at the same level, duration has more impact on the listener accuracy than intensity for all the target positions.

Combining fundamental frequency and intensity also significantly improve listener accuracy. This is illustrated in Figure 7-12. 75% increment in F0 and 25% increment in intensity produced 8%, 20% and 20% improvements in listener accuracy respectively in the target positions. Further increment in F0 or intensity beyond this point did not significantly improve the listener accuracy.

F0 and duration were also combined during this experiment as presented in Figure 7-12. A 25% increase in duration and a 75% increase in F0 gave a significant improvement in listener accuracy. Keeping the duration increment at 25% and increasing the F0 by 100% gives a further improvement in listener accuracy. On the other hand, when the F0 increment is kept at 75% and the duration increased by 50%, T1 and T3 utterances produced further improvement but the improvement reduced in T3. The highest overall improvement was noticeable in D25P100. This implies that F0 has more effect than duration in improving listener accuracy.





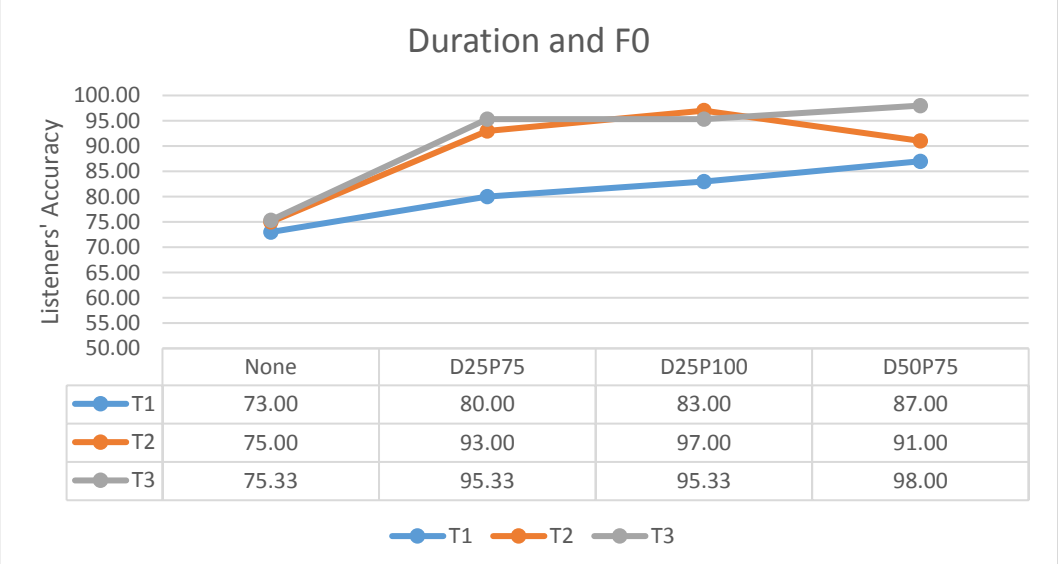
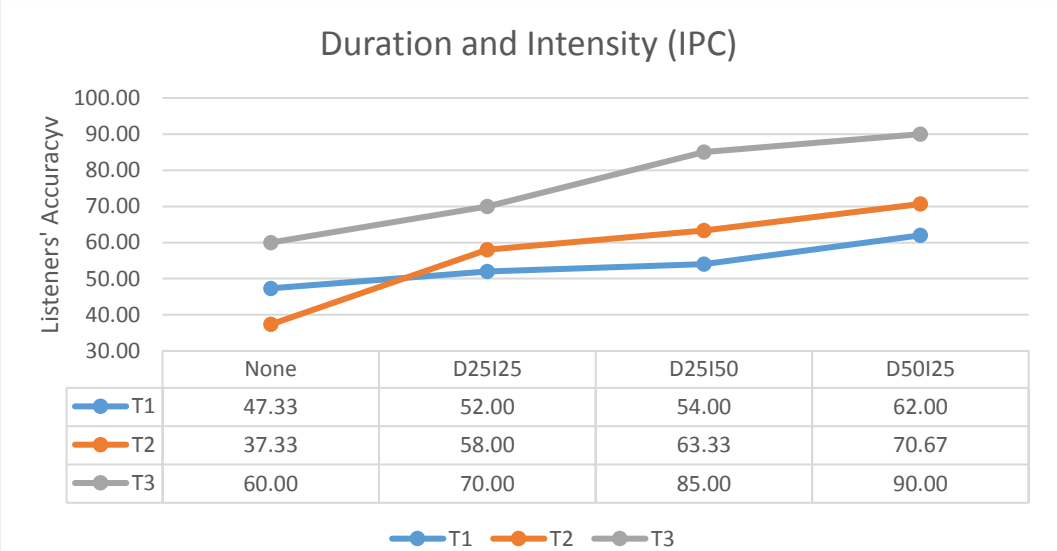
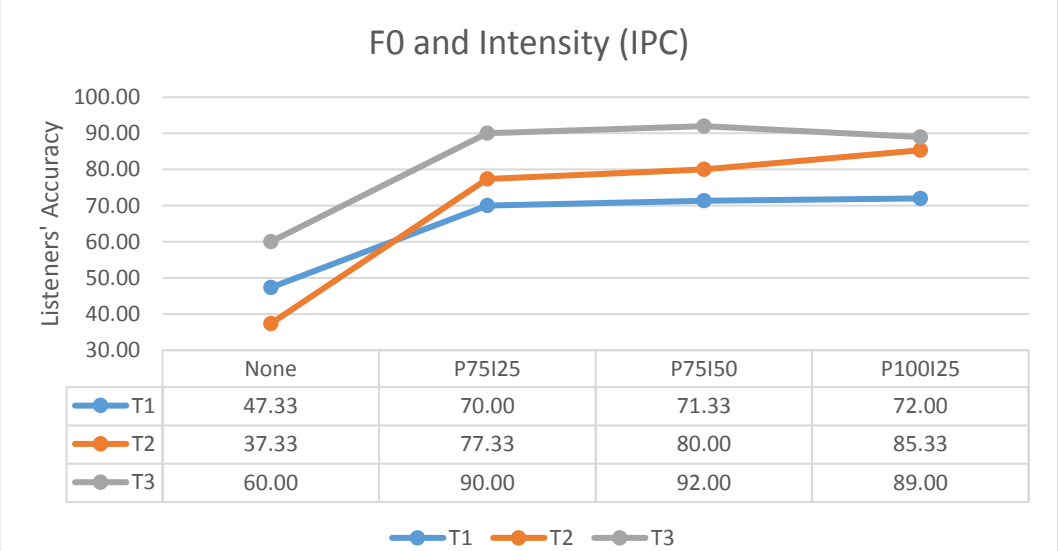


Figure 7-12. Effects of Combination of Two Features on Listener Accuracy in AMP Utterances





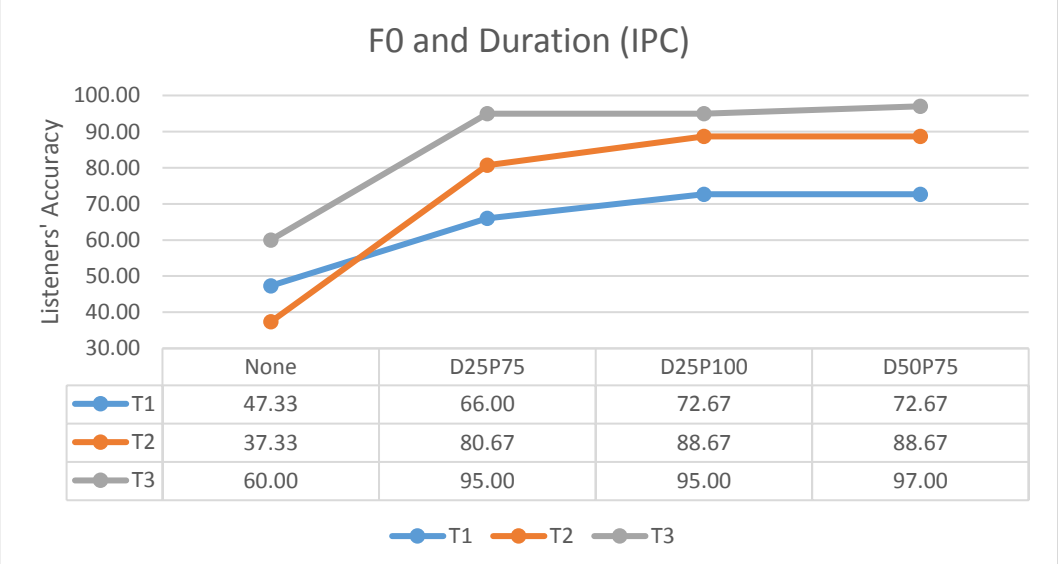


Figure 7-13. Effects of Combination of Two Features on Listener Accuracy in IPC Utterances

Combining the three features also gave significant improvements but not as high as when just two features are combined. The corresponding improvements in listener accuracy for T1, T2, and T3 utterances were 14%, 16% and 21% in AMP utterances and 22%, 41% and 33% in IPC utterances as shown in Figure 7-14.

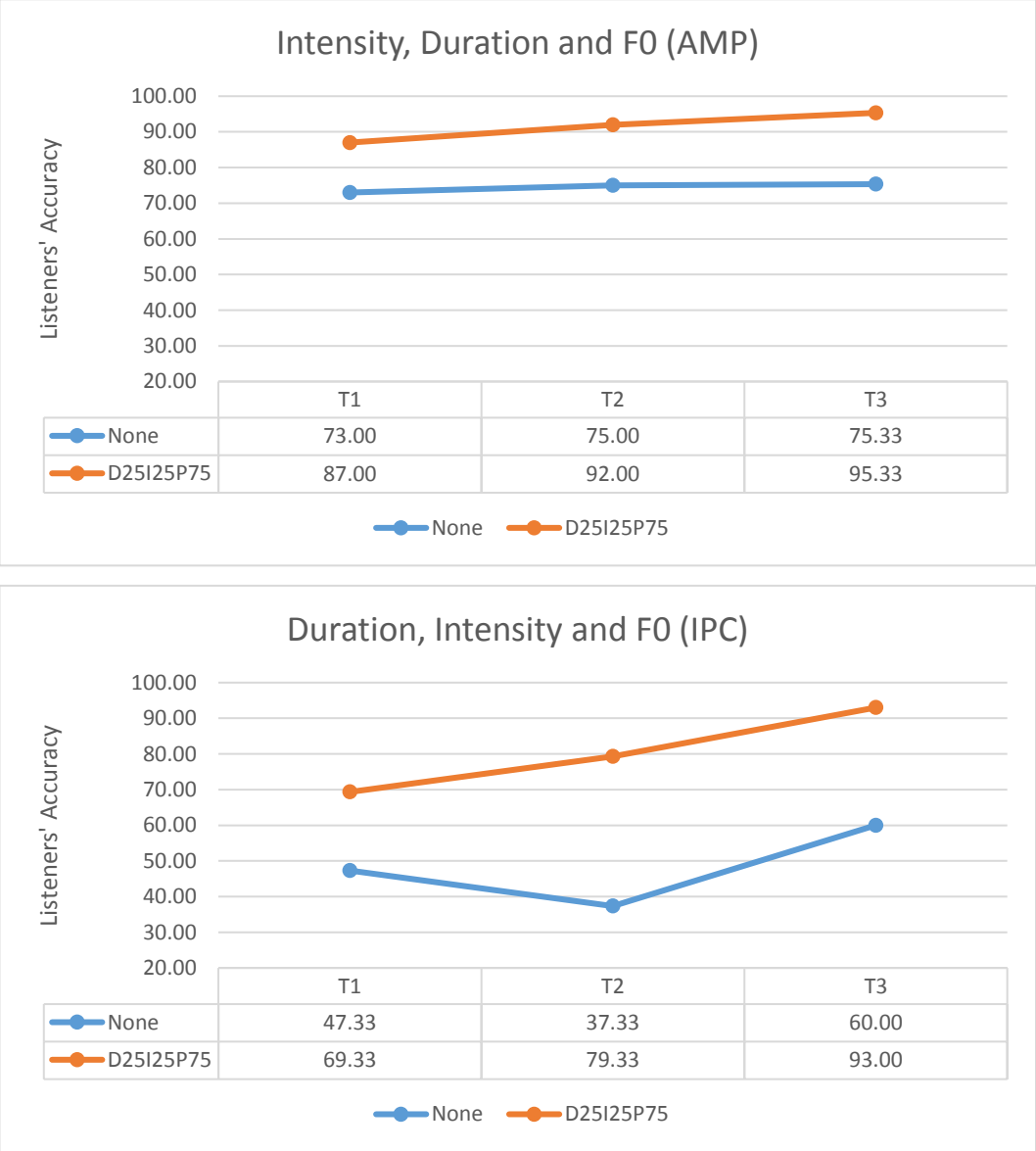


Figure 7-14. Effects of Combination of Three Features on Listener Accuracy in AMP & IPC Utterances

Based on these results, it has been shown that combining 3 features do not give significant improvements when compared with individual and combination of two features. It is also difficult for dysarthric speakers to control these three features simultaneously. The results of combining the three features in AMP and IPC

utterances are very similar but IPC utterances showed the highest improvements in all the combinations as shown in Figure 7-14.

7.7.4 Pitch Contour Modifications and Intensity & Durational Manipulations

The modified pitch contours for IPC utterances described in Section 7.7.2 were also combined with individual combinations with intensity and duration increments. The resulting accuracies are illustrated in Figure 7-15 and Figure 7-16 respectively. Increasing the intensity of the target word for utterances where the pitch contours have been corrected gave a consistent significant rise in listener accuracy as the increments are progressively increased from 25% to 100%. Duration increments (Figure 7-16) also improved the listener accuracy but the relationship is less linear than that seen in intensity increments (Figure 7-15). Increments beyond 50% in duration for T2 and T3 utterances and beyond 75% in duration for T1 utterances did not result in significant improvement.

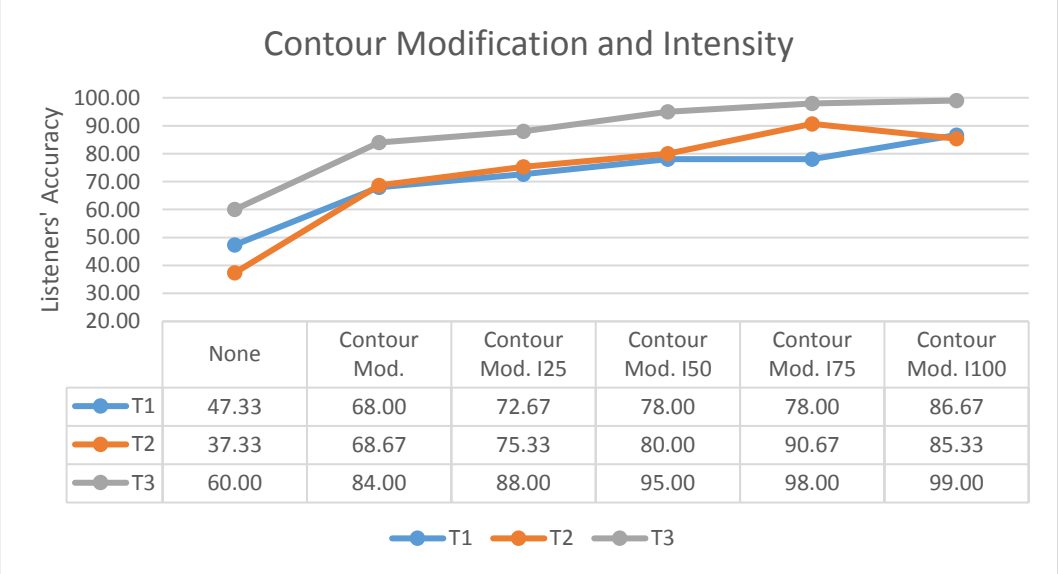


Figure 7-15. Effects of Pitch Contour Modification and Intensity on Listener Accuracy in IPC Utterances

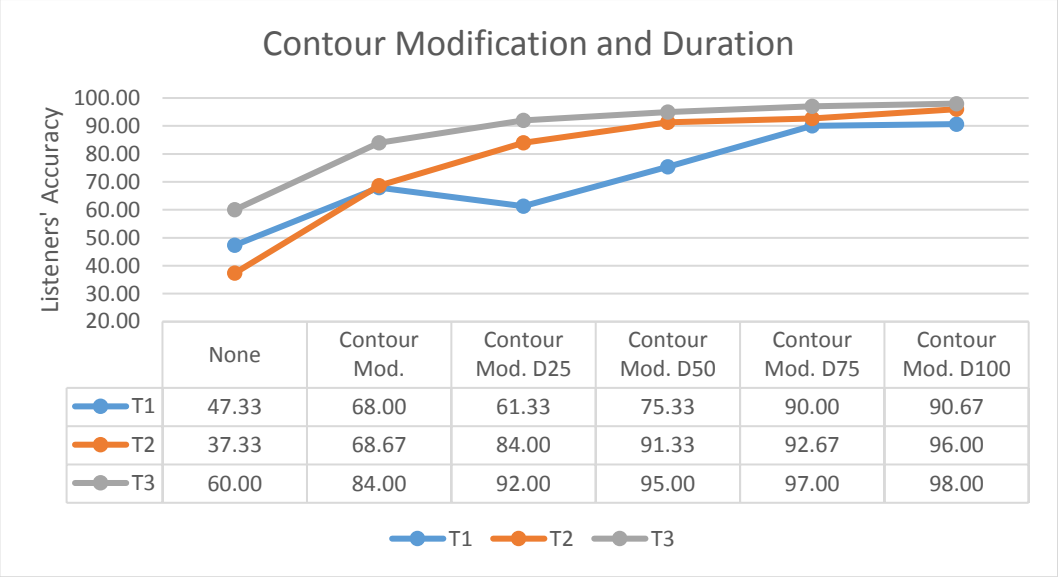


Figure 7-16. Effects of Pitch Contour Modification and Duration on Listener Accuracy in IPC Utterances

7.7.5 Effects of Addition of Pauses

The last sets of modifications included the addition of pauses before and after the target word. This was introduced because in the initial study it was discovered that healthy controls also signalled stress by adding pauses before or after the target word. The average pause used by healthy controls was 250ms. The effects of adding a pause before or after the target word are illustrated in Figure 7-17 and Figure 7-18 for AMP and IPC utterances respectively.

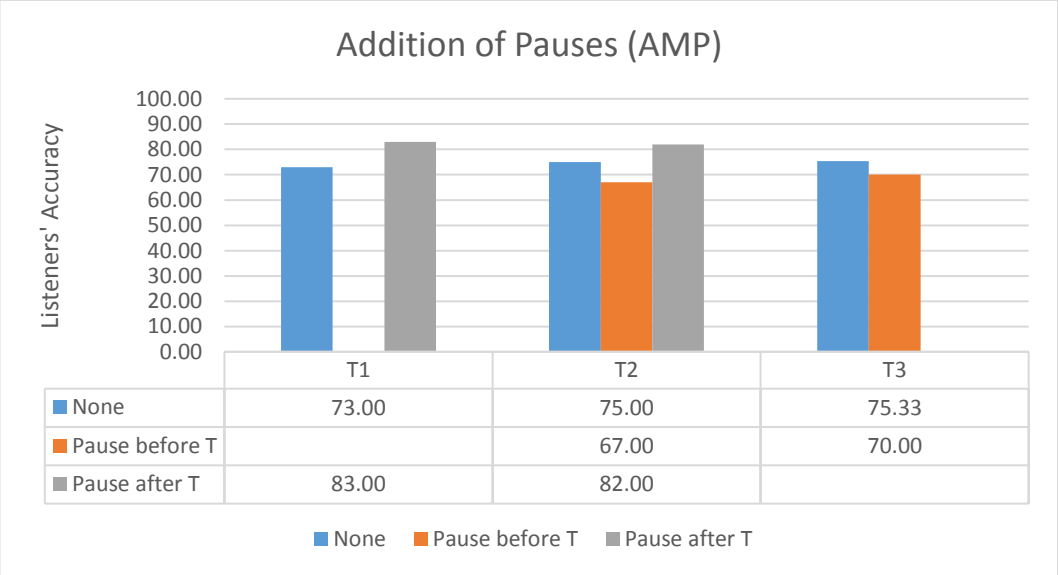


Figure 7-17. Effects of Addition of Pauses on Listener Accuracy in AMP Utterances

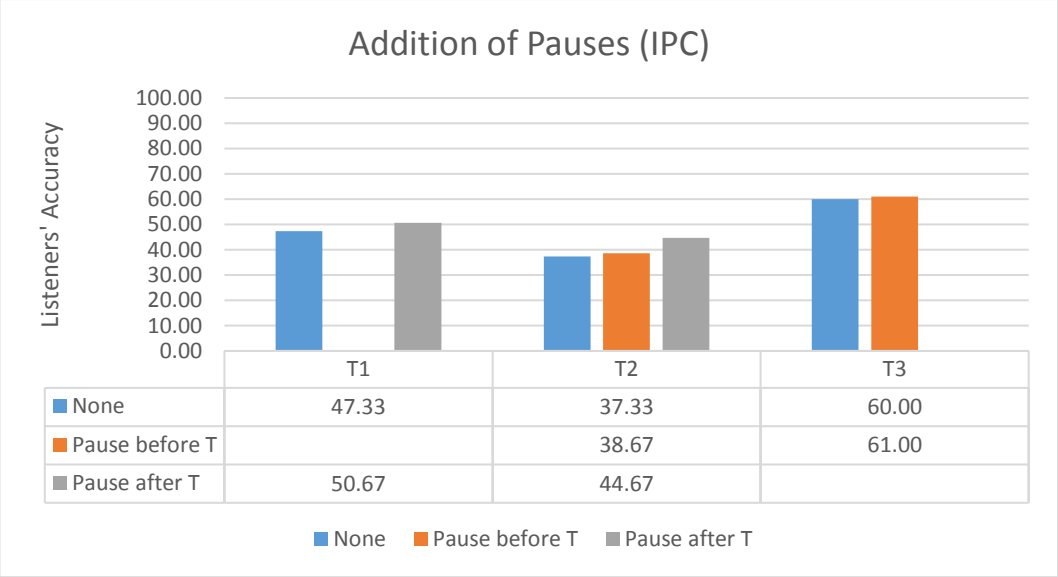


Figure 7-18. Effects of Addition of Pauses on Listener Accuracy in IPC Utterances

For AMP utterances, the addition of a pause before the target word reduced the listener accuracy by 7% in T2 utterances and 5% in T3 utterances. However, the addition of pauses after the target word increased the listener accuracy by 10% in T1 utterances and 7% in T2 utterances. On the other hand, the addition of pauses before the target word in IPC utterance did not have any effect on the listener accuracy while the addition of pauses after the target word improves the listener accuracy by 4% in T1 utterances and 7% in T2 utterances. This implies that dysarthric speakers who are not able to amplify the intensity, duration or F0 of the target word can add a pause after the target to emphasise stress.

7.8 Clinical Implications

From this study, some clinical implications of manipulating intensity, F0 and duration in stress marking exercises during therapy have been identified. The identified clinical implications of this study are as follows:

A. Different degrees of manipulations are necessary for different acoustic parameters: intensity – 50%, F0 – 100%, duration – 50% (that is 100% of Healthy Control speakers' duration increment).
B. Manipulation of any of the 3 stress markers, as well as changes in pitch contour, resulted in improved perceptual outcomes – treatment can focus on areas of strength rather than rehabilitating aspects in deficit.

C. Combining parameters did not improve listener accuracy further – therapy can be simplified by focusing on just one parameter.
D. Changes to pitch contours might improve stress marking in some speakers who are unable to increase their intensity, duration or F0 on targets.
E. Insertions of pauses can also bring some improvement without the need for amplification of any other aspects.

These are recommendations for the treatment of dysarthria during stress marking exercises. These recommendations will be incorporated in the dysarthria management tool presented in Chapter 8 of this thesis.

7.9 Summary

Perceptual analyses of how healthy control and dysarthric speakers mark stress have been presented in this chapter. The extent to which the two speaker groups use three acoustic features, intensity, F0 and duration, to mark stress was also investigated. The deficiencies in utterances from dysarthric speakers were identified and shown in the increments of the peak intensity, peak F0 and duration of the target. Two listening experiments were set up in this chapter, where the effects of modifications of these acoustic features on the ability of 50 untrained listeners were examined. The results of these experiments indicated that therapists can focus on a single feature during therapy sessions and the different degrees of manipulation is needed for different features. Clinical recommendations for therapy have been made at the end of this chapter on how intensity, F0 and duration can be used to improve dysarthric speakers' intelligibility in stress marking exercises.

Chapter 8

8 Assistive Technology Tools Developed for Dysarthria Management

8.1 Introduction

In this chapter, novel MATLAB-based assistive technology tools for the assessment and treatment of dysarthria will be presented. The tools are developed to assist both therapists and patients in assessing the patient's speech and in tracking their progress during and after therapy. The parent-tool, called DySATTOOL, comprises of four assessment tools and one treatment tool. The first assessment tool, called SETool, was developed to assist the clinicians analyse and extract relevant speech features from the dysarthric speech. The SETool can also be used to analyse other types of speech as its functionalities are not limited to dysarthric speech only. The second assessment tool, called DDKTool, will automatically analyse DDK utterances and track the progress of the patients during the DDK task. The third and fourth assessment tools will be developed to automatically detect dysarthria in speech and classify the speech samples based on the severity. The treatment-related tool presented in this chapter will focus on the use of prosodic features in treatment of dysarthria during stress marking exercise. The functionalities, controls, requirements, as well as limitations of these tools, will also be discussed in this chapter.

8.2 Dysarthria Assessment and Treatment Tool (DySATTOOL)

8.2.1 Overview

DySATTOOL is a tool for the management (assessment and treatment) of dysarthria which involves processing, analysis and classification of audio signals using signal processing technologies. The tool also provides useful information to the users, through visual and non-visual feedback, regarding the state of the

analysed audio signal. The tool is designed to be used by both the patients and the clinicians before, during and after therapy sessions.

The management of dysarthria using instrumental methods has been a major topic of discussion among researchers in recent years. These instrumental methods focus on developing tools for processing speech signal and providing the clinician with feedback to allow them make informed decisions during assessment and therapy. Ability to fully interact with these tools and get valuable feedbacks (both visual and non-visual) is paramount whilst developing such tools. Interpretability and relevance of results provided by these tools are also important. These factors were considered when developing DySATTOOL and the other tools presented in this chapter. A screenshot of DySATTOOL is shown in Figure 8-1. DySATTOL was designed and developed as a parent tool to provide easy access to the other dysarthria assessment and treatment tools discussed in Sections 8.3 to 8.6.

Figure 8-1. Screenshot of Dysarthria Assessment and Treatment Tool (DySATTOOL)

8.2.2 DySATTOOL Functionalities

The DySATTOOL is designed to give both patients and clinicians access to five tools for the management of dysarthria. The tool allows users to save and download user-specific information collated when using the five dysarthria management tools. The five dysarthria management tools are Speech Examination Tool (SET), DDK Analysis Tool (DDKTool), Dysarthria Detection Tool (DyTECTOOL), Stress Marking Task (SMAT), and Dysarthria Severity Classification Tool (DySECTOOL). User's information such as name, reference number and results from assessment and treatment tasks can be safely stored using this tool. The tool makes it easy for clinicians to monitor the progress of patients across to generate the progress report for patients and make an informed decision on the type of treatment to be recommended to such patients.

8.2.3 Controls

Start Session: Starts a new session for DYSATTOOL. User information is collected which includes name and biodata. A reference number is generated for new users which can later be used to access user's progress report. Also, the reference number is transferred to other functions after the session has been started. Therefore, this function has to be selected before all other functions become active. The reference number is stored in the format: Name_Age_Date as a string.

Speech Examination Tool: Launches the SET tool. The Speech Examination Tool analyses audio samples by extracting time-domain and frequency-domain speech features. It also allows users to visualize extracted features in an interactive graphical user interface. Audio signals to be analysed can be recorded directly using this tool,

DDK Analysis Tool: Launches the DDKTool. The DDK Analysis Tool performs an automatic analysis of DDK audio signal using a novel technique. The measured parameters include mean DDK rate, minimum DDK rate, and maximum DDK rate, coefficient of variation of DDK rate, mean Peak Intensity, minimum Peak Intensity, maximum Peak Intensity, and coefficient of variation of Peak Intensity. Audio signals to be analysed can be recorded directly using this tool.

Dysarthria Detection Tool: Launches the DyTECTOOL. The Dysarthria Detection Tool automatically detects dysarthria in speech using 29 speech features and Neural Networks Classification. Audio signals to be analysed can be recorded directly using this tool.

Stress Marking Tool: Launches the SMAT. The Stress Marking Task (SMAT) analyses dysarthric speech during stress marking task. The intensity and fundamental frequency targets during the task are based on research findings. Audio signals to be analysed can be recorded directly using this tool.

Dysarthria Severity Classification Tool: Launches the DySECTOOL for the classification of dysarthric speech into 3 severity levels using 29 speech features and Neural Networks Classification.

Close Session: Closes the current session. It reset all parameters and makes the other functions inactive until a new session is started. Before a session is closed, a message box pops up to confirm if the current session should be closed or not.

8.3 Speech Examination Tool (SETool)

8.3.1 Overview

Analysing speech signals is very important in managing speech-related disorders. Certain speech features need to be extracted in order to effectively analyse these speech signals. SETool allows users to record, analyse and visualise speech signals. The extracted features are displayed and compared in the tool. A screenshot of the tool is shown in Figure 8-2 showing its functionality and controls.

8.3.2 SETool Functionalities

The SETool is designed to allow users to analyse speech signals and extract speech features. The extracted features include fundamental frequency, intensity, ZCR, formants, voice-unvoiced segmentation and time-domain representation (waveform) of the signal.

The SETool also allows users to compare three speech features simultaneously using three plot spaces provided. As shown in Figure 8-2, the waveform, ZCR and STE are displayed simultaneously. In addition, audio signals can be recorded

directly using this tool. The SETool allows users to playback a specified segment of the audio signal being analysed.

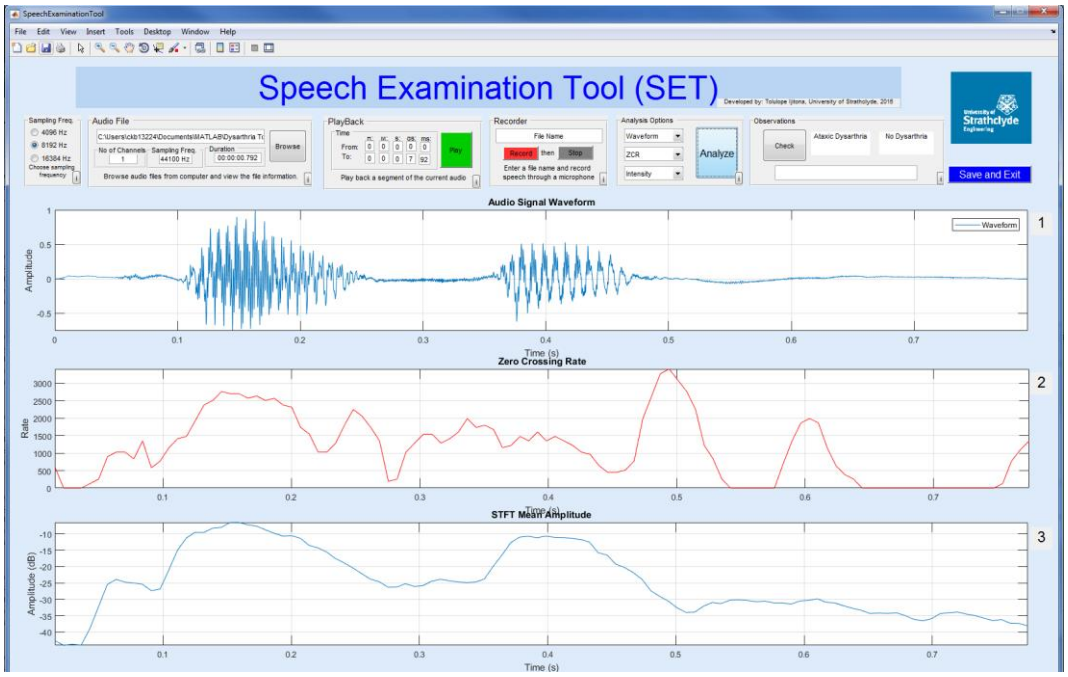


Figure 8-2. Screenshot of Speech Examination Tool (SETool)

8.3.3 Controls

Sampling Frequency: Allows users to choose the preferred sampling frequency. The users choose one out of four available options; 4096 Hz, 8192 Hz, 16384 Hz and 32768 Hz. The default is 8192 Hz.

Audio File: Allows the users to navigate through the computer's folders to select the audio signal to be analysed. The selected path and file name of the signal is displayed after the selection has been made.

Playback: Allows users to playback a specific segment of the audio signal being analysed. The user can specify the start time and end time of the segment to be played back.

Recorder: Records audio signal through the computer's default microphone (or pre-setup microphone) and saves the recorded signal in the current folder using the file name pre-defined by users.

Analysis Options: Contains three drop-down functions which allow the users to choose the feature to extract and display on the three plots. The default function is

the “waveform”. Once the “Analyse” button is clicked the selected audio signal is analysed and the selected features are extracted and displayed on the plot spaces.

Observations: Performs an automatic assessment of the audio signal based on Neural Network classification using MFCC features only. The audio signals are classified as dysarthric or healthy.

Plots 1, 2 and 3: Displays the extracted feature based on the user’s choice in “Analysis Options”. This allows the users to compare up to three extracted features. An illustration of the 3-plot display is shown in Figure 8-2.

Save and Exit: Saves the current extracted data in an excel file and closes the SETool. The file is saved with the reference number assigned to the patient which can be accessed later.

8.4 Automatic DDK Analysis Tool (DDKTool)

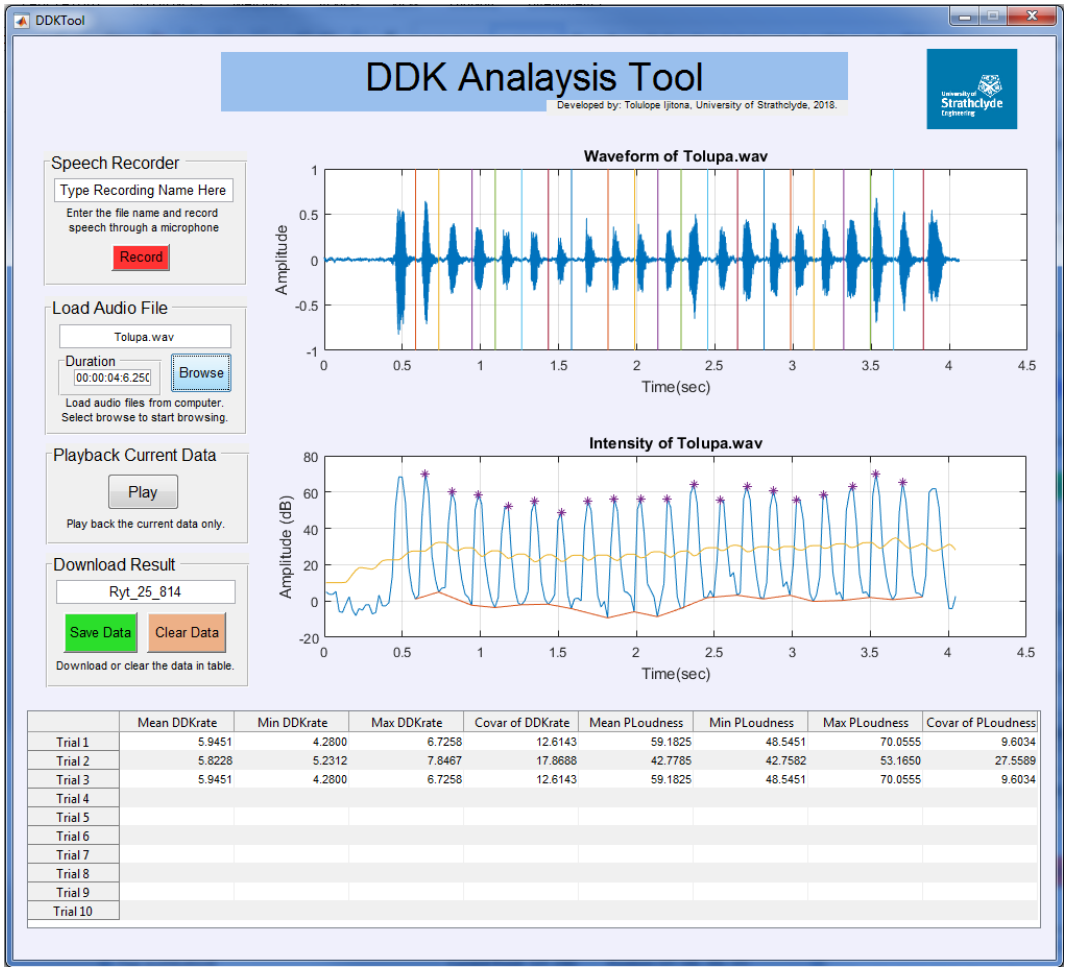
8.4.1 Overview

The automatic DDK analysis tool (DDKTool) analyses and automatically calculates the rate of DDK repetitions in an audio signal. DDKTool is based on the novel automatic DDK analysis tool presented in Chapter 5 of this thesis. This tool records and analyses DDK signals while providing feedback to the users. This tool is very useful in the assessment of speakers with potential DDK difficulty or inconsistency. The DDK task is carried out by allowing the users to produce a repetition of one of the DDK syllables (that is, pa, ta, ka or pataka) as fast as they can. This DDK tasks can be repeated up to 10 times and the clinicians can assess the variability and consistency of the results across various trials. This provides a piece of valuable information on how progressive or otherwise, the speaker’s disorder is. The tool is also useful when checking for the patient’s progress during and after therapy sessions.

8.4.2 Functionalities

The DDKTool allows users to record, save and analyse DDK signals. The tool automatically segments the DDK signals into individual syllables and estimates the duration and peak intensity of each syllable. The tool then estimates the DDKrate

which is the number of DDK syllables per second. The average DDKrate, minimum DDKrate, maximum DDKrate and the covariance of the DDKrate are estimated for each DDK signal. Also, the average, minimum, maximum and the covariance of the peak loudness are also estimated. The screenshot of the tool is illustrated in Figure 8-3. The tool allows users to analyse and compare up to 10 DDK signals while displaying the results of the 10 trials in a table as shown in Figure 8-3. The tool is therefore useful not only in analysing DDK signals but also in examining the speaker's consistency and the variability of the measured parameters.



	Mean DDKrate	Min DDKrate	Max DDKrate	Covar of DDKrate	Mean PLoudness	Min PLoudness	Max PLoudness	Covar of PLoudness
Trial 1	5.9451	4.2800	6.7258	12.6143	59.1825	48.5451	70.0555	9.6034
Trial 2	5.8228	5.2312	7.8467	17.8688	42.7785	42.7582	53.1650	27.5589
Trial 3	5.9451	4.2800	6.7258	12.6143	59.1825	48.5451	70.0555	9.6034
Trial 4								
Trial 5								
Trial 6								
Trial 7								
Trial 8								
Trial 9								
Trial 10								

Figure 8-3. Screenshot of the Automatic DDK Analysis Tool (DDKTool)

8.4.3 Controls

Speech Recorder: Records DDK audio signals and stores them as .wav files in the current folder. The users are allowed to specify the preferred file name. The recorded DDK audio signal is automatically analysed and the results of the segmentation are displayed on the two plots.

Load Audio File: Allows the users to navigate through the computer's folders to select the audio signal to be analysed. The selected path, file name and the duration of the selected audio signal are displayed after the selection has been made.

Playback Current Data: Plays back the current audio signal being analysed. This could be the recorded audio signal or the loaded audio signal.

Plots: The DDKTool consists of two plots arranged vertically. The first plot shows the waveform of the segmented audio signal and the second plot shows the peak intensities of the individual DDK syllable.

Table: The value of eight extracted features are displayed in the table for up to 10 trials. It allows users to compare the results from up to 10 trials. Variation across trials can be accessed across the eight measured parameters.

Download Result: Consists of two functions; save data and clear data. The save data function allows users to save the data in the table in an excel file using the patient's reference number as discussed in Section 8.2. The clear function, on the other hand, clears all the data in the table and clears the plot areas.

8.5 Automatic Dysarthria Detection Tool (DyDECTOOL)

8.5.1 Overview

In Chapter 6, novel techniques for the detection of dysarthria in speech signal using extended feature extraction and machine learning classification techniques were presented. An automatic detection tool called DyDECTOOL was designed and developed to give patients and clinicians access to these novel techniques developed in MATLAB. The DyDECTOOL automatically analyses and classifies speech signals by extracting 29 prosodic, voice quality, pronunciation and wavelets features as described in Section 6.4.2. The extracted features are then classified using machine learning classification techniques as presented in Section 6.4.4. The classification output (result), as well as the extracted features, are visually presented to the users as shown in the screenshot in Figure 8-4.

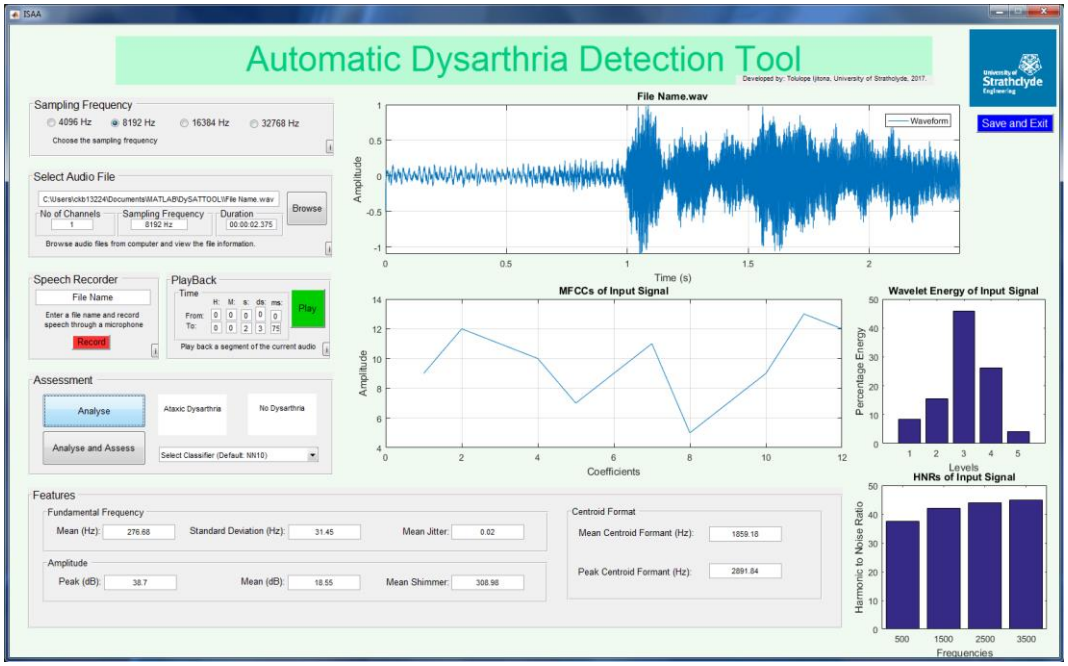


Figure 8-4. Screenshot of Automatic Dysarthria Detection Tool (DyDECTOOL)

8.5.2 Tool Functionalities

DyDECTOOL analyses and classifies speech signal into two classes (dysarthric and healthy classes) using the classification technique chosen by the user. The extracted features including Mel-Frequency Cepstrum Coefficients as well as the 4th level wavelet energies and the harmonic-to-noise ratios of the signal are visually presented as illustrated in Figure 8-4. This allows the users to further examine and assess the extracted speech features. The output of the assessment can be saved in an excel file with a file name corresponding to the patient's reference number.

8.5.3 Controls

The "Sampling Frequency", "Select Audio File", "Playback" and "Speech Recorder" controls perform the same functions as those in SETool (as described in Section 8.3.3).

Assessment: this function allows users to either analyse or analyse & assess the speech signal. It also allows users to choose the preferred classification technique for the assessment.

Features & Plots: The 29 extracted features are either presented pictorially using plots or in a textbox, as shown in Figure 8-4. Fundamental frequency, intensity and

formant-based features are presented in textbox while others are illustrated for interpretation and comparison.

Save and Exit: Enables the users to save the extracted features as well as the result of the assessment in an excel file using the patient's reference number before exiting the tool.

8.5.4 Dysarthria Severity Classification Tool (DySECTOOL)

The dysarthria severity classification tool (DySECTOOL) is quite similar to the DyDECTOOL except that the speech signals are classified into three severity levels (mild, moderate and severe). The screenshot of DySECTOOL is illustrated in Figure 8-5. It is expected that the DySECTOOL will only be used in assessing the speech of a patient after the speech signal has been classified as dysarthric using the DyDECTOOL.

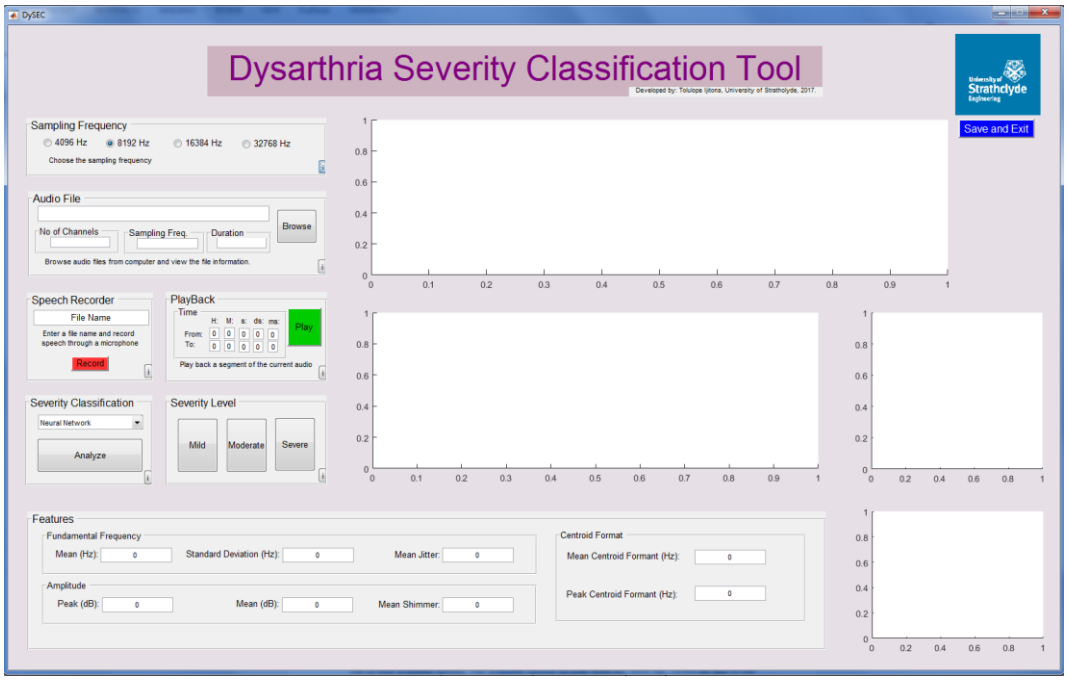


Figure 8-5. Screenshot of Dysarthria Severity Classification Tool (DySECTOOL)

The DySECTOOL provides the additional function of classifying dysarthric speech signals into 3 severity levels based on the novel severity classification technique presented in Section 6.5 of this thesis. Apart from the similar "Sampling Frequency", "Select Audio File", "Playback" and "Speech Recorder" and "Save

and Exit" controls, DySECTOOL has two additional controls namely; "Severity Classification" and "Severity Level".

Severity Classification: Enables the users to choose the preferred classification technique to be used in classifying the dysarthric speech signal.

Severity Level: Shows the output of the severity classification. The corresponding severity level (mild, moderate or severe) is highlighted.

8.6 Stress Marking Task (SMAT)

8.6.1 Overview

The stress marking task (SMAT) is a behavioural treatment tool that is designed to assist in improving speakers' intelligibility during therapy sessions. This tool is developed based on the research findings, presented in Chapter 7 of this thesis, on the effects of modification of intensity, fundamental frequency and duration on stress marking in dysarthric speech. The results have shown that listeners are more likely to identify the stressed word in a sentence if the intensity is increased by 50% or the fundamental frequency is increased by 100% or the duration is increased by 50%. Prior to the use of this tool, clinicians are advised to carry out an initial assessment of the patient to determine the most effective feature to work on based on their severity and ability to speak louder, raise the pitch or elongate words.

8.6.2 Tool Functionalities

SMAT consists of 120 exercises (that is, 40 exercises per feature) involving 10 sentences with SVOA (Subject-Verb-Object-Adverbial) structure. There are four target positions per sentence. These are no target/stress, stress at the beginning of the sentence (that is, on the subject, T_1), stress in the middle of the sentence (that is, on the object, T_2) and stress at the end of the sentence (that is, on the adverbial T_3). As illustrated in Figure 8-6, the exercises are selected by specifying the preferred feature, the sentence number and the target position. For each exercise, the texts of the sentence are displayed on the top right side of the interface and the speakers are required to stress the highlighted word. In the plot area, the target feature-profile is displayed on top and the speaker feature-profile is displayed

below. (Please note that the feature-profile can either be intensity profile, fundamental frequency profile or duration profile). Each exercise is repeated multiple times until the speaker-profile matched the target-profile.

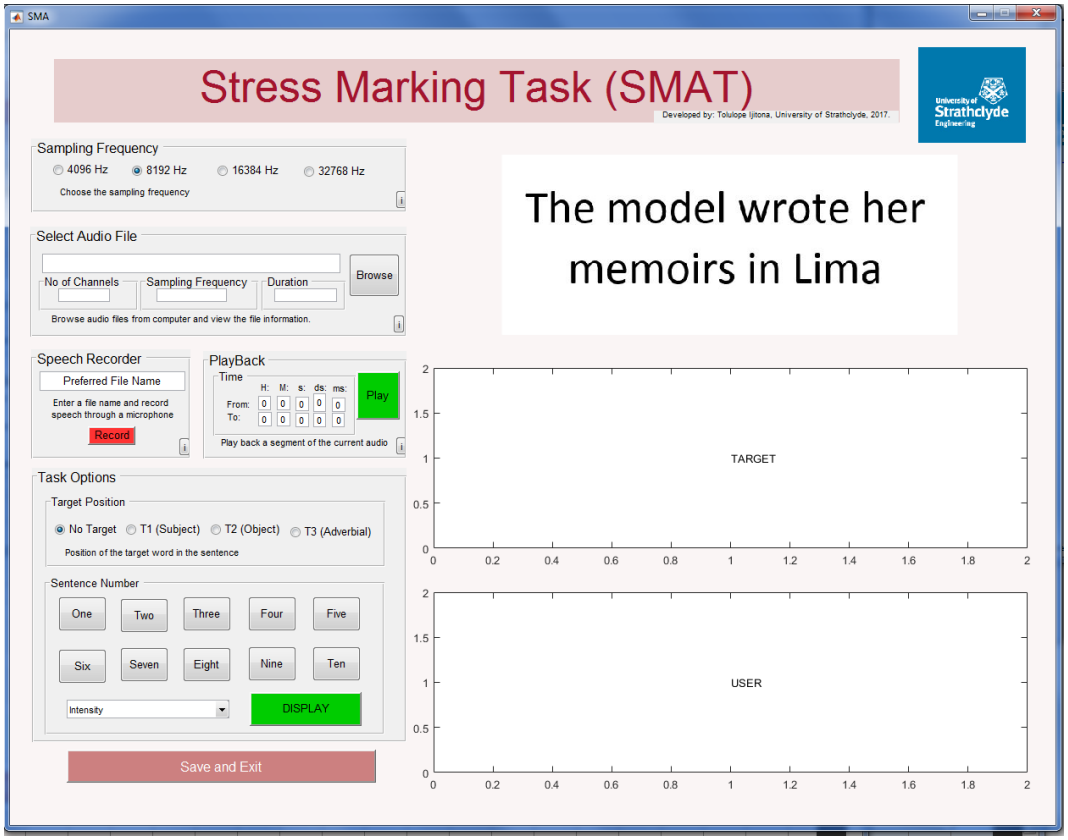


Figure 8-6. Screenshot of the Stress Marking Task Tool (SMAT)

8.6.3 Controls

Apart from the common controls "Sampling Frequency", "Select Audio File", "Playback" and "Speech Recorder", the additional controls include:

Target Position: Allows users to select one of the four target positions.

Sentence Number: Allows users to select one of the 10 SVOA sentences

Feature Selection: Allows users to select the preferred feature to modify (that is, intensity, F0 or duration).

8.7 Summary

In this chapter, novel assistive technology tools for the assessment and treatment of dysarthria through the use of digital signal processing principles and machine

learning techniques have been presented. These tools developed, in MATLAB, will assist users (clinicians and patients) in assessing dysarthric speech by extracting relevant speech features usefully in the management of this speech disorder.

Chapter 9

9 Conclusions and Future Works

9.1 Conclusion

In this thesis, a variety of signal processing techniques have been developed and used in the analysis, assessment and treatment of a neurological speech disorder called dysarthria which include Short-Time Fourier Transform, Wavelet Transform, Formant Analysis, Syllable Segmentation, Prosody (intensity, duration and pitch) Modification, Speech Segmentation, Harmonic–to-Noise Ratio Analysis, Mel Frequency Cepstral Analysis, and other Speech Processing techniques. The application of these techniques as well as Machine Learning Classification techniques has been researched and proposed for the management of dysarthria.

In Chapter 2, an extensive literature review was presented, focused on various techniques used by clinicians as well as techniques proposed by researchers over the years. Key areas such as how the severity of dysarthria is measured and classified with respect to the speech intelligibility, perceptual and acoustic techniques used in dysarthria assessment, strategies used in dysarthria management and techniques used in the treatment of dysarthria are reviewed. Specifically, current techniques for dysarthria management including the Lee Silverman Voice Treatment, Dysarthria Treatment Programme, Computerised Assessment and Treatment of Rate, Intonation and Stress, and Music Therapy were reviewed. The major gaps identified in these approaches include robustness (most methods are focused on one of two features), need for improved accuracy, ease of application and need for an automatic system with little or no human interference.

In Chapter 3, speech features that differentiate dysarthric speech from healthy controlled speech were reviewed. Techniques used in pre-processing the speech signals to a form where the speech features are easily and effectively extracted were presented in this chapter. The extracted features include time-domain features (STE, ZCR and durational features), spectral features (fundamental frequency and formants), cepstral features (Mel Frequency Cepstral Coefficients) and extended

features (jitter, shimmer, harmonic-to-noise ratio, and wavelets). Various feature extraction techniques were analysed and their effects on dysarthric speech were discussed in this chapter. A review of methods used in Silence-Unvoiced-Voiced segmentation was also presented as well as a review of machine learning techniques used in the classification of dysarthric speech. Research gaps in these techniques were identified and novel dysarthria management methods proposed in Chapters 4 to 8 were aimed at addressing these gaps.

The first novel contribution of this thesis, an algorithm for the automatic silence-unvoiced-voiced segmentation of dysarthric speech, was presented in Chapter 4. This method is an improved segmentation approach which makes use of STE, LPEV and ZCR. This method uses a two-layer approach to reduce segmentation errors due to reduced loudness in dysarthric speech. The first layer of segmentation combines LPEV and STE to separate silence segments whilst the second layer uses the ZCR to distinguish between voiced and unvoiced segments. Experimental results showed that the use of speaker-specific thresholding helps in improving the segmentation performance despite the speaker's voice quality and severity.

In Chapter 5, a novel scheme for the automatic analysis of DDK productions for the assessment of dysarthria was presented. The algorithm is an enhanced segmentation method based on a speaker-specific threshold which varies with the speaker's intensity and voice quality. The automatic DDK analysis algorithm comprises of three steps which include DDK syllable segmentation, minimum duration merging and DDK metrics estimation. In the first step, the segmentation threshold is calculated as a function of the moving average of the DDK signals. This is followed by rectification of syllable over-segmentation by merging pseudo-syllables based on the minimum DDK syllable duration. In the last step, the DDK metrics which includes DDK rates and DDK covariance for each production is calculated as the number of DDK syllables per second. Experimental results showed that this algorithm gives a better segmentation performance than manually labelled syllables and makes the DDK metrics easier and faster to compute.

Novel methods for the detection of dysarthria and classification of speakers' severity using machine learning techniques were presented in Chapter 6. The first method presented in this chapter consists of an algorithm for the automatic detection

ataxic dysarthria using extended speech features called Centroid Formants. This method results in an accuracy of 75.6% using neural network classification technique. The second method presented in Section 6.4 consists of a novel robust automatic dysarthria detection algorithm which combines prosody, voice quality, pronunciation and wavelet features in the development of the classification feature vector comprising of 23 features. Automatic detection was carried out using multiple machine learning classifiers and the ANN classifier gives the best performance with an accuracy of 99.4%. The method was further developed to classify the dysarthric speech signals into various severity levels based on the speakers' intelligibility scores. The neural network classifier gives the highest performance with an accuracy of 99.7% in classifying the speech signals into 3 severity levels.

In Chapter 7 of this thesis, the stress production deficits in the dysarthric speech were investigated with the aim to help clinicians make informed decisions when managing dysarthria using stress production exercise. This investigation includes the identification of deficits by analysing both dysarthric and healthy controlled speech samples. The analysis reveals that all the three prosodic cues (intensity, fundamental frequency and duration) used by healthy controlled speakers in marking stress are impacted in dysarthric speech. The intensity, fundamental frequency and duration of the stressed word in the dysarthric utterances were lower than that of healthy controlled speakers. The effects of modifying these three prosodic features in the ability of listeners to correctly identify the stressed word in dysarthric utterances were further investigated. The experimental results from the investigation reveal that modification of at least one of the prosodic features was sufficient to increase the listeners' ability to correctly identify the stressed word in dysarthric utterances. Even though the required modifications differ for the three features (50 % increment for intensity, 100% increment for fundamental frequency and 50% increment for the duration), the results of the investigation show that clinicians can focus on improving any one of the features to get similar results. This would also be useful when speakers are unable to modify multiple features simultaneously.

The algorithms presented in Chapters 4 to 7 were developed into interactive tools in MATLAB and these tools were presented in Chapter 8 of this thesis. The first tool presented was the parent tool called DySATTOOL from which the other tools which include SETool, DDKTool, DyDECTOOL, DySECTOOL and SMAT can be accessed. The SETool allows users to analyse speech samples and extract speech features such as Short-Time Energy, Fundamental Frequency, Formants, Zero Crossing Rate and MFCCs. The DDKTool was developed to allow users to record and analyse DDK samples by automatically segmenting the DDK samples into individual syllables and calculating the average DDK rate as well as the covariance of the DDK rates. The DyDECTOOL and DySECTOOL allow users to perform automatic detection and severity classification of dysarthric speech respectively using the novel algorithms presented in Chapter 6. The sixth tool called SMAT was developed to assist users in extracting and measuring the three prosodic features required to mark stress during a stress production exercise.

9.2 Future Work

In the work presented in this thesis, there are various research topics that can be explored for future research. In Chapter 4, the use of linear prediction error variance (LPEV) in the segmentation of dysarthric speech has been proposed. The inclusion of the LPEV feature in silence unvoiced voiced segmentation of dysarthric speech has shown a better performance when compared with the traditional methods and its application in dysarthric speech recognition needs to be explored since the LPEV is not influenced by varying intensity experienced in dysarthric speech.

Moreover, one of the notable contributions of this research proposed in Chapter 6 (Section 6.3) was the introduction of an extended feature called the Centroid Formant. The application of the Centroid Formants in the detection of Ataxic dysarthria shows promising results but its application in the detection of other types of dysarthria needs to be further explored. Also, the Centroid Formants as a representation of the energy distribution in the frequency domain contains unique information about the speaker which can be used in speech recognition applications. Another interesting research area where the Centroid Formants can be applied is in the detection of emotions in human speech. These and other potential applications of the Centroid Formants have been left for future research.

Furthermore, one of the questions yet to be answered is the ability to differentiate the six types of dysarthria using speech processing techniques. Considering the novel technique for the detection of dysarthria using prosody, voice quality, pronunciation and wavelets features presented in Chapter 6, future research may include the application of these features in the classification of the various types of dysarthria.

Finally, the algorithms presented in this research work have shown promising results in the detection, classification and clinical management of dysarthria, there is, however, a need to further explore the application of these algorithms in other speech-related and clinical applications. There is also a need to further test their performances using datasets relating to other motor speech disorders.

References

[1] A. Hernandez and M. Chung, "Dysarthria Classification Using Acoustic Properties of Fricatives," *Proceedings of SICSS*, vol. 2019, p. 16, 2019.

[2] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

[3] P. McCaffrey. "Dysarthria Characteristics." California State University. http://www.csuchico.edu/~pmccaffrey/syllabi/SPPA342/342unit14.html (accessed 7 January, 2016).

[4] P. McCaffrey. "Dysarthria: Definition and Description; Etiology." California State University. http://www.csuchico.edu/~pmccaffrey/syllabi/SPPA342/342unit11.html (accessed 13 March, 2016).

[5] D. B. Freed, *Motor Speech Disorders: Diagnosis & Treatment*. Cengage Learning, 2011.

[6] E. C. Guerra and D. F. Lovey, "A modern approach to dysarthria classification" in *Proceedings of the 25th Annual International Conference of the IEEE* in *Engineering in Medicine and Biology Society*, pp. 2257-2260, Sept. 2003.

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

[8] K. Gurugubelli and A. K. Vuppala, "Perceptually Enhanced Single Frequency Filtering for Dysarthric Speech Detection and Intelligibility Assessment," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6410-6414, 2019.

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

[10] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential Diagnostic Patterns of Dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 246-269, 1969.

[11] 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 *2015 International Symposium on Signals, Circuits and Systems*, pp. 1-4, July 2015.

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

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

[14] A. Rueda, J. C. Vásquez-Correa, C. D. Rios-Urrego, J. R. Orozco-Arroyave, S. Krishnan, and E. Nöth, "Feature representation of pathophysiology of Parkinsonian dysarthria," in *Proceedings of INTERSPEECH*, pp. 1-5, 2019.

[15] J. P. Hosom, A. B. Kain, T. Mishra, J. P. H. v. Santen, M. Fried-Oken, and J. Staehely, "Intelligibility of modifications to dysarthric speech," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings*, pp. I-924-I-927, April 2003.

[16] K. Johnston, P. Morrow, B. Scotney, and O. Duffy, "Lip Contour Identification in Texture Data of 3D Face Mesh Sequences," in *2011 Irish Machine Vision and Image Processing Conference*, pp. 20-25, Sept. 2011.

[17] M. Pacula, T. Meltzer, M. Crystal, A. Srivastava, and B. Marx, "Automatic detection of psychological distress indicators and severity assessment in crisis hotline conversations," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4863-4867, May 2014.

[18] H. Martens, T. Dekens, G. Van Nuffelen, L. Latacz, W. Verhelst, and M. De Bodt, "Automated Speech Rate Measurement in Dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 3, pp. 698-712, 2015.

[19] J. Carmichael, "Dysarthria diagnosis via respiration and phonation," in *2015 International Conference and Workshop on Computing and Communication*, pp. 1-5, Oct. 2015.

[20] C. Zhang, J. Dang, J. Zhang, and J. Wei, "Investigation on articulatory and acoustic characteristics of dysarthria," in *9th International Symposium on Chinese Spoken Language Processing*, pp. 326-330, Sept. 2014.

[21] R. Palmer and P. Enderby, "Methods of speech therapy treatment for stable dysarthria: A review," *Advances in Speech Language Pathology*, vol. 9, no. 2, pp. 140-153, 2007.

[22] L. A. Mahler and L. O. Ramig, "Intensive treatment of dysarthria secondary to stroke," *Clinical Linguistics & Phonetics*, vol. 26, no. 8, pp. 681-694, 2012.

[23] C. Atkinson-Clement *et al.*, "Psychosocial Impact of Dysarthria: The Patient-Reported Outcome as Part of the Clinical Management," *Neurodegenerative Diseases*, pp. 1-10, 2019.

[24] T. Arias-Vergara, J. C. Vasquez-Correa, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Noeth, "Gender-dependent GMM-UBM for tracking Parkinson's disease progression from speech," in *12th ITG Symposium in Speech Communication*, pp. 1-5, 2016.

[25] M. Jalil, F. A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *2013 International Conference on Technological Advances in Electrical, Electronics and Computer Engineering*, 2013, pp. 208-212, May 2013.

[26] M. V. Mujumdar and R. F. Kubichek, "Design of a dysarthria classifier using global statistics of speech features," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing* pp. 582-585, March 2010.

[27] A. Lowit, A. Kuschmann, and K. Kavanagh, "Phonological markers of sentence stress in ataxic dysarthria and their relationship to perceptual cues," *Journal of communication disorders*, vol. 50, pp. 8-18, 2014.

[28] M. Dhanalakshmi and P. Vijayalakshmi, "Intelligibility modification of dysarthric speech using HMM-based adaptive synthesis system," in *2nd International Conference on Biomedical Engineering*, pp. 1-5, March 2015.

[29] R. Cardoso *et al.*, "Frenchay dysarthria assessment (FDA-2) in Parkinson's disease: cross-cultural adaptation and psychometric properties of the

European Portuguese version*Journal of neurology,* vol. 264, no. 1, pp. 21-31, 2017.

[30] G. Defazio, M. Guerrieri, D. Liuzzi, A. F. Gigante, and V. Di Nicola, "Assessment of voice and speech symptoms in early Parkinson's disease by the Robertson dysarthria profile," *Neurological Sciences,* vol. 37, no. 3, pp. 443-449, 2016.

[31] S. L. Christina, P. Vijayalakshmi, and T. Nagarajan, "HMM-based speech recognition system for the dysarthric speech evaluation of articulatory subsystem," in *2012 International Conference on Recent Trends In Information Technology*, pp. 54-59, April 2012.

[32] R. Patel and P. Campellone, "Acoustic and perceptual cues to contrastive stress in dysarthria" *Journal of Speech, Language, and Hearing Research,* vol. 52, no. 1, pp. 206-222, 2009.

[33] P. Enderby, "Frenchay dysarthria assessment," *British Journal of Disorders of Communication,* vol. 15, no. 3, pp. 165-173, 1980.

[34] S. Robertson, "The efficacy of oro-facial and articulation exercises in dysarthria following stroke," *International journal of language & communication disorders,* vol. 36, no. sup1, pp. 292-297, 2001.

[35] R. D. Kent, *Intelligibility in speech disorders: Theory, measurement and management*. John Benjamins Publishing, 1992.

[36] T. H. Falk, R. Hummel, and W. Y. Chan, "Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4480-4483, May 2011.

[37] C. Fangxin and A. Kostov, "Optimization of dysarthric speech recognition," in *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1997. P*, 1997, pp. 1436-1439, Nov 1997.

[38] M. S. De Bodt, M. E. H.-D. a. Huici, and P. H. Van De Heyning, "Intelligibility as a linear combination of dimensions in dysarthric speech," *Journal of communication disorders,* vol. 35, no. 3, pp. 283-292, 2002.

[39] K. M. Yorkston, E. A. Strand, and M. R. Kennedy, "Comprehensibility of dysarthric speech: Implications for assessment and treatment planning," *American Journal of Speech-Language Pathology*, vol. 5, no. 1, pp. 55-66, 1996.

[40] A. D. Dykstra, M. E. Hakel, and S. G. Adams, "Application of the ICF in Reduced Speech Intelligibility in Dysarthria,", *Semin Speech Lang,* vol. 28, no. 04, pp. 301-311, 2007.

[41] K. M. Yorkston, M. Hakel, D. R. Beukelman, and S. Fager, "Evidence for effectiveness of treatment of loudness, rate, or prosody in dysarthria: A systematic review," *Journal of Medical Speech-Language Pathology*, vol. 15, no. 2, pp. xi-xi, 2007.

[42] K. Yorkston, D. Beukelman, and R. Tice, "Phoneme Intelligibility Test," *Lincoln, NE: Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital*, 1999.

[43] K. M. Yorkston and D. R. Beukelman, "A comparison of techniques for measuring intelligibility of dysarthric speech," *Journal of communication disorders,* vol. 11, no. 6, pp. 499-512, 1978.

[44] M. Rousseaux, P. Krystkowiak, O. Kozlowski, C. Özsancak, S. Blond, and A. Destée, "Effects of subthalamic nucleus stimulation on parkinsonian dysarthria and speech intelligibility," *Journal of neurology*, vol. 251, no. 3, pp. 327-334, 2004.

[45] V. Hazan and D. Markham, "Acoustic-phonetic correlates of talker intelligibility for adults and children," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3108-3118, 2004.

[46] T. Khan, J. Westin, and M. Dougherty, "Classification of speech intelligibility in Parkinson's disease," *Biocybernetics and Biomedical Engineering*, vol. 34, no. 1, pp. 35-45, 2014.

[47] G. Weismer and J. S. Laures, "Direct magnitude estimates of speech intelligibility in dysarthria: Effects of a chosen standard," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 3, pp. 421-433, 2002.

[48] T. B. Lagerberg, J. Å. Johnels, L. Hartelius, and C. Persson, "Effect of the number of presentations on listener transcriptions and reliability in the assessment of speech intelligibility in children," *International journal of language & communication disorders*, vol. 50, no. 4, pp. 476-487, 2015.

[49] J. K.-Y. Ma, C. B. Schneider, R. Hoffmann, and A. Storch, "Speech prosody across stimulus types for individuals with Parkinson's Disease," *Journal of Parkinson's disease*, vol. 5, no. 2, pp. 291-299, 2015.

[50] K. M. Yorkston, V. L. Hammen, D. R. Beukelman, and C. D. Traynor, "The effect of rate control on the intelligibility and naturalness of dysarthric speech," *Journal of Speech and Hearing Disorders*, vol. 55, no. 3, pp. 550-560, 1990.

[51] F. Chen, T. Guan, and L. L. N. Wong, "Effect of temporal fine structure on speech intelligibility modeling," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4199-4202, July 2013.

[52] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4924-4927, May 2011.

[53] M. J. Kim, Y. Kim, and H. Kim, "Automatic Intelligibility Assessment of Dysarthric Speech Using Phonologically-Structured Sparse Linear Model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694-704, 2015.

[54] R. Ivers and N. Goldstein, "MULTIPLE SCLEROSIS: A CURRENT APPRAISAL OF SYMPTOMS AND SIGNS," in *Proceedings of the staff meetings. Mayo Clinic*, vol. 38, p. 457, 1963.

[55] P. Green and J. Carmichael, "Revisiting dysarthria assessment intelligibility metrics," in *Eighth International Conference on Spoken Language Processing*, 2004.

[56] S. Robertson, "Robertson Dysarthria Profile," *Buckinghamshire: Winslow*, 1982.

[57] S. J. Robertson and F. E. Thomson, *Working with Dysarthric Clients: Practical Guide to Therapy for Dysarthria*. Communication Skill Builders, 1987.

[58] S. J. Robertson and F. Thomson, "Speech therapy in Parkinson's disease: a study of the efficacy and long term effects of intensive treatment," *International Journal of Language & Communication Disorders*, vol. 19, no. 3, pp. 213-224, 1984.

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

[60] S. R. Shahamiri and S. K. Ray, "On the use of array learners towards Automatic Speech Recognition for dysarthria," in *2015 IEEE 10th Conference on Industrial Electronics and Applications*, pp. 1283-1287, June 2015.

[61] R. Sriranjani, M. R. Reddy, and S. Umesh, "Improved acoustic modeling for automatic dysarthric speech recognition," in *2015 Twenty First National Conference on Communications*, pp. 1-6, March 2015.

[62] H. Tolba and A. S. El, "Towards the improvement of automatic recognition of dysarthric speech," in *2nd IEEE International Conference on Computer Science and Information Technolog*, pp. 277-281, Aug. 2009.

[63] R. J. Wenke, P. Cornwell, and D. G. Theodoros, "Changes to articulation following LSVT® and traditional dysarthria therapy in non-progressive dysarthria," *International Journal of Speech-Language Pathology*, vol. 12, no. 3, pp. 203-220, 2010.

[64] M. S. E. Langarani and J. v. Santen, "Modeling fundamental frequency dynamics in hypokinetic dysarthria," in *IEEE Spoken Language Technology Workshop*, pp. 272-276, Dec. 2014.

[65] L. Baghai-Ravary and S. W. Beet, *Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders*. Springer Science & Business Media, 2012.

[66] G. Vyas, M. K. Dutta, J. Prinosil, and P. Harár, "An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features," in *39th IEEE International Conference on Telecommunications and Signal Processing*, pp. 515-518, 2016.

[67] E. A. Belalcazar-Bolaños, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, "Automatic detection of Parkinson's disease using noise measures of speech," in *Symposium of Signals, Images and Artificial Vision*, pp. 1-5, Sept. 2013.

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

[69] K. U. R and M. S. Holi, "Automatic detection of neurological disordered voices using mel cepstral coefficients and neural networks," in *2013 IEEE Point-of-Care Healthcare Technologies*, pp. 76-79, Jan. 2013.

[70] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Computer speech & language*, vol. 29, no. 1, pp. 132-144, 2015.

[71] 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.

[72] J. Kim *et al.*, "Automatic estimation of Parkinson's disease severity from diverse speech tasks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[73] K. L. Kadi, S. A. Selouani, B. Boudraa, and M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease

using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 233-247, 2016.

[74] H. Dahmani, S.-A. Selouani, D. O'shaughnessy, M. Chetouani, and N. Doghmane, "Assessment of dysarthric speech through rhythm metrics," *Journal of King Saud University-Computer and Information Sciences*, vol. 25, no. 1, pp. 43-49, 2013.

[75] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully automated assessment of the severity of Parkinson's disease from speech," *Computer speech & language*, vol. 29, no. 1, pp. 172-185, 2015.

[76] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, p. 629030, 2009.

[77] R. Norel, M. Pietrowicz, C. Agurto, S. Rishoni, and G. Cecchi, "Detection of amyotrophic lateral sclerosis (ALS) via acoustic analysis", 2018.

[78] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. Narayanan, "Intelligibility classification of pathological speech using fusion of multiple high level descriptors," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[79] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5070-5074, March 2017.

[80] T. B. Ijitona, J. J. Soraghan, A. Lowit, G. Di-Caterina, and H. Yue, "Automatic detection of speech disorder in dysarthria using extended speech feature extraction and neural networks classification," in *IET 3rd International Conference on Intelligent Signal Processing*, pp. 1-6, Dec. 2017.

[81] X. Wang, J. Zhang, and Y. Yan, "Automatic Detection of Pathological Voices Using GMM-SVM Method," in *2009 2nd International Conference on Biomedical Engineering and Informatics*, pp. 1-4, Oct. 2009.

[82] N. Souissi and A. Cherif, "Dimensionality reduction for voice disorders identification system based on Mel Frequency Cepstral Coefficients and Support Vector Machine," in *7th International Conference on Modelling, Identification and Control*, pp. 1-6, Dec. 2015.

[83] T. Kinnunen *et al.*, "Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1990-2001, 2012.

[84] S. Gupta, S. Sharanyan, and A. Mukherjee, "Performance analysis of Support Vector Machine as classifier for voiced and unvoiced speech," in *2010 International Conference on Computer and Communication Technology*, pp. 397-401, Sept. 2010.

[85] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "Hmm-Based and Svm-Based Recognition of the Speech of Talkers With Spastic Dysarthria," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. III-III, May 2006.

[86] M. N. Stolar, M. Lech, and N. B. Allen, "Detection of depression in adolescents based on statistical modeling of emotional influences in parent-adolescent conversations," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 987-991, April 2015.

[87] H. Ackermann and I. Hertrich, "Speech rate and rhythm in cerebellar dysarthria: An acoustic analysis of syllabic timing," *Folia Phoniatrica et Logopaedica,* vol. 46, no. 2, pp. 70-78, 1994.

[88] V. L. Hammen and K. M. Yorkston, "Speech and pause characteristics following speech rate reduction in hypokinetic dysarthria," *Journal of communication disorders,* vol. 29, no. 6, pp. 429-445, 1996.

[89] K. C. Hustad, T. Jones, and S. Dailey, "Implementing speech supplementation strategies: effects on intelligibility and speech rate of individuals with chronic severe dysarthria," *Journal of Speech, Language, and Hearing Research,* vol. 46, no. 2, pp. 462-474, 2003.

[90] H. Martens, G. V. Nuffelen, M. D. Bodt, T. Dekens, L. Latacz, and W. Verhelst, "Automated assessment and treatment of speech rate and intonation in dysarthria," in *7th International Conference on Pervasive Computing Technologies for Healthcare*, pp. 382-384, May 2013.

[91] H. M. Clark, "Neuromuscular treatments for speech and swallowing: A tutorial," *American Journal of Speech-Language Pathology,* vol. 12, no. 4, pp. 400-415, 2003.

[92] D. P. Kuehn, "The development of a new technique for treating hypernasality: CPAP," *American Journal of Speech-Language Pathology,* vol. 6, no. 4, pp. 5-8, 1997.

[93] K. M. Yorkston and D. R. Beukelman, "Ataxic DysarthriaTreatment Sequences Based on Intelligibility and Prosodic Considerations," *Journal of Speech and Hearing Disorders,* vol. 46, no. 4, pp. 398-404, 1981.

[94] P. A. Dagenais, G. R. Brown, and R. E. Moore, "Speech rate effects upon intelligibility and acceptability of dysarthric speech," *Clinical Linguistics & Phonetics,* vol. 20, no. 2-3, pp. 141-148, 2006.

[95] M. A. Pilon, K. W. McIntosh, and M. H. Thaut, "Auditory vs visual speech timing cues as external rate control to enhance verbal intelligibility in mixed spastic ataxic dysarthric speakers: a pilot study," *Brain Injury,* vol. 12, no. 9, pp. 793-803, 1998.

[96] M. Nishio and S. Niimi, "Comparison of speaking rate, articulation rate and alternating motion rate in dysarthric speakers," *Folia Phoniatrica et Logopaedica,* vol. 58, no. 2, pp. 114-131, 2006.

[97] C. M. Fox, L. O. Ramig, M. R. Ciucci, S. Sapir, D. H. McFarland, and B. G. Farley, "The science and practice of LSVT/LOUD: neural plasticity-principled approach to treating individuals with Parkinson disease and other neurological disorders," in *Seminars in speech and language*, 2006, vol. 27, no. 4, pp. 283-299.

[98] S. Borrie, M. McAuliffe, G. Tillard, T. Ormond, T. Anderson, and J. Hornibrook, "Effect of Lee Silverman Voice Treatment (LSVT®) on articulation in speakers with Parkinson's Disease," *New Zealand Journal of Speech-Language Therapy,* vol. 62, pp. 29-36, 2007.

[99] A. Lowit, C. Dobinson, C. Timmins, P. Howell, and B. Kröger, "The effectiveness of traditional methods and altered auditory feedback in improving speech rate and intelligibility in speakers with Parkinson's disease," *International Journal of Speech-Language Pathology,* vol. 12, no. 5, pp. 426-436, 2010.

[100] N. P. Solomon, A. S. McKee, and S. Garcia-Barry, "Intensive voice treatment and respiration treatment for hypokinetic-spastic dysarthria after

traumatic brain injury," *American Journal of Speech-Language Pathology*, vol. 10, no. 1, pp. 51-64, 2001.

[101] Drummond, S., L. Worley, and A. Watson, "Description and implementation of a dysarthria treatment program.," ed. ASHA Convention, Chicago, 2003.

[102] H. Martens *et al.*, "The effect of intensive speech rate and intonation therapy on intelligibility in Parkinson's disease," *Journal of communication disorders*, vol. 58, pp. 91-105, 2015.

[103] J. Tamplin and D. Grocke, "A music therapy treatment protocol for acquired dysarthria rehabilitation," *Music Therapy Perspectives*, vol. 26, no. 1, pp. 23-29, 2008.

[104] J. Tamplin and F. A. Baker, "Therapeutic singing protocols for addressing acquired and degenerative speech disorders in adults," *Music Therapy Perspectives*, vol. 35, no. 2, pp. 113-123, 2017.

[105] S. Park, D. Theodoros, E. Finch, and E. Cardell, "Be clear: A new intensive speech treatment for adults with nonprogressive dysarthria," *American Journal of Speech-Language Pathology*, vol. 25, no. 1, pp. 97-110, 2016.

[106] J. Tamplin, "A pilot study into the effect of vocal exercises and singing on dysarthric speech," *NeuroRehabilitation*, vol. 23, no. 3, pp. 207-216, 2008.

[107] G. Castellanos, G. Daza, L. Sanchez, O. Castrillon, and J. Suarez, "Acoustic Speech Analysis for Hypernasality Detection in Children," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5507-5510, Sept. 2006.

[108] A. Kumar, H. Hemani, N. Sakthivel, and S. Chaturvedi, "Effective preprocessing of speech and acoustic features extraction for spoken language identification," in *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials*, pp. 81-88, May 2015.

[109] X. Li, "SPEech Feature Toolbox (SPEFT) Design and Emotional Speech Feature Extraction," Marquette University, 2007.

[110] V. Mitra, A. Tsiartas, and E. Shriberg, "Noise and reverberation effects on depression detection from speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5795-5799, March 2016.

[111] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[112] A. Biswas, P. K. Sahu, A. Bhowmick, and M. Chandra, "Acoustic feature extraction using ERB like wavelet sub-band perceptual Wiener filtering for noisy speech recognition," in *2014 Annual IEEE India Conference*, pp. 1-6, Dec. 2014.

[113] D. Huijun, I. Y. Soon, S. N. Koh, and C. K. Yeo, "A post-processing technique for regeneration of over-attenuated speech," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3889-3892, April 2009.

[114] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach* (Wiener Filters Chapter 5). John Wiley & Sons, p. 444, 2005.

[115] U. Sharma, S. Maheshkar, and A. N. Mishra, "Study of robust feature extraction techniques for speech recognition system," in *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management*, pp. 654-658, Feb. 2015.

[116] E. Loweimi, S. M. Ahadi, T. Drugman, and S. Loveymi, "On the Importance of Pre-emphasis and Window Shape in Phase-Based Speech Recognition," in *Advances in Nonlinear Speech Processing*: Springer, pp. 160-167, 2013.
[117] A. V. Oppenheim, J. R. Buck, and R. W. Schafer, *Discrete-time signal processing. Vol. 2*. Upper Saddle River, NJ: Prentice Hall, 2001.
[118] N. Alcaraz Meseguer, "Speech analysis for automatic speech recognition," *Institutt for elektronikk og telekommunikasjon*, 2009.
[119] S. Gautam and L. Singh, "Developmental pattern analysis and age prediction by extracting speech features and applying various classification techniques," in *2015 International Conference on Computing, Communication & Automation*, pp. 83-87, 2015.
[120] M. S. Yakcoub, S. A. Selouani, and D. OShaughnessy, "Speech assistive technology to improve the interaction of dysarthric speakers with machines," in *3rd International Symposium on Communications, Control and Signal Processing*, pp. 1150-1154, March 2008.
[121] M. A. Wahed, "Computer aided recognition of pathological voice," in *2014 31st National Radio Science Conference*, pp. 349-354, April 2014.
[122] T. Dekens, H. Martens, G. V. Nuffelen, M. D. Bodt, and W. Verhelst, "Speech rate determination by vowel detection on the modulated energy envelope," in *2014 Proceedings of the 22nd European Signal Processing Conference*, pp. 1252-1256, Sept. 2014.
[123] D. Enqing, L. Guizhong, Z. Yatong, and C. Yu, "Voice activity detection based on short-time energy and noise spectrum adaptation," in *6th International Conference on Signal Processing*, pp. 464-467, 2002.
[124] Y. Y. Aye, "Speech Recognition Using Zero-Crossing Features," in *2009 International Conference on Electronic Computer Technology*, pp. 689-692, Feb. 2009.
[125] M. Greenwood and A. Kinghorn, "SUVing: automatic silence/unvoiced/voiced classification of speech," *Undergraduate Coursework, Department of Computer Science, The University of Sheffield, UK*, 1999.
[126] N. Erdol, C. Castelluccia, and A. Zilouchian, "Recovery of missing speech packets using the short-time energy and zero-crossing measurements," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 3, pp. 295-303, 1993.
[127] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146-157, 2002.
[128] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830-1847, 2004.
[129] R. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in *Advanced Techniques in Computing Sciences and Software Engineering*: Springer, pp. 279-282, 2010.
[130] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Pearson, 2011.
[131] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in

American Society for Engineering Education (ASEE) Zone Conference Proceedings, pp. 1-7, 2008.

[132] A. Manjare Chandraprabha and S. D. Shirbahadurkar, "Pitch and duration modification for expressive speech synthesis in Marathi TTS system," in *2015 International Conference on Pervasive Computing*, pp. 1-4, Jan. 2015.

[133] A. Lowit, A. Kuschmann, J. M. MacLeod, F. Schaeffler, and I. Mennen, "Sentence stress in ataxic dysarthria: a perceptual and acoustic study," *Journal of Medical Speech Language Pathology,* vol. 18, no. 4, pp. 77-82, 2010.

[134] R. Tripathy and H. K. Tripathy, "Unalike methodologies of feature extraction & feature matching in Speech Recognition," in *2014 International Conference on High Performance Computing and Applications*, pp. 1-6, Dec. 2014.

[135] A. V. Oppenheim and A. S. Willsky, *Signals and Systems: Pearson New International Edition*. Pearson Education Limited, 2013.

[136] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems*. Prentice Hall, 1997.

[137] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on neural networks,* vol. 15, no. 5, pp. 1135-1150, 2004.

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

[139] D. Gerhard, *Pitch extraction and fundamental frequency: History and current techniques*. Department of Computer Science, University of Regina Regina, 2003.

[140] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 25, no. 1, pp. 24-33, 1977.

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

[142] M. Sondhi, "New methods of pitch extraction," *IEEE Transactions on audio and electroacoustics,* vol. 16, no. 2, pp. 262-266, 1968.

[143] J. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics,* vol. 20, no. 5, pp. 367-377, 1972.

[144] N. Kamaruddin, A. W. A. Rahman, and N. S. Abdullah, "Speech emotion identification analysis based on different spectral feature extraction methods," in *The 5th International Conference on Information and Communication Technology for The Muslim World*, pp. 1-5, Nov. 2014.

[145] U. N. Wisesty, Adiwijaya, and W. Astuti, "Feature extraction analysis on Indonesian speech recognition system," in *3rd International Conference on Information and Communication Technology*, pp. 54-58, May 2015.

[146] V. S. Selvam, V. Thulasibai, and R. Rohini, "Speech training system based on resonant frequencies of vocal tract," in *13th International Conference on Advanced Communication Technology*, pp. 674-679, Feb. 2011.

[147] J. M. Elvira, F. J. Dickin, and R. A. Carrasco, "A comparison of speech feature extraction employing autonomous neural network topologies," in *IEEE Colloquium on Systems and Applications of Man-Machine Interaction Using Speech*, pp. 9/1-9/5, Mar. 1991.

[148] S. B. Magre and R. R. Deshmukh, "A Review on Feature Extraction and Noise Reduction Technique," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 4, no. 2, pp. 352 -356, 2014.

[149] M. Chetouani, A. Hussain, B. Gas, and J. L. Zarader, "Non-Linear Predictors based on the Functionally Expanded Neural Networks for Speech Feature Extraction," in *2006 IEEE International Conference on Engineering of Intelligent Systems*, pp. 1-5, 2006.

[150] L. R. Rabiner and B. Gold, "Theory and application of digital signal processing," *Englewood Cliffs, NJ, Prentice-Hall, Inc., pp. 777,* 1975.

[151] J. D. Markel and A. J. Gray, *Linear prediction of speech*. Springer Science & Business Media, 2013.

[152] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 28, no. 4, pp. 357-366, 1980.

[153] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task.", *Proceedings of the SPECOM, pp.191-194,* 2005.

[154] N. Moritz, J. Anem, x00Fc, ller, and B. Kollmeier, "An Auditory Inspired Amplitude Modulation Filter Bank for Robust Feature Extraction in Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 11, pp. 1926-1937, 2015.

[155] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology,* journal article vol. 16, no. 6, pp. 582-589, 2001.

[156] M. R. Devi and T. Ravichandran, "A novel approach for speech feature extraction by Cubic-Log compression in MFCC," in *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, pp. 182-186, Feb. 2013.

[157] H. Trang, L. Tran Hoang, and N. Huynh Bui Hoang, "Proposed combination of PCA and MFCC feature extraction in speech recognition system," in *2014 International Conference on Advanced Technologies for Communications*, pp. 697-702, Oct. 2014.

[158] S. Gaikwad, B. Gawali, P. Yannawar, and S. Mehrotra, "Feature extraction using fusion MFCC for continuous marathi speech recognition," in *2011 Annual IEEE India Conference*, pp. 1-5, Dec. 2011.

[159] M. K. Pichora-Fuller, B. A. Schneider, E. MacDonald, H. E. Pass, and S. Brown, "Temporal jitter disrupts speech intelligibility: A simulation of auditory aging," *Hearing research,* vol. 223, no. 1-2, pp. 114-121, 2007.

[160] X. Li *et al.*, "Stress and emotion classification using jitter and shimmer features," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. IV-1081-IV-1084, 2007.

[161] 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.

[162] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP Journal on Advances in Signal Processing*, p. 9, 2009.

[163] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4749-4753, 2015.
[164] C. T. Ferrand, "Harmonics-to-noise ratio: an index of vocal aging," *Journal of voice*, vol. 16, no. 4, pp. 480-487, 2002.
[165] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE signal processing magazine*, vol. 8, no. 4, pp. 14-38, 1991.
[166] H. G. Stark, *Wavelets and Signal Processing: An Application-Based Introduction*. Springer Berlin Heidelberg, 2005.
[167] C. K. Chui, *An Introduction to Wavelets*. Elsevier Science, 2016.
[168] L. Chun-Lin, "A tutorial of the wavelet transform," *NTUEE, Taiwan*, 2010.
[169] P. J. Van Fleet, *Discrete Wavelet Transformations: An Elementary Approach with Applications*. Wiley, 2011.
[170] S. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier Science, 1999.
[171] P. Ong, W. K. Lee, and R. J. H. Lau, "Tool condition monitoring in CNC end milling using wavelet neural network based on machine vision," *The International Journal of Advanced Manufacturing Technology*, vol. 104, no. 1-4, pp. 1369-1379, 2019.
[172] R. Graf, S. Zhu, and B. Sivakumar, "Forecasting river water temperature time series using a wavelet–neural network hybrid modelling approach," *Journal of Hydrology*, vol. 578, p. 124115, 2019.
[173] M. Mandler and M. Scharnagl, "Financial cycles across G7 economies: A view from wavelet analysis," 2019.
[174] J. Nobre and R. F. Neves, "Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets," *Expert Systems with Applications*, vol. 125, pp. 181-194, 2019.
[175] P. S. Addison, *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC press, 2017.
[176] R. Senthilkumar and R. Gnanamurthy, "A robust wavelet based decomposition of facial images to improve recognition accuracy in standard appearance based statistical face recognition methods," *Cluster Computing*, vol. 22, no. 5, pp. 12785-12794, 2019.
[177] T. Ramalingam and P. Dhanalakshmi, "Speech/music classification using wavelet based feature extraction techniques," *Journal of Computer Science*, vol. 10, no. 1, p. 34, 2014.
[178] Y. Huang, A. Wu, G. Zhang, and Y. Li, "Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition," *IET Signal Processing*, vol. 9, no. 4, pp. 341-348, 2015.
[179] E. S. Sazonov*, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. L. Melanson, and M. R. Neuman, "Automatic Detection of Swallowing Events by Acoustical Means for Applications of Monitoring of Ingestive Behavior," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 626-633, 2010.
[180] C. M. Vikram and K. Umarani, "A Wavelet Based MFCC Approach for the Phoneme Independent Pathological Voice Detection," in *2013 Third International Conference on Advances in Computing and Communications*, pp. 153-156, Aug. 2013.
[181] S. Akbarzadeh, M. Heydarzadeh, F. Chen, S. Lee, and C.-T. Tan, "Reducing the variability in auditory evoked response to pitch matched

stimuli using Wavelet Scattering Transform," in *2018 IEEE 23rd International Conference on Digital Signal Processing*, pp. 1-4, 2018.

[182] H. Deng and D. O. Shaughnessy, "Voiced-Unvoiced-Silence Speech Sound Classification Based on Unsupervised Learning," in *2007 IEEE International Conference on Multimedia and Expo*, pp. 176-179, July 2007.

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

[184] B. Uslu and H. Tora, "The use of cumulants for voiced-unvoiced segments identification in speech signals," in *2014 22nd Signal Processing and Communications Applications Conference*, pp. 971-974, April 2014.

[185] S. Mondal and A. D. Barman, "Clustering based voiced-unvoiced-silence detection in speech using temporal and spectral parameters," in *2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks*, pp. 390-394, Nov. 2015.

[186] D. Arifianto, "Dual Parameters for Voiced-Unvoiced Speech Signal Determination," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. IV-749-IV-752, April 2007.

[187] A. DeMino and G. Dynamics, "Assessing Dysarthria severity using global statistics and boosting," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers*, pp. 1103-1106, Nov. 2011.

[188] H. Hermansky, "The purpose, history, current state, and some evolving trends in feature extraction for speech recognition," in *Proceedings of the Fifth International Symposium on Signal Processing and Its Applications*, p. 6 1999.

[189] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4605-4608, April 2009.

[190] D. Korzekwa, R. Barra-Chicote, B. Kostek, T. Drugman, and M. Lajszczak, "Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech," in *2019 Proceedings of INTERSPEECH*, 2019.

[191] D. Mulfari, G. Meoni, and L. Fanucci, "Machine Learning in Assistive Technology: a Solution for People with Dysarthria," in *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, pp. 308-309, 2018.

[192] V. Despotovic, O. Walter, and R. Haeb-Umbach, "Machine learning techniques for semantic analysis of dysarthric speech: An experimental study," *Speech Communication,* vol. 99, pp. 242-251, 2018.

[193] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in neural information processing systems*, pp. 231-238, 1995.

[194] U. N. Wisesty and A. T. H. Liong, "Indonesian speech recognition system using Discriminant Feature Extraction — Neural Predictive Coding (DFE-NPC) and Probabilistic Neural Network," in *2012 IEEE International Conference on Computational Intelligence and Cybernetics*, pp. 158-162, July 2012.

[195] K. Saeed and M. K. Nammous, "A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-

Signal Image," *IEEE Transactions on Industrial Electronics*, vol. 54, no. 2, pp. 887-897, 2007.

[196] G. Hinton *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 82-97, 2012.

[197] T. Ghiselli-Crippa and A. El-Jaroudi, "Voiced-unvoiced-silence classification of speech using neural nets," in *International Joint Conference on Neural Networks*, vol. ii, pp. 851-856, Jul 1991.

[198] T. Hori *et al.*, "The MERL/SRI system for the 3RD CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 475-481, Dec. 2015.

[199] E. Chandra and C. Sunitha, "A review on Speech and Speaker Authentication System using Voice Signal feature selection and extraction," in *IEEE International Advance Computing Conference*, pp. 1341-1346 March 2009.

[200] L.-c. Jiao and Springerlink (Online service), *Advances in natural computation : second international conference, ICNC 2006, Xi'an, China, September 24-28, 2006 : proceedings* (Lecture notes in computer science,, no. 4221-4222). Berlin ; New York: Springer, 2006.

[201] T. Fukuda, W. Yamamoto, and T. Nitta, "Distinctive phonetic feature extraction for robust speech recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.*, pp. II-25-8, April 2003.

[202] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144-152, 1992.

[203] L. Wang, *Support vector machines: theory and applications*. Springer Science & Business Media, 2005.

[204] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," in *Data mining techniques for the life sciences*: Springer, pp. 223-239, 2010.

[205] M. Awad and R. Khanna, *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Apress, 2015.

[206] Q. Yu, Y. Li, and P. Jia, "Speech emotion recognition using supervised manifold learning based on all-class and pairwise-class feature extraction," in *2013 IEEE Conference Anthology*, pp. 1-5, Jan. 2013.

[207] P. Cunningham and S. J. Delany, "k-Nearest neighbour classifiers," *Multiple Classifier Systems*, vol. 34, pp. 1-17, 2007.

[208] S. A. Rieger, R. Muraleedharan, and R. P. Ramachandran, "Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers," in *9th International Symposium on Chinese Spoken Language Processing*, pp. 589-593, Sept. 2014.

[209] T. L. Pao, W. Y. Liao, T. N. Wu, and C. Y. Lin, "Automatic visual feature extraction for Mandarin audio-visual speech recognition," in *IEEE International Conference on Systems, Man and Cybernetics*, pp. 2936-2940, Oct. 2009.

[210] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob, "Automatic detection of prolongations and repetitions using LPCC," in *2009 International Conference for Technical Postgraduates*, pp. 1-4, Dec. 2009.

[211] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 216-221, 2013.

[212] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[213] J. Schmidhuber, "Deep learning in neural networks: An overview," *Journal of the International Neural Networks Society*, vol. 61, pp. 85-117, 2015.

[214] L. Peng, M. Peng, B. Liao, G. Huang, W. Li, and D. Xie, "The advances and challenges of deep learning application in biological big data processing," *Current Bioinformatics*, vol. 13, no. 4, pp. 352-359, 2018.

[215] J. Berry, C. North, and M. T. Johnson, "Sensorimotor adaptation of speech using real-time articulatory resynthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3196-3200, May 2014.

[216] E. Półrolniczak and M. Kramarczyk, "Analysis of the dependencies between parameters of the voice at the context of the succession of sung vowels," in *2016 Signal Processing: Algorithms, Architectures, Arrangements, and Applications*, pp. 72-77, Sept. 2016.

[217] Y.-T. Wang, R. D. Kent, J. R. Duffy, and J. E. Thomas, "Analysis of diadochokinesis in ataxic dysarthria using the motor speech profile program™," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 1-11, 2009.

[218] M. Pawar and R. Kokate, "A Robust Wavelet Based Decomposition and Multilayer Neural Network for Speaker Identification," in *Innovations in Electronics and Communication Engineering*: Springer, pp. 197-209, 2019.

[219] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 121-133, 1979.

[220] J. Bonada, "Automatic technique in frequency domain for near-lossless time-scale modification of audio," in 2000 *International Computer Music Conference*, 2000.

[221] J. L. Flanagan and R. M. Golden, "Phase vocoder," *The Bell System Technical Journal*, vol. 45, no. 9, pp. 1493-1509, 1966.

[222] A. T. Cemgil and S. J. Godsill, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals," in *2005 13th European Signal Processing Conference*, pp. 1-4, Sept. 2005.

[223] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323-332, 1999.

Appendices

Appendix A Comparison of State-of-the-art Dysarthria Assessment Techniques

Techniques	Year	Treatment aim	Type of study	Key findings	Limitations
Lee Silverman Voice Treatment (LSVT)	2001	Increase Loudness	Traumatic Brain Injury, ataxic dysarthria	Loudness increased, improved intelligibility	Prosody, resonance, respiration and phonation not considered.
Altered Auditory Feedback (AAF)	2010	Improve Speech rate	Parkinson's disease	Improved speech rate	Other speech features not considered
Computerised Frenchay's Dysarthria Assessment (CFDA)	2015	Respiration and Phonation assessment	Spastic Dysarthria	Gives similar results with the traditional method	Prosody and resonance not accounted for.
Computerized Assessment and Treatment of Rate, Intonation and Stress (CATRIS)	2010	Automated speech assessment & therapy	Parkinson's Disease	Improved speech rate and intonation	Testing limited to Parkinson's disease. Lack of rhythm, respiration & phonation features.