AN INTRA-SEVERITY CLASSIFICATION AND ADAPTATION TECHNIQUE TO IMPROVE DYSARTHRIC SPEECH RECOGNITION ACCURACY

BASSAM ALI QASEM AL-QATAB

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

AN INTRA-SEVERITY CLASSIFICATION AND ADAPTATION TECHNIQUE TO IMPROVE DYSARTHRIC SPEECH RECOGNITION ACCURACY

BASSAM ALI QASEM AL-QATAB

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **BASSAM ALI QASEM AL-QATAB**

Matric No: **WHA110045**

Name of Degree: **DOCTOR OF PHILOSOPHY**

AN INTRA-SEVERITY CLASSIFICATION AND ADAPTATION TECHNIQUE TO IMPROVE DYSARTHRIC SPEECH RECOGNITION ACCURACY

Field of Study: **Computer Human Interaction and Signal Processing**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

AN INTRA-SEVERITY CLASSIFICATION AND ADAPTATION TECHNIQUE TO IMPROVE DYSARTHRIC SPEECH RECOGNITION ACCURACY

ABSTRACT

Dysarthria is a motor speech impairment at the neurological and/or muscular levels that caused difficulty in pronouncing words clearly. Automatic speech recognition (ASR) system is increasingly applied as assistive technology to aid an individual with physical disability particularly the speech impaired community such as dysarthria speakers. However, the development of an effective ASR system is hindered by the data sparsity, either in the coverage of the language or the size of the existing speech databases. The speaker adaptation (SA) technique is one of the solutions to overcome the data sparsity issue of ASR for dysarthric speakers. Our proposed method introduces the intra-severity classification and adaptation techniques which are applied sequentially in two stages of system development. Firstly, intra-severity classification intended to identify the level of severity of the dysarthric speakers. Secondly, the identified severity level of a particular dysarthric speaker in the first stage is applied to the corresponding intra-severity adaptation of dysarthric speech. For the classification part, there are six algorithms used to classify the intra-severity of dysarthric speakers. The algorithms include Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Support Vector Machine (SVM), Naive Bayes (NB), Classification And Regression Tree (CART), Random Forest (RF). The Random Forest (RF) algorithm was proposed as a classifier for the intra-severity classification of the dysarthric speaker which has the lowest average ranking score as compared to other benchmark classifiers. The intra-severity adaptation of the ASR system was developed using two well-known adaptation techniques which are the Maximum Likelihood Linear Regression (MLLR) and Maximum A Posterior

(MAP) as well as a combination of them. The results showed that the combination of MLLR+MAP adaptation outperforms all adaptation techniques with total improvement in Word Error Rate (WER) from 39.84% to 18.48% with 53.61% improvement from the baseline WER in the overall performance of the system. The total improvement of the WER based on severity level were 66.32%, 52.35%, and 45.20% for mild, moderate, and severe severity level respectively for the hybrid MLLR+MAP adaptation technique. The combination of the adaptation techniques in sequential order helps to take advantage of each adaptation technique and avoid the flaws of each technique in relation to adaptation data size.

Keywords: Dysarthria, Automatic Dysarhric Speech Recognition System, Classification Algorithms, Adaptation Techniques, Feature Selection Methods.

PENGURUSAN DAN TEKNIK PENDIDIKAN INTRA-KEBERHASILAN UNTUK MENINGKATKAN KETERANGAN PENGIKTIRAFAN DYSARTHRIC ACKURASI

ABSTRAK

Dysarthria adalah kekurangupayaan pertuturan motor pada peringkat neurologi dan / atau otot yang menyebabkan kesukaran pertuturan dengan jelas. Sistem pengecaman ucapan automatik (ASR) semakin meningkat penggunaanya sebagai teknologi bantuan untuk membantu individu yang mengalami kecacatan fizikal terutamanya masyarakat kurang upaya pertuturan seperti penutur dysarthria. Walau bagaimanapun, perkembangan sistem ASR yang berkesan dihalang oleh sparsiti data, sama ada dalam liputan bahasa atau saiz pangkalan data ucapan yang sedia ada. Teknik penyesuaian penutur (SA) adalah salah satu daripada penyelesaian untuk mengatasi isu sparsiti data ASR untuk penutur dysarthric. Kaedah yang kami cadangkan memperkenalkan intra-keterukan yang digunakan secara berurutan dalam dua peringkat pembangunan sistem. Pertama, klasifikasi intra-keterukan digunakan untuk mengenal pasti tahap keterukan untuk penutur dysarthric. Kedua, tahap keterukan yang dikenal pasti bagi penutur dysarthric tertentu pada peringkat pertama digunakan pada penyesuaian intra-keterasan yang sepadan dengan penutur dysartric sistem ASR yang dibangunkan sebelum peringkat pembangunan. Dalam bahagian klasifikasi, terdapat enam algoritma yang digunakan untuk mengklasifikasikan intra-keterukan penutur dysarthric. Algoritma tersebut termasuk Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Support Vector Machine (SVM), Naive Bayes (NB), Classification And Regression Tree (CART), Random Forest (RF). Algoritma Random Forest dicadangkan sebagai classifier untuk klasifikasi intra severity penturur dysathria kerana ia menunjukkan skor klasifikasi purata yang terendah bila dibandingkan dengan classifier yang lain. Penyesuaian intra-

keterukan sistem ASR telah dibangunkan dengan menggunakan dua teknik penyesuaian yang terkenal iaitu Regresi Linear Maksimum (MLLR) dan Maksimum A Posterior (MAP) serta gabungannya. Keputusan menunjukkan bahawa gabungan penyesuaian MLLR + MAP mengatasi semua teknik penyesuaian dengan peningkatan jumlah kadar ralat perkataan (WER) dari 39.84% kepada 18.48% dengan peningkatan 53.61% daripada WER asas dalam prestasi keseluruhan sistem. Peningkatan jumlah WER berdasarkan tahap keterukan adalah 66.32%, 52.35%, dan 45.20%, untuk tahap keterukan ringan, sederhana dan teruk untuk teknik penyesuaian MLLR + MAP hibrid. Kombinasi teknik penyesuaian dalam susunan berurutan membantu untuk memanfaatkan kelebihan setiap teknik penyesuaian dan mengelakkan kelemahan setiap teknik berkaitan dengan saiz data penyesuaian

Kata kunci: Dysarthria, Sistem Pengiktirafan Ucapan Dysarhric Automatik, Pengkelasan Algoritma, Teknik Penyesuaian, Kaedah Pemilihan Ciri.

ACKNOWLEDGEMENTS

First, I am thankful to ALLAH SWT (the Most Merciful and Most Beneficent) for his guidance, inspiration, and help to successfully complete this thesis.

Following, I would like to express my deepest appreciation to my supervisors **Prof. Dr. Siti Salwah Salim** and **Dr. Mumtaz Begum Mustafa**. They have been there providing their heartfelt and guidance at all times and have given me invaluable guidance, inspiration, and suggestion in my quest for knowledge. They have given me all the freedom to pursue my research, while silently and non-obtrusively ensuring that I stay on course and do not deviate from the core of my research. Without guidance and persistent help of theirs, this thesis would not have been possible.

Most important, I would like to express my gratitude to my family and parents for being a source of support and encouragement. My mother who always supplicate ALLAH for guidance. My wife, who has always been with me, she has always pushed me when I was down. My lovely children Aseel, Shahd, Ali, Amgad, Younes, and Yousef for their innocent smile, may ALLAH save and bless them.

My deepest appreciation with grateful thanks goes to my brothers Fahmi, Hakim, Khaled, and Zakaria and sisters Ebtisam, and Rahma. I am really proud of them. They persistently support me, incorporate and encourage on my path towards achieving my Ph.D.

Special thanks to my fellow friends in the Human-Computer Interaction (HCI) Lab and Multimodal Interaction Lab, University of Malaya as well as those who involved directly or indirectly upon completion of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

ADSR	:	Automatic Dysarthric Speech Recognition
AHC	:	Agglomerative Hierarchical Clustering
ALS	:	Amyotrophic Lateral Sclerosis
ANN	:	Artificial Neural Networks
AR	:	Articulation Rate
ASR	:	Automatic Speech Recognition
BAM	:	Baseline Speech Acoustic Model
CAIDS	:	Computerized Assessment of Intelligibility of Dysarthric Speech
CART	:	Classification and Regression Tree
CAS	:	Childhood Apraxia of Speech
CIFE	:	Conditional Information Feature Extraction
CMIM	:	Conditional Mutual Information Maximization
CMLLR	:	Constrained MLLR
ComParE	:	Computational Paralinguistic Challenge
Condred	:	Conditional redundancy
CP	:	Cerebral Palsy
DAB	:	Darley, Aronson and Brown
DIA	:	Dutch Intelligibility Assessment
DISR	:	Double Input Symmetrical Relevance
DMOS	:	Degradation Mean Opinion Score
DNN	:	Deep Neural Network
DS	:	Design Science
DSRM	:	Design Science Research Methodology
EM	:	Expectation-Maximization

F0	:	Fundamental Frequency
F1		First Formant Frequency
F2	:	Second Formant Frequency
fMLLR	:	feature space MLLR
GLR	:	Generalization Likelihood Ratio
GMM	:	Gaussian Mixture Model
GNE	:	Glottal-to-Noise Excitation ratio
GUI	:	Graphical User Interface
HMM	:	Hidden Markov Model
HNR	:	Harmonics-to-Noise Ratio
HTK	:	Hidden Markov Model Toolkit
ICAP	:	Interaction Capping
IDEA	:	Individual with Disability Education Act
IS	:	Information System
JMI	:	Joint Mutual Information
KNN	:	K-Nearest Neighbor
LDA	:	Linear Discriminant Analysis
LHMR	:	Low-to-High Modulation Energy Ratio
LLD	:	Low Level Descriptors
LogHNR	:	Logarithmic Harmonics-to-Noise Ratio
LP	:	Linear Prediction
LSP	:	Line Spectrum Pair
LVCSR	:	Large Vocabulary Continuous Speech Recognition
MAP	:	Maximum A Posteriori
MDA	:	Mahalanobis Discriminant Analysis

MFCC	:	Mel Frequency Cepstral Coefficients
ML	:	Maximum Likelihood
MLLR	:	Maximum Likelihood Leaner Regression
NAT	:	Number of Adaptation Technique
NB	:	Naive Bayes
NICO	:	Neural Interface Computation
NKI CCRT	:	Advanced Head and Neck cancer - Concomitant, Chemo-Radiation Treatment
NOE	:	Number of experiments performed
NOF	:	Number of Features
NS	:	Number of Datasets
NST	:	Number of Severity Level
PCC	:	Percentage of Correct Consonants
PD	:	Parkinson disease
PJ	:	Perceptual Judgments
PLF	:	Phonological Features
PMF	:	Phonemic Features
PPC	:	Perceptual Percentage of Correct
Prob. Of Voicing	:	Probability of Voicing
PVI	:	Pairwise Variability analysis
PWD	:	Persons With Disabilities
QDA	:	Quadratic Discriminant Analysis
RASTA	:	Relative Spectral Transform
RF	:	Random Forest

RMS	:	Root-Mean-Square
RMS	:	Root Mean Square Energy
ROI	:	Rank-Order Inconsistency
SA	:	Speaker Adaptation
SAT	:	Speaker Adaptive Training
SD	:	Speaker Dependent
SI	:	Speaker Independent
SOM	:	Self Organizing Map
STARDUST	:	Speech Training And Recognition for Dysarthric Users of ASsistive Technology
SVM	:	Support Vector Machine
TD-PSOLA	:	Time-Domain Pitch Synchronous Overlap Add
TNCF	:	Total Number of Correctly-testing Features
TNF	:	Total Number of Features used
UA-Speech	:	Universal Access Speech
VI	:	Variability Index
VID	:	Voiceless Interval Duration
VOT	:	Voice Onset Time
VS	:	Vowel Space
WER	:	Word Error Rate
ZCR	:	Zero-Crossing Rate

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

Speech is an amazing achievement of the human motor system whereby sound is produced at rates of up to 30 segments per second in an action of precise coordination. This requires the work of more muscle fibers than any other mechanical performance by humans (Ray Kent, Kent, Weismer, & Duffy, 2000).

1.1 Definition and Understanding of Speech Impairment

According to the Individual with Disability Education Act (IDEA), impairment of speech and language is defined as a means of communication disorder that negatively affects the educational performance of a child. IDEA classified speech and language impairment into four core areas, which are articulation, fluency, voice and language impairments (Center for Parent Information And Resources, 2011).

A communication disorder is defined as the inability to receive, comprehend, process, and send concepts (symbols system which is verbal, nonverbal and graphical). According to the American Speech-Language-Hearing Association, speech and language impairment is a form of communication disorder that affects the individual (American Speech-Language-Hearing Association, 1993).

Speech impairment may be acquired or developed. Individuals may suffer from one or more of speech impairments. It ranges from mild to profoundly severe with regards to its severity (American Speech-Language-Hearing Association, 1993; Center for Parent Information And Resources, 2011; Prelock, Hutchins, & Glascoe, 2008).

According to Colorado Department of Education (2017), "a child with impairment of speech or language may have a communication disorder, which prevents the child from receiving reasonable educational benefits from regular education." The speech or language impairment criteria consist of a barrier to proper communication, in writing and/or orally in one's primary language, in social and academic interactions. There can also be demonstrations of inappropriate or undesirable behavior due to limited communication skills, as the person may be unable to communicate without using assistive, augmentative/alternative communication devices or systems (Colorado 2017).

1.2 Terms and Categories of Speech Impairment

There are three main types of speech impairment with regards to articulation, fluency and, voice. The articulation impairment is related to producing the speech sound in an incorrect manner. The substitution, addition, omission, or distortion are the characteristics of the articulation impairment that may lead to poor clarity of the sound (American Speech-Language-Hearing Association, 1993; Center for Parent Information And Resources, 2011). For example, difficulty in articulating certain sounds, such as "i" or "r".

Fluency refers to the disruption of the flow of speaking. The characteristics of the fluency impairment include typical rhythm, rate and repetition of the sound, phrases, words, and, syllables. The struggled behavior and inappropriate mannerism such as inhalation, exhalation, or phonation pattern may accompany this kind of impairment. The speech sound which is not suitable for an individual's age or gender is defined as a voice disorder. The characteristics of this kind of speech impairment are the poor generation and the lack of loudness, resonance, pitch, vocal quality, and duration (American Speech-

Language-Hearing Association, 1993; Center for Parent Information And Resources, 2011).

In relation to these three main types of speech impairment, there are many terms that are used in reference to speech impairment, such as childhood apraxia of speech (CAS), dysarthria, stuttering, voice, etc. This study focuses only on those suffering from speech impairment due to dysarthria. The term childhood apraxia of speech (CAS) included just to compare with dysarthria of speech, which is related to each other and classified according to the level of impairment:

1.2.1 Childhood Apraxia of Speech (CAS)

Apraxia means without action in which "a" means absent of, and "praxia" means the performance of the action, from the Greek word "praxis" (Freed, 2012). Childhood Apraxia of Speech (CAS) is a neurological childhood speech sound disorder whereby the control for the muscles of speech movements are affected due to unknown neuromuscular defects (for example, abnormal tone and reflexes) (American Speech-Language-Hearing Association, 2007; Strand & McCauley, 2008). CAS is motor speech impairment, due to the damage in the areas in the brain that is involved with speaking. A person suffering from speech apraxia is unable to coordinate the sounds in syllables and words. The extent of severity relies on the nature of brain damage and range, from mild to severe. Individuals with CAS know what words to say, but their brains have difficulty in controlling the muscle actions precisely to pronounce the words. The consensus of the investigation is that the apraxia of speech in children has the following segmental and suprasegmental features: (a) repeated productions of syllables or words caused by inconsistent errors on consonants and vowels (b) coarticulatory transitions between sounds and syllables that are lengthened and disrupted, and (c) inappropriate prosody,

especially in the realization of lexical or phrasal stress (American Speech-Language-Hearing Association, 2007).

1.2.2 Dysarthric Speech

The literal definition of dysarthria is "disorder utterance" in which "dys" means disorder or abnormal and "arthria" means to utter distinctly, from the Greek word, "arthroun" (Freed, 2012). Dysarthria is an impairment of the neuron-motor speech, whereby the muscles controlling the speech organs are weak, move slowly, or not move at all. The causes of dysarthria include muscle dystrophy, cerebral palsy, head injury, and stroke (Green et al., 2003). The severity and classification of dysarthria depend on the area of the nervous system affected as well as the site and degree of neurological damage. Clinically, dysarthria is assessed according to the articulation and speech intelligibility, in accordance with the measures of human perception (Kayasith, Theeramunkong, & Thubthong, 2006a). A common assessment tool is the Frenchay dysarthria assessment (Enderby, 1980b). It is a clinical-based tool that has the latest enhanced version which was released in 2008. Another tool used is called the Computerized Assessment of Intelligibility of Dysarthric Speech (CAIDS) (Yorkston, Beukelman, & Traynor, 1984). Symptoms of dysarthria include slurred speech, weak or imprecise articulator contact, low volume, and hypernasality.

One of the functions of the central nervous system is to control speech production. Any lesions in this system can cause different perturbations of speech and is based on the location and type of lesion. The term dysarthria is associated with this type of lesions. Therefore, speech analysis of the patients suffering from this pathology can reveal important information for assessment and treatment, thus increasing the effectiveness and reliability of the diagnosis process.

The main goals of the motor speech evaluation used in the assessment tool, as reported in (Roth, 2011) includes, to describe the speech characteristics, to differentially diagnose dysarthria types, to confirm the presence of neurologic disease and the level of nervous system involvement, to determine speech impairment severity, to determine the presence of other neurologic communication disorders, to determine recovery prognosis, and to define an intervention plan.

Childhood Apraxia of Speech (CAS) and dysarthria are very closely related to each other and are classified under articulation impairment. To differentiate between them, Table 1.1 shows the relationship between specific functions and possible speech impairment, and Figure 1.1 gives specific terms related to the problems of the oral motor (Strand & McCauley, 2008).

Table 1.1 and Figure 1.1 show that for dysarthric speakers, the function of the execution of movement is affected, which may include weakness, the law in motion, and, speed. In other words, these deficits can be improved by encouraging dysarthric speakers to keep practicing the pronunciation of the words without fear and hesitation.

Table 1.1 Relationship between specific function and possible speech disorders (Strand & McCauley, 2008)

Function	Neural Process	Possible speech impairment
Specifying range of motion, direction, speed and force of movement	Motor planning and programming	Childhood apraxia of speech(CAS)
Executing of movement resulting in acoustic output	Motor execution	Dysarthria

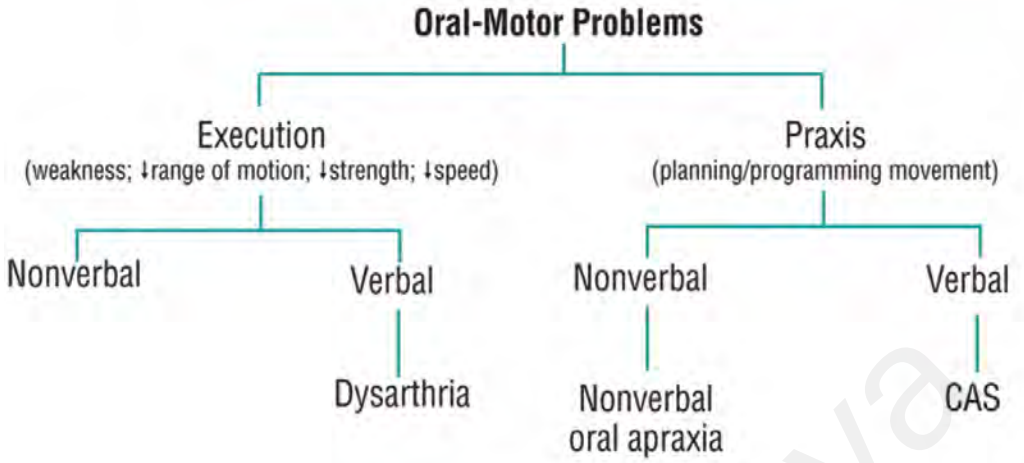

Figure 1.1: Specific terms related to oral-motor problems (Strand & McCauley, 2008).

1.3 Automatic Speech Recognition (ASR)

Human communication is considered as one of the most crucial and essential forms of the human social interdependence and exchange of information throughout their lives. The primacy of spoken communication in human psychology has been extended through the technological platform such as the internet, television, radio, movies, and telephony.

Besides human-human communication, human-machine interaction has gained an important preference among humans as it is based on interface objects and functions that are graphically-represented, called graphical user interface (GUI), which is utilized in most computers such as menus, icons, pointers, and windows. Most computer operating systems and applications also depend on users' mouse clicks and keyboard strokes, with a display monitor or to show feedbacks. However, today's computers are still in the primary stages in comparison to the fundamentals of the human's abilities: to listen, understand, speak, and learn.

The spoken language system includes both the components of speech recognition and speech synthesis. Furthermore, understanding the dialog component, and domain knowledge are required to manage the interaction with the user and accurately interpret the speech for the necessary action to be taken. Challenges, like the flexibility, robustness, ease of integration, and, engineering efficiency might be present in each component of the spoken language system (Huang, Acero, & Hon, 2001).

Automatic speech recognition (ASR) is the process and related technology for obtaining the word or words sequence from a given speech signal. Speech recognition applications that have appeared over the last decade include voice dialing, interactive voice response, call routing, data entry and dictation, voice search, control and command (voice user interface with the computer), hands-free computing (automotive applications), creation of structured documents (e.g., legal and medical transcriptions), control of appliance by voice, learning of languages with the aid of computers, robotics, and, search of content-based spoken audio (He & Deng, 2008; Huang et al., 2001).

The automatic speech recognition system (ASR) consists of basic components such as the acoustic model, language model, dictionary, signal processing, and decoder as shown in the blue dotted box in Figure 1.2. The acoustic model includes the representation knowledge about acoustics, phonetics, microphone, environmental variability, gender, and the speakers' dialect differences. The language model refers to the knowledge of the system on the likeness of the word sequences. In some ASR system, the semantics and functions desired to be performed may also include the language model. Dictionary refers to the phonetical representation of the words, whether in single or multiple representations. Signal processing refers to the process of speech signal through the processing module to extract the related features for the decoder. The adaptation refers to

the modification of the acoustic model or a language model to improve the performance in terms of recognition accuracy. The decoder uses the above components to produce the word sequence that almost matches the input feature vectors with the stored acoustic features (Huang et al., 2001).

The development process includes the understanding of human speech and language, recognizing the languages in a way that it can be used to automate it, using time alignment for word boundaries, developing the system to easily manipulate speech processing, developing a statistical approach to estimate the likelihood appearance of the words in sentences, and finally working with Large Vocabulary Continuous Speech Recognition (LVCSR). Therefore, the challenge is to develop a machine that can understand speech, make decisions for desirable actions to be taken, and respond like how humans do (Furui, 2005).

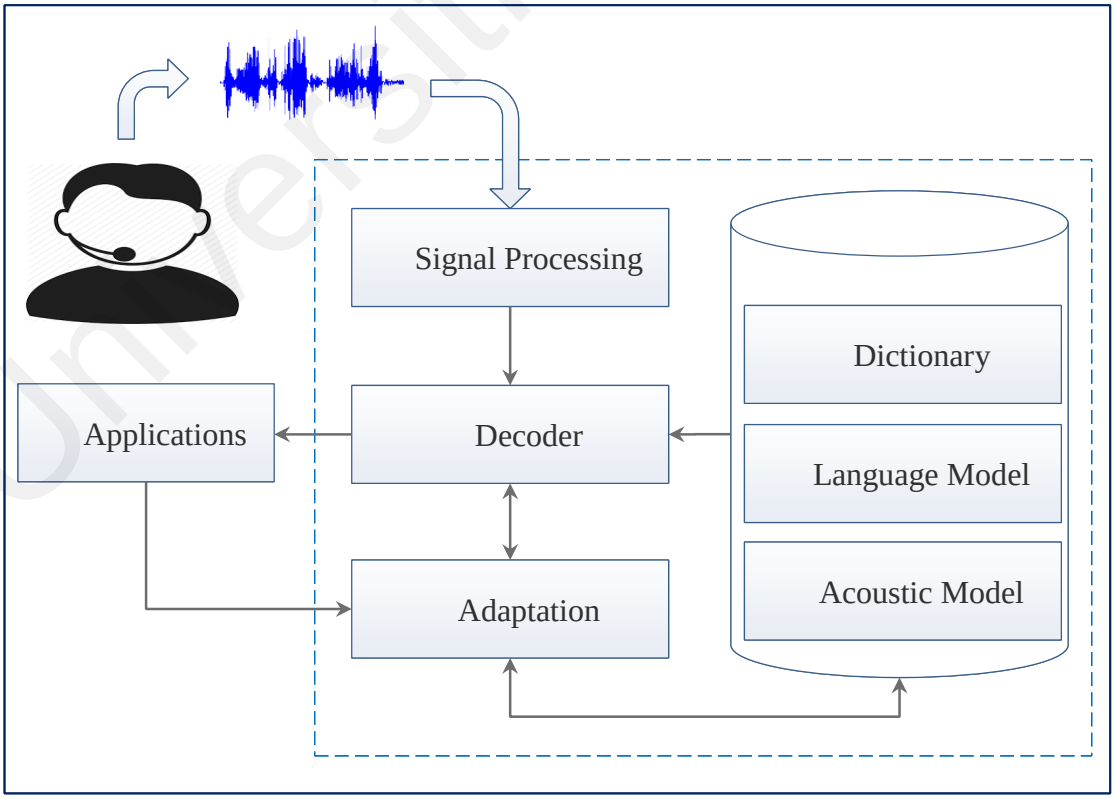

Figure 1.2: Basic system architecture of an automatic speech recognition (ASR) system

The speech recognition system is being improved progressively for the past few years. The speech recognition system yields high recognition result when the system is tested on the same speaker that was used during the training stage of the system development. In fact, the speaker-dependent speech recognition model performs better than the speaker-independent speech recognition models (Woodland, 2001). In (Gauvain & Lee, 1994; Gauvain & Lee, 1994; Kuhn, Junqua, Nguyen, & Niedzielski, 2000; Leggetter & Woodland, 1995; Woodland, 2001) the adaptation techniques were proposed to reduce the mismatches between the classified parameters of training conditions and testing conditions using some adaptation data.

There are many approaches used to perform speaker adaptation in the area of speech recognition technology. The speaker adaptation aims to obtain the final system that has a desirable speaker-dependent (SD)-like properties that require a small portion of the speaker-specific (parameters) training data to develop the full SD system. Generally, the speaker adaptation (SA) system is designed to improve the overall performance for all speaker interaction with the ASR systems.

1.4 Research Motivation

An automatic dysarthric speech recognition system has many potential applications. Remarkably, it is revealed from the previous studies that the automatic speech recognition (ASR) system has been used to help dysarthric speakers in various fields. The most applied domains are:

- *Early and self-diagnoses:*

  One of the most essential motivations for developing the ASR system for dysarthria is the convenience of using a home-based assessment program for easy automatic determination of the severity level of dysarthric speakers. Furthermore,

the home-based assessment can also be used for early diagnoses of dysarthria which leads to providing a suitable treatment at an early stage of the disability (Bowen et al., 2012).

- *Computer games:*

  Automatic speech recognition can be used to develop a game application for dysarthric speakers. The benefit of this game application is to give feedback to the parents, instructors, or pathological therapists of the degree, types, and severity of the dysarthric speakers. The game application can also be used as a tool to improve the speech difficulties for dysarthric speakers (Kitzing, Maier, & Åhlander, 2009; Parker, Cunningham, Enderby, Hawley, & Green, 2006).

- *Computer-Assisted interaction:*

  The dysarthric speakers particularly the adults have shown to be more interested in interacting with the ASR system (human-computer interaction) than using the traditional interaction (using keyboard). Some dysarthric speakers are unable to type with their hands or find it too tiring (Hamidi, Baljko, Livingston, & Spalteholz, 2010; Hux, Rankin-Erickson, Manasse, & Lauritzen, 2000; Thomas-Stonell, Kotler, Leeper, & Doyle, 1998) thus providing a motivation to develop the automatic dysarthric speech recognition system.

On the other hand, there is some practical motivation to do this research which includes:

- There is plenty of unimpaired speech corpus as compared to the very few corpora for impaired speech (which is attributed to the difficulties in collecting the data from impaired speakers). This motivates the researcher to use adaptation techniques to reduce the mismatches between the features of the impaired speaker

with the unimpaired trained acoustic model in automatic speech recognition techniques.

- The improvement to the recognition accuracy of the automatic speech recognition system over the last decades for unimpaired speakers (Xiong, Barker, & Christensen, 2018) motivates the researcher to investigate the acoustic features and adaptation techniques that helps to improve the recognition accuracy off the automatic dysarthric speech recognition system that can improves the life quality of dysarthric speaker.
- One of the aims of ASR research is to develop an application for dysarthric speakers which might be used in assistive technology or used in web applications with the low computational cost. This motivates the researcher to use severity level classification and adaptation techniques that help to limit the classification and adaptive acoustic model to the total number of the severity level of dysarthric speakers.

1.5 Problem Statement

The classification of dysarthria has gained importance among the researchers. First is to fully understand the types of impairment which result in empirical (sound) features that help the development of programs, to easily identify the disorder and its characteristics (Kim, Kent, & Weismer, 2011). Secondly, classifications are needed to compare the types of dysarthria with each other or with controlled speech, resulting in accurate identification of impairment (Kim et al., 2011; Liss et al., 2009). There is no standard measurement for speech severity in dysarthria, thought the speech intelligibility is often used to determine the level of the speech mechanism that is affected by the neurological disease (Kent et al., 1989). One of the main challenges in differentiating between the effects of severity and type of dysarthria is the lack of relevant analysis from a sufficiently large number of

speakers with different types of dysarthria and with various levels of severity (Kim et al., 2011).

When analyzing the symptoms of dysarthria, both at the articulatory and acoustic level, types of dysarthria is used to identify speakers with dysarthria (Weismer, Kim, Maassen, & van Lieshout, 2010). Furthermore, (Kim et al., 2011; Weismer et al., 2010) concludes that the large inter-speaker variability associated with dysarthria type is caused by the variation in the severity of the speech involved. In other words, the classification of speakers with dysarthria according to severity is highly similar to the type classification (Kim et al., 2011). However, the use of severity level classification is not thoroughly investigated in dysarthria speech recognition despite some research that focuses on Spastic Severity Disorder Classification (Paja & Falk, 2012). Furthermore, the common feature selection techniques are the forward selection procedure and the backward selection procedure, which is time-consuming and needs a pre-defined justification to obtain the desired feature selection in order to obtain the highest classification accuracy with less features (Kim, Kumar, Tsiartas, Li, & Narayanan, 2015; Middag, Martens, Van Nuffelen, & De Bodt, 2009).

Dysarthric speakers generally produce speech that is difficult to be understood by those unfamiliar with the speakers (Christensen, Casanueva, Cunningham, Green, & Hain, 2014). This physical disability results in the need for a system that can understand the spoken language, and become more appealing compared to the traditional method interfaces, such as using a keyboard and mouse. As such, the ASR system was developed as a means of communication aid for dysarthric speakers.

Due to the tiredness and frustration of the dysarthric speakers to speak for long periods of time, there is a lack of speech databases that are available, to provide sufficient speech

samples to train a speaker-dependent (SD) ASR system (Gale, Chen, Dolata, van Santen, & Asgari, 2019; Sharma & Hasegawa-Johnson, 2013; Xiong et al., 2018) as well as to use the standard speaker-independent ASR system (Despotovic, Walter, & Haeb-Umbach, 2018; Gale et al., 2019; Mengistu & Rudzicz, 2011), as such giving rise to the term data sparsity. This occurs when a number of data samples cannot produce enough parameters that can identify the presented data samples. Thus, the recognition accuracy will not be as high as expected or may even be worse than the original recognition accuracy of the system (Shinoda, 2011).

To overcome the problem of data sparseness, two approaches can be considered. The first is to recognize impaired (dysarthric) speech by using the unimpaired (normal) speech acoustic model (Stern & Lasry, 1987). However, the intelligibility of a dysarthric speaker's speech is very low. This results in typical measures of speech acoustics to have values with ranges that differ significantly from those for unimpaired speech (Liu, Tsao, & Kuhl, 2005). Thus the acoustic models trained in unimpaired speech will not be able to adjust to this mismatch (Morales & Cox, 2009).

The alternative solution is Speaker Adaptation (SA). It has the ability to learn the acoustic characteristics of individual speakers and adapt them to specific speakers. The SA ASR system helps to compensate the inconsistencies in the speech production and to reduce the irregularities between the features of the speaker with the trained acoustic model of the ASR system (Kotler & Thomas-Stonell, 1997; Rudzicz, Namasivayam, & Wolff, 2011; Sharma & Hasegawa-Johnson; Stern & Lasry, 1987).

1.6 Research Objectives

The objectives of this research are as follow:

1. To identify the suitable classification algorithms and acoustic features of dysarthria for automatic classification of dysarthric speech severity level.
2. To identify the suitable adaptation techniques in relation to data size and level of severity of the dysarthric speech towards improvement in recognition accuracy of dysarthric speech recognition.
3. To design and develop the intra-severity automatic dysarthric speech recognition method using the identified classification and adaptation techniques in objectives 1 and 2.
4. To evaluate the accuracy of the developed automatic dysarthric speech recognition method by comparing it with the baseline acoustic model.

1.7 Research Questions

The following research questions are suggested as a guide for conducting this research at the different phases to accomplish the research objectives. The questions are listed based on the objectives in the previous section:

Objective #1:

1. What is the importance of classifying the dysarthric speech severity level?
2. What are the acoustic features that affect the classification of the severity level of dysarthric speech?
3. What are the statistical functions that can be used to determine the dimensions of the vector for each acoustic feature?
4. What is the effect of the reduction of the statistical function per acoustic feature?

Objective #2:

1. What is the best adaptation technique that obtains the highest recognition accuracy?
2. What is the effect of increasing the data adaptation amount to improve ADSR's recognition accuracy?

Objective #3:

1. What are the best classifier and adaptation techniques that can be used to design and implement the proposed system to improve the recognition accuracy of the automatic dysarthric speech recognition system (ADSR)?

Objective #4:

1. What are the measurements used to evaluate the classification accuracy and the recognition accuracy of the severity level of the automatic dysarthric speech recognition system?
2. How are the results of the proposed system compared to other related methods in terms of classification accuracy, recognition accuracy and the combination of both?

1.8 Significance of the Research

This research is focusing on the classification and adaptation techniques to help dysarthric speakers overcome their disability and be more involved with the society. There are several significance of this research as the following:

- Introduce an intra-severity classification and adaptation technique where each severity consists of utterances from speakers belonging to the same severity

level. For example, the dataset for mild level includes the utterances from the mild speakers. Thus, in the classification phase, it will help to automatically identify the severity level of dysarthric speakers, which is required during the adaptation phase. In other words, the classification phase will classify the dysarthric speakers based on the total number of severity level which will be similar to the total number of the adaptation model built for the ADSR system.

- The lack of corpora for dysarthric speakers will be overcome as this research will use the adaptation techniques to reduce the mismatches between the unimpaired speakers (mostly available corpora) and dysarthric speakers based on severity level.
- Using the hybrid adaptation techniques based on the standard adaptation techniques, which are MLLR and MAP adaptation techniques helps to improve the recognition accuracy of the automatic dysarthric speech recognition system as the hybrid method is aimed at obtaining the benefits of both techniques and at the same time reduce the disadvantage of those techniques.

The researchers in the field of dysarthria might benefits from this research as this research is focusing on improving the recognition accuracy of the automatic dysarthric speech recognition system. The dysarthric speakers, as well as the therapist, is one of the beneficiaries of this research as they will get feedback when this research will be used as a base for a complete system.

1.9 Research Scope

This research focuses on the improvements of the ASR system's recognition accuracy of dysarthric speakers, which includes three main areas:

- *Dysarthria of speech:*

  This field of research includes the understanding of dysarthria, its causes, and the solutions proposed by the therapists. The intelligibility of the dysarthric speakers, types of dysarthria, disease types and the severity level of dysarthria has been explored to fully understand in this field of study. This research also focuses on the techniques used to help dysarthric speakers, using classifications of dysarthric speakers based on type, disease, and severity level.

- *Automatic Speech Recognition:*

  This field of research includes the understanding of speech production and mimics the acoustic system of the human being. It includes the databases (corpora), which features the extraction and coding of the language model used to build the acoustic model for the ASR system. The different types of acoustic models in ASR, such as speaker-independent, speaker-dependent and speaker adaptive has been investigated in this research.

- *Classification Algorithm:*

  The classification algorithm is investigated to gain extra knowledge on the basic methodology of its design which includes feature selection, feature extraction, and type of classification method, like statistical or machine learning.

The three areas of research have been explored in general. However, the scope of this research focuses on the intersection of the three areas as shown in Figure 1.3. The severity level (intra-severity) based classification and adaptation methods are used to enhance the accuracy of the automatic speech recognition for dysarthric speakers. To be more specific, the focus is on the severity level, the adaptation techniques, and the machine learning techniques for the classification of dysarthric speech.

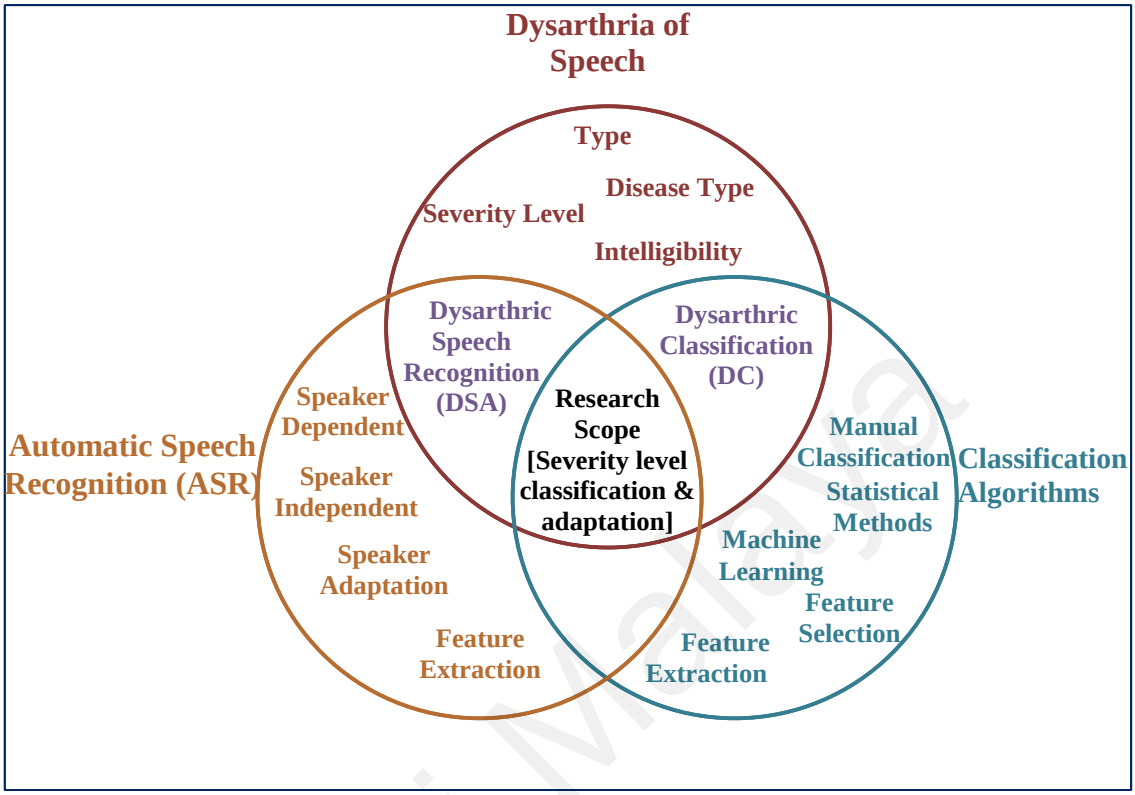

Figure 1.3: Research Scope

1.10 Thesis Organization

This thesis is organized as follows:

Chapter 1 provides a general introduction to the research with regards to the background and problem statement, introduction of the research objectives, and highlighting the scope and contributions of the research topic. The introduction of the speech impairment with regards to speech dysarthria and its deficiencies in the speech production system is described in more detail. A basic concept of automatic speech recognition, its types, its acoustic model types, and its performance are explored in this chapter.

Chapter 2 presents the literature reviews; it covers the dysarthric speech disorder and automatic speech recognition for dysarthric speech. The perceptual and objective assessment, classification, and type of dysarthria for dysarthric speakers are also described in this chapter. On top of that, this chapter provides an overview of the classification techniques used to classify the dysarthric speakers and to measure the degree of intelligibility and level of severity of dysarthric speakers. The automatic speech recognition for dysarthric speech is explored and described as part of a technique used to help those affected to overcome their disabilities.

Chapter 3 describes the proposed technique in both the classification and automatic speech recognition for dysarthric speech. The classification phase consists of corpus selection, feature extraction for certain sets of identified features, and feature selection for obtaining optimal sets of features using different feature ranking algorithms that were covered in the first part of the chapter. The proposed automatic speech recognition part of dysarthric speakers is presented as a second part section of this chapter, which consists of corpora used for training of the acoustic model, and the adaptation techniques which aim to minimize the mismatch between speakers' variability in the acoustic model.

Chapter 4 discusses the results obtained from applying the classification and adaptation techniques used in ASR for dysarthric speakers. The comparison of the results with other related works in classification and also in ASR for dysarthric speakers is described to evaluate our proposed method.

Finally, Chapter 5 ends this research by describing the capabilities and features of the proposed methodologies. Also included in this chapter are some promising directions which can be used as guidelines for further research in the future.

CHAPTER 2: LITERATURE REVIEW

2.1 Overview of this Chapter

This chapter presents the outcome of the intensive literature review of the related domain so as to answer the research questions and to meet the research aim. It covers the major component of the research as presented in section 1.8 above.

2.2 Gauging and Assessing the Severity of Dysarthric Speech

Speech robustness is one of the most important aspects of speech production, which is the speaker's ability to reduce a variety of interior and exterior noises (Kent et al., 2000). The speech motor control has the ability to distinguish between normal and neurologically disordered speech depending on the articulator's internal models, a motor-sensory combination which is based on rhythm, and with the dynamic articulations' specification, contained inside motor score or program (Kent et al., 2000). One major advantage of the study of speech disorder is to enhance the speech motor control's theories regarding the development of speech, standard regulation of speech, and identifying disorders of speech caused by diseases of the neurology (Kent et al., 2000).

The impairment of articulation, voice, and prosody are the standard characterized factors for dysarthria (described in chapter 1 section 1.2). On the other hand, the type and severity of dysarthria state the nature of impairment of speech (Kent et al., 2000). Dysarthria involves disorders in motor execution, with disruptions at different levels depending on the type of dysarthria. According to (Liss et al., 2009), no study has been conducted for dysarthric speaker classification as a large number of dysarthric speakers include dysarthria type, the severity of impairment, and disease types. The study also concludes that the measurement used for the classification of any type of dysarthria that can directly distinguish the dysarthria type is rare. Also, a combination of several existing

measurements has not been considered in the existing literature. An example is the research on speaking rate by (Niimi, 2001). However, this measurement could not be used to distinguish all types of dysarthria. Formant frequencies and formant transition rates have shown no particular differences between dysarthria type (Weismer et al., 2010).

Dysarthria type is related to the presentation of symptoms, which is divided into hypokinetic, hyperkinetic, ataxic, flaccid-spastic mix, spastic, and flaccid (Darley, Aronson, & Brown, 1969a, 1969b; Kent, Vorperian, Kent, & Duffy, 2003) Dysarthria disease is more complicated and corresponds with dysarthria type because one disease can be affected by one or more dysarthria type (Duffy, 2006). Dysarthria severity level concerns the degree of dysarthric impairment which is not a common research area among researchers (Young & Mihailidis, 2010).

2.2.1 Perceptual Judgments

The Perceptual Judgments (PJ) of speech is one of the most common traditional methods for dysarthria assessment (Murdoch, 1998). The first differential diagnosis of dysarthria was initially proposed by Darley et al. (1969b). The study was performed on seven different neurologic diseases including cerebellar ataxia, pseudobulbar palsy, bulbar palsy, amyotrophic lateral sclerosis, Parkinson's disease (PD), dystonia, and chorea. Figure 2.1 shows the category of each affected speech production mechanism.

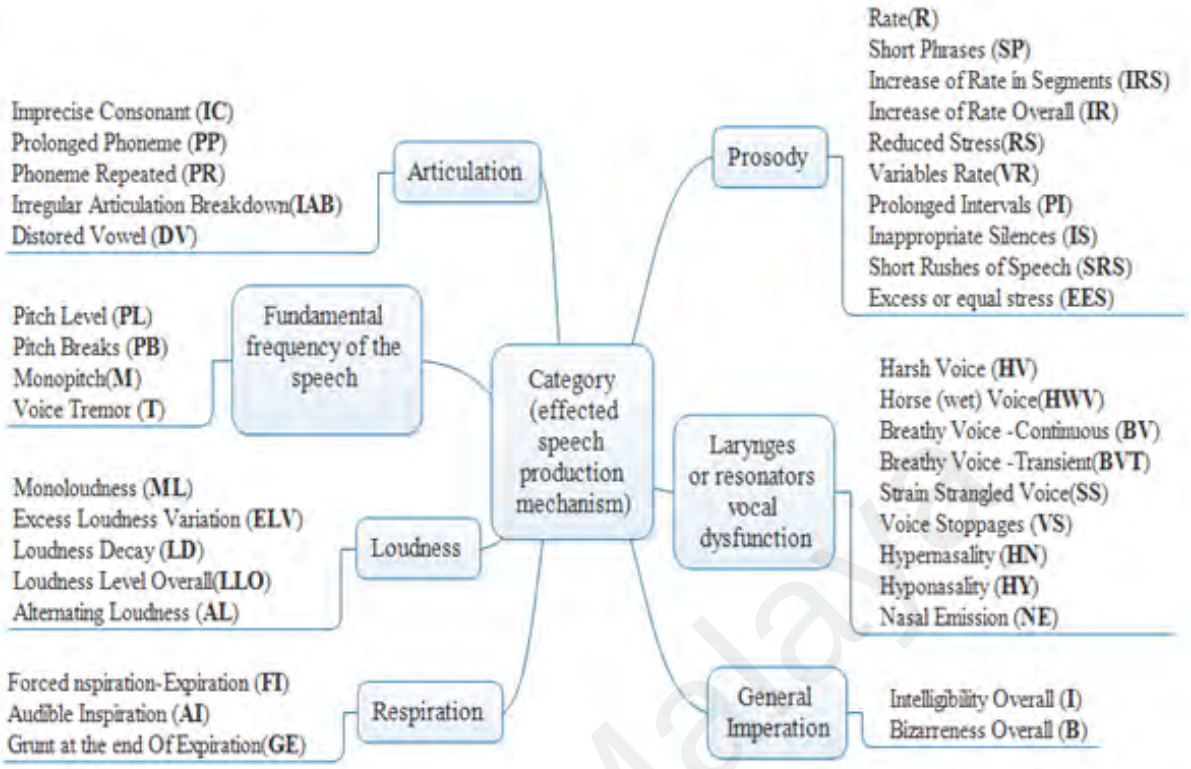

Figure 2.1: Speech production and it's related features used for perceptual judgment of dysarthric classification (Darley et al., 1969b).

In Darley et al. (1969b), samples that were collected for the study were rated on a seven-point scale by qualified speech-language pathologists perceptually. The results were in a group of features and were divided into eight clusters. Table 2.1 shows the eight clusters of features and their causes.

Table 2.1: Articulation clusters and their related features and caused

No	Cluster	Features related	Caused
1	Articulatory Inaccuracy	**IC**,**IAB**,**DV**	Disruption of the coordinator activity
2	Prosodic Excess	**R**,**PI**,**IS**,**EES**,**PP**	Neuromuscular defect
3	Prosodic Insufficiency	**M**,**ML**,**SP**,**RS**	Muscle Movement restriction
4	Articulatory-Resonatory Incompetence	**HN**,**IC**,**DV**	Previous and reduced power contraction
5	Phonatory Stenosis	**PL**,**PB**,**ELV**,**HV**,**SS**,**VS**,**R**,**SP**	Laryngeal Physiologic Narrowness
6	Phonatory Incompetence	**BV**,**BVT**,**AI**,**SP**	Reduced Muscular Contraction
7	Resonatory Incompetence	**HN**,**NE**,**SP**,**IC**	Previous and velo-pharyngeal port failure
8	Prosodic-Phonatory Insufficiency	**M**,**ML**,**HV**	Hypotonia

The speech perceptual analysis has two types of speech assessment tests, both of which are carried out manually. First, the articulatory test which is used as a clinical tool and is based on the perceptions of clinicians, speech therapists, or pathologists (Sommers, Logsdon, & Wright, 1992). The articulatory test concerns the severity level and diagnoses the errors of dysarthric speech. The knowledge, training, and experiences obtained by the clinicians guide the organization to set up a common standard and methods to help clinicians to make easy decisions regarding the disorder of the speaker (Ray Kent, 1996). An early standard test for dysarthric speech was "Frenchay Dysarthria Assessment", which used a different speech profile pattern that included 11 sections to address the types of dysarthria (Enderby, 1980b), and is used by many other researchers (Green et al., 2003; Mark Hawley et al., 2007). Some researchers combined the articulatory features with the spectrum features to precisely diagnose their results (Markov, Dang, & Nakamura, 2006). An A-score is used by Kayasith & Theermukong

(2009) for articulatory assessment. Two experts judged each speaker's speech and assigned the proportion of phonemes pronounced correctly over the total phonemes pronounced (modification of the Percentage of Correct Consonants (PCC)).

The intelligibility test is the second test of perceptual analysis for dysarthric speech. It measures the level of understanding between speakers and listeners (Kayasith & Theeramunkong, 2009). Compared to the articulatory test which depends on clinicians' knowledge, training and experiences, the intelligibility test is based on a group of listeners without any hearing impairment (Ray Kent, Miolo, & Bloedel, 1994). Some researchers used the combination of speech features like the quality of voice, articulation, prosody, and nasality as a factor to assess the comprehensibility (De Bodt, Hernández-Díaz Huici, & Van De Heyning, 2002; Narendra & Alku, 2019). Similar to the articulatory test, the intelligibility test needs subjective judgment. Moreover, both tests are painstaking and require human comprehension and appreciation (Kayasith & Theeramunkong, 2009). Twelve listeners with no hearing impairment have been assessed for their speech for three sessions, which consist of scales of rating, multiple choices, and transcription of the word (Kayasith & Theeramunkong, 2009). The final intelligibility score (I-score) is calculated by adding up and averaging evaluations from all listeners' for the three sessions.

The criteria for the classification of speech severity are reported in Kayasith & Theermukong (2009) which is used to classify the articulatory and intelligibility tests. Table 2.2 shows the classification of speech severity for articulation and intelligibility according to the severity score (Kayasith & Theeramunkong, 2009).

Table 2.2: severity level and severity speech for articulatory and intelligibility test

Severity level	Severity score
Very severe	0.00-0.40
Severe	0.41-0.60
Moderate	0.61-0.80
Mild	0.81-0.95
Normal	0.96-1.00

2.2.2 Objective and Perceptual Classification

The challenge of the Perceptual Judgments (PJ) method is to invariably distinguish the individuals and their existing speech disorders, more so with the existence of many different types of the intervention program. Many researchers have stated that judgment according to human perception alone is insufficient and finding of automatic features is necessary to obtain reasonable results in relation to disorder classifications (Fonville et al., 2008; Van der Graaff et al., 2009). Thus, the objective measure is used to improve the perceptions' precision and to utilize alternative comparison reference between different studies and subjects (Guerra & Lovey, 2003).

The combination of perturbation of speech's judgment of objective and perception is the trend of the dysarthria assessment (Ray Kent, Weismer, Kent, Vorperian, & Duffy, 1999). The main purpose of the perceptual combination and objective judgment is to receive the benefit of the automatic algorithms, to process the speech to automatically assess the level of severity for the dysarthric speakers, which was earlier perceived as less reliable by clinicians (Guerra, 2002).

In Guerra & Lovey (2003) a novel approach has been proposed in the form of features that are perceptual and objective, from speech signals that are impaired (pathological). The methodology: first, look at the features are extracted by using the digital signal processing algorithm. Then the clinician compares the less reliably judged combination with the remaining features which are directly taken from medical records or judgments of perception. The classifiers used are Linear Discriminant Analysis (LDA) and non-linear, based on Self-Organizing Map (SOM). The non-linear classifier has performed better than the linear classifier as well as providing data that are represented bi-dimensionally. This helps in a better understanding of the relationship between deviations of speech and the damage location in the central or peripheral nervous system.

2.2.3 Objective Classification

Currently, perceptual assessment (subjective intelligibility assessment tests) is used by speech-language pathologists to characterize speech disorder severity, as well as plan and monitor the treatment, and document intelligibility change over time (Klopfenstein, 2009). However, this type of assessment is expensive, strenuous, and depends on many biases and variables, e.g., being familiar with the patients and their pathologies of speech (De Bodt et al., 2002). On the other hand, objective measurement is economic and dependable. It ables to help in pharmacological and/or surgical evaluation of treatment (Constantinescu et al., 2010). Meanwhile, the automated machine-based system is becoming the preference of the clinicians to help predict the decision for the treatments (Hill et al., 2006; Maier et al., 2009).

With regards to the objective intelligibility measurement, the force alignment of the speech acoustic features to the phonemic and phonological features for a given unknown speech is one of the approaches used to predict corresponding hypothesis speech (Falk,

Chan, & Shein, 2012). In Middag et al (2009) an example was proposed for the systems to use the objective intelligibility measurement for dysarthric speech.

2.2.4 Dysarthric Speech Intelligibility

It is important to characterize a particular speech disorder's effects on intelligibility (Paja & Falk, 2012). The severity of speech in dysarthria has no standard, whereas the intelligibility of speech is frequently used to determine the level of speech mechanism affected by a neurological disease (Kent et al., 1989). Even though the number of speakers with different types of dysarthria at various severity levels is sufficiently large, the low number of associated analysis poses a challenge in differentiating severity effects and dysarthria type (Kim et al., 2011).

Each severity level has its own characteristics and can be used to classify speech impairment. It has been reported that both severity level and disorder types are used to measure the intelligibility of speakers (Falk et al., 2012). The direct measure of dysarthric severity has been proposed in Falk et al (2012). Meaning, intelligibility is one of the approaches for the classification of severity of dysarthria.

To identify the intelligibility deficits in dysarthria using the signal properties, and to determine acoustic features for each dysarthria type, are the two goals of the acoustic method in classification of dysarthria. The acoustic method has been conducted to determine the acoustic measure that predicts the level of intelligibility and physical correlation of perceptual features, to determine the dysarthria type (Rayd Kent & Kim, 2008).

Some researchers reported that speech intelligibility can be assessed using phoneme level intelligibility. The benefits of using the phoneme level intelligibility are: (1) the

speech intelligibility core is equivalent to the phoneme level intelligibility, (2) phonemes (consonants, vowels, and consonant words) are used as stimuli to observe the phoneme intelligibility and is relevant for dysarthria of all severities, and (3) adequately built assessment of phoneme intelligibility can lead to obtaining associated segmental information (Ray Kent, Weismer, Kent, & Rosenbek, 1989). An investigation of both phonemic features (acoustic model of phoneme and phonological features) and acoustic model of dimensions of articulation (articulation manner and place, voice, etc.) has been addressed in (Van Nuffelen, Middag, De Bodt, & Martens, 2009).

The reliability of measures of phoneme intelligibility for dysarthric speakers has been evaluated using Phonemic Features (PMF) model, Phonological Features (PLF) model, and combination of both phonemic and phonological features (PMF +PLF) model (Van Nuffelen et al., 2009). The computed scores using intelligibility models of the three types are matched to scores of perceptual intelligibility from the assessment of standardized intelligibility (Van Nuffelen et al., 2009). The calculation of the phonemic feature model uses the triphone model that is generated by comparing the pronounced speech with the transcription of the typical phoneme, according to what the speaker says. The final feature represents an average of a good pronunciation by a given speaker. On the other hand, the phonological feature generation uses the speaker context-independent acoustic model, which works on phonological space counted to 24 phonological features, like voice, fricative, nasal, etc.

Dysarthric speech intelligibility classification has been performed on a binary intelligibility label (Kim et al., 2015) as a score of intelligibility on at least five scales, or percentage (Frank Rudzicz, Namasivayam, & Wolff, 2012). Binary intelligibility is useful

for automation of intelligibility classification assessment, while the intelligibility score and percentage are useful for clinical and perceptual assessment.

In Dhanalakshmi & Vijayalakshmi (2015) a Gaussian Mixture Model (GMM)-based speaker identification system was used to evaluate the speaker identity, while Degradation Mean Opinion Score (DMOS) evaluates the naturalness, and the Word Error Rate (WER) assesses the intelligibility evaluation. The evaluation of the intelligibility carried out using the adaptive synthesis system is used to generate speech from the dysarthric speakers for the purpose of enhancing communication of the dysarthric speakers and normal speakers. Time-Domain Pitch Synchronous Overlap Add (TD-PSOLA) modifies the rate of synthesized speech. Table 2.3 shows the factor that alters the duration of specific dysarthric speakers (Dhanalakshmi & Vijayalakshmi, 2015).

2.2.5 Features of Dysarthric Speech

Guerra (2002) used 20 features from the 38 features listed in figure 2.1 to classify dysarthric speakers. The study uses a hybrid approach in which the automatic features have been extracted automatically using the speech algorithms and perceptual judgment by experts. The study used eleven of twenty features, to be applied on automatic feature extraction and the remaining nine features are for perceptual judgments by experts.

Table 2.3: Rate of Speech and Duration Factor of TD-PSOLA

Severity Level	Speaker ID	Speech rate (Phonemes/second)	Duration factor
Mild	BB	8.5	0.75
	FB	6.4	0.7
	MH	7	0.8
Moderate	RK	7	0.6
	JF	5.5	0.6
	RL	3	0.55
Severe	BK	3	0.5
	SC	4.5	0.5
	BV	2.3	0.65

The Kurtosis of Linear Prediction (LP) residual (κ_{LP}) signal is used to distinguish the excitation of the atypical vocal source (referring to vocal breathing and harshness). The rate-of-change of the signal in log-energy is used to characterize the speech temporal impairments with regards to the short-term (60ms) temporal dynamics. Delta zeroth-order cepstral coefficient's ($\sigma\Delta c_0$) standard deviation is used to compute the short term temporal dynamics. Temporal of speech is concentrated on the unclear distinction between the adjacent phonemes caused by articulation's inaccurate placement. The Low-to-High Modulation Energy Ratio (LHMR) is used to characterize the speech temporal impairment associated with the long term (256ms) temporal dynamics. Representation of the modulation spectral signal, which is auditory-inspired is used to represent the modulation spectral energy's ratio from frequencies lower than 4 Hz to frequencies greater than 4 Hz (Falk et al., 2012; Falk & Wai-Yip, 2010; Paja & Falk, 2012). Prosody

features are used as parameters to identify speech impairment such as standard deviation of the fundamental frequency ($\sigma\ f_0$), the range of fundamental frequency f_0 (Δ_{f0}), and percentage of segments of voice in words uttered (%V) (Paja & Falk, 2012). In Paja & Falk (2012) the combination of these features in one dimension feature is used for automatic prediction of intelligibility.

Practical applications are used to help dysarthric speakers. These applications use blind methods such as the second-formant slope transitions, duration of tone unit, and variation of the fundamental frequency (F0) (Kent et al., 1989; Schlenck, Bettrich, & Willmes, 1993). One advantage of using blind methods as a gauge for the intelligibility of dysarthric speech is that they don't need to prioritize information for the given word to be uttered (Klopfenstein, 2009). Moreover, dysarthric speech's rhythmic disturbances are characterized by the spectrum of modulation, which is the power spectrum of the speech signal envelope (LeGendre, Liss, & Lotto, 2009).

Harmonics-to-Noise Ratio (HNR), Glottal-to-Noise Excitation ratio (GNE), and Mel Frequency Cepstral Coefficients (MFCCs) are speech features used for the classification of dysarthric speech especially it's severity (Godino-Llorente, Gómez-Vilda, & Blanco-Velasc, 2006; Paja & Falk, 2012). MFCCs have the capability to capture the movement of irregular vocal fold or the lack of closure of vocal-fold caused by the change in mass/tissue (Godino-Llorente et al., 2006). GNE quantifies the excitation ratio, due to vocal fold oscillation versus turbulent noise (Godino-Llorente et al., 2010). HNR uses the difference in the ratio between the component of the periodic signal's energy and the component of the aperiodic signal's energy (Teixeira & Fernandes, 2014). The combination of all these features in one dimension is proposed in Paja & Falk (2012).

One measure used to identify severity type is the Low-to-High Modulation energy Ratio (LHMR) (Falk et al., 2012). The higher LHMR values are affected by the intelligibility level according to how the modulation spectral frequency contents are (greater or lower than 4 Hz). Frequency below 4 Hz causes less ineligibility (Falk et al., 2012) and above 4 Hz increases the intelligibility (Doh-Suk, 2004).

Some of the features like perturbations in temporal dynamics (long and short term), atypical excitation of the vocal source, separation of information of vocal tract and source, nasality, prosody, and composite measure (Falk et al., 2012). According to De Bodt et al., (2002) and Falk et al., (2012), a linear combination of the characteristics of dysarthric speakers, the performance was better than its individual measures. However, the use of composite measures proved to overcome standalone measures.

The Variability Index (VI) is defined as the average syllable variability for a given utterance, after comparing the duration of neighboring syllables with the normalized duration of each syllable (Deterding, 2001). Compared to a control group of speakers, VI values were lower for a group of ataxic dysarthria. Furthermore, control speech and ataxic dysarthria have different intersubjective variability in VI values (Stuntebeck, 2002). The calculation of VI is as the following:

$$VI = \frac{1}{n-2}\sum_{n=1}^{k-2}|d_{k+1} - d_k| \qquad (2.1)$$

Where d_k is the normalized duration of the kth; and n is the number of syllables in the utterance.

Nowadays**,** speakers with mild and moderate dysarthria are using automatic speech recognition (ASR) as an objective measure to identify intelligibility (Doyle et al., 1997;

Maier et al., 2009; Sharma, Hasegawa-Johnson, Gunderson, & Perlman, 2009). For severe dysarthric speakers, the difficulty is for the speakers to make the speech more understandable, to be reflected by the ASR technologies. Pre-processing approaches are used to enhance the ASR in identifying the type of speakers with dysarthria (Middag et al., 2009; Frank Rudzicz, 2007). One limitation of using ASR for dysarthric speakers is the limitation of the vocabulary size (Doyle et al., 1997). By being speaker-dependent (Frank Rudzicz, 2007) and having data availability sparseness (Green et al., 2003), accurately trained models are needed to overcome this limitation.

With regards to the pitch, mild dysarthria is associated with low pitch variation (monotonicity), whereas severe dysarthria is associated with high pitch variation (Falk et al., 2012; Schlenck et al., 1993).

2.2.5.1 Acoustic features of dysarthric speech

An acoustic analysis provides data on dysarthria as accompanied by several diseases and may also include speech behavior over time (Ray Kent et al., 1989). There are many speech parameters that play an important part in decreasing the speech intelligibility, like voicing contrasts, nasalization, and vowel height (Weismer, Martin, & Kent, 1992).

Voice Onset Time (VOT) (Liu, Tseng, & Tsao, 2000), second formant frequency (F2) slope (Kent et al., 1992; Y. Kim, Weismer, Kent, & Duffy, 2009), and acoustic vowel space (McRae, Tjaden, & Schoonings, 2002; Tjaden & Wilding, 2004; Weismer, Jeng, Laures, Kent, & Kent, 2001) are some of the acoustic features used to determine speech intelligibility of speakers with dysarthria. Dysarthria type is characterized according to acoustic measurements such as slow rate of speaking, VOT with high variability, almost similar duration of utterance with regards to vowel/syllable, and fundamental frequency (F0) range across utterances which are abnormally large, has been associated to ataxic

dysarthria (Kent et al., 2000). Hypokinetic dysarthria is associated with normal or faster rate of speaking, high mean F0, decrease F0 variability, and decrease in extents and slopes of F2 (Goberman, Coelho, & Robb, 2005; Solomon & Hixon, 1993). Spastic dysarthria has been studied in (Ozawa, Shiromoto, Ishizaki, & Watamori, 2001; Özsancak, Auzou, Jan, & Hannequin, 2001). Hyperkinetic, flaccid, and mixed dysarthria have been studied (Liss et al., 2009; Wang, Kent, Duffy, & Thomas, 2005).

The Root-Mean-Square (RMS) intensity contour, F0 contour, F2 transitions extent and duration, M1 for fricatives (/s/ and /ʃ/) during three 50-ms-long windows approaching the vocalic nucleus (25-ms overlap between adjacent windows), first and second formant frequencies from four corner vowels, voiceless interval duration, and vowel and sentence duration, were acoustic measurements studied in (Kim et al., 2011). These measurements were required to derive these variables for analysis: RMS intensity range of utterance, F0 range (maximum-minimum) of utterance, F2 slope, M1 difference between /s/ and /ʃ/, acoustic vowel space, Pairwise Variability analysis (PVI), and rate of articulation.

According to the Kim et al. (2011), the following acoustic measures were significantly correlated with speech intelligibility according to disease group: F2 slope, vowel space, the difference of M1 for /s/ and /ʃ/, rate of articulation, Voiceless Interval Duration, and range of F0 interquartile. All clinical groups, except for Parkinson's disease (PD), had a significant rate of articulation, and scores of speech intelligibility for all four disease groups showed significant regression of F2 slope.

Regarding the classification type, the study conducted in Kim et al (2011) showed that there are acoustic variables that contribute to the classification of dysarthria. The intensity ring (dB), Voiceless Interval Duration (VID), and Articulation Rate (AR) have a significant effect on etiology classification. Furthermore, the range of F0 and Articulation

Rate (AR) affect the classification of type, whereas ranges of F0, the slope of F2, and vowel space (VS) have more effect on severity classification.

2.2.5.2 Vowels of dysarthric speech

Vowel-related phonological contrasts like tense-lax, high-low, and front-back are used to predicate the intelligibility of words among dysarthria speakers (Ansel & Kent, 1992). Vowel space area has an important effect on speech intelligibility as found in (Liu et al., 2005). The reduction of displacements in the articulation of dysarthric speakers resulted in squeezing of vowel space area which is affected by the speech intelligibility (Neel, 2008; Tjaden & Wilding, 2004; Weismer et al., 2001). The vowel space area is measured in F1 and F2. F1 measures (dimension) jaw opening and tongue height while the F2 measures (dimension) the tongue position (Neel, 2008).

The features that contribute to intelligibility variability found in the cerebral palsy dysarthric adult are contrasts in a vowel (tense-lax, high-low, and front-back) (Ansel & Kent, 1992). The errors were: short vowels recognized as long vowels and vice versa (called short-long pair error), high vowels recognized as low vowels (called tongue height error), front vowels recognized as back vowels (called tongue advancement error), and target monophthong recognized as diphthong, or a diphthong recognized as a different diphthong (Ansel & Kent, 1992; Thubthong, Kayasith, Manochiopinig, Leelasiriwong, & Rukkharangsarit, 2005).

Some researchers have categorized the error pattern according to misrecognized types, which is considered as recognition error. The misrecognized types are substitution (manner, placement, the combination of manner and placement, height, short-long, pair, and others), distortion, addition, omission, and reduction (Manochiopinig, Thubthong, & Kayasith, 2007; Manochiopinig, Thubthong, & Kayasith, 2008). The results showed that

substitution has the most number of error patterns with all kinds of dysarthric speech characteristics as well as the only error pattern to occur in vowels and tones (Manochiopinig et al., 2007; Manochiopinig et al., 2008; Thubthong et al., 2005). It is followed by reduction, distortion, omission and addition, in that order (Manochiopinig et al., 2007; Manochiopinig et al., 2008).

2.2.5.3 Consonants of dysarthric speech

The consonant is one feature that plays an important role in dysarthria. It applies in almost all perceptual types of dysarthria. However, there is no consensus on the acoustic measure to cover all consonants (Ray Kent & Kim, 2003). The Voice Onset Times (VOT) is one of the consonant features (consonant stops) which is mostly used to distinguish between impaired and non-impaired speech (Auzou et al., 2000). Another factor in the consonants feature is the fricatives, where controlling articulation precisely for speech production makes this feature interesting for researchers. Many researchers focused on the fricative /s/, which happens frequently in multiple languages, as well as its distinct spectral pattern (Ray Kent & Kim, 2003). A study by Chen and Stevens to acoustically measure the fricative /s/ as pronounced by normal and dysarthric speakers show that both the perceptual and spectral analysis have been compared based on spectrographic observation (Chen & Stevens, 2001). The analysis criteria were the initial sound, proper tongue position, and change from the current fricative to the following vowel (Chen & Stevens, 2001). The study, as obtained by the judges, concludes that between intelligibility and fricative /s/ ratings of the speakers, a correlation exists (Chen & Stevens, 2001).

With regards to phonemes category, dysarthric speakers were reported to have difficulty in pronouncing alveolar phonemes for the initial and final consonant. Whereas,

the high recognition accuracy, labial phonemes of final consonant and glottal phonemes of initial consonant were obtained. In the initial consonant, the approximant phonemes have more recognition accuracy than affricate phonemes (Manochiopinig et al., 2007; Manochiopinig et al., 2008). As for the error pattern in the consonant, most errors were reported to be manner errors (place, voicing, or both) (Thubthong et al., 2005; Whitehill & Ciocca, 2000).

2.2.5.4 Prosody features of dysarthric speech

With regards to the correlation between the prosody and assessment of dysarthric speech, some studies reported that the speakers with severe dysarthria had higher mean F0 and shorter tone units while the speakers with mild dysarthria or neurological disease had lower mean F0 and longer tone unit (Schlenck et al., 1993). The prosody parameters, such as the unit of tone (word, ratio, duration), the variation of F0, and mean F0 had participated in distinguishing the dysarthric speech and had interaction with speech intelligibility (Bunton, Kent, Kent, & Rosenbek, 2000). Fundamental frequency contour parameters are one of the prosody features that had interaction with dysarthric speech intelligibility and dysarthric types (Bunton, Kent, Kent, & Duffy, 2001).

The prosody features for the non-intelligible dysarthric speech was obtained using the utterance level features and phonetic level features. The utterance level features elicit the following [0.1 0.25 0.5 0.75 0.9] quantiles, interquartile pitch range and its delta, variance in pitch, Z-score of each phoneme duration, normalized L0-norm ratio and the normalized utterance duration and their sums (Kim et al., 2015). Features for the phonetic level involve the variance of pitch contour and stylization parameter. They are calculated by fitting quadratic polynomials for each phonetics (Kim et al., 2015).

2.2.5.5 Nasality features of dysarthric speech

Hypernasality feature is one of the nasality indices which has gained importance for dysarthric speech classification. It is because hypernasality is a feature for dysarthric speech classification type and severity level. Hypernasality is also a factor that diminishes intelligibility (Ray Kent & Kim, 2003).

2.2.5.6 Distance measure of dysarthric speech

There are some approaches that are used to compare the features from both the dysarthric speakers' speech and non-impaired speakers' speech. An example proposed by Gu, Harris, Shrivastav, & Sapienza (2005) is by computing the distance measure (e.g., Itakura-Saito distortion) between the given speech utterance samples for dysarthric speakers and equally the same speech of non-impaired speakers. Dynamic time warping is applied to calculate the differences of the speech feature, like utterance durations (Gu et al., 2005).

A combination of the possible source of variability features with the feature selection for feature dimensionality is one of the easy and common fusion methods for solving issues of dimensionality. This is called the feature-level fusion (Kim et al., 2015).

2.2.5.7 Other features of dysarthric speech

Speech quality features such as ratio of harmonics to noise, shimmer and jitter, spectral features (such as formants and mel-frequency cepstral coefficients), scores of automatic speech recognition (such as word recognition or confidence score of phoneme, prosodic, phonemic and perceptual features), and features of estimated speech parameter (such as phonological features), are features that provides indicators for speech variabilities in general and to enhance the speech recognition rate (Dibazar, Berger, & Narayanan, 2006; Kodrasi & Bourlard, 2019; Andreas Maier et al., 2009; Middag, Bocklet, Martens, &

Nöth, 2011; Van Nuffelen et al., 2009). Researchers have studied those features at the word level. Some researchers studied a simple sentence or passage level, which has the following advantages: 1. Stimuli used during the data collection 2. The simplicity of the pronunciation of segmentation which is composed of short duration and intelligibility information. 3. Data collected exhibit real-world communication scenarios. The sentence-level speech production data has more variability and complexity with regards to the robustness and characteristics of the feature, as compared to word-level or single phonetics data (Kim et al., 2015).

The features of voice quality, like Harmonics to Noise Ratio (HNR), and shimmer and jitter are used for intelligibility classification tests (Kim et al., 2015; Narendra & Alku, 2018, 2019). Features of voice quality are proposed to be effective in speech intelligibility (Kim et al., 2015). Preminger and Van Tasell (1995) reported that speech intelligibility is less affected by perceptual speech quality. Speech quality dimensions, like total impression, the effort to listening, and loudness have been suggested to be more predictable than speech intelligibility scores (Preminger & Van Tasell, 1995).

The pronunciation features, like the duration of phonetics, cepstral mean normalized 39-dimension Mel-Frequency Cepstral Coefficients (MFCCs), and formants, have been used for pronunciation variation (Witt, 1999). Some researchers used statistical spectral features to address pronunciation features. This feature includes [.05 1.25 5.75 9.95] quantiles, interquartile range, and third-order polynomial coefficients of the first, second, and third formats of their bandwidths and their derivatives for each segment of the vowel in each utterance (Kim et al., 2015). Some pronunciation features extracted from waveforms of speech at the utterance-level excluded the silence where the maximum and standard deviation of cepstral mean normalized 39 MFCCs are estimated (Kim et al.,

2015). The temporal features contain the duration of pause and average syllable, and exclude beginning and ending silence to the duration of average vowel and syllable ratio number, calculated from the phonetic transcription (Kim et al., 2015).

According to Kim, Martin, Hasegawa-Johnson, & Perlman (2010) articulatory manner change is accompanied by the speech of dysarthria. In contrast, the variability of articulatory place can affect all normal and dysarthric speech. Moreover, for complex phonetics production, there is an increase in articulatory errors (Kim, Martin, Hasegawa-Johnson, & Perlman, 2010). The utterance-level prosodic feature variation also increases for dysarthric speech (Kim et al., 2010). The super-Gaussianity of the speech spectral coefficients arises due to the pauses between the phonemes and due to formant transitions in voiced sounds used to classify the healthy and dysarthric speakers (Kodrasi & Bourlard, 2019).

2.2.6 The Techniques for the Classification of Dysarthric Speech

The techniques used for the classification of dysarthric speech in (Guerra, 2002; Guerra & Lovey, 2003) are the differential diagnosis of dysarthrias proposed by Darley, Aronson, and Brown (DAB), where a Linear Discriminant Analysis (LDA) and a non-linear method using Self Organizing Map (SOM) are used.

DAB was implemented using average PJ of clinicians on 38 dimensions, grouped into 8 clusters. A minimal Euclidean distance between the combination of clusters reported for each type of dysarthria and the vector formed by each cluster's occurrence form the base for the decision to be made (Darley et al., 1969b).

LDA method was an alternative approach where the final dataset (combined data) is separated into different groups using linear surfaces. In this method, the input vector is

classified into a group, with the existence of a minimal squared distance between it and the observation. LDA, therefore, finds a unique discriminant equation for patients of each group or class. Features that positively contribute to the linear equation's final magnitude is the source in the relevancy of the classifier's decision (Guerra & Lovey, 2003).

SOM method in a non-linear approach where an unsupervised artificial neural network based on SOM is used (Fritzke, 1994; Kohonen, 1990). One advantage of using the SOM method is that the group's spatial distribution can be represented bi-dimensionally. The idea of the SOM method is to extract the features and analyze according to vectors such as weights associated with the neurons' centroid from the relevant groups (Kohonen, 1990).

According to Guerra (2002), the LDA improved its classification recognition rate. On the other hand, the SOM classifier provided almost a 5% better classification ratio than LDA and almost 20% better than DAB. Additionally, the SOM provided a better perceptual percentage of correct classification (PPC), backward relevancy analysis that is more precise and produces a map containing each group's detailed information.

Linear, quadratic and Mahalanobis distance-based discriminant functions are used as discriminant analysis classifiers for automatic detection of dysarthria speech severity (Paja & Falk, 2012). In Paja & Falk (2012) cross-validation of 15-fold was used with validation use consuming 30% and training use taking up 70% of the recorded data. The results in Paja & Falk (2012) show that Mahalanobis distance classifier obtained the best accuracy of 95% when used with combined features (the nasality-related features and prosody features totaling 6 features on one side and the remaining 28 consisting of salient acoustic features, making up 34 features in total). Paja & Falk (2012) proved that the combined features perform better than stand-alone features.

The utterances have been clustered according to the speech characteristics of the subjects (Kim, Kumar, Tsiartas, Li, & Narayanan, 2012). The bottom-up Agglomerative Hierarchical Clustering (AHC) with a single Gaussian was used to group together utterances of similar speech. By using a majority voting rule, the ad-hoc scheme that jointly classifies all utterances inside a cluster enforces smoothness constraint (Kim et al., 2012).

Line Spectrum Pair (LSP) is one of the linear prediction parametric representation for spectral information of the speech space and is related to the formant of speech sound or natural resonances (Qian, Soong, Chen, & Chu, 2006). LSP feature from each utterance with Generalization Likelihood Ratio (GLR) distance is used to perform the AHC clustering and to perform a posterior smoothing for the test set (Kim et al., 2015).

Table 2.4 shows the related classification of dysarthric speakers, which includes the classification algorithms, evaluation method, data set, extracted features, feature selection used, and results obtained from studies in classifying dysarthric speech.

Table 2.4: Summary of classification and feature selection methods used for classifying the dysarthric of speech

Author(s)	Classifier Algorithm	Evaluation Method	Data Set	Extracted Features	Feature Selectio	Results	Note
Frederic L. Darley et al., (1969)	Perceptual judgment (three judges used for this assessment)	Minimal Euclidean Distance	212 dysarthric subjects	38 features as presented in Figure 2.1	No	66.1 % accuracy of classification	Seven groups of dysarthria were studied - it considers first studied regarding dysarthric speech
Guerra & Lovey (2003)	Linear discriminant analysis(LDA) and non-linear based on self-organizing map (SOM) (artificial neural network based on SOM network)	Percentage of correct classification(PPC)	62 dysarthric subjects (database collected by Aronson and colleagues, 1993)	20 speech features (11 extracted using computer algorithm and 9 extracted using perceptual judgment)	No	The non-linear classifier (85.83%) and linear classifier (81.1%)	Dysarthria type classification
Tiago H Falk, Chan, & Shein (2012)	Class-based linear estimators	Pearson (R) and Spearman rank (RS) correlation coefficients, along with their corresponding p-values	Universal Access Speech database (10 spastic dysarthric speakers)	A Typical vocal source excitation (1 feature), temporal dynamics (4 features), Nasality (8 features), prosody (3 features), and composite (3 composite features)	No	combine many features have an effect on the objective measurement of dysarthric word intelligibility	The atypical vocal source excitation, temporal dynamics and prosodic as features to dysarthric intelligibility assessment

Author(s)	Classifier Algorithm	Evaluation Method	Data Set	Extracted Features	Feature Selectio	Results	Note
Middag et al (2009a)	Linear regression models	Fivefold cross-validation (CV) and root mean squared error (RMSE) for performance criterion	Dutch Intelligibility Assessment (DIA), 211 speakers (51 control speakers)	55 Phonemic Features (PMFs), 24 Phonological Features (PLFs), and 768 Context-Dependent Phonological Features (CD-PLFs)	Yes	Combine many features have an effect on automating intelligibility assessment	Three groups of the feature were studies in order to predict the intelligibility (PMF, PLF, and CD-PLF)
Schlenck, Bettrich, & Willmes (1993)	Discriminant Analysis(ALLOC80)	Leaving-one-out strategy	154 normal subjects and 84 dysarthric subjects	Mean Fundamental Frequency(F0), its standard deviation, and the highest and lowest F0 measurement	No	84% of male speakers and 100% of the female speakers were correctly classified	distinguish between dysarthria and normal speakers
Godino-Llorente, Gómez-Vilda, & Blanco-Velasc (2006)	Gaussian mixture model (GMM)	K-fold cross-validation	53 control subjects and 173 pathological subjects	MFCC	Yes	94%	distinguish between pathological and normal speakers (used F-Ratio and Fisher's discriminant ratio as a feature selection)

Author(s)	Classifier Algorithm	Evaluation Method	Data Set	Extracted Features	Feature Selectio	Results	Note
Fonville et al. (2008)	KAPPA statistics	Confidence Interval	100 dysarthric subjects	neurologists and neurological trainees(manual assessment of type off dysarthria)	No	35%	Dysarthria type classification
Y. Kim, Kent, & Weismer (2011)	Discriminant Function Analysis (Statistical Analysis)	SPSS Version 16.0 for every single Acoustic variable and for all eight acoustic variables	107 dysarthric subjects	8 segmental/ suprasegmental features: 2nd formant frequency slope, articulation rate, voiceless interval duration, 1st-moment analysis for fricatives, vowel space, F0, intensity range, and Pairwise Variability Index	No	The classification accuracy of dysarthria using disease type or severity level outperform classification using dysarthria type	Comparison of the classification of dysarthria based on type, disease, and severity
Liss et al. (2009)	Discriminant Function Analysis	A cross-validation method	55 subjects including control subjects	10 features of rhythm metrics	No	80% successful in classifying speakers into their appropriate group	the effectiveness of the rhythm metrics in dysarthria type classification

Author(s)	Classifier Algorithm	Evaluation Method	Data Set	Extracted Features	Feature Selectio	Results	Note
De Bodt et al (2002)	Linear Regression Analysis	The correlation between the four dimensions and the overall judged intelligibility	79 dysarthric subjects	Voice quality, articulation, nasality, and prosody	No	95% prediction interval of Judged and calculating rating were in agreement of 75% of the patients	The effectiveness of the linear combination of voice quality, articulation, nasality, and prosody on the overall intelligibility of dysarthric speakers
Kim, et al (2010)	Listeners Classification	Two-way ANOVA analysis of the correct percentage of consonant categories	7 dysarthric subjects	-	No	more intelligible speakers produce more correctly articulated consonant	Consonants classification to three types of articulations which are articulatory complexity, place of articulation and manner of articulation. (The classification perform manually by listeners and compared to the pronunciation dictionary

Author(s)	Classifier Algorithm	Evaluation Method	Data Set	Extracted Features	Feature Selectio	Results	Note
Middag et al (2011)	Ensemble Linear Regression (ELR) Support Vector Regression (SVR)	Pearson Correlation Coefficient (PCC) and the Root Mean Squared Error (RMSE)	85 subjects suffered from cancer in different regions of the larynx (German) 122 subjects as Flemish Pathologic al Speech (Dutch)	Acoustical features and phonological features	Yes	SVR classifier outperformed ELR classifier in both data sets	The effectiveness of using ASR in intelligibility prediction model- the combination of acoustical and phonological features
Kim et al. (2015)	Linear Discriminant Analysis (LDA) classifier, K-Nearest Neighbor (KNN) classifier and Support Vector Machine (SVM)	leave-one-subject-out for testing, and used random cross-validation for parameter tuning	55 subjects of NKI CCRT Speech Corpus and 10 subjects(6 dysarthric and 4 control) of the TORGO database	Prosody, Pronunciation, Voice quality, and All	Yes	73.5% unweighted classification and 72.8% Weighted classification	Sentence level features of pathological speech for automatic intelligibility classification (binary classification - intelligible and not intelligible)

Author(s)	Classifier Algorithm	Evaluation Method	Data Set	Extracted Features	Feature Selectio	Results	Note
Kim et al. (2012)	Naïve Bayes, Noisy-Majority and joint both	-	NKI CCRT Speech Corpus	multiple language phoneme probability(MLPP), prosodic and international features, voice quality, and pronunciation features	-	76.8% accuracy on a test set when joint classification is applied (Naïve Bayes+Noisy-Majority)	Intelligibility classification (binary classification (intelligible and not intelligible)
Paja & Falk (2012)	Linear Quadratic A Mahalanobis distance-based discriminant analysis classifier	Randomized bootstrap (15-fold) cross-validation was used with 70% of the input data recordings kept for system training and 30% left for validation.	Universal Access Speech database (10 spastic dysarthric speakers)	Atypical vocal source excitation, temp-oral dynamics, nasality, and prosody. A subset of six of these features was shown to be useful for speech intelligible-ility prediction, as well as the alternate, features Mel-frequency cepstral coeffici-ents (MFCCs), glottal-to-noise excitation ratio (GNE), and	Yes Nine top featu res	combination of features make Mahalanobis distance-based discriminant outperform with 95% the linear and quadratic classifier	Spastic Severity Disorder Classification

Author(s)	Classifier Algorithm	Evaluation Method	Data Set	Extracted Features	Feature Selectio	Results	Note
				harmonics-to-noise ratio (HNR)			
Narendra & Alku (2018)	SVM	Leave-one-subject-out cross-validation (Classification Accuracy)	TORGO database	Glottal features log-energy, MFCCs (13), Mel-spectrum (26), zero-crossing rate, pitch, jitter, shimmer, voicing probability, spectral flux, roll-off points, spectral centroid, the position of spectral maximum and minimum	Yes	Almost 94% with feature selection algorithms	examines the effectiveness of glottal source parameters in dysarthric speech classification from three categories of speech signals(non-words, words, and sentences)
Narendra & Alku (2019)	SVM	Leave-one-subject-out cross-validation (Classification Accuracy)	TORGO database and UA-Speech database	Glottal features log-energy, MFCCs (13), Mel-spectrum (26), zero-crossing rate, pitch, jitter, shimmer, voicing probability, spectral flux, roll-off points, spectral centroid	Yes	89 % on TORGO database and 96 % on UA-Speech database	Identifying the Dysarthric speakers from coded telephone speech

2.2.7 Classification of Dysarthric Speaker Summary

After discussing and reviewing several topics related to the classification of dysarthric speech which is summarized in Table 2.4. Some of the key points are discussed below:

2.2.7.1 Dysarthric features and feature selection

There are many features extracted from dysarthric speech signal and used for the classification of dysarthric speech including the following: A typical vocal source excitation, prosodic features, nasality features, phonemic features, pathological features, fundamental frequency, MFCC, articulation features, voiceless interval duration, rhythm features, voice quality features, acoustic features, phonological features, g Harmonics-to-Noise Ratio (HNR), and Glottal-to-Noise Excitation ratio (GNE). Some of these features are very large, so feature selections are used and summarized as the following:

- Features selection (F-Ratio, Fisher's discriminant ratio, and forward selection procedure).
- From a set of input variables, the selection of a subset of variables is the focus of feature selection. The subset variables have the ability to describe the input data and provide good prediction results as well as reduce the computational time while reducing the effect of noise or irrelevant variables (Chandrashekar & Sahin, 2014, Guyon & Elisseeff, 2003).
- For dysarthric speech, the use of feature selection is to reduce the number of a feature used to predict the intelligibility or severity level of speech.

2.2.7.2 Classification of dysarthric speech

- The classification of dysarthric speakers is performed to identify the followings:

- (Neurologic diseases, dysarthria type, the intelligibility of the speech, differentiation between non-impaired speakers and dysarthric speakers, severity level).

- Classification algorithm used are as follow:

  - Linear Discriminant Analysis (LDA).
  - Artificial Neural Network (ANN).
  - Discriminant Analysis (ALLOC80).
  - Gaussian Mixture Model (GMM).
  - Statistical Analysis (Discriminant function analysis, linear regression analysis).
  - K-Nearest Neighbor (KNN).
  - Support Vector Machine (SVM).
  - Quadratic Discriminant Analysis (QDA).
  - Mahalanobis Discriminant Analysis (MDA).

2.2.7.3 Corpora used in dysarthric speech classification

- NKI CCRT (Advanced head and neck cancer - Concomitant chemo-radiation treatment) used for binary intelligibility classification of pathological speech.
- TORGO used for binary intelligibility classification (Intelligible or not intelligible).
- DIA used to study the effect of phonemic and phonological features on the automatic intelligibility assessment.
- UA-Speech used for spastic severity classification for dysarthric speakers.

2.3 Automatic Speech Recognition System for Dysarthric Speakers

For individuals severely disabled physically and with dysarthria, their educational and vocational opportunities have been expanded significantly by the speech recognition technology (Ferrier, Shane, Ballard, Carpenter, & Benoit, 1995). Neurologic impairment affects manual motor control and speech intelligibility (Ferrier et al., 1995; Mathew, Jacob, Sajeev, Joy, & Rajan, 2018; Wilson, Abbeduto, Camarata, & Shriberg, 2019). The issues that are faced by the pathologies to enhance the speech for dysarthric speakers are the poor control of motor function and consequently, a production rate that is slow. Also, the techniques used as assistive devices need an extensive form of training (Ferrier, 1991; Takashima, Takiguchi, & Ariki, 2019). Even for extremely poor speech intelligibility, communication mode often preferred is speech (Ferrier et al., 1995).

Joy & Umesh (2018) explored multiple ways to improve the recognition accuracy of GMM-HMM and DNN-HMM acoustic models for the TORGO dysarthric speech database. The results showed that trained speaker-specific acoustic models that incorporate various acoustic model parameters, speaker normalized cepstral features, and complex DNN-HMM models improved the recognition accuracy for dysarthric speakers (Joy & Umesh, 2018).

Doyle et al. (1997) used the perceptual assessment in his experiments to compare two types of recognition for dysarthria, which are automatic recognition and human listeners (Doyle et al., 1997). The results presented in Doyle et al. (1997) showed that the automatic recognition for the control speakers is more consistent than dysarthric speakers. The results obtained are according to the gender and it showed that female speakers performed better as compared to the male speakers. The experiment included six speakers who were divided into three severity levels (mild, moderate, and severe) with every level consisting of two speakers, one male, and one female. According to Doyle et al. (1997), the amount of training data affects the accuracy of the speech recognition for dysarthric speakers.

ASR system for speech dysarthria relies on speech intelligibility and consistency (Doyle et al., 1997).

The perceptual stage in Doyle et al. (1997) consists of ten young adults without hearing impairment with ages ranging from 22 to 27 years. They served as listeners to evaluate impaired and controlled speakers. All listeners were first-year students in communication sciences and disorders, and none of them had clinical experience with dysarthric speakers. According to Doyle et al. (1997), the recognition accuracy for dysarthric speakers had overlapped with the control speakers and between severity type itself. On the other hand, the perceptual assessment showed that the score of the listeners for all of the controlled and impaired (including severity classification of dysarthria) speakers was better than the score by automatic speech.

Modern approaches use automatic speech recognition to assess speakers with dysarthria (Green et al., 2003; Mark S. Hawley et al., 2007; Mark Hawley, Enderby, Green, Cunningham, & Palmer, 2006; Kayasith & Theeramunkong, 2009; Kayasith et al., 2006a; Kayasith, Theeramunkong, & Thubthong, 2006b; Takashima et al., 2019; Zaidi, Boudraa, Selouani, Addou, & Yakoub, 2019). Speech Training And Recognition for Dysarthric Users of ASsistive Technology (STARDUST) is one of the projects developed based on Hidden Markov Model (HMM) algorithms that work on command words to help severely dysarthric speakers (Green et al., 2003). A prototype ASR system to facilitate speakers with dysarthria, to be treated with equipment (electronic assistive technology) is reported in (Mark Hawley et al., 2007; Mark Hawley et al., 2006). A small data has been used to develop the system with reasonable recognition achievements. Some researchers focused on the automatic prediction of speech recognition for children with dysarthria using speech indices (Kayasith & Theeramunkong, 2009; Kayasith et al.,

2006a, 2006b). For the speech prediction indices, the tests have been performed in a limited environment, and the results needed more exploring.

2.3.1 Speaker Adaptation

The main aim of the speaker-dependent system is for the successful recognition and consistency of sound production. The speaker adaptable system has the ability to learn the acoustic characteristics of the individual speaker and adapt them to the specific speaker. This ability of the speaker adaptable system helps to compensate inconsistencies in speech production (Stern & Lasry, 1987).

The improvement of the recognition accuracy for dysarthric speakers using the adaptation techniques based on the severity level of dysarthria is developed in (Bhat, Vachhani, & Kopparapu, 2016). In the adaptation of tempo, a pre-determined adaptation parameter α is used for the temporal reduction of the sonorant regions of an utterance. The severity based adaptation of tempo for Indian language is developed from both Hidden Markov Model (HMM) and Deep Neural Network (DNN). The results show that using tempo adaptation improves the recognition accuracy of the dysarthric speakers from both HMM and DNN based acoustic models. The improvement includes the speaker-independent model and speaker adaptive model.

2.3.1.1 MLLR adaptation technique

Among the linear mapping strategies, the Maximum Likelihood Leaner Regression (MLLR) is used among the acoustic feature spaces of many speakers. It is an adaptation technique popularly used in ASR (Gales & Woodland, 1996; Leggetter & Woodland, 1995). In MLLR, representation of HMMs' mean vectors of the Gaussian distribution is as

$$\mu = (\mu_1, \ldots\ldots, \mu_n)' \qquad (2.2)$$

Where the dimension of a feature vector is represented by n. The following transformation is used to update the mean vector in equation (2.2):

$$\hat{\mu} = A\boldsymbol{\mu} + \boldsymbol{b}, \qquad (2.3)$$

Where, $n \times n$ matrix is represented by A, and -dimensional vector is represented by $\boldsymbol{b}$.. The following shows how Equation (2.3) can be written into a linear mapping:

$$\hat{\mu} = W\boldsymbol{\xi}, \qquad (2.4)$$

Where $\boldsymbol{\xi} = (1, \mu_1, \ldots\ldots, \mu_n)'$. W is a $n \times (n+1)$ matrix. Its first column is identical to $\boldsymbol{b}$.

Maximum Likelihood (ML) can be estimated by using the Expectation-Maximization (EM) algorithm to calculate W. In the following, the feature vector's sequence is X:

$$X = \{x_1, \ldots\ldots, x_T\} \qquad (2.5)$$

The following is a rewritten auxiliary function:

$$Q(W, \bar{W}) = K - \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)[K_m + \log|\sum_m| + (x_t - W\boldsymbol{\xi})'\sum_m^{-1}(x_t - W\boldsymbol{\xi})] \quad (2.6)$$

Where, at time t, the posterior probability of being in mixture component m is represented as $\gamma_m(t)$. K is a term independent from the output probability, and for mixture component m, normalization factor is represented by K_m. The following equation uses ML estimation to estimate the $\bar{W}$ of W:

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\sum_m^{-1}x_t\xi'_m = \sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\sum_m^{-1}\widetilde{W}\xi_m\xi'_m \qquad (2.7)$$

Equation (2.7) can be solved when the covariance matrix for each mixture component is diagonal. When Z compensates equation (2.7)'s left-hand side:

$$Z = \sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\sum_m^{-1}x_t\xi'_m \qquad (2.8)$$

Furthermore, defining the matrix $G^{(i)}$ whose (j,q)-th the element g_{iq}, is

$$g_{iq} = \sum_{m=1}^{m} v_{ij}^{(m)} d_{qj}^{(m)} \qquad (2.9)$$

Where v_{ij} is the (i,j)-th element of matrix V, d_{ij} is the (i,j)-th element of matrix D, V and D are

$$V^m = \sum_{m=1}^{M}\gamma_m(t)\sum_m^{-1}, \qquad (2.10)$$

$$D^m = \xi_m\xi'_m \qquad (2.11)$$

Using these equations, $\widetilde{W}$ is obtained as follows.

$$\widetilde{w}'_i = G^{(i)-1}z'_i, \qquad (2.12)$$

Where $\widetilde{w}_i$is the i-th column vector of $\widetilde{w}_m$, and z_i is the i-th column vector Z.

2.3.1.2 MAP adaptation technique

Maximum A Posteriori (MAP) adaptation is one of the well-known approaches used for automatic speech recognition. It is one of the approaches for statistical modeling

(Shinoda, 2011). Its estimation is known as Bayes estimation of the vector parameter, as loss function is not specified (DeGroot, 2005; Gauvain & Lee, 1994; Reynolds, Quatieri, & Dunn, 2000). MAP is valuable when dealing with problems occurring as a result of sparse training data for which Maximum Likelihood (ML) estimation produces inexact expected estimates, by providing consolidated prior knowledge in the training data (Gauvain & Lee, 1994).

MAP estimation is applied to two groups of applications stated as model adaptation and parameter smoothing, which is relevant to the same problem of parameter estimation for sparse training data (Gauvain & Lee, 1994). With regard to the amount of data, the MAP is more efficient as compared to the ML estimation approach when the amount of data is low. Whereas, as the amount of data increases, the estimation of the parameter for MAP and ML converges (Shinoda, 2011).

Let $f(x|\theta)$ denote the probability density function (pdf) of x variable, and the sample $\chi = \{x_1, \ldots\ldots, x_T\}$ denote the given set of T observation vectors. The parameters are to be estimated from θ (a random vector having values in the speakers' space) by using T samples of x with a probability density function (pdf). The parameters estimated in ML estimation are as follows:

$$\tilde{\theta} = \arg\max_{\theta} f(x|\theta) \qquad (2.13)$$

Where $\tilde{\theta}$ is the maximum likelihood estimator of θ. In MAP, θ is increased as more data samples are observed. The prior distribution is the parameter distribution before data observation. In this regards, let the prior distribution for θ be $g(\theta)$. The parameter's pdf, after χ, $g(\theta|\chi)$ is observed, can be formulated based on Bayes' Theorem as the following which is called a posterior distribution,

$$g(\theta|\chi) = \frac{f(\chi|\theta)g(\theta)}{\int f(\chi|\theta)g(\theta)d\theta} \qquad (2.14)$$

The maximum value of the posterior distribution is provided by MAP estimation. In other words, the maximum of a posterior distribution is the value where posterior distribution mode is maximized and can be expressed as $\tilde{\theta}$ and is computed as the following:

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmax}}\, g(\theta|\chi)$$

$$= \underset{\theta}{\operatorname{argmax}}\, f(\chi|\theta)g(\theta) \qquad (2.15)$$

According to the last equation (2.15), both ML and MAP performed almost similarly in the case of lack of knowledge about θ.

2.3.2 Dysarthric Speech Corpora

The lack of speech data available for dysarthric speakers is one of the major stumbling blocks in the development of dysarthric ASR systems. There are a few corpora used by researchers in which some are available for free and some of them are payable. This section will focus on some of the corpora used in developing a dysarthric ASR system.

2.3.2.1 Whitaker corpus

The Whitaker corpus is a collection of isolated word utterances spoken by six-person suffering from cerebral palsy. The total utterances of Whitaker corpus are 19,275 isolated words. The corpus contains utterances of one non-impaired speaker which is used as a reference. The isolated word is divided into two sets, the first set consists of alphabets, digits, and 10 control words with a total of 46 words (referring as "TI-46"). The second set consists of phonetically diverse words (referring as "Grandfather"). The corpus is

available for researchers to study different aspects of speech disorder (Deller, Liu, Ferrier, & Robichaud, 1993).

2.3.2.2 Nemours speech corpus

NEMOURS speech corpus comprises of 814 short nonsense sentences. 11 male speakers have spoken these sentences. 74 sentences are required to be spoken by each speaker. The form of the sentence is "The X is Ying the Z" where X ≠ Z (Menendez-Pidal, Polikoff, Peters, Leonzio, & Bunnell, 1996). Closed-set phonetic contrasts (e.g. voice, manner, and place) are provided by the constraints of the target words X, Y, and Z (Ray Kent et al., 1989). More details about this corpus will be explained in the next chapter.

2.3.2.3 UA-Speech Corpus

UA-Speech corpus is universal access to the database of audiovisual. It is publicly-available at the University of Illinois and is described in more detail in Kim et al. (2008). The dataset was recorded using a 7-channel microphone array. A wide range of impairment's severity, ranging from 2% to 95% word intelligibility was covered by using the data from 10 spastic dysarthric speakers. The corpus is an isolated word-level transcription. 765 isolated word utterances were read by each speaker. These included 100 common words in the Brown corpus of written English (e.g., it, is, you), 19 computer commands (e.g., backspace, delete), 300 uncommon words selected from children's novel, 26 radio alphabet letters (e.g., Alpha, Bravo), and 10 digits (zero to nine) (Kim et al., 2008).

2.3.2.4 Madison Mayo Clinic dysarthria database

The dysarthria database of Madison Mayo Clinic contains speech samples in digital form. It is recorded at the Mayo Clinic in Rochester, Minnesota (Kim et al., 2011). The

database is represented by 107 dysarthric subjects in total, who suffer from, Parkinson's disease (Males =29, Females = 10), stroke (Males = 21, Female=18), traumatic brain injury (Males =7, Females=5), and multiple system atrophy (Males=11, Females=6) (Kim et al., 2011). Six words were required to be uttered by every speaker (wax, ship, sip, sight, shoot, and hail) for 10 times. The criteria for the chosen words are the acoustic characteristics and vocalic nuclei of the words, which require obvious vocal tract change (Kent et al., 1989; Kim et al., 2009). The participants were asked to utter five sentences (The boiling tornado clouds moved swiftly, The potato stew is in the pot, The blue spot in on the key, Combine all the gradients in a large bowl, and Put the high stack of cards on the table). Acoustic and intelligibility data are derived by choosing these sentences (Kim et al., 2011). The speech samples were collected in a quiet room using a digital audiotape recorder (DAT; TASCAM DA-P1) with a high-quality microphone (SHURE SM 58) with 16-bit quantization at a sampling rate of 44.1 kHz. After recording, a TF32 program was used to analyze the speech (Kim et al., 2011).

2.3.2.5 TORGO speech corpus

Fifteen subjects, classified into 7 control and 8 dysarthric speakers (3 females and 5 males, aged from 16 to 50 years old) make up the TORGO speech corpus that has good coverage of intelligibility range. Participants suffering from Amyotrophic Lateral Sclerosis (ALS) and Cerebral Palsy (CP) are included in the speech corpus. For CP participants, some of the examples of their impairment are ataxic, athetoid, and spastic (Rudzicz et al., 2011). A speech-language pathologist based on the Frenchay Dysarthria Assessment diagnosed The TORGO speech corpus (Enderby, 1980a).

2.3.3 Dysarthria and Quality of Life

Damage of the neuromuscular systems that regulate speech is often accompanied by a variable rate of speech, imprecise articulation, disordered speech prosody, and excessive nasalization (Frederic Darley et al., 1969). Deficits in speech physiology may result in the deformation of the acoustic signal and reduced speech intelligibility. Difficulties with social interaction, vocational placement, and academic performance are mainly associated with deficits of intelligibility which reduces the quality aspect of life. Voice-operating software and voice-command assistive devices are speech technologies to help assist the Persons with Disabilities (PWDs) (Mark Hawley, 2002; Mark Hawley et al., 2006).

2.3.4 Related Work on Dysarthric Speakers

In a seminal work by Ferrier et al. (1995) for developing ASR system for dysarthric speaker, their test subjects consist of ten adults with spastic cerebral palsy with five males and five females as well as one male and one female nondisabled subject (Ferrier et al., 1995). The age of the disabled speakers ranged from 12 to 62 while for the nondisabled speakers, the age of males was 33, and 30 for the female. All subjects were not suffering from any hearing disability. One noticeable feature for the subjects is that they all had received five years or more of speech therapy (Ferrier et al., 1995).

For the experimental setting, they have used the realistic Highball 33-984c omnidirectional microphone and the TEAC W-450R cassette recorder was used to record testing samples (Ferrier et al., 1995). Using a 486 Gateway computer, the DragonDictate System was operated. For this system, the subject wore a standard headset microphone which was placed half-inch from the mouth (Ferrier et al., 1995). The subject was asked to speak a Grandfather passage where global intelligibility as measured by CAIDS was

assessed by three graduate students of speech-language pathology. Both transcription and multiple-choice formats were used for a single word, and transcription only format was used for sentences. Recognition levels were then compared to nondisabled speakers. The text used for recording is one of the most memorable texts that helped to produce the quality in the sound, which is the Pledge of Allegiance. Some of the subjects had visual tracking problem which caused the dictation process to be carried out at a slower rate. One subject had a reading problem, so the words in the choice list were read out by the researcher (Ferrier et al., 1995).

For all subjects, 200 msec was the level of pause set. There is some human intervention to dictate words to the DragonDictate system. To achieve high recognition accuracy, human intervention provides the speech recognizer with an accurate model of the user (Ferrier et al., 1995). The score given to the word recognized is according to one of the following choices: the word was correctly recognized, the word has been missed but the correct word is in the list of possible word recognition, or the word was wrongly recognized with no possible word recognition in the estimated list (Ferrier et al., 1995). Another score that has been recorded was for word fluency features, which included disfluencies, non-speech sound, and intra-word pauses.

Disfluency has been defined as any word, syllable, or sound repetition (Ferrier et al., 1995). Non-speech sounds are determined as noises, like laughing, coughing, smacking of the lip, or sounds associated with control of saliva (Ferrier et al., 1995). Intra-word includes discernible pauses between phonemes caused by poor articulation (phonatory control), slowed rate, or breathing during word production (Ferrier et al., 1995).

Ferrier et al. (1995) found that dysarthric speakers with high-intelligibility speech had more recognition accuracy as compared to low-intelligibility speech. One of the

variations affecting the low-intelligibility speaker is fatigue, which makes constructing a speech corpus to train a speech recognition system a laborious task (Ferrier et al., 1995; Kayasith & Theeramunkong, 2009). The study concluded that if there were more dysarthric voice and fluency features like non-speech sound and intra-word pauses, achieving 80% recognition takes a longer time. The decoding done by human listeners is a complex task which uses structured knowledge of words, besides the acoustic and linguistic knowledge (Ferrier et al., 1995). Human listeners automatically normalize speech patterns that are different, like dysarthric speech or nonstandard dialects (Ferrier et al., 1995).

The corpus used in Kayasith & Theeramunkong (2009) consists of 67 words of the Thai language. The corpus was phonetically balanced which was set by a speech therapist at Siriraj Hospital. The children who pronounced the words suffered from literacy limitations, therefore, every word was accompanied by pictures. Recording of 67 words was done in a controlled environment (quiet room with the door closed). Every word was recorded five times and used for evaluation (Kayasith & Theeramunkong, 2009). The speech stimuli used in this experiment is the picture represents by the targeted word and displayed on a computer screen. In case of faulty pronunciation or if the speaker has difficulty in pronunciation, the system will provide an example of the word pronunciation (Kayasith & Theeramunkong, 2009).

There were two groups of the subject, the dysarthric (cerebral palsy) and controlled speakers in equal numbers with regards to its proportion and gender (Kayasith & Theeramunkong, 2009). For system development in Kayasith & Theeramunkong (2009) the HTK toolkits were used in developing the speech recognition system. Also used were the Neural Interface Computation (NICO) toolkit to develop the Artificial Neural

Networks (ANN) speech recognition system. By using a random initial weight of between -1.0 and 1.0 for the standard back-propagation method, the network was trained.

Three different measures were used for the evaluation. They are the *root-mean-square of difference* (Δ_{rms}), the *Pearson's correlation coefficient* (R^2) between Φ (also the I- and A-score) and the recognition rate (of ANN or HMM), and the *Rank-Order Inconsistency* (ROI). Thus, to predict recognition performance, a function is generated. Recognition performance from Φ (also the I- and the A-score) is predicted using these functions. The calculation of Δ_{rms} is done using the actual and predicted recognition rates (from ANN or HMM). By considering the differences between the performances of ANN and HMM, the margin for the different results of each measure was calculated (an acceptable bound). This margin is then used to determine if the difference between the actual and the predicted recognition rates from Φ (also the I- and the A-score) is acceptable or not.

According to Kayasith & Theeramunkong (2009), the initial and cluster consonants are the toughest to pronounce among the five classifications, which are tone, vowel, initial, final, and cluster consonants. Phoneme levels had high confusion from the signal of the dysarthric speakers, as shown by the experiments.

An articulation test was used to study Thai stroke patients' dysarthric speech characteristics (Manochiopinig et al., 2007; Manochiopinig et al., 2008). The characteristics of speech comprise of tones, vowels, initial, final and cluster consonants (Manochiopinig et al., 2007; Manochiopinig et al., 2008). There were 14 subjects who suffered from a stroke and with speech dysarthria, divided into 5 females and 9 males (Manochiopinig et al., 2007; Manochiopinig et al., 2008). The types of dysarthria used in this experiment are flaccid (11 cases) and spastic (3 cases) (Manochiopinig et al., 2007; Manochiopinig et al., 2008). There are 68 words used to assess the characteristics of

speech for dysarthric speakers. Each word is of a single-syllable level and covers all the phonemes of the Thai language (Manochiopinig et al., 2007; Manochiopinig et al., 2008). The stimuli used to encourage the participants to pronounce the words is in picture form. The results showed that the highest recognition rate was obtained from the vowels and tone characteristics (Manochiopinig et al., 2007; Manochiopinig et al., 2008).

Three classifiers were used to classify prosody, voice quality, and pronunciation features, all of which are unweighted. They are the average recall of Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) (Kim et al., 2015). The results have been conducted on 2 of every feature type separately, and three features combined together. The pronunciation features with the SVM classifier with 3rd-order polynomial kernel function had the best performance for intelligibility classification as compared with LDA and KNN. Feature-level fusion combines all types of features and obtained the highest level of accuracy (Kim et al., 2015). For the TORGO dataset, the classification results for the feature selection showed that the LDA classifier had the best performance as compared to the SVM classifier, while KNN is not reported because of the instability of the results data (Kim et al., 2015).

The number of subjects used in Van Nuffelen et al (2009) was 211 speakers, divided into 60 speakers with dysarthria, 42 with pathological speech secondary to hearing impairment, 12 children with cleft lip, 7 diagnosed with dysphonia, 37 with a laryngectomy, 2 with glossectomy, and lastly 51 control speakers. They recorded 10,550 consonant-vowel-consonant words taken from the Dutch Intelligibility Assessment (DIA) (Van Nuffelen et al., 2009). The percentages of the severity type of dysarthric speakers are 53%, 39% and 8% for mild, moderate and severe severity respectively.

The training for developing both phonemic and phonological feature models included both the pathological and control speakers. The intelligibility scores for dysarthric speakers were compared with the perceptual intelligibility measured in which the high scores were recorded for every speaker after five repetitions (Van Nuffelen et al., 2009). The five-fold cross-validation experiment with some restriction with regards to the number of features in the five-fold has been applied to simplify the model (Van Nuffelen et al., 2009).

The results showed in Van Nuffelen et al (2009) confirmed that the combined phonemic and phonological features model for dysarthric speakers has the highest correlation between the computed intelligibility scores (objective intelligibility) and perceptual intelligibility scores. The study concluded that to overcome the large deviations between perceptual and computed scores of intelligibility, the severe dysarthric intelligibility speakers, which had fewer participations in the training should be increased to be used for training of the model (Van Nuffelen et al., 2009). The features group, like vowel-related phonemic and phonological features, lateral-, silence-, fricative-, velar-, and plosive-related features are observed to be useful for the clinical point of view. One more observation from this study is that vowel-related is the most important feature that contribute to the dysarthric speech's intelligibility.

Table 2.5 listed the studies related to the automatic speech recognition system for dysarthric speakers. The adaptation techniques, type of speech, model type, data set, and results obtained are the criteria used to summaries the studies in ASR for dysarthric speakers.

Table 2.5: Automatic Speech Recognition System for Dysarthric Speakers

Author(s)	Adaptation techniques	Type of speech	Model Type	Data Set	Results	Note
Dhanalakshmi & Vijayalakshmi (2015)	adaptation CMLLR+MAP	Continuous	Word Model speaker independent	NEMOURS Corpus(10 Males) as for recognition system CMU Arctic database for synthesis system	Both speech recognition and speech synthesis can be used to assist the intelligibility of dysarthric speakers as well as the speech rate affected speech intelligibility	used speech recognition and speech synthesis technique to improve intelligibility
Doyle et al. (1997)	-	Isolated words	speaker-independent Model	6 subjects dysarthric speakers(3 male and 3 female) and 6 subjects as normal speakers	Commercial ASR system can help in predicting the intelligibility of the dysarthric speakers (more practice give more accurate results) while for perceptual recognition it is steady	compare the results of the commercial IBM Voice type with the nonbearing impaired adult listeners regarding the intelligibility
Sharma et al (2009)	no adaptation	Isolated words	Speaker dependent model	7 subjects from UA-Speech database	In word level with the small size of vocabulary the recognition accuracy is higher than their respective intelligibility ratings while for medium-size one, monophone and triphones are less	The research concerns about the effectiveness of the speaker-dependent model for dysarthric speakers for small vocabulary size (recognize word level and phone level)

Author(s)	Adaptation techniques	Type of speech	Model Type	Data Set	Results	Note
Green et al. (2003)	No adaptation	Isolate words	Speaker dependent model	8 subject as paper wrote	The goal is to allow these clients to control assistive technology by voice	small vocabulary, speaker-dependent, isolated-word application, the speech material more variable than normal, and only a small amount of data is available for training
Ferrier et al. (1995)	No adaptation	Continuous(read Grandfath-er passage)	Speaker independent model	10 dysarthric speakers	90% recognition within eight dictation sessions for people with good intelligibility	Examined speech recognition accuracy using Dragon Dictate for adults with cerebral palsy compared with control subjects. people with good intelligibility were more successful at using speech recognition
Kayasith & Theeramun kong (2009)	No adaptation	Isolated words	Speaker dependent model	7 dysarthric subjects and 8 control subjects	HMM reference, Φ achieved low rank-order inconsistency of 18%, compared to 36% for the articulatory test and 25% for the intelligibility test. ANN reference, Φ had a low inconsistency of 7% while the articulatory test and the intelligibility test gained high inconsistency of 54% and 43%, respectively	both Hidden Markov Model (HMM) and Artificial Neural Network (ANN) speech recognition system was developed to recognize words to predict the speech recognition rate for the dysarthric speaker with the perceptual calculated A-score and I-Score

Author(s)	Adaptation techniques	Type of speech	Model Type	Data Set	Results	Note
Mark Hawley (2002)	No adaptation	Isolated words	Speaker independent model	1 subjects	46% of recognition rate within 14 sessions, and 64% recognition rate for correct words in the choice list of 5 words	study the effects of using Dragon Dictate ASR system in the improvement of recognition rate after a certain session of repeated words
Mark Hawley et al. (2006)	No adaptation	Isolated words	Speaker dependent	The data build for every subject willing to participate in the system	95% of word recognition rate in the test conditions and 87% in every usage in the uncontrolled noise condition	Development of a voice-input voice-output communication aid (VIVOCA) for people with disordered or unintelligible speech (Severe dysarthric speakers)- use both ASR and TTS systems
Mark Hawley et al. (2007)	No adaptation	Isolated words(command set)	Speaker dependent	The data build for every subject willing to participate in the system(17 subjects participated for evaluation of the system)	88.5 % of accuracy for pre-training stage (before user training) and 95.4 of accuracy after user training. 86.9% of accuracy for word recognition when subjects used the system at home	This paper studies the effectiveness of the user training stage for improving the performance of the recognition accuracy (user training provides more training data for ASR system).
Kayasith, et al (2006a)	No adaptation	Isolated words	speaker-independent Model	16 dysarthric subjects and 8 control subjects	The root means square error between the prediction rates and recognition rates is less than 7.0% and 2.5% for HMM and ANN ASR respectively.	introduce new indicator called speech consistency score (SCS) for dysarthric speech recognition rate prediction

Author(s)	Adaptation techniques	Type of speech	Model Type	Data Set	Results	Note
Kayasith, et al (2006b)	No adaptation	Isolated words	Speaker independent Model	16 dysarthric subjects and 8 control subjects	The system an average improvement of 9.56% and 7.86% of prediction for both articulatory and intelligibility tests respectively	introduce new indicator called speech confusion index (Ø) for dysarthric speech recognition rate prediction
Dibazar et al (2006)	MAP Adaptation	Phoneme level	Speaker dependent	657 impaired speech subjects and 53 control subjects	Above 76% of recognition using training samples with multiple labels	This paper focuses on the recognition of five specific pathologies
Maier et al. (2009)	No adaptation	Continuous speech (read the passage "The North Wind and the Sun")	Speaker independent Model	90 subjects with head and neck cancer	50% of the mean of the recognition rate for the laryngectomees (LE) and 48% mean of word accuracy for oral cancer(OC)	Study the effects of speech recognition on the objectify and quantify the most important aspect of pathologic speech (the intelligibility)
Middag et al (2009b)	No adaptation	Isolated words	Speaker independent Model	-	90% using Person correlation coefficient between mean professional listeners' scores and the objective scores	Using ASR as a tool for objective intelligibility assessment for pathological speech
Bhat et al (2016)	feature space MLLR (fMLLR) based speaker adaptive training (SAT) adaptation	Isolated word	Speaker independent and speaker adaptive model	Universal Access Speech Corpus (UA-Speech) 13 subjects control speakers and 15 subjects of dysarthric speakers	The proposed speaker-independent and speaker-adapted systems provide an improvement of 47.11% and 55.81% respectively, for GMM-HMM-TA and 48.44% and 63.67% for DNN-HMM-TA respectively	Using the severity level based tempo adaptation of sonorants (vowels, glides, liquids, and nasals) in dysarthric speakers to improve the speech recognition for dysarthric speakers. Two models developed which are Gaussian Mixture

Author(s)	Adaptation techniques	Type of speech	Model Type	Data Set	Results	Note
						Model (GMM)- HMM and Deep Neural Network (DNN)-HMM.
Sriranjani, Ramasubba Reddy, & Umesh (2015)	MLLR and feature space MLLR (fMLLR) based speaker adaptive training SAT	Continuous speech	Speaker Independent model	Nemours Corpus and UA-Speech corpus for dysarthric subjects and Wall Street Journal (WSJ0) corpus and TI digits for control subjects	improvement of 18.09% and 50.00% over baseline system for Nemours database and Universal Access speech (digit set) database respectively	Used the unimpaired speech data to pooled the SI acoustic model with the adaptation data of the dysarthric speech to improve the recognition accuracy of dysarthric speakers
Mathew et al (2018)	No adaptation	Isolated word	Speaker Independent model	TORGO database	Word level accuracy of 63.27% with PLP Feature set and 61.69% with MFCC Feature set	Comparing the feature extraction types to recognizing the word level of Dysarthric speech with the human listener assessment
Joy & Umesh (2018)	No Adaptation	Continuous speech	Speaker Independent model	TORGO database	The results show significant improvements over previous attempts at building ASR systems for TORGO for sever and sever-moderate dysarthric speech	Explored multiple ways to improve The recognition accuracies of GMM-HMM and DNN-HMM acoustic models for the TORGO dysarthric speech database.
Zaidi et al. (2019)	No Adaptation	Continuous speech	Speaker Independent model	Nemours Corpus	54.78 % using MFCC's, JITTER and SHIMMER 52.41 % using PNCC's, JITTER and SHIMMER	Improve the recognition accuracy of ASR for dysarthric speakers using the Power Normalized Cepstral

Author(s)	Adaptation techniques	Type of speech	Model Type	Data Set	Results	Note
						Coefficients (PNCC) and FMCC feature extraction in concatenation with several variances of JITTER and SHIMMER

2.3.5 Automatic Dysarthric Speech Recognition Summary

After discussing and reviewing several topics related to the ADSR for dysarthric speech which is summarized in Table 2.5. Some of the key points are discussed below

- Adaptation techniques (MAP, MLLR , fMLLR, SAT, CMLLR).
- Acoustic model (SD, SI, SA).
- Type of speech (isolated words (mostly), phoneme, continuous speech (rarely)).
- Corpora used (NEMOURS, TORGO, UA-Speech).
- SA model was applied to the commercial ASR, like DragonDictate.

2.4 Findings of the Literature Review

As discussed earlier in this chapter, the existing classification algorithms used in the classification of dysarthric speech are based on type and disease of dysarthria (Kim et al., 2011). However, the use of the severity level classification is not thoroughly investigated in dysarthria of speech classification and recognition despite some research that focuses on Spastic Disorder Classification (Paja & Falk, 2012). From literature review, the common feature selection methods are the forward selection procedures and the backward selection procedures which are time-consuming and need a predefined justification to obtain the desired feature selection, and to accomplish the highest classification accuracy with only few features (Kim et al., 2015; Middag et al., 2009). The common practice is to select one feature selection method to rank the features according to their relevance to the type of dysarthric speech severity level, which can be achieved using the feature ranking methods. In fact, using more than one feature ranking method is likely to produce more different ranking sets, and presenting only one set given by a particular method can be misleading (Kuncheva, 2007). Most algorithms used for dysarthric speech are with the small dimensions of features. However, the large dimension of acoustic features is not

fully investigated with different feature ranking methods and classification algorithms. Furthermore, using the speech tools that used to obtain a set of features which are required many adjustment and settings, and in some cases obtaining features from different speech tools containing only a small dimension of features that results in difficulties in the use of the classification algorithms.

In the Automatic Dysarthric Speech Recognition system, the accuracy of recognition depends on the acoustic model, which includes speaker-independent, speaker-dependent, and speaker adaptation (Hamidi et al., 2010; Shinoda, 2011) with speaker adaptation being the subset of the speaker-independent model.

To address the above-mentioned issue of the existing classification and ASR for dysarthric speakers, the large dimensional acoustic feature is extracted using the openSMILE tool (Eyben et al., 2013) which is based on, the severity level of dysarthric speech. Feature ranking methods were used to rank the feature based on their relation to the severity level of dysarthric speech, which is mild, moderate, and severe. The feature ranking methods are fast and do not suffer from some limitations, such as classifier-dependency, or the lack of interpretability (Saeys, Abeel, & Van de Peer, 2008; Santana & de Paula Canuto, 2014). When using the severity level of dysarthric speech as the classification base, the performance of the different classification algorithms are compared by applying several classification techniques.

In the ADSR, the acoustic model enriched with speech data from the normal speakers and the adaptation data were used to improve the recognition accuracy of dysarthric speech as in (Mustafa, Salim, Mohamed, Al-Qatab, & Siong, 2014). The combination of some adaptation techniques results in the improvement of the recognition accuracy of the

ADSR (Sriranjani, Ramasubba Reddy, & Umesh, 2015). It is applied in this study based on the severity level of dysarthric speech, to improve recognition accuracy.

2.5 Classification Algorithms, Feature and Adaptation Identification for Intra-Severity ADSR

In this section, the identification of classification algorithms and features for the proposed intra-severity classification and adaptation ADSR are described. Section 2.5.1 and section 2.5.2 discusses the classification algorithms and features for intra-severity ADSR in fulfilling the first objective of this research. In Section 2.5.3, the adaptation techniques applying for intra-severity ADSR are described to achieve the second objective of this research.

2.5.1 Classification Algorithms Identification

The classification algorithms used in the previous works are Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Discriminant Analysis (ALLOC80), Gaussian Mixture Model (GMM), Statistical Analysis (Discriminant function analysis, linear regression analysis), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Quadratic Discriminant Analysis (QDA), and Mahalanobis Discriminant Analysis (MDA). Furthermore, those algorithms are used in the classification of dysarthric speech to identify the neurologic diseases, dysarthria type, the intelligibility of the speech, and differentiation between non-impaired speakers and dysarthric speakers.

This research used machine learning algorithms rather than deep learning algorithms. The reasons for using machine learning are (1) machine learning almost always require structured data, whereas deep learning networks rely on layers of the ANN (artificial neural networks) (Deng & Yu, 2014; LeCun, Bengio, & Hinton, 2015). In this research, the classification algorithms used to identify the severity level of dysarthric speakers that

are categorized into three levels of severity level which is mild, moderate and severe severity level. (2) Machine learning algorithms are built to "learn" to do things by understanding labeled data, then use it to produce further outputs with more sets of data. However, they need to be retrained through human intervention when the actual output isn't the desired one (Deng & Yu, 2014; LeCun et al., 2015). (3) The main concern of this research is to improve the recognition accuracy for ADSR by applying adaptation techniques wither standalone or a combination of standalone adaptation techniques; thus, the classification techniques used to automatically identify the suitable adaptive acoustic model. (4) The huge number of acoustic features extraction and feature selection also investigate to enhance the classification accuracy as the feature selection method used to rank the most relevant acoustic features for each severity level of dysarthric speakers.

For classification algorithms, the common algorithms used for dysarthric speech are the small dimension of features. In this research, algorithms used in include SVM, LDA, and ANN as well as some of the well-known algorithms used for comparison with algorithms used for previous research like Classification and Regression Tree (CART), Naive Bayes (NB), and Random Forest (RF) which describes as the followings:

2.5.1.1 Linear Discriminant Analysis (LDA)

In 1936, Fisher originally developed the Linear Discriminant Analysis (LDA), which is a classification method and has been used effectively in a wide variety of problems.

In statistical pattern classification, LDA is a well-known technique to compress the information contents (with respect to classification) and to improve the discrimination of a feature vector by a linear transformation. Improvement in the recognition performance for small-vocabulary systems is the result of supplying LDA to automatic

speech recognition (Haeb-Umbach & Ney, 1992). In addition to this, LDA easily handles cases of unequal within-class frequencies with performances examined on randomly generated test data. Maximal separability is guaranteed as the ratio of between-class to within-class variance in any particular data set is maximized by this method. The distribution of the feature data is better understood with the help of this method (Balakrishnama & Ganapathiraju, 1998).

LDA's main objective is to separate data samples into distinct groups called classes. LDA transforms the data into a different space, usually with a dimension that is lower, maximizing the between-class separability while minimizing variability of within-class. Optimal distinguishing between the classes is by this transformation called the feature projection. In any particular data set, the ratio of between-class to within-class variance is maximized by using the LDA method. Therefore, maximum separability is guaranteed (McLachlan, 2004).

LDA is commonly used in learning problems by the machine, like data dimensionality reduction, pattern and face recognition, and feature extraction. It is a simple and mathematically robust method which usually generates models whose accuracy is similar to complicated methods (Guerreiro, 2008).

2.5.1.2 Classification and Regression Tree (CART)

Breiman et al., (1984), developed Classification and Regression Trees (CART), a method of classification which uses past data to build the decision tree classification model and then use it to classify new data sample.

Prediction or classification of cases is permitted by tree-building algorithms (set of if-then (split) conditions), used by Classification and Regression Tree (CART) models.

The regression-type model refers to a CART model that predicts the value of continuous variables from a set of continuous and/or categorical predictor variables. The classification-type CART model is used for the prediction of the value of the categorical variables from a set of continuous and/or categorical predictor variables. Like CART, one noticeable advantage of decision tree-based models is that they can handle smaller data, though scalable to large problems (Balakrishnama & Ganapathiraju, 1998).

CART training comprises of four processes. The first step is making the decision tree, where the recursive splitting of nodes is used. Based on the distribution of classes in the dataset, each derived node is assigned to a predicted class. At the second step, "maximal tree" is produced and building the decision tree is stopped. The produced tree is often large that it probably over-fits the information from the learning samples. The third step is tree "pruning," which is a sequence of making simpler trees by amputation of increasingly important nodes. The last is the optimum tree selection. At this stage, only the tree that fits the information in the learning dataset is selected from the pruned trees (Roger & Lewis, 2000).

2.5.1.3 Artificial Neural Network (ANN)

Artificial neural networks (Teuvo Kohonen, 1982) or neural networks are usually considered as a simulation of the nervous system's information-processing. By studying the system of neurons and learning rules derived from biological models, early work in this field was inspired (Depenau, 1995).

There are two different kinds of neural networks. Feed-forward network with a simple perceptron and its extension, and the multi-layer perceptron which is one of the most commonly used neural networks for classification. Neurons or nodes are interconnected computing units that make up all neural networks. Inputs from different

units in the network, or from the outside world are received by each unit. Output based on these inputs is calculated. Units are organized into layers in the feed-forward network. It is made up of L processing layers, with the first layer (1=0) being the input layer, and the last layer (1=L) being the output layer. Through these two layers, all communications with the outside world are done. Hidden layers are the intermediate layers of units which cannot be reached from the outside, and hidden units are the units within them. The layer closest to input is the first hidden layer and the one closest to the output is the last hidden layer. A simple perceptron is what the network is called if there are not any hidden layers. On the other hand, it is called a multi-layer perceptron, or X-layer perceptron, where X is the number of hidden layers sum by 1. In a feed-forward network, starting from the input, passing of information is from a lower layer toward a higher layer and not the reverse of that. This indicates that lower layer units may only be connected to higher layer units, without allowing feedback or interconnections (Depenau, 1995).

2.5.1.4 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a powerful algorithm learning machine originating from the theory of statistical learning, first introduced by (Vapnik, 1998). It has been used successfully in a wide variety of problems like face detection, malware detection, handwriting recognition, and many others (Witten et al., 2016). This method is also popular because of its high level of generalizability and its capability in handling high dimensional input data relative to neural networks and decision trees (Theodoridis & Koutroumbas, 2006).

SVM can employ a small training set for creating generalizable nonlinear classifiers which are the main advantages of this classifier in high-dimensional feature space. In the case of having large training sets, SVM chooses a small set of support vector

that are required for designing the classifier. It can significantly decrease the computational cost of testing (Jain et al., 1999). Because of the above-mentioned advantages of SVM, it is one of the most popular classification techniques ineffective computing. SVM classifiers propose competitive performance results for emotion recognition compared to other classification techniques.

2.5.1.5 Naive Bayes (NB)

For machine learning and data mining, one of the most efficient and effective inductive learning algorithms is Naive Bayes (NB) (Zhang, 2004).

2.5.1.6 Random Forest (RF)

An ensemble of the machine learning algorithm, Random Forest (RF) is best defined as a "combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" (Lebedev et al., 2014). To date, this algorithm produces one of the best classification accuracies in many applications. Compared to other techniques, it has important advantages in terms of opportunity for efficient parallel processing, tuning simplicity, robustness to noise, and the ability to handle highly non-linear biological data. For handling of high-dimensional problems, with often redundant number of features, RF is the ideal candidate due to these factors. Several approaches for feature set reduction within and outside the context of RF have been proposed, although RF can itself be considered as an effective feature selection algorithm, to further improve its performance (Tuv et al., 2009).

2.5.2 Acoustic Features Identification

There are many features extracted from dysarthric speech signal and used for the classification of dysarthric speech including the following: A typical vocal source

excitation, prosodic features, nasality features, phonemic features, pathological features, fundamental frequency, MFCC, articulation features, voiceless interval duration, rhythm features, voice quality features, acoustic features, phonological features, log Harmonics-to-Noise Ratio (HNR), and Glottal-to-Noise Excitation ratio (GNE) groups based on the acoustic features. (Eyben, 2015) defined the acoustic features to be used for real-time speech and music analysis. In this research, some of the features introduced in (Eyben, 2015) were used, which is divided into four groups based on the acoustic features. These are the prosodic, voice quality, spectral and cepstral groups. These features have been used in the classification of the dysarthric speech, to differentiate between non-impaired speakers and dysarthric speakers, as in (Godino-Llorente et al., 2006; Schlenck et al., 1993), to differentiate based on type of dysarthria as in (Fonville et al., 2008; Guerra & Lovey, 2003; Liss et al., 2009), to differentiate based on severity of dysarthric speech (Kim et al., 2011; Paja & Falk, 2012), or to measure the intelligibility of dysarthric speakers (Kim et al., 2015; Middag et al., 2011). For each feature, there are parameters computed for a short time frame of an audio signal at a given time, called the acoustic Low Level Descriptors (LLD) (Eyben, 2015; Schuller, 2013), more details are given in chapter 3 section 3.2.1.2.

2.5.3 Adaptation Techniques Identification for Intra-Severity ADSR

In (Al-Qatab, Mustafa, & Salim, 2014; Mustafa et al., 2014), the acoustic model enriched with speech data from the normal speakers and the adaptation data were used to improve the recognition accuracy of dysarthric speech. The combination of some adaptation techniques results in the improvement of the recognition accuracy of the ADSR (Sriranjani et al., 2015) which is applied in this study based on the severity level of dysarthric speech, to improve its recognition accuracy.

Two of the well-known adaptation techniques for the ASR system are the maximum a posterior (MAP) (Gauvain & Lee, 1994) and the parameter transformation-based adaptation using the maximum likelihood linear regression (MLLR) (Leggetter & Woodland, 1995). These techniques have been proven to be effective for developing the ASR system with data sparsity for dysarthric speech. As a transformation based approach, MLLR has no further improvement at a certain point although there is more adaptation data available (Shinoda, 2011). MLLR usually requires recorded speech of a new speaker with the use of the same text or sentences recorded from the reference speaker, which is referred to as text-dependent (Digalakis & Neumeyer, 1996). On the other hand, MAP is more efficient as compared to the ML estimation technique when the data size is small. However, as the size of the data increases, the estimation of the parameter for MAP and ML is converging towards an equilibrium point (Kotler & Thomas-Stonell, 1997).

The adaptation techniques used for ADSR are MLLR, MAP, fMLLR, SAT, and CMLLR. This study has sequentially applied MLLR and MAP and vice versa. In the first hybrid approach (MLLR+MAP), the overall regression classes were estimated using the global regression classes, which is then switched to MAP adaptation using the prior transformation regression classes as an input to update each phoneme in the acoustic model updated firstly by MLLR. On the other hand, the MAP+MLLR hybrid approach estimates the parameter and updates each Gaussian of all phonemes in the acoustic model, then switches to MLLR adaptation using the updated acoustic model as an input to construct estimated global regression classes.

2.6 Summary

This chapter describes the literature review of the current research. The classification of dysarthric speech has been investigated. The summary of the findings from the related works listed in Table 2.4 and discussed in section 2.2.7. The automatic dysarthric speech

recognition described in this chapter which includes the adaptation techniques, dysarthric speech corpora, and the acoustic model used which is summaries in section 2.3.5. The finding from the literature review presented in this chapter as well as the identification of the classification algorithms, feature selection method, and adaptation techniques.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Overview

This chapter is devoted to the discussion of the methodology carried out throughout this research. The research methodology adopted in this research is designed based on the patterns of the design science (DS), to accomplish the research objectives. The design science research methodology process is described in section 3.2. The following sections described each process of the DSRM in more detail. The proposed Intra-Severity ADSR architecture, the speech corpus, acoustic features, features extraction tools and techniques, selection of feature methods, and classification algorithms used are described in section 3.5 of this chapter. Section 3.6 describes the data analysis carried out in chapter 4 as well as measuring the performance of the classification and ADSR for the proposed Intra-Severity ADSR. Section 3.8 summarizes this chapter.

3.2 The Design Science Research Methodology Process

The structure of this research is adopted from the design science research methodology (DSRM) process model that is normally used for an information system (IS) proposed by (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). DSRM is carried out in this research to meet three objectives, which are: (1) it is consistent with previous literature, (2) it provides a nominal process model to conduct the design science (DS) research, and (3) it provides a mental model to present and evaluate DS research in IS. The characteristics of this research outcome are considered as an artifact to be delivered as one of the main aims of the DSRM. In this research, the artifact is an intra-severity classification and adaptation technique to improve dysarthric speech recognition accuracy.

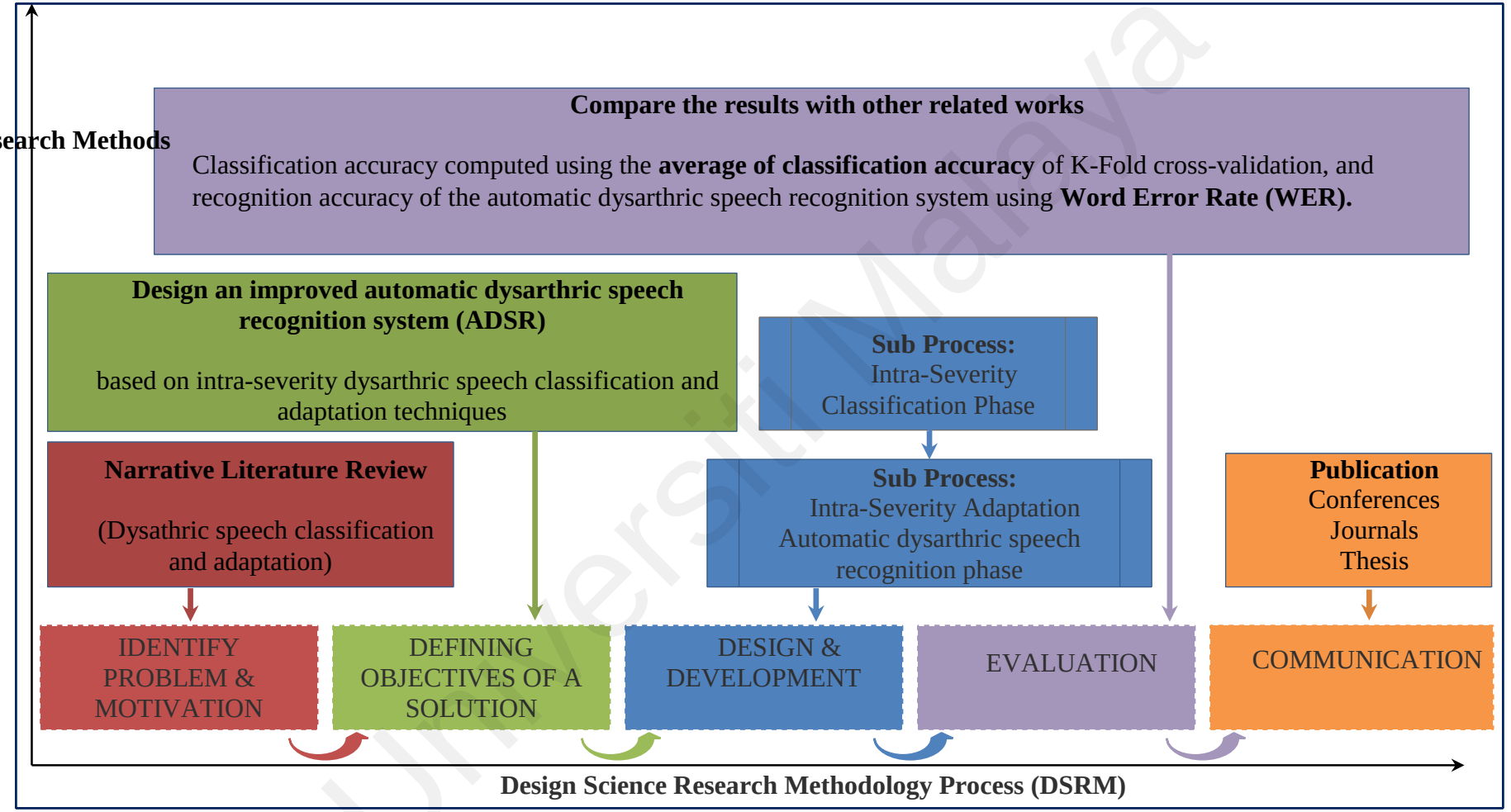

Figure 3.1: Adopted DSRM process model (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007)

Figure 3.1 depicted the DSRM adopted for this research which includes the five activities. These activities are described in the following sections and an adequate description of each activity is explained throughout the thesis.

3.3 Problem Identification and Motivation

A narrative literature review is one of the most essential steps in every research. It helps to define the problems that require seeking solutions. Chapter 2 which includes the literature review described the state of the science related to both, the classification of the dysarthric speech, and the automatic speech recognition for dysarthric speakers for the contextual point of view. Conduction of the literature review or narrative literature review helps in two aspects. First is to discover the potential problem/s in the classification and automatic speech recognition for the dysarthric speaker, and second is to propose a solution for the identified problem (/s) which will help in the automatic classification and automatic speech recognition for the dysarthric speakers. Thus, the problem statement and the questions for this research are presented in the first chapter which includes also the motivation of this research.

By maintaining a solution's value to the identified problem, two things are achieved. First, the researcher and the targeted readers of the research are motivated to seek a solution and be satisfied with the results. Second, it settles the argumentation with the understanding of the problem by the researcher.

3.4 Defining the Objectives for a Solution

This research proposed a solution for the design and improvement of the classification and automatic recognition accuracy for dysarthric speakers using the severity based classification and adaptation. In chapter 2 section 2.4 and section 2.5 are included the

identification of the classification algorithms, feature selection and adaptation techniques used as a solution to improve the recognition accuracy of the ADSR system.

This solution has the ability to enhance the accuracy of the classification and automatic speech recognition for dysarthric speakers by addressing the problems identified in the existing systems.

3.5 Design and Development of the Proposed Intra-Severity Automatic Dysarthric Speech Recognition

In this research, the databases' (Corpora) are identified to include the dysarthric severity level with an equal number of speakers. The features are extracted from each dysarthric severity level to perform the severity based classification of dysarthric speech. The classification of dysarthric speech-based severity levels obtained using a well-known classification method using the feature selection method to reduce the computational cost of the system. The adaptation techniques are also performed for automatic dysarthric speech recognition systems based on the severity levels of dysarthric speech.

The design of the proposed intra-severity automatic dysarthric speech recognition has three phases. Figure 3.2 shows the overall design used in this research. The three phases are classification, adaptation, and development of automatic recognition for the unknown speech. In the first phase, the automatic classification of dysarthric speech based on severity level is applied which helps the system to automatically know the adaptation model to apply. The adaptation models, which is the second phase of the proposed ADSR, include three models similar to the severity level of dysarthric speakers which are Mild, Moderate, and Severe adaptation models. The third phase and the final phase is the automatic dysarthric speech recognition to automatically predict the unknown text based on the speech file.

The proposed Intra-Severity ADSR uses the severity level of dysarthric speech to classify the dysarthric speech into three severity levels which are Mild, Moderate, and Severe levels. Furthermore, the adaptation models created using the adaptation techniques for the proposed Intra-Severity ADSR also used the severity levels which are Mild, Moderate, and Sever severity levels.

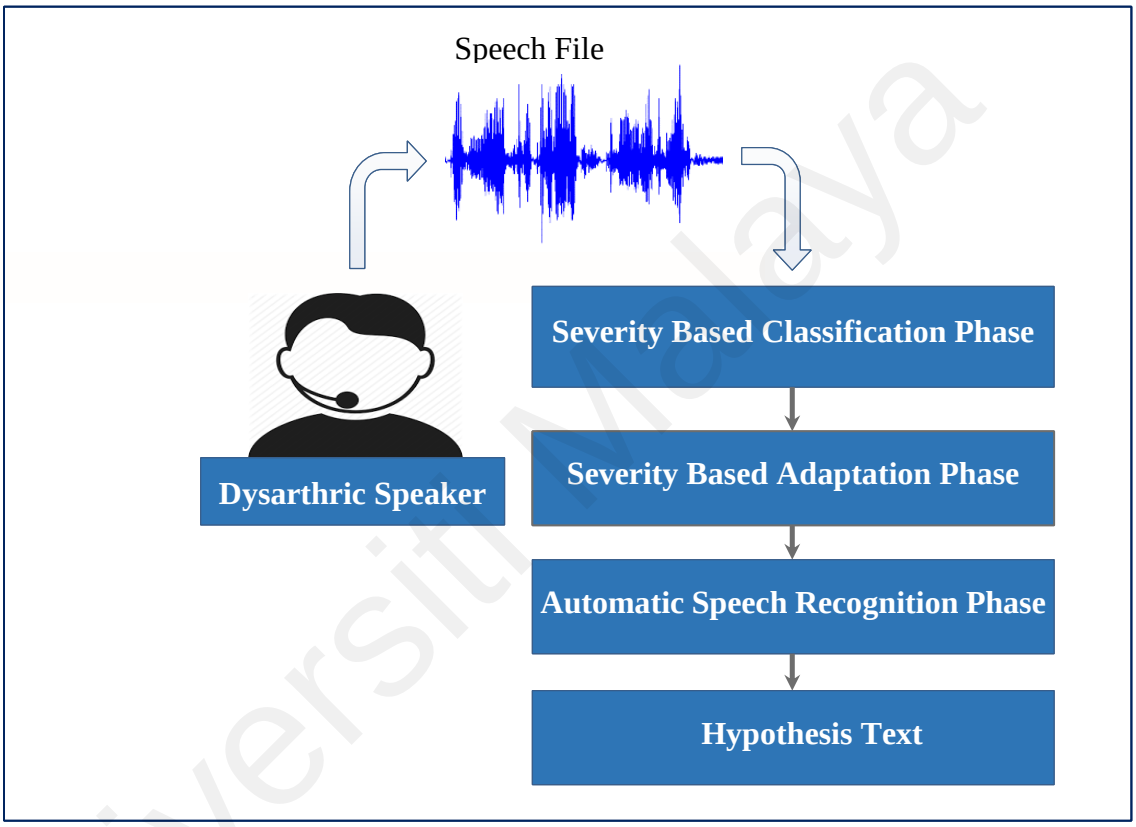

Figure 3.2: Overall architecture of the proposed Intra-Severity automatic dysarthric speech recognition

3.5.1 Classification Phase

This phase includes several steps, which include identification of database, feature extraction, feature selection, and classifier algorithms, as shown in Figure 3.3.

3.5.1.1 Speech Corpus

As the main focus of the classification phase of this experiment is to predict the correct severity level of dysarthric speech, the corpus should meet these two criteria. First, the

speech is continuous rather than isolated words, and second, the corpus includes all levels of severity with an equal number of participants for each severity level. The NEMOURS database (Menendez-Pidal, Polikoff, Peters, Leonzio, & Bunnell, 1996) is the dysarthric speech database that meets these criteria and was selected as the database for training. To maintain equal participants per severity level, two speakers were excluded from this experiment as data from one speaker was missing, and so another speaker was left out to maintain that balance.

As for the TORGO dysarthric speech database (Frank Rudzicz et al., 2012), severity level distribution is not equal among all participants, which include two subjects who are mildly dysarthric, one subject who is moderately dysarthric, one subject who is moderate-to-severely dysarthric, and four subjects who are severely dysarthric (Mengistu & Rudzicz, 2011). The Universal Access Speech Corpus (UA-Speech) (Kim et al., 2008) is also one of the open-access corpora for dysarthric speakers that collect the data from participants in a word-level speech consisting of 765 isolated words per speaker.

The NEMOURS speech corpus is used to extract the feature vectors for the feature selection stage, and development and testing of the classifiers.

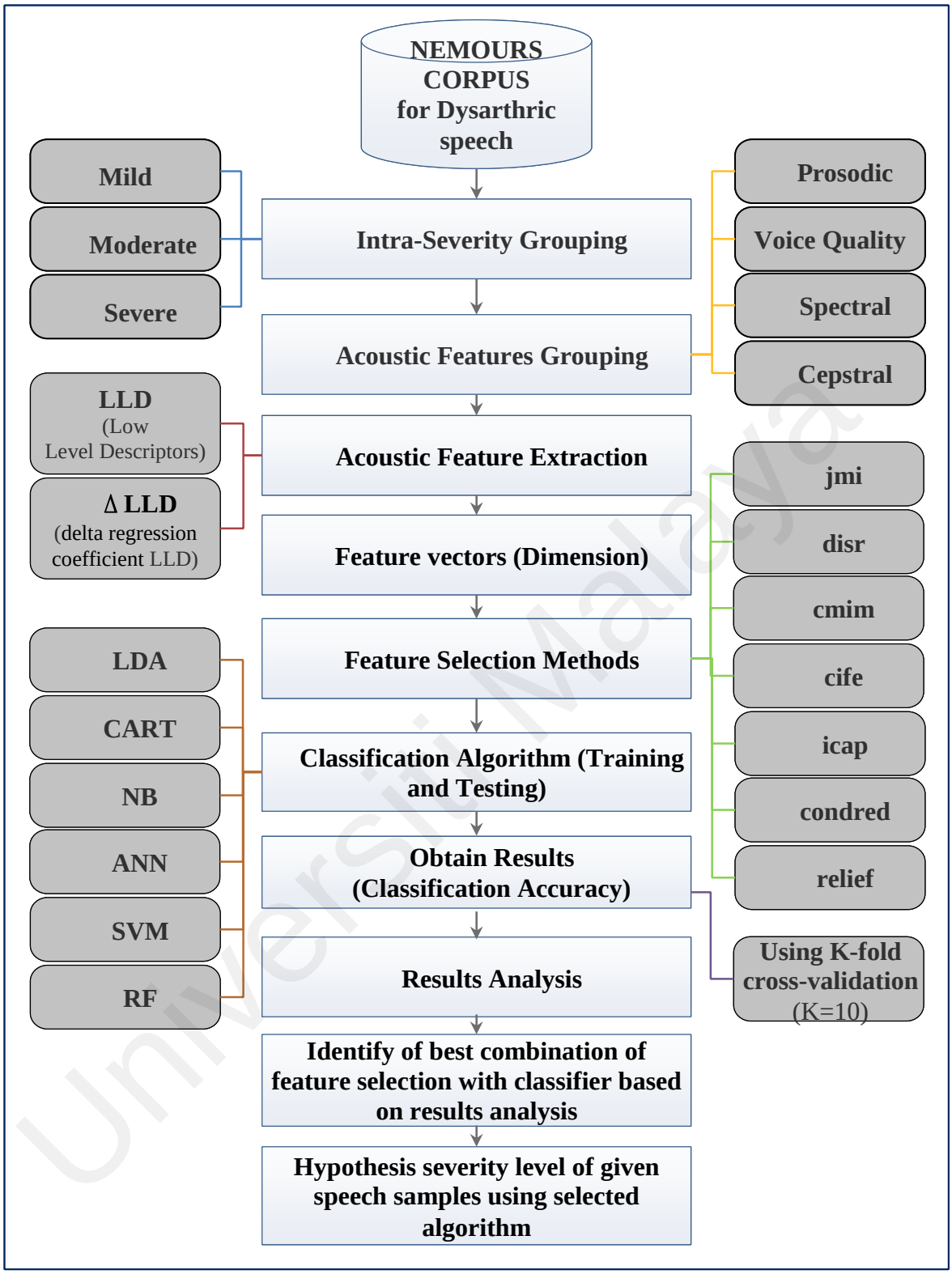

Figure 3.3: The classification phase diagram

3.5.1.2 Acoustic features extraction

As described in chapter 2 section 2.5.2, there are four acoustic features have been identified for the classification of the severity level of dysarthric speakers. This section provided more details about the features extracted from each group.

Figure 3.4 below shows the features extracted from dysarthric speech wave files. For each feature, there are parameters computed for a short time frame of an audio signal at a given time, called the acoustic Low Level Descriptors (LLD) (Eyben, 2015; Schuller, 2013). For instance, the prosodic feature includes six acoustic LLD, which are loudness (sum of the audio spectrum), Relative Spectral Transform (RASTA), fundamental frequency (F0), Root Mean Square Energy (RMS Energy), Zero-Crossing Rate (ZCR), and Probability of Voicing (Prob. Of Voicing).

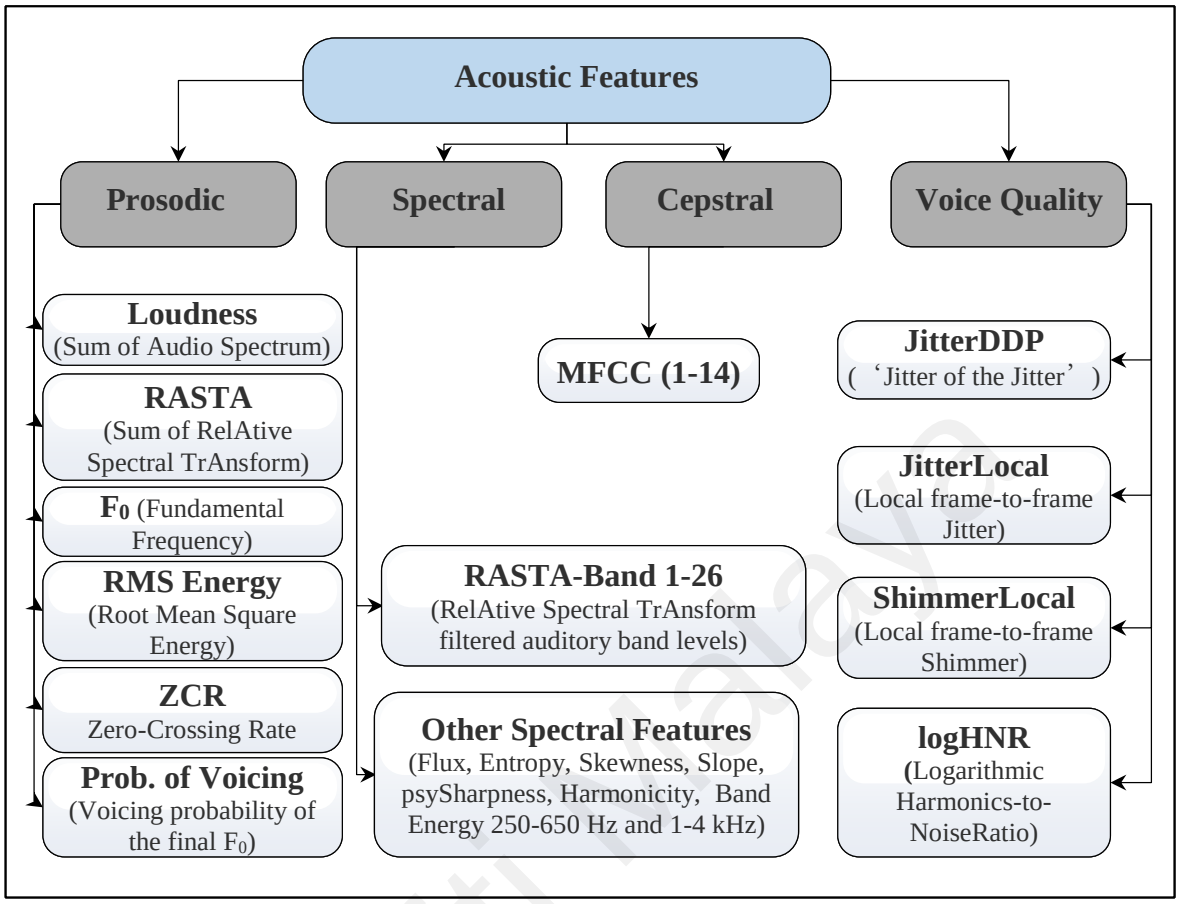

Figure 3.4: The acoustic features and related LLD acoustic parameters

There is a number of statistical functional computed for each acoustic LLD. The values of LLD computed are either smoothed by a window length 3 moving average filter as a standard, referred here as LLD or smoothed by delta regression coefficient referred here as Δ LLD (Eyben et al., 2013). Figure 3.5 shows the structural diagram of features, in which each feature has a set of LLD and each LLD has a set of statistical functions that is smoothed in different ways to obtain the final feature vectors (Dimension). The feature vectors will be combined for all acoustic features and used later as a parameter for feature selection and classifier.

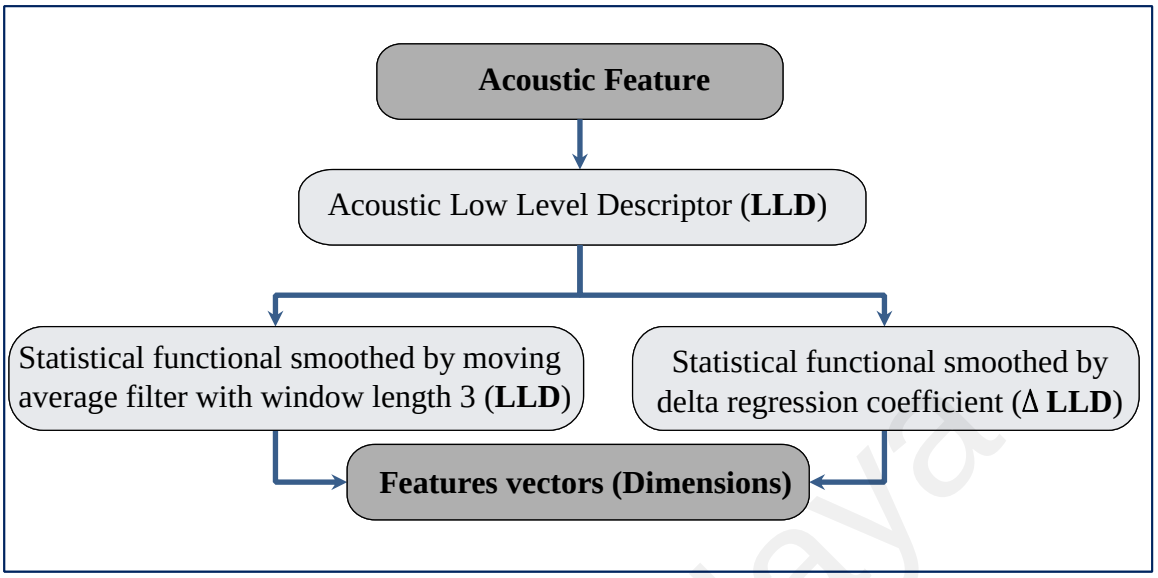

Figure 3.5: Structural diagram for acoustic feature

Based on the number of statistical functional applied to the LLD, there are two groups in this experiment. Table 3.1 shows the related features of each group and the number of statistical functional applied to each LLD.

Table 3.1: Acoustic feature of LLD groups and number of statistical functional applied for each group

LLD Group ID	Features	# LDD	# Statistical Functional per LLD	Total
A	Prosodic (Loudness, RASTA, RMS Energy, ZCR)	4	100	400
	Spectral	34	100	3400
	Cepstral	14	100	1400
B	Prosodic (Prob. Of Voicing)	1	78	78
	Voice Quality	4	78	312
	Prosodic (F_0)	1	83	83
Total				5673

Table 3.2 depicts the statistical functional applied to each group of LLD, group of statistical functional, and the number of statistical functional applied.

Table 3.2: Statistical Functional, groups, number, and applied LLD group used in the experiment

No.	Statistical Functional	SF Group	# SF	LLD Group
1	Quartiles 1–3, 3 inter-quartile ranges	Percentiles	6	A,B
2	1% Percentile (≈min), 99% percentile (≈max)	Percentiles	2	A,B
3	Percentile range 1–99%	Percentiles	1	A,B
4	Position of min/max, range (max − min)	Temporal	3	A,B
5	Arithmetic mean[1], root quadratic mean	Moments	2	A,B
6	Contour flatness	Temporal	1	A,B
7	Standard deviation, skewness, kurtosis	Moments	3	A,B
8	Rel. duration LLD is above 25/50/75/90% range	Temporal	4	A,B
9	Rel. duration LLD is rising	Temporal	1	A,B
10	Rel. duration LLD has positive curvature	Temporal	1	A,B
11	Gain of linear prediction (LP), LP coefficients 1-5	Modulation	6	A,B
12	Mean, max, min, SD of segment length	Temporal	4	A,B[2]
13	Mean value of peaks, Mean value of peaks – arithmetic mean	Peaks	3	A
14	Mean/SD of inter peak distances	Peaks	2	A
15	Amplitude mean of peaks, of minima	Peaks	2	A
16	Amplitude range of peaks	Peaks	1	A
17	Mean/SD of rising/falling slopes	Peaks	4	A
18	Contour centroid	Temporal	1	A[3],B
19	Linear regression slope, offset, quadratic error3	Regression	3	A[3],B
20	Quadratic regression a, b, offset, quadratic error3	Regression	4	A[3],B
21	Percentage of non-zero frames	Temporal	1	B[2]

[1] Arithmetic mean of LLD and positive arithmetic mean of Δ LLD. [2] applied to F_0 only in group B (LLD, not Δ LLD). [3] Applied to LLD not Δ LLD in group A. SF- Statistical functional

3.5.1.3 Feature selection dimensions

The large numbers of features extracted in this research make it difficult to perform the classification of severity type. As a result, feature selection is a possible solution to create different training sets in order to identify the most significant features related to the specific type of severity level. Alternatively, reducing the number of features used in the classification algorithm can also solve this problem.

There is no method used in the literature to select the optimal number of features. However, the equation below is suggested to better the computation cost (Khoshgoftaar, Golawala, & Van Hulse, 2007; Samsudin, Shafri, Hamedianfar, & Mansor, 2015):

$$NOF = \log_2 n \qquad (3.1)$$

Where NOF is the number of features to be picked up for classification algorithms, and the total number of extracted features is n.

Using the above equation, the number of features extracted as shown in Table 3.1 is 5673 ($n = 5673$), and so the number of features that will be used in classification algorithms are 13 features:

$$NOF = \log_2 5673$$

$$NOF = 12.47 \approx 13\ Features$$

These 13 features will be selected for the classification algorithms after applying the feature selection method to identify the most significant features among all features extracted for this experiment.

3.5.1.4 Feature selection methods

In this research, the feature selection methods applied prior to the running of the classification algorithms. The objective of using different feature selection methods is to create different training sets and to increase the diversity among the classifiers, which is a key feature in improving the performance of the multi-classifiers system. In addition to this, selection methods of two different features may give rise to two different sets of features.

Thus, presenting only one feature set can be misleading and may produce suboptimal results (Kuncheva, 2007). The seven feature selection methods used in this study, namely, Conditional redundancy (Condred) and Relief, Interaction Capping (ICAP), Conditional Information Feature Extraction (CIFE), Conditional Mutual Information Maximization (CMIM), Double Input Symmetrical Relevance (DISR), and Joint Mutual Information (JMI) (Parmar et al., 2015). The following subsections are briefly explained about each selection method:

(a) ***Joint Mutual Information (JMI)***

It was proposed by (Gao, Hu, & Zhang, 2018) where increasing the complementary information between features reduces redundancy (Brown, Pocock, Zhao, & Luján, 2012). The class for evaluating the importance of features, the already-selected feature, and the joint mutual information between the candidate features are employed by JMI, unlike other feature selection methods (Gao et al., 2018).

(b) ***Double Input Symmetrical Relevance (DISR)***

To reduce redundancy, (Meyer & Bontempi, 2006) symmetric relevance criterion is used, where the concept of complementary information between the feature is promoted.

Symmetrical relevance on all combinations of two features is measured by this criterion (Meyer & Bontempi, 2006).

(c) ***Conditional Mutual Information Maximisation (CMIM)***

This method searches for the most discriminative features by finding the optimal trade-off between relevance and redundancy of the feature (Fleuret, 2004). In this case, the feature selection only maximizes the mutual information of the feature while adding additional information to the already selected feature set.

(d) ***Conditional Information Feature Extraction (CIFE)***

This method was proposed by (Lin & Tang, 2006), where, by reducing the class-relevant redundancies among features maximizes the class-relevant information aimed.

(e) ***Interaction Capping (ICAP)***

The ICAP (Jakuline, 2005), uses interaction gain measures to detect the relevant feature. In this method, any feature if not relevant to the class on its own, it can be relevant when combined with another feature.

(f) ***Conditional Redundancy (Condred)***

(Brown et al., 2012) proposed this method, and is used for comparison purposes.

(g) ***Relief***

This was introduced by (Kira & Rendell, 1992). It is a feature grading algorithm method. The objective of this method is the quality estimation of features to differentiate samples that are near to each other in a dataset. Original Relief can only handle Boolean concept problems, but extensions have been developed to work in classification problems and in regression.

3.5.1.5 Classification algorithms

In this research, six classification algorithms were used, which includes SVM, LDA, and ANN as well as some of the well-known algorithms used for comparison with algorithms used for previous research like Classification and Regression Tree (CART), Naive Bayes (NB), and Random Forest (RF).

The classification algorithms classify the severity level of a given dysarthric speaker based on the acoustic features extracted into specific severity level which are mild, moderate and severe. This step is the final step of the classification phase of the proposed intra-severity automatic dysarthric speech recognition.

3.5.1.6 Procedures and tools

The tool used for feature extraction is openSMILE version 2.3.0, while the configuration for feature extraction used from the wave files of the severity level of the dysarthric speakers is the standard INTERSPEECH 2016 Computational Paralinguistic Challenge (INTERSPEECH 2016 ComParE Set) (Eyben et al., 2013).

In speech analysis, the typical frame lengths range from 20 to 60 milliseconds (ms), with the most commonly chosen frame period is 10ms (Rabiner, 1989; Young et al., 2009). For the proposed solution, 60ms were used as frame length, with 10ms as frame period. According to (Eyben, 2015) to compute LLD, the frame must contain enough data and the quasi-stationary of the signal is ensured to be within the frame with respect to LLD of interest by identifying the length of the frame.

The procedures for feature extraction is in three steps. First, the samples pronounced by each speaker are listed into one individual file for each speaker. This file is used as an input to the openSMILE tool to produce the features for each separate file

(the total number of sample files per speaker is 74). Second, each file generated in the first step is combined in three separate files according to their severity level. Third, the three separate files produced in the second step are combined in one feature file, including the class types which are severe, moderate, and mild. This file is then used as an input for the feature selection step for the classification algorithm.

MATLAB software version R2014b is used to combine all feature selection with classification algorithms in a single coding file. MATLAB is a well-known tool widely used by developers and researchers. The feature selection was obtained from FEAST toolbox version 2.0 (downloadable from https://github.com/Craigacp/MIToolbox/). There are seven feature selection methods used in this experiment described previously in this chapter.

The various toolbox used for the classification algorithms include: statistical toolbox used to build LDA, and CART classification methods, Neural network toolbox used to build the ANN models, libvm version 3.22 developed by Chih-Chung Chang and Chih-Jen Lin to build the SVM classification model (can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm) (Chang & Lin, 2011), Naive Bayes code which uses the default algorithms developed in MATLAB program, and the code for the Random Forest (can be downloaded from https://code.google.com/archive/p/randomforest-matlab/downloads).

3.5.2 Automatic Dysarthric Speech Recognition Phases

This phase is designed to automate the speech recognition of dysarthric speech. The ADSR involves the severity based adaptation phase and automatic speech recognition phase as described previously in the proposed intra-severity automatic speech recognition. Figure 3.6 showed the diagram for building the ADSR which starts with the

building process of the baseline speech acoustic model using the Wall Street Journal (WSJ1) corpus (Linguistic Data Consortium, 1994), TIMIT corpus (Garofolo & Consortium, 1993), and TORGO corpus. Following this is the model adaptation of the NEMOURS database to build the SA (adapted) speech acoustic model. The following section describes the related ADSR components.

3.5.2.1 Speech corpora

The speech corpuses used in this experiment are the Wall Street Journal (WSJ1) corpus (Linguistic Data Consortium, 1994), TIMIT corpus (Garofolo & Consortium, 1993), TORGO corpus, and NEMOURS corpus. The TORGO corpus and NEMOURS corpus were described in more detail in the previous chapter. WSJ1 corpus and TIMIT corpus are described below.

TIMIT speech corpus: at Texas Instruments and MIT, the TIMIT acoustic-phonetic continuous speech corpus was developed and distributed by the US National Institute of Standards and Technology. Eight major dialect division of American English is represented by it, comprising of 438 male speakers and 192 female speakers making up a total of 630 speakers (Garofolo & Consortium, 1993).

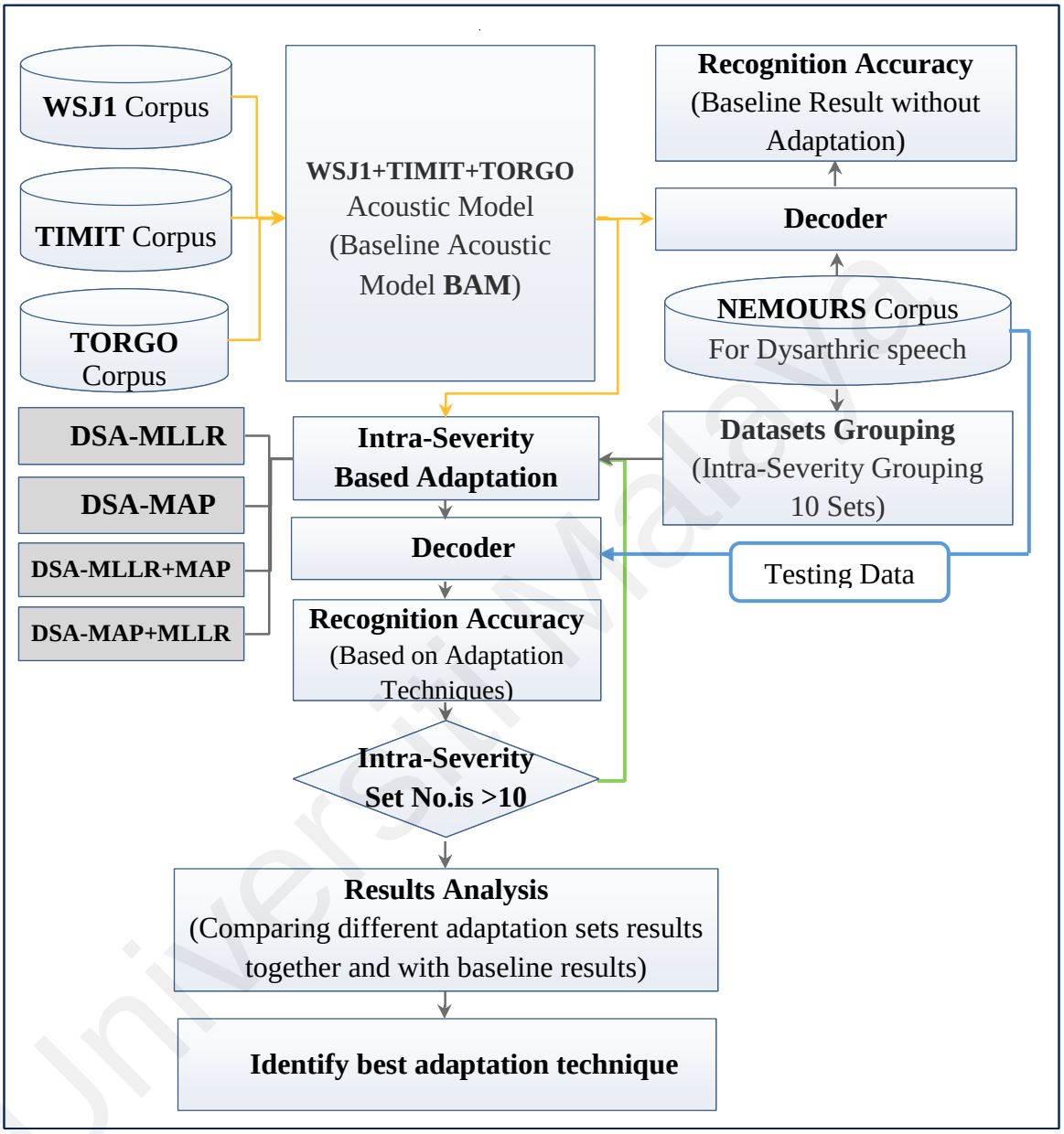

Figure 3.6: The automatic dysarthric speech recognition phases diagram

Wall Street Journal (WSJ1) corpus is collected for the development and evaluation of the Large Vocabulary Continuous Speech Recognition System (LVCSR). The WSJ1 corpus includes nearly 78,000 utterances (~73 hours of speech) labeled as training utterances. There were 245 subjects who participated in the recording the training set. The subjects were made up of journalists (20) and non-journalists (225) (Linguistic Data Consortium, 1994). Some training data has been spontaneously dictated by journalists (4,000 utterances of the training set). Nevertheless, the level of experience in dictation among journalists varies. On the other hand, the testing set labeled in the WSJ1 corpus has nearly 8,200 utterances (~8 hours of speech). There were 30 subjects that participated in recording the test set. The subjects were made up of journalists (20) and non-journalists (10) (Linguistic Data Consortium, 1994). Similar to the training set, some utterances have been spontaneously dictated (6,800 utterances of the testing set), producing more dictated data than the training set. The recording of the corpus used two microphones, a Sennheiser close-taking head-mounted microphone and microphones of varying types.

3.5.2.2 Speech corpus selection

For developing the speaker-independent (baseline) speech acoustic model, the WSJ1, TIMIT and TORGO speech databases were used as described in Table 3.3. On the other hand, for developing the adaptation model, NEMOURS databased is used after the removal of the two speakers as mentioned earlier in section 3.5.1.1. Table 3.4 shows the severity levels of the speakers of the NEMOURS database. More information about the speakers' severity levels and their intelligibility score can be found in (Menendez-Pidal et al., 1996). The intelligibility score is computed as the average of scores for three sessions by 12 non-hearing impaired listeners.

Table 3.3: Database used in this experiment for the training, adaptation and testing stages

Corpus Name	Size (sentences or recording time)	Number of Speakers	Vocabulary Size (number of words)	Level of Severity (did they cover all three levels)	Usage of the data corpus
WSJ1 (Linguistic Data Consortium, 1994)	77800 utterance (~73 hours of speech)	245 speakers	Less than 20K	non-impaired speech	Training
TIMIT (Garofolo & Consortium, 1993)	4620 training utterances-approximately 3.14 hours for training	462(326 males and 136 females)	Less than 6k of words	non-impaired speech	Training
TORGO (Frank Rudzicz et al., 2012) (Enderby, 1980b) (Mengistu & Rudzicz, 2011)	500 utterances per speaker (Approximately 3 hours of speech for each speaker)	8(five males-3 females)	Less than 1000 words	4–Severely,1-Moderate to Severely,1-Moderately, 2-very Mild.	Training
NEMOURS (Menendez-Pidal et al., 1996)	74 sentence(4.06 Hours)	9(9 males)	Less than 200 words	Mild-Moderate-Severe	Adaptation and testing

Table 3.4: The intelligibility scores and classification of dysarthric speech of NEMOURS database according to human assessment

Severity Level	Speaker ID	Intelligibility score
Mild	BB	89.7
	FB	92.9
	MH	92.1
Moderate	RK	68.6
	JF	78.5
	RL	73.3
Severe	BK	58.2
	SC	51.5
	BV	57.5

3.5.2.3 Experimental Procedures

(a) ***Development of the baseline speech acoustic model (BAM)***

The Baseline Acoustic Model (BAM) is trained using the data from different corpora: WSJ1 (Linguistic Data Consortium, 1994), TIMIT (Garofolo & Consortium, 1993), and TORGO (Frank Rudzicz et al., 2012) databases. The main purpose of using these three speech databases is to enrich the acoustic model, and for the ASR system's accuracy of recognition to improve (Al-Qatab et al., 2014; Paul & Baker, 1992). The speech used for developing the BAM includes both the non-dysarthric speech (Control Speakers) and dysarthric speech (Dysarthric Speakers).

The baseline acoustic model developed in this research did not make use of the adaptation technique and was used as benchmark comparison (in term of recognition accuracy) with the proposed automatic dysarthric speech recognition accuracy (intra-severity based adaptation). Furthermore, the baseline acoustic model also used to show that using the non-impaired speech corpora can help to overcome the lack of corpora for impaired speech (dysarthric speech).

(b) ***Intra-severity based adaptation***

For Speaker Adaptation (SA) of BAM, the MLLR and MAP techniques were used. These techniques were used individually and in combination (hybrid). The four SA experiments are made up of Dysarthric Speaker Adaptation (DSA) using MAP, MLLR, and a hybrid of these two techniques, namely DSA-MLLR, DSA-MAP, DSA-MLLR+MAP, and DSA-MAP+MLLR respectively.

In this study, the adaptation is performed individually for a different level of severity, adaptation techniques, and the amount of data used, which are computed using the following equation:

$$NOE = NST * NAT * NS \qquad (3.2)$$

Where NOE is the number of experiments performed, NST denotes the number of severity levels, NAT denotes the number of adaptation techniques, and NS denotes the number of datasets used in this experiment. As such, the number of experiments conducted in this study is 120 (3 x 4 x 10). The adaptation models: MILD (mild dysarthric speech), MODR (moderate dysarthric speech), and SEVR (severe dysarthric speech) are trained accordingly to obtain severity adapted models.

(c) ***Testing data***

The NEMOURS corpus is used for the testing stage in this research as shown in Figure 3.5. The complete samples include in the NEMOURS corpus used for the testing stage. The testing data are applied to the acoustic models built as described in subsection b of section 3.2.2.3. The intra-severity based testing performs to the intra-severity based adaptation. The differences between the adaptation data set and testing data set are that the adaptation stage is divided into the ten sets which every set contains speech from all speakers in the specific severity level as described in section 3.2.2.4 while in the testing

stage the complete samples are used to obtain the results based on the severity level. More information about the NEMOURS corpus described in section 3.5.2.2 as well as in section 2.6.2 in chapter 3.

(d) ***Speech data coding***

12 Mel-Frequency Cepstral Coefficients (MFCCs) were extracted. This includes C0 as an energy component for every 10 ms analysis frame using a 25-ms Hamming window, and their first and second derivatives computed to obtain a 39-dimensional feature vector. The cepstral mean and energy normalization was applied to the feature vectors during training and testing.

(e) ***HMM topology and tools used***

To build the HMM topology and to train the acoustic model, the 3-states left to right context-dependent triphones were used. 41 monophones (which contained silence and short pauses) were used to construct all the triphone models. By applying decision tree clustering, the context-dependent triphones acoustic model was tied so that the acoustic performance is enhanced and the common features among the states are shared. Additionally, to gain extra acoustic performance, the 16 mixture Gaussians per state was performed. This results in the utterances used, states, and the number of triphones to build a trained acoustic model for BAM to be at 86,547, 8,108, and 9,423 respectively. The HTK (version 3.4.1) toolkit (Young et al. 2009) is used to perform the speech coding and training of the baseline speech acoustic model. The word network constructed from the sentences of the test data includes the same form of constricting the sentence (see section 2.6.2 in chapter 2). The dictionary used for training and testing is extracted from the Carnegie Mellon University (CMU) pronouncing dictionary. The dictionary was used for words, silence, and space included the training and testing sentences.

3.5.2.4 Adaptation dataset

Adaptation datasets used for this experiment have been divided into ten sets based on the duration of the utterance (30 sets in total). Table 3.5 depicts the duration of the adaptation dataset used to adapt the acoustic model. The adaptation dataset was grouped based on the minimum optimal speech duration, which is about one minute of data (Shinoda, 2011). Each dataset is an intra-severity data, where each severity consists of utterances from speakers belonging to the same severity level. For example, the dataset for mild severity level includes the utterances from the three mild severity speakers. The test dataset covers the speeches of all the speakers with the same severity level. The data is divided randomly from the database based on the amount of data needed for adaptation, with the aim of increasing recognition accuracy.

Table 3.5: Duration in seconds of the adaptation datasets

Adaptation set	Duration in seconds for each level of severity		
	Mild (MILD)	Moderate (MODR)	Sever (SEVR)
Set-1	63.52	92.14	143.93
Set-2	118.61	181.62	240.47
Set-3	183.84	277.75	350.55
Set-4	241.60	363.13	458.02
Set-5	299.77	455.09	561.59
Set-6	363.92	550.40	691.01
Set-7	421.98	631.43	791.26
Set-8	491.15	730.41	907.21
Set-9	551.64	823.40	1008.84
Set-10	608.18	908.55	1103.42
TOTAL	3344.21	5013.92	6256.30

3.6 Evaluation

The fourth step of the DSRM process is the evaluation which is described in more detail in chapter 4. In this research, a series of a comparative analysis based on the classification accuracy of dysarthric speakers as well as the recognition accuracy of ASR for dysarthric speakers are conducted to compare the accuracy rate for both the classification and the recognition rate of different related work in the field of classification and recognition techniques. This section has explained the analysis of the data used to evaluate the proposed Intra-Severity ADSR in both classification and adaptation phases. The data analysis used in the proposed Intra-Severity ADSR based on the classification of dysarthric speech and the automatic dysarthric speech recognition, which is discussed in the following sections:

3.6.1 Classification of Dysarthric Speech

The acoustic feature analysis is used to identify the best acoustic features that can be used in the classification phase. The analysis of the acoustic features includes the sub-acoustic feature that will help to identify the best performance of the sub-acoustic features. It also includes the analysis of the performance of all the sub-acoustic features and the combination of each sub-acoustic features as well as the acoustic features and the combination of all the acoustic features.

The classification algorithms analysis is used to identify the classification algorithms' best performance. The combination of each selection of feature and classification algorithms is considered as an independent classification algorithm which has its own performance.

Identifying the best acoustic features and the best classification algorithms help to select the acoustic features and the classification algorithms to develop the automatic dysarthric speech recognition systems.

3.6.2 Automatic Dysarthric Speech Recognition

The data analysis in this phase includes the performance of the ADSR when using the acoustic model with the data from the normal speakers. It also includes the analysis of each adaptation techniques: when it is applied to the baseline acoustic model, and when it is not applied.

The results of all the adaptation techniques proposed in this study were also analyzed to obtain the best performance of the adaptation techniques, to improve the recognition accuracy of the automatic dysarthric speech recognition system.

3.6.3 Performance Measure

Measuring the performance of the proposed solution performs based on the two parts of the solution which are classification and adaptation of dysarthric speech based on the severity level. Each performance measure of the classification and automatic dysarthric speech recognition describe in the following subsections.

3.6.3.1 Classification accuracy

To calculate the classification accuracy for each classifier algorithm, the k- fold cross-validation, where k is assigned to 10 (Ishibuchi & Nojima, 2013; McLachlan, Do, & Ambroise, 2005) is commonly used to calculate the rate of accuracy of the classifier algorithm for severity level of dysarthric speakers. In this method, the extracted features from dysarthric speakers (including all severity levels) are randomly divided into 10 equal sizes of set samples, where nine partitions are assigned for model training, and the

remaining one is used as the test set for model evaluation. For each run, one partition is used as a test data and the remaining partitions are used as training data. To ensure all 10 partitions are used as test data, this procedure is repeated 10 times. To produce a single estimation, the mean score of all 10 runs was calculated. Compared to a repeated random sub-sampling, the advantage of this method is that for both training and validation, all observations are used, and each observation is used for validation for exactly once only. The average classification accuracy rate is calculated using the equation below:

$$\text{Average Classification Accuracy Rate} = 100 \times (TNCF/TNF) \qquad (3.3)$$

where TNCF is the Total Number of Correctly-testing Features, and TNF is the Total Number of Features used.

For the selection of the best classifier or best feature selection method, the raking method proposed by Friedman's M Statistic (Neave & Worthington, 1988) is used (Brazdil & Soares, 2000). In this method, each classifier received a rank based on the measured accuracy rates on each feature group, where the classifier with the highest accuracy rate on the features group is assigned rank 1 and the classifier with second highest accuracy rate is assigned rank 2 and so on. In the case of two classifiers achieving equal accuracy rates, then the rank is divided between them. For example, considering the accuracies of 50%, 60%, 62%, 62%, and 67% achieved from five different classifiers on different group features, their ranking score would be 5, 4, 2.5, 2.5, and 1 respectively. The performance of the classifier algorithms is evaluated using the ranking method represented by the following equation:

$$Ranking\ (x_1^n) = \begin{cases} Ranking\ base\ on\ highest, & x_i\ is\ identical\ value \\ \frac{n}{2}, & x_i\ for\ each\ equal\ value \end{cases} \qquad (3.4)$$

Where x_1^n is the set of accuracy rates for the classification algorithms used, the number of classification algorithms used is n, and the current value in the x set is i.

For calculating the final ranking of a classifier on different features groups, the mean score of each classifier is calculated. Therefore, the lowest average ranking score is considered the best classifier. The following equation is used to calculate the best classifier based on average ranking score:

$$Best\ Classifier(X_1^n) = \text{Min}(Average(Ranking(x_1^n))) \quad (3.5)$$

Where X_1^n is the set of classification algorithms used, n is the total number of the classifier, and $Ranking\ (x_1^n)$ is the ranking score of the accuracy rate of different feature groups.

3.6.3.2 Recognition accuracy

Word Error Rate (WER) is generally the way to measure the effectiveness of an ASR system. In a total recognition task, global and incorrect word recognition are measured by WER. Alternatively, measuring an error rate may also be done in smaller units, such as detailed errors, syllables, or phonemes. These include deletion rates, substitution, and insertion of phoneme (Mokbel et al., 1996) as follows:

$$Word\ Error\ Rate\ (WER) = \frac{Insertion + Substitution + Deletion}{Number\ Of\ Words} * 100\% \quad (3.6)$$

where:

- Phoneme insertion: Due to the slow speaking rate of a dysarthric speaker, an extra sound or sounds is/are added to the intended word. This causes a monosyllabic word to be interpreted as two syllables.
- Phoneme substitution: People suffering from dysarthria make pronunciation errors (e.g., twee instead of tree), thus one phoneme is substituted with another.
- Phoneme deletion: People suffering from dysarthria do not produce certain sounds, causing the omission of all the syllables or specific sounds.

The final recognition accuracy of the proposed technique is determined using the Average of Word Error Rate used in this to obtain final recognition accuracy. The Average Word Error Rate calculated by summing up all the word error rate for results from each adaptation data sets divided by the number of adaptation sets which is 10 data sets (details of the WER for each adaptation set found in Appendix E, Table E.2). The following equation used to calculate the final WER:

$$Word\ Error\ Rate\ (WER) = \frac{Sum(WER\ for\ 10\ Adaptation\ Sets)}{10} \qquad (3.7)$$

Equation 3.7 used to evaluate the recognition accuracy in chapter 4.

3.7 Communication

In finalizing this research, the documentation of all activities and processes of this research is performed in the form of a thesis. Additionally, some of the major findings are published and some are submitted for possible publication to the related journal and conferences.

3.8 Summary

This chapter provides the DSRM developed for the current research. The process of DSRM is described throughout the sections of this chapter. More details about the design and development of the proposed system are presented. The overall system architecture, which includes classification and automatic speech recognition for dysarthric speech is explained. The experiment design for each part of the system development is explained in detail, which includes the acoustic features, speech corpus, and techniques used for each part of the proposed system. The analysis of the data and the performance measures for each part of the proposed system are also explained in this chapter.

CHAPTER 4: ANALYSIS, RESULTS, AND DISCUSSION

4.1 Overview

In this chapter, the results obtained from the experiments will be analyzed. Discussions on the results will be explained in detail to show the system's accuracy.

The results are analyzed in two parts. The first part will focus on the analysis of the results on the classification of the dysarthric severity level of the system. This part will report the best acoustic features and the best classification algorithms to classify the speech of the dysarthric speakers to its level of severity, which is mild, moderate, or severe. In the second part, the analysis of the results from the Automatic Dysarthric Speech Recognition (ADSR) will be illustrated. This chapter includes the best adaptation techniques that could be used to help obtain a high recognition accuracy for dysarthric speakers.

4.2 The Dysarthric Severity Level Classification

The acoustic features used to obtain the set of features from the dysarthric speech audio files (samples) which are classified according to the severity level (mild, moderate, and severe) are the prosodic, voice quality, spectral and cepstral features. For those features that have sub-features, the results will be obtained using each sub-feature and the combination of all sub features separately, and will be included in the comparison of all sub features, which will be called "All". For example, the voice quality acoustic feature has four sub-features which are jitterDDP, jitterLocal, shimmerLocal, and logHNR, so the fifth sub-feature is the combination of all the four sub-features, and will be labeled as "All". This combination will also be used for the main acoustic features, which will include the combination of all acoustic features, and will also be called as "All".

The number of acoustic features for every sub-feature will be 13, as shown in chapter 3. For the combination of sub-features, the 13 features will be selected after combining all the sub-features. The selection of these features will be according to the feature selection algorithms.

The results are analyzed based on the classification algorithms. There are six classification algorithms used in this experiment, which are LDA, CART, NB, ANN, SVM, and RF. The results will show the effectiveness of each classification algorithm based on the feature selection method and acoustic feature.

In both acoustic feature and classification algorithms analysis, the average ranking method is used to obtain the best performance, whether on the acoustic feature or classification algorithms analysis.

The classification accuracy of the dysarthric severity level is depicted in Table 4.1. It will be used as the base to various forms of analysis such as the average ranking score, either for acoustic features' performance or classification algorithms' performance, which is discussed below.

Table 4.1: The classification accuracy based on classification algorithms, feature selection, and acoustic features

Classifier algorithm	Feature Selection Algorithm	Acoustic Features																
		Prosodic							Voice Quality					Spectral			Cepstral	All
		audspec (Loudness)	audspecRasta-Sum	F0final	PCM-RMS Energy	PCM-Zero-Crossing Rate (ZCR)	voicingFinalUnclipped	All	jitterDDP	JitterLocal	shimmerLocal	logHNR	All	audspecRasta-Band 1-26	PCM- Oher Spectral Features	All	MFCC	
LDA	jmi	77.16	71.94	65.77	72.98	57.70	66.37	65.14	44.62	49.50	52.24	71.00	66.35	77.98	74.47	78.10	70.72	75.22
	disr	85.00	72.38	66.22	75.38	56.62	67.13	65.49	43.54	56.11	56.91	70.71	68.62	76.42	74.65	80.32	72.39	76.59
	cmim	77.64	70.56	66.53	76.29	57.06	66.33	81.82	46.70	51.94	55.06	71.80	69.04	78.51	77.79	80.17	69.37	75.06
	cife	69.64	69.53	71.02	63.40	58.25	63.94	63.65	44.30	46.86	50.57	72.22	66.38	70.87	71.48	65.77	60.22	62.62
	icap	75.53	74.33	66.67	76.57	56.78	66.80	81.25	40.83	49.55	52.41	71.65	68.80	80.32	78.06	81.51	61.12	74.15
	condred	74.02	69.21	60.80	62.90	58.39	65.02	62.92	40.41	48.36	50.47	68.92	66.82	73.12	71.79	67.70	68.46	76.23
	relief	86.61	65.92	62.32	85.30	77.02	79.90	87.55	53.76	54.07	69.98	68.00	69.04	81.55	95.64	93.86	81.09	75.22
CART	jmi	70.71	59.90	83.18	73.55	55.88	62.49	82.13	49.70	54.64	54.95	74.77	71.78	71.02	68.92	72.51	64.72	74.48
	disr	75.08	62.60	81.67	75.38	53.60	63.83	80.65	52.69	59.60	62.33	75.21	74.61	72.96	69.84	75.66	63.86	76.57
	cmim	74.15	61.24	82.90	73.89	56.90	63.52	84.82	52.10	54.19	62.91	73.89	75.82	75.09	77.04	73.10	65.14	76.57
	cife	71.87	59.78	79.75	63.09	55.41	55.11	79.85	49.43	55.10	52.81	72.98	66.37	68.78	62.92	60.65	61.71	68.20
	icap	71.79	60.50	81.85	70.74	53.28	65.20	83.34	49.53	54.07	52.70	77.95	76.01	71.77	75.53	75.08	62.88	79.41
	condred	69.06	59.28	80.47	60.65	51.18	57.35	80.17	50.29	54.65	54.36	73.11	72.06	66.81	65.77	59.16	64.40	77.02
	relief	85.26	68.02	60.52	78.40	66.22	75.52	86.20	59.18	54.95	64.86	69.65	64.84	74.48	90.85	90.09	75.69	77.77
NB	jmi	72.22	59.92	72.52	63.81	62.49	64.43	71.31	52.42	54.02	57.36	74.02	70.69	73.58	70.88	75.38	65.76	79.89
	disr	76.29	62.61	72.35	70.26	61.25	68.04	69.68	52.39	57.05	62.61	73.40	73.72	72.68	72.07	77.16	66.99	79.89
	cmim	74.15	62.47	72.53	61.81	62.61	66.33	79.58	52.86	55.70	63.50	75.99	72.67	75.97	76.58	81.07	67.27	80.77
	cife	67.86	60.52	79.57	56.05	59.90	59.88	73.72	53.74	50.17	55.97	69.81	68.77	66.20	69.23	65.92	60.95	66.55
	icap	71.31	62.15	71.81	59.31	61.89	65.19	78.96	54.03	56.34	57.09	74.03	72.52	75.84	77.32	80.91	65.92	83.18
	condred	68.45	58.84	71.29	62.31	57.65	58.70	68.76	52.70	53.49	54.98	71.17	67.14	63.06	67.74	62.31	66.33	77.46

Classifier algorithm	Feature Selection Algorithm	Acoustic Features																
		Prosodic							Voice Quality					Spectral			Cepstral	All
		audspec (Loudness)	audspecRasta-Sum	F0final	PCM-RMS Energy	PCM-Zero-Crossing Rate (ZCR)	voicingFinalUnclipped	All	jitterDDP	JitterLocal	shimmerLocal	logHNR	All	audspecRasta-Band 1-26	PCM- Oher Spectral Features	All	MFCC	
	relief	83.15	66.70	56.89	70.59	65.05	80.20	82.14	54.03	52.25	67.71	65.45	64.38	75.38	85.28	80.62	78.99	77.34
ANN	jmi	80.01	67.72	74.62	77.78	53.76	63.05	76.27	43.36	45.17	46.07	77.49	71.87	74.40	77.17	77.81	70.86	77.47
	disr	79.12	72.21	73.07	75.39	55.10	67.13	74.74	44.31	54.80	55.20	78.80	79.73	71.49	76.88	81.20	70.27	75.06
	cmim	77.91	65.77	73.99	75.55	58.11	65.72	83.48	45.94	49.68	54.27	78.97	76.87	73.13	79.58	80.63	72.08	78.67
	cife	71.60	67.11	72.10	64.57	53.16	64.40	72.50	41.01	48.79	46.67	73.86	67.31	72.21	72.38	63.35	60.37	64.31
	icap	73.42	69.08	71.69	75.97	54.38	65.62	85.27	42.30	47.45	50.01	77.61	79.74	75.82	81.22	81.52	64.14	81.96
	condred	78.80	66.22	75.80	64.53	52.27	62.00	74.89	42.02	45.17	44.80	74.30	71.47	73.73	73.29	67.85	70.26	78.21
	relief	89.01	73.58	58.71	80.94	75.40	80.37	87.10	51.94	51.51	68.63	70.26	64.40	82.29	95.64	94.45	79.17	74.91
SVM	jmi	79.42	71.48	69.38	77.79	58.45	66.20	64.42	43.26	48.01	49.70	73.25	69.35	80.21	76.27	79.45	70.41	76.58
	disr	78.68	75.83	65.32	76.72	58.28	65.93	68.93	43.55	48.02	48.20	73.72	72.07	76.57	77.04	81.81	67.43	78.97
	cmim	78.66	70.86	67.27	75.84	60.22	64.55	82.12	44.44	48.35	48.92	73.00	69.63	78.81	81.39	82.42	69.37	74.45
	cife	71.89	70.88	74.31	62.18	55.55	64.23	65.30	43.72	46.56	49.52	72.22	69.38	72.21	69.99	67.44	60.38	59.31
	icap	75.83	74.17	69.85	77.93	56.45	65.76	81.53	43.25	48.36	50.01	74.21	69.54	78.68	81.82	81.06	63.51	77.02
	condred	76.88	71.32	66.10	65.60	56.91	64.56	64.41	43.38	44.45	47.92	70.12	69.24	76.13	72.69	69.35	68.61	74.31
	relief	76.09	62.35	58.86	62.34	50.00	56.61	80.03	50.90	49.86	50.58	61.38	49.83	75.24	92.95	89.92	81.11	63.07
RF	jmi	81.52	69.98	87.83	82.13	68.37	71.04	87.97	55.25	63.50	61.43	81.38	80.63	79.01	78.83	81.99	74.31	83.19
	disr	82.30	71.91	86.63	82.58	66.84	73.73	88.74	58.54	66.05	70.41	80.14	83.03	80.17	77.79	79.72	73.59	82.58
	cmim	82.25	70.87	87.53	81.97	66.97	72.48	91.14	58.56	61.40	69.94	82.29	81.67	79.57	83.94	84.23	74.63	87.82
	cife	80.00	68.63	86.91	71.32	63.81	66.02	86.31	55.86	59.46	61.24	78.53	75.97	77.33	69.97	71.04	72.83	78.38
	icap	81.68	72.67	88.15	80.33	64.43	73.30	91.15	58.40	63.06	60.65	82.00	81.56	80.18	81.08	81.98	70.88	89.64
	condred	77.02	69.52	86.93	72.35	62.46	67.39	87.52	57.06	62.03	64.12	80.47	79.74	74.03	71.78	69.68	72.36	85.27
	relief	92.01	75.09	69.99	84.68	75.55	86.07	89.66	64.69	60.08	71.03	75.51	72.51	83.93	95.80	95.79	83.50	83.48

4.2.1 Acoustic Feature Analysis

The intra-acoustic features include the analysis of each acoustic feature independently. The main goal of this analysis is to show the effectiveness of the sub acoustic feature in classifying the severity levels of dysarthric speakers.

4.2.1.1 Intra feature analysis

(a) ***Prosodic acoustic features***

Table 4.2 shows the ranking score obtained from the classification accuracy in Table 4.1. The ranking score varies from 1 to 7 based on the highest classification accuracy of the number of sub acoustic features used in this experiment. The calculated average ranking score is shown at the end of Table 4.2.

Figure 4.1 shows the graphic chart for the average ranking score for the sub acoustic features of the prosodic acoustic features. As seen from Table 4.2 and Figure 4.1, the best classification performance for the severity level of dysarthric speakers is the audspec (Loudness) prosodic acoustic features, with the lowest average ranking score of 2.24. The results listed in Table 4.2 showed that the combination of the prosodic acoustic features has the second-highest score with 2.26 average ranking score. The F0Final ranked third with an average ranking score of 3.38.

Table 4.2: Average ranking score for Prosodic Acoustic Features

Classifier algorithm	Feature Selection Algorithm	Acoustic Features						
		Prosodic						
		audspec (Loudness)	audspecRasta-Sum	F0final	PCM-RMS Energy	PCM-Zero-Crossing Rate (ZCR)	voicingFinalUnclipped	All
LDA	jmi	1	3	5	2	7	4	6
	disr	1	3	5	2	7	4	6
	cmim	2	4	5	3	7	6	1
	cife	2	3	1	6	7	4	5
	icap	3	4	6	2	7	5	1
	condred	1	2	6	5	7	3	4
	relief	2	6	7	3	5	4	1
CART	jmi	4	6	1	3	7	5	2
	disr	4	6	1	3	7	5	2
	cmim	3	6	2	4	7	5	1
	cife	3	5	2	4	6	7	1
	icap	3	6	2	4	7	5	1
	condred	3	5	1	4	7	6	2
	relief	2	5	7	3	6	4	1
NB	jmi	2	7	1	5	6	4	3
	disr	1	6	2	3	7	5	4
	cmim	2	6	3	7	5	4	1
	cife	3	4	1	7	5	6	2
	icap	3	5	2	7	6	4	1
	condred	3	5	1	4	7	6	2
	relief	1	5	7	4	6	3	2
ANN	jmi	1	5	4	2	7	6	3
	disr	1	5	4	2	7	6	3
	cmim	2	5	4	3	7	6	1
	cife	3	4	2	5	7	6	1
	icap	3	5	4	2	7	6	1
	condred	1	4	2	5	7	6	3
	relief	1	6	7	3	5	4	2
SVM	jmi	1	3	4	2	7	5	6
	disr	1	3	6	2	7	5	4
	cmim	2	4	5	3	7	6	1
	cife	2	3	1	6	7	5	4
	icap	3	4	5	2	7	6	1
	condred	1	2	3	4	7	5	6
	relief	2	3	5	4	7	6	1

Classifier algorithm	Feature Selection Algorithm	Acoustic Features						
		Prosodic						
		audspec (Loudness)	audspecRasta-Sum	F0final	PCM-RMS Energy	PCM-Zero-Crossing Rate (ZCR)	voicingFinalUnclipped	All
RF	jmi	4	6	2	3	7	5	1
	disr	4	6	2	3	7	5	1
	cmim	3	6	2	4	7	5	1
	cife	3	5	1	4	7	6	2
	icap	3	6	2	4	7	5	1
	condred	3	5	2	4	7	6	1
	Relief	1	6	7	4	5	3	2
Average Ranking		**2.24**	**4.71**	**3.38**	**3.71**	**6.64**	**5.05**	**2.26**

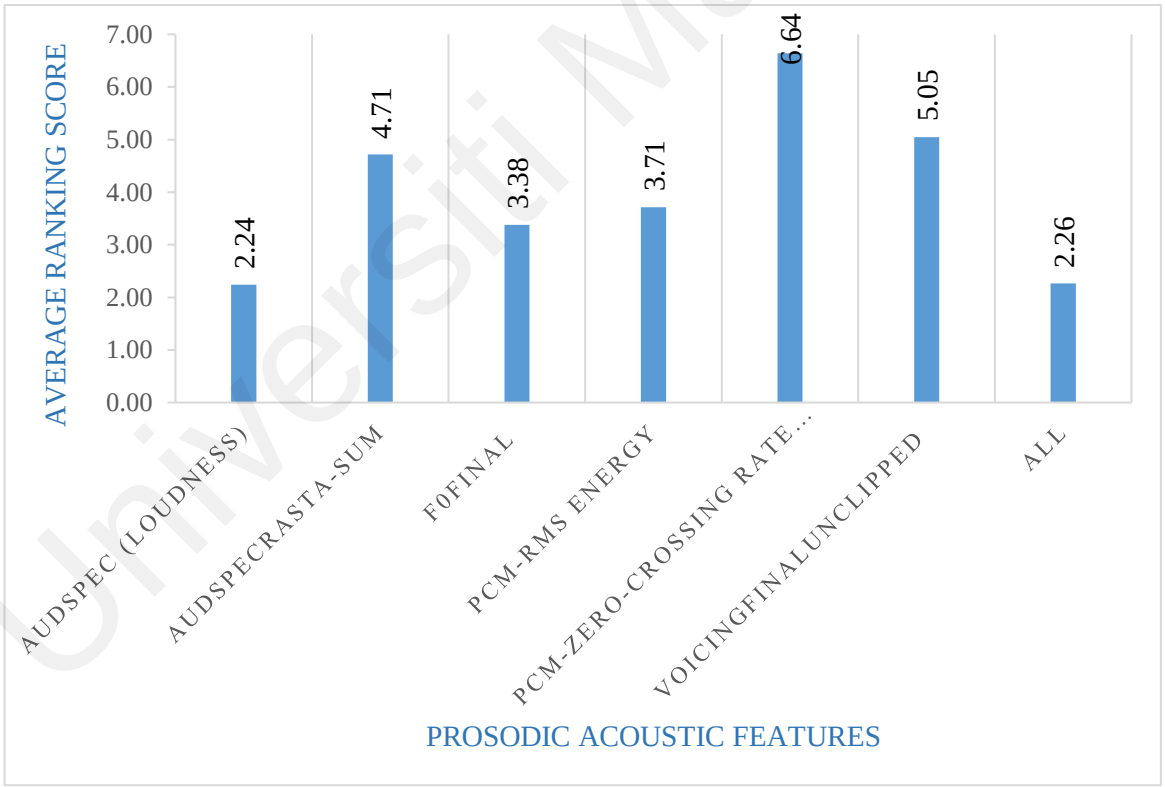

Figure 4.1: The graph chart for Average ranking Score for Prosodic Acoustic Features

(b) ***Voice quality acoustic features***

Table 4.3 shows the ranking score obtained from the classification accuracy in Table 4.1. The ranking score varies from 1 to 5 based on the highest classification accuracy of the voice quality sub acoustic features. The calculated average ranking score is shown at the end of Table 4.3.

Figure 4.2 shows the graphic chart for the average ranking score of the voice quality sub acoustic features. As shown in Table 4.3 and Figure 4.2, the best classification performance for the severity level of dysarthric speakers is the logHNR voice quality acoustic features with the lowest average ranking score of 1.19. The results listed in Table 4.3 showed that the combination of the voice quality acoustic features can be a competitor to sub voice quality acoustic features as it is ranked second, followed by lorHNR with 2.00 average ranking score. The shimmerLocal ranked third with an average ranking score of 3.02. The average ranking score is computed as the average of the ranking score obtained based on the classification accuracy for each classification algorithm and feature selection method used in this experiment.

Table 4.3: Average Ranking Score for Voice Quality Acoustic Features

Classifier algorithm	Feature Selection Algorithm	Acoustic Features				
		Voice Quality				
		jitterDDP	JitterLocal	shimmerLocal	logHNR	All
LDA	jmi	5	4	3	1	2
	disr	5	4	3	1	2
	cmim	5	4	3	1	2
	cife	5	4	3	1	2
	icap	5	4	3	1	2
	condred	5	4	3	1	2
	relief	5	4	1	3	2
CART	jmi	5	4	3	1	2
	disr	5	4	3	1	2
	cmim	5	4	3	2	1
	cife	5	3	4	1	2
	icap	5	3	4	1	2
	condred	5	3	4	1	2
	relief	4	5	2	1	3
NB	jmi	5	4	3	1	2
	disr	5	4	3	2	1
	cmim	5	4	3	1	2
	cife	4	5	3	1	2
	icap	5	4	3	1	2
	condred	5	4	3	1	2
	relief	4	5	1	2	3
ANN	jmi	5	4	3	1	2
	disr	5	4	3	2	1
	cmim	5	4	3	1	2
	cife	5	3	4	1	2
	icap	5	4	3	2	1
	condred	5	3	4	1	2
	relief	4	5	2	1	3
SVM	jmi	5	4	3	1	2
	disr	5	4	3	1	2
	cmim	5	4	3	1	2
	cife	5	4	3	1	2
	icap	5	4	3	1	2
	condred	5	4	3	1	2
	relief	2	4	3	1	5

Classifier algorithm	Feature Selection Algorithm	Acoustic Features				
		Voice Quality				
		jitterDDP	JitterLocal	shimmerLocal	logHNR	All
RF	jmi	5	3	4	1	2
	disr	5	4	3	2	1
	cmim	5	4	3	1	2
	cife	5	4	3	1	2
	icap	5	3	4	1	2
	condred	5	4	3	1	2
	Relief	4	5	3	1	2
Average Ranking		**4.81**	**3.95**	**3.02**	**1.19**	**2.02**

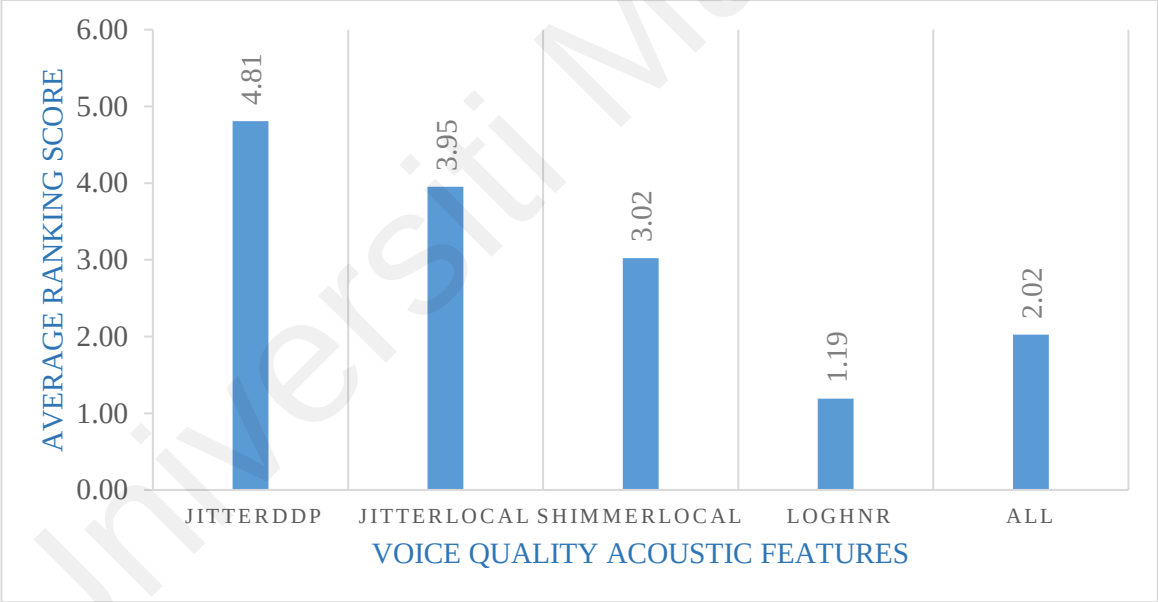

Figure 4.2: The graph chart for Average Ranking Score for Voice Quality Acoustic Features

(c) ***Spectral acoustic features***

There are two spectral acoustic features included in this section as well as the combination of two acoustic features. The performance of the spectral acoustic features are reported in Table 4.4 and Figure 4.3.

Table 4.4 shows the ranking scores obtained from the classification accuracy for the spectral acoustic features depicted in Table 4.1.

Table 4.4 and Figure 4.3 depict the best classification performance for the severity level of dysarthric speakers, where the combination of all spectral acoustic features with the lowest average ranking score of 1.83. The results listed in Table 4.4 show that the PCM-other spectral features of the spectral acoustic features is ranked second, followed by the combination of spectral acoustic features with 1.98 average ranking score. The audspecRasta-band 1-26 was third with an average ranking score of 2.19.

Table 4.4: Average Ranking Score for Spectral Acoustic Features

Classifier algorithm	Feature Selection Algorithm	Acoustic Features		
		Spectral		
		audspecRasta-Band 1-26	PCM- Oher Spectral Features	All
LDA	jmi	2	3	1
	disr	2	3	1
	cmim	2	3	1
	cife	2	1	3
	icap	2	3	1
	condred	1	2	3
	relief	3	1	2
CART	jmi	2	3	1
	disr	2	3	1
	cmim	2	1	3
	cife	1	2	3
	icap	3	1	2
	condred	1	2	3
	relief	3	1	2

Classifier algorithm	Feature Selection Algorithm	Acoustic Features		
		Spectral		
		audspecRasta-Band 1-26	PCM- Oher Spectral Features	All
NB	jmi	2	3	1
	disr	2	3	1
	cmim	3	2	1
	cife	2	1	3
	icap	3	2	1
	condred	2	1	3
	relief	3	1	2
ANN	jmi	3	2	1
	disr	3	2	1
	cmim	3	2	1
	cife	2	1	3
	icap	3	2	1
	condred	1	2	3
	relief	3	1	2
SVM	jmi	1	3	2
	disr	3	2	1
	cmim	3	2	1
	cife	1	2	3
	icap	3	1	2
	condred	1	2	3
	relief	3	1	2
RF	jmi	2	3	1
	disr	1	3	2
	cmim	3	2	1
	cife	1	3	2
	icap	3	2	1
	condred	1	2	3
	Relief	3	1	2
Average Ranking		**2.19**	**1.98**	**1.83**

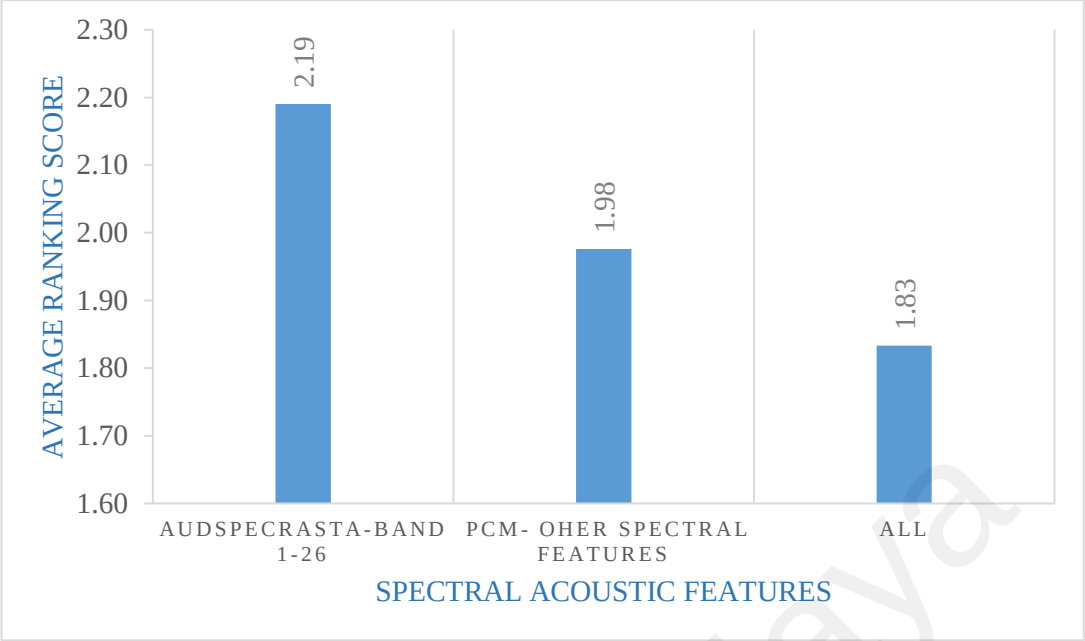

Figure 4.3: The graphic chart for Average Ranking Score for Spectral Acoustic Features

4.2.1.2 The acoustic feature analysis

This part of the analysis is on the performance of four acoustic features, which are prosodic, voice quality, spectral, and cepstral. In each feature, the combination of sub-features is selected to be used for comparison in this experiment.

The best performance to classify the severity level of dysarthric speakers is achieved by the prosodic acoustic features, as shown in Table 4.5, with the average ranking score of 2.21. The ranking score shown in Table 4.5 is calculated from the classification accuracy depicted in Table 4.1. Figure 4.4 depicts the graph chart for the average ranking score of four acoustic features as well as the combination of these acoustic features.

Table 4.5 shows that the combination of the acoustic features have significant achievement as it is placed second, followed by the prosodic acoustic features, with an average ranking score of 2.40. The third-ranking feature is the spectral acoustic features with an average ranking score of 2.50.

Comparing these results showed in Table 4.5 with the results showed in Kim et al. (2015) which used the prosodic and voice quality features for binary classification of dysarthric speakers. The features dimensions used were 6 and 5 features for prosodic and voice quality acoustic features respectively.

The binary classification of intelligibility based on prosodic acoustic features is 71. 3% and 75.5% respectively for unweighted and weighted average recall using the SVM classification algorithms. The LDA classification algorithm obtained 65.3% for unweighted and 69.1% for weighted average recall. The results from this study as listed in Table 4.1 above shows that the prosodic acoustic features obtained 72.39 % and 72.55% average classification accuracy respectively using the SVM and LDA classification algorithms.

For voice quality features, the binary classification of intelligibility is 66.3% and 66.0% respectively for unweighted and weighted average recall using the SVM classification algorithms. The LDA classification algorithms obtained 68.9% for unweighted and 71.7% for weighted average recall. The results from this study as listed in Table 4.1 above show that the voice quality acoustic features obtained 67.39 % and 67.86% average classification accuracy using the SVM and LDA classification algorithms respectively.

Table 4.5: Average Ranking Score for Acoustic Features

Classifier algorithm	Feature Selection Algorithm	Acoustic Features				
		Prosodic	**Voice Quality**	**Spectral**	**Cepstral**	**All**
LDA	jmi	5	4	1	3	2
	disr	5	4	1	3	2
	cmim	1	5	2	4	3
	cife	3	1	2	5	4
	icap	2	4	1	5	3
	condred	5	4	3	2	1
	relief	2	5	1	3	4

Classifier algorithm	Feature Selection Algorithm	Acoustic Features				
		Prosodic	**Voice Quality**	**Spectral**	**Cepstral**	**All**
CART	jmi	1	4	3	5	2
	disr	1	4	3	5	2
	cmim	1	3	4	5	2
	cife	1	3	5	4	2
	icap	1	3	4	5	2
	condred	1	3	5	4	2
	relief	2	5	1	4	3
NB	jmi	3	4	2	5	1
	disr	4	3	2	5	1
	cmim	3	4	1	5	2
	cife	1	2	4	5	3
	icap	3	4	2	5	1
	condred	2	3	5	4	1
	relief	1	5	2	3	4
ANN	jmi	3	4	1	5	2
	disr	4	2	1	5	3
	cmim	1	4	2	5	3
	cife	1	2	4	5	3
	icap	1	4	3	5	2
	condred	2	3	5	4	1
	relief	2	5	1	3	4
SVM	jmi	5	4	1	3	2
	disr	4	3	1	5	2
	cmim	2	4	1	5	3
	cife	3	1	2	4	5
	icap	1	4	2	5	3
	condred	5	3	2	4	1
	relief	3	5	1	2	4
RF	jmi	1	4	3	5	2
	disr	1	2	4	5	3
	cmim	1	4	3	5	2
	cife	1	3	5	4	2
	icap	1	4	3	5	2
	condred	1	3	5	4	2
	Relief	2	5	1	3	4
Average Ranking		2.21	3.57	2.50	4.29	2.43

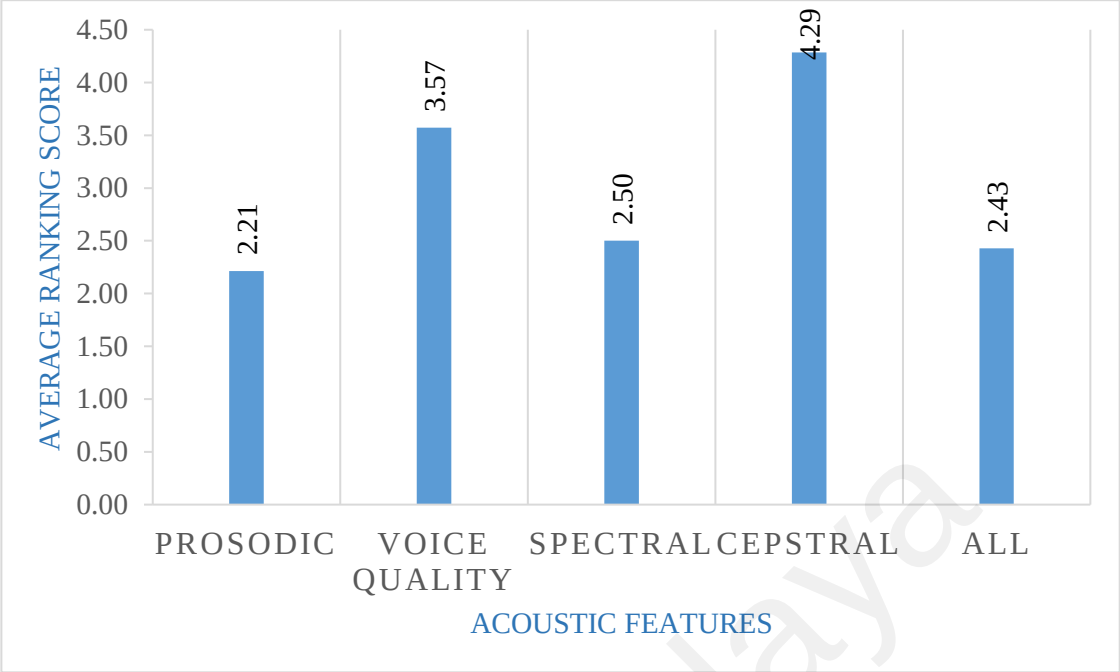

Figure 4.4: Graphic chart for all Acoustic Features Groups

4.2.1.3 Overall acoustic features

The overall acoustic features are listed together to show the comparison of all the acoustic features in classifying the severity level of dysarthric speakers. The comparison includes all of the acoustic features used in this experiment. The sub-features, as well as the combination of the sub-features, are also included. The main objective of this analysis is to report the best performance of the acoustic features for classifying the severity level of dysarthric speakers. The total number of acoustic features used is 17, which includes all features that were discussed in the previous section of this chapter.

The best performance of the overall acoustic features used to classify the severity level of dysarthric speakers is the combination of prosodic acoustic features as it obtained the lowest average ranking score among the overall features as shown in Table 4.6. Figure 4.5 also depicts the graph chart for the average ranking score for all the acoustic features which shows that the lowest ranking score is obtained by the combination of prosodic acoustic features with an average ranking score of 4.48. The second-best performance

was obtained by the sub-features of prosodic acoustic features, named the audspec(Loudness) with an average ranking score of 4.83. The loudness acoustic feature also considers as one of the acoustic features used in the perceptual (subjective) studies to identify the quality of the voice in dysarthric speakers (Hartelius, Runmarker, & Andersen, 2000). Both the combination of all the features of spectral and the overall combination of all the features obtained an average score of 4.95, which ranks them as the third-highest performance for classifying the severity level of dysarthric speakers.

As such, the prosodic, voice quality, spectral and cepstral acoustic features have a significant impact on the classification of dysarthric speakers' severity level. The combination of all the acoustic features achieved a high average ranking score in classifying the severity level of acoustic features. For example, the combination of acoustic features achieved the third-highest average ranking score in all the previous results as well as in the overall acoustic feature analysis where it achieved the first and third highest performance among all of the acoustic features used, as can be seen in Table 4.6 and Figure 4.5.

Paja & Falk (2012) also reported that the combination of the features obtained the highest performance to classifying the spastic severity for dysarthric speakers. The acoustic features selection algorithms used by Paja & Falk (2012) has only 9 acoustic features to use for three classification algorithms, while this study used 13 acoustic features among overall of 5,673 acoustic features (Based on the most related acoustic features).

Table 4.6: Average Ranking Score of overall features

Verification Method	Acoustic Features																
	Prosodic							Voice Quality					Spectral			Cepstral	All
	audspec (Loudness)	audspecRasta-Sum	F0final	PCM-RMS Energy	PCM-Zero-Crossing Rate (ZCR)	voicingFinalUnclipped	All	jitterDDP	JitterLocal	shimmerLocal	logHNR	All	audspecRasta-Band 1-26	PCM- Oher Spectral Features	All	MFCC	
Average Ranking	4.83	10.57	7.24	8.00	13.76	11.19	4.48	16.79	15.88	14.48	6.17	8.40	6.17	5.17	4.95	9.98	4.95

Figure 4.5: The graph chart for Average Ranking Score of overall features

4.2.2 Classification Algorithms Analysis

In the previous sections, the acoustic features are analyzed to identify the best classifier to classify the severity level of dysarthric speakers. In this section, the classification algorithms will be analyzed to report the best classification algorithms for classifying the severity level of the dysarthric speakers.

There are six classification algorithms used to report the best performance to classify the severity level of the dysarthric speakers. The classification algorithms used are LDA, CART, NB, ANN, SVM, and RF Algorithms.

Table 4.7 reports the ranking score obtained from the classification accuracy depicted in Table 4.1. The number of ranking varies from 1 to 42 score according to the number of classification algorithms. There are six classification algorithms and seven features selection method for each classification algorithms, totaling to 42 ranking score (six classification algorithms x seven features selection methods = 42 ranking score).

The average ranking score depicted in Table 4.7 and Figure 4.6 shows that the Random Forest (RF) algorithms with the "relief" features selection method obtain the highest performance for classifying the severity level of dysarthric speakers with the average ranking score of 4.88. The second and third highest performance for classifying the severity level for dysarthric speakers is obtained by RF algorithms with the "cmim" and "icap" features selection algorithms, with an average range of scores of 5.29 and 6.41 respectively. It can also be seen from Figure 4.6 that the RF algorithms obtained the highest performance for classifying the severity level for dysarthric speakers. The Random Forest algorithms are used to identify the most relevant features for the pathophysiology of parkinsonian dysarthria, which also obtains the highest classification accuracy to classify the Parkinson's disease from healthy speakers (Rueda et al., 2019).

Comparing with the results showed in Kim et al. (2015) pronunciation and voice quality for binary classification of dysarthric speakers was varied based on the acoustic features.

The binary classification of intelligibility is 73.5% for unweighted and 72.8% for weighted average recall for the SVM classification which was the best performance for the classifier. The results from this study as listed in Table 4.1 above shows that the SVM classification algorithms obtained 71.96% average classification accuracy. The results in this study were computed as an average classification accuracy rather than the best recognition accuracy as there are seven feature selection algorithms used for each classification algorithms with the highest classification accuracy was 78.97%. The results showed that the RF algorithms obtained high performance as described previously.

Narendra & Alku (2019) used almost the same acoustic features used in this study including the glottal features to classifying dysarthric and non-impaired speakers. The classification accuracy was 94.29% using SVM classification algorithms and 89.64% classification accuracy using RF classification accuracy when comparing the classification accuracy obtained from Narendra & Alku (2019) and this study respectively. The difference between the current study and (Narendra & Alku, 2019), is that the current research classifies the dysarthric speech and non-impaired speech categorized into word, non-word, and sentences.

Table 4.7: Average Ranking Score for all classification algorithms

Classifier	Feature Selection Algorithm	Acoustic Features																	Average Ranking
		Prosodic							Voice Quality					Spectral			Cepstral	All	
		audspec (Loudness)	audspecRasta-Sum	F0final	PCM-RMS Energy	PCM-Zero-Crossing Rate (ZCR)	voicingFinal Unclipped	All	jitterDDP	JitterLocal	shimmerLocal	logHNR	All	audspecRasta-Band 1-26	PCM- Oher Spectral Features	All	MFCC		
LDA	jmi	20	9	35	23	23	15	38	27	30	29	33	38	13	26	23	16	29	25.12
	disr	5	7	33	20	29	13	36	33	11	18	34	32	16	25	19	11	23	21.47
	cmim	19	16	32	14	25	17	18	25	24	21	30	29	12	15	20	21	31	21.71
	cife	39	18	26	33	21	32	41	30	38	31	28	36	38	33	38	42	41	33.24
	icap	26	3	31	13	28	14	20	41	29	28	31	30	4	14	12	38	36	23.41
	condred	30	20	38	35	19	26	42	42	33	32	39	35	30	31	35	23	27	31.59
	relief	3	28	37	1	1	4	6	11	19	3	40	28	3	3	3	3	28	13.00
CART	jmi	38	39	7	22	31	36	16	22	17	23	15	20	37	39	30	31	33	26.82
	disr	27	31	10	19	37	33	21	15	7	12	14	12	31	37	26	34	25	23.00
	cmim	28	35	8	21	27	34	12	18	18	10	20	11	23	20	29	30	26	21.76
	cife	34	40	12	34	33	42	24	24	13	26	27	37	39	42	41	37	37	31.88
	icap	35	37	9	26	38	24	14	23	20	27	9	9	35	24	28	36	12	23.88
	condred	40	41	11	40	41	40	22	21	16	24	25	18	40	41	42	32	22	30.35
	relief	4	23	39	8	7	5	10	2	14	7	38	39	24	5	4	6	17	14.82
NB	jmi	32	38	20	32	12	29	32	16	21	16	19	22	28	34	27	29	11	24.59
	disr	23	30	21	28	15	10	33	17	9	11	23	13	32	30	25	26	10	20.94
	cmim	29	32	19	39	11	16	25	13	12	9	12	14	18	22	14	25	9	18.76
	cife	42	36	13	42	17	38	30	12	26	19	37	31	41	38	37	39	38	31.53
	icap	37	34	23	41	14	25	26	9	10	17	18	15	19	17	16	28	6	20.88
	condred	41	42	25	37	24	39	35	14	22	22	32	34	42	40	40	27	19	31.47

Classifier	Feature Selection Algorithm	Acoustic Features																	Average Ranking
		Prosodic							Voice Quality					Spectral			Cepstral	All	
		audspec (Loudness)	audspecRast a-Sum	F0final	PCM-RMS Energy	PCM-Zero-Crossing Rate (ZCR)	voicingFinal Unclipped	All	jitterDDP	JitterLocal	shimmerLocal	logHNR	All	audspecRast a-Band 1-26	PCM- Oher Spectral Features	All	MFCC		
	relief	6	26	42	27	8	3	15	10	23	6	41	41	21	6	18	5	20	18.71
ANN	jmi	11	24	15	11	36	35	27	35	41	41	11	19	25	18	24	15	18	23.88
	disr	14	8	18	18	34	12	29	29	15	20	7	7	36	21	13	18	30	19.35
	cmim	18	29	17	17	22	22	13	26	28	25	6	8	29	12	17	13	14	18.59
	cife	36	25	22	30	39	30	31	40	31	40	21	33	34	29	39	41	39	32.94
	icap	31	21	24	15	35	23	11	38	37	33	10	6	20	10	11	33	8	21.53
	condred	15	27	14	31	40	37	28	39	40	42	16	21	27	27	34	19	16	27.82
	relief	2	5	41	6	3	2	8	19	25	5	35	40	2	2	2	4	32	13.71
SVM	jmi	13	11	29	10	18	18	39	36	36	35	24	26	5	23	22	17	24	22.71
	disr	16	1	36	12	20	20	34	32	35	38	22	17	15	19	10	24	13	21.41
	cmim	17	15	30	16	16	28	17	28	34	37	26	23	10	9	7	20	34	21.59
	cife	33	13	16	38	32	31	37	31	39	36	29	25	33	35	36	40	42	32.12
	icap	25	4	28	9	30	21	19	37	32	34	17	24	11	8	15	35	21	21.76
	condred	22	12	34	29	26	27	40	34	42	39	36	27	17	28	33	22	35	29.59
	relief	24	33	40	36	42	41	23	20	27	30	42	42	22	4	5	2	40	27.82
RF	jmi	10	17	2	4	4	9	5	8	2	13	3	4	9	13	8	8	5	7.29
	disr	7	10	6	3	6	6	4	4	1	2	5	1	7	16	21	9	7	6.76
	cmim	8	14	3	5	5	8	2	3	5	4	1	2	8	7	6	7	2	5.29
	cife	12	22	5	25	10	19	9	7	8	14	8	10	14	36	31	10	15	15.00
	icap	9	6	1	7	9	7	1	5	3	15	2	3	6	11	9	14	1	6.41
	condred	21	19	4	24	13	11	7	6	4	8	4	5	26	32	32	12	3	13.59
	relief	1	2	27	2	2	1	3	1	6	1	13	16	1	1	1	1	4	4.88

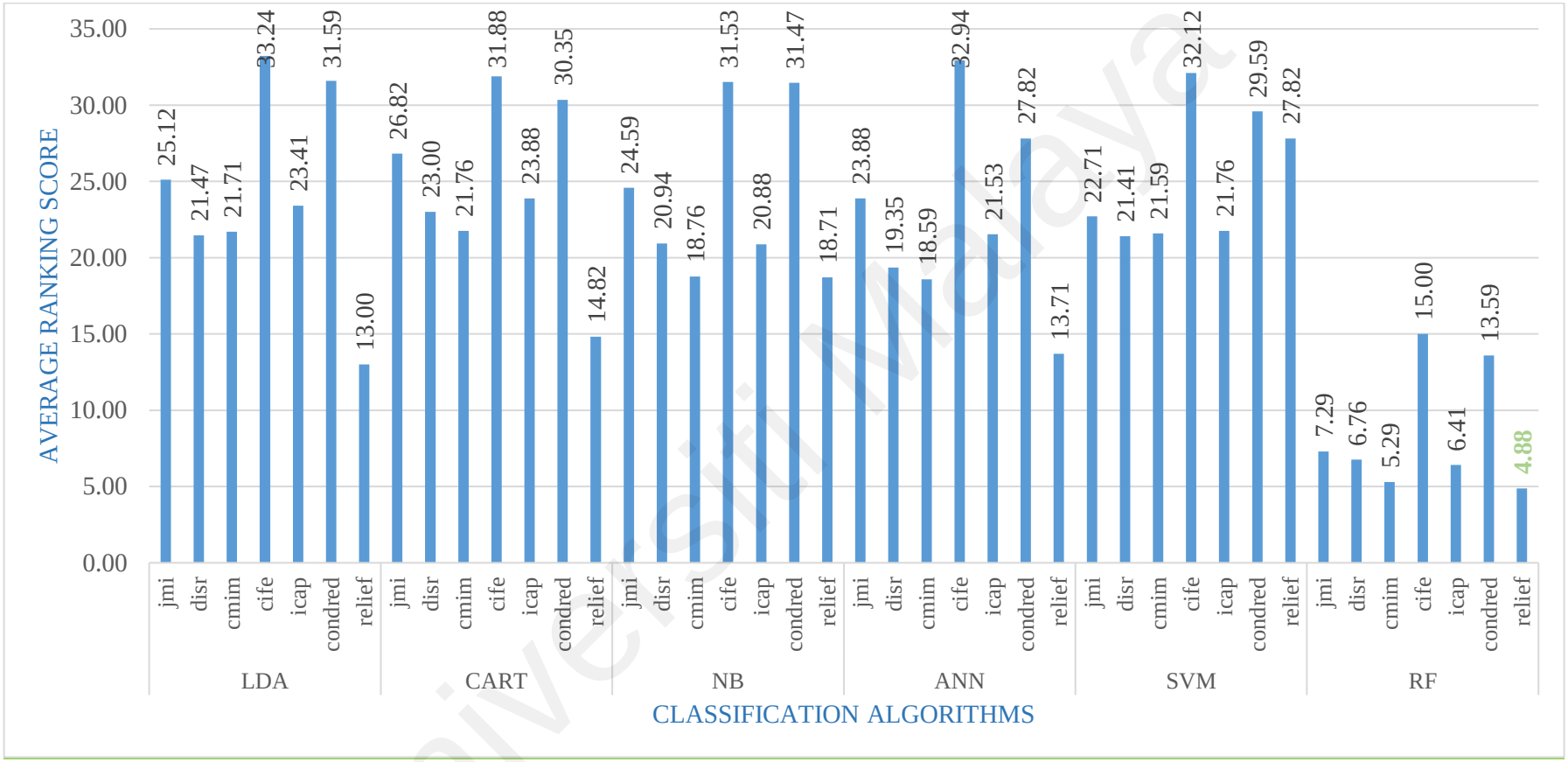

Figure 4.6: The graph chart for Average Ranking Score of all classification algorithms

4.2.3 Statistical Analysis

In this research, the two way ANOVA statistical analysis is used to determine if there is an interaction between the two independent variables on the dependent variable. Thus, it used to compare the means differences between groups that have been split into two independent variables (called factors). The two independents variables in this study are the classification algorithms and the acoustic features, while the dependent variable is the classification accuracy.

Figure 4.7 shows the two way ANOVA analysis obtained based on the classification accuracy listed in Table 4.1. The focus rows in Figure 4.7 are the "CA stands for Classification algorithms", "AF stands for Acoustic Features" and “CA*AF”. These rows inform us whether our independent variables (the "Classification Algorithms CA" and "Acoustic Features AF" rows) and their interaction (the "CA*AF" row) have a statistically significant effect on the dependent variable "classification accuracy ". As can see from the "Sig." column of the "CA*AF" row that there is a statistically significant interaction between classification algorithms CA and acoustic features AF at the $p = .016$ level ($p < 0.5$). Furthermore, both classification algorithms (CA) and acoustic features (AF) have statistically significant as the $p=0.00$ ($p < 0.001$).

Tests of Between-Subjects Effects

Dependent Variable: Classification Accuracy

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	64909.419[a]	101	642.668	18.052	.000	.749
Intercept	3203396.441	1	3203396.441	89978.657	.000	.993
CA	5888.285	5	1177.657	33.079	.000	.213
AF	53037.336	16	3314.833	93.109	.000	.709
CA * AF	3995.090	80	49.939	1.403	.016	.155
Error	21788.263	612	35.602			
Total	3496597.528	714				
Corrected Total	86697.681	713				

a. R Squared = .749 (Adjusted R Squared = .707)

Figure 4.7 : Statistical significance of the two-way ANOVA

4.3 Performance of Automatic Dysarthric Speech Recognition (ADSR)

The results are analyzed in several sections, starting with the effectiveness of using a controlled speech corpus to build the automatic speech recognition engine for dysarthric speech. The adaptation techniques that help improve the performance of the ADSR will be discussed in more detail. The overall average improvement of using the adaptation techniques are compared with the results when no adaptation techniques are used.

4.3.1 The Effectiveness of Using Controlled Speech Corpus for Dysarthric Speech Recognition

Table 4.8 shows the results of increasing data size to the word error rate of the dysarthric speakers. It was found that as the corpus size increases, the word error rate decreases for the dysarthric speakers. Figure 4.8 shows the graphic chart for the results reported in Table 4.8. For example, the word error rate for the mild severity level has decreased from 37.27% when using WSJ1Model to only 34.46%, and 34.01% when using

the WSJ1TimitModel and WSJ1TimitTorgoModel respectively. This improvement in recognition accuracy is also applied to a moderate severity level.

Table 4.8: The Word Error Rate (WER) of using increased data for building the ASR system for dysarthric speakers

Acoustic Model	Size of data	Mild (WER%)	Moderate (WER%)	Severe (WER%)
WSJ1Model	77800 utterances (~73 hours of speech)	37.24	43.99	45.12
WSJ1TimitModel	82420 utterances (~76.14 hours of speech)	34.46	44.89	45.04
WSJ1TimitTorgoModel	86420 utterances (~100.14 hours of speech)	34.01	40.24	45.27

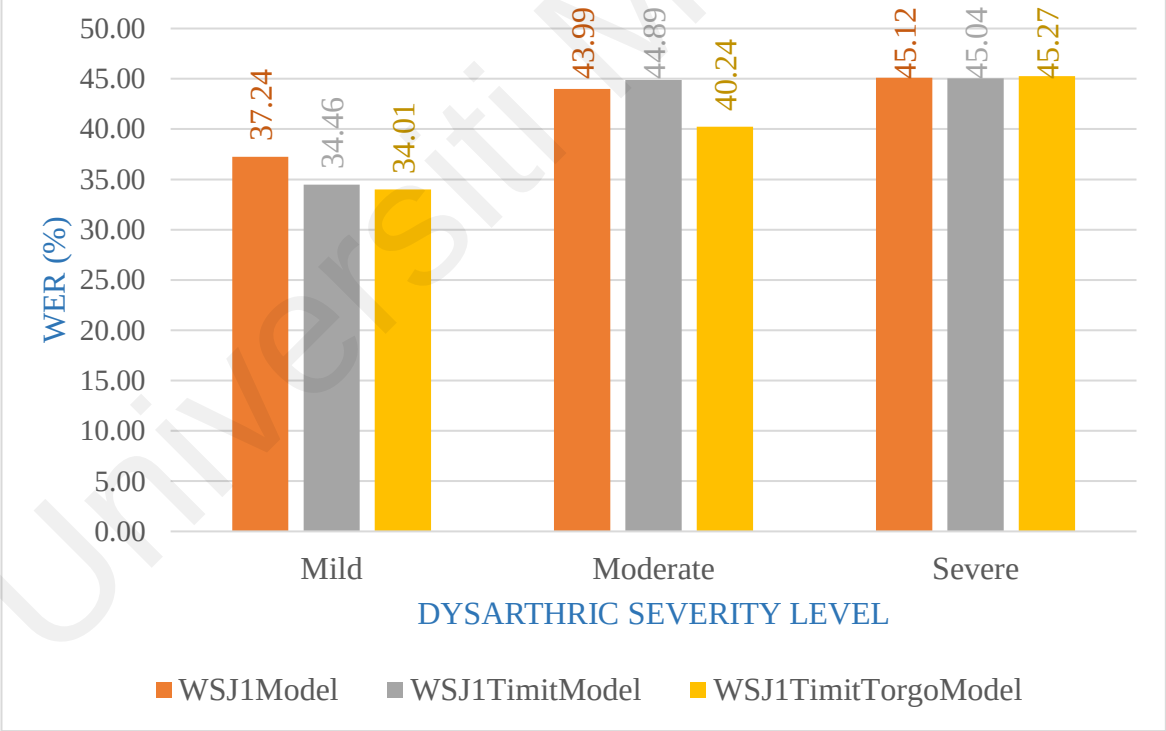

Figure 4.8: Graphic chart of WER of using increasing data for building the ASR system for dysarthric speakers

Table 4.9 shows the overall results of using more corpus to the Word Error Rate (WER) of the dysarthric speakers. It shows from the results depicted in Table 4.9 that the WER dropped from 42.12% when using WSJ1Model (less sample for training the acoustic model), and 41.47% when using the MSJ1TimitModel, to 39.84% when using the WSJ1TimitTorgoModel, with a total improvement of recognition accuracy of 2.28%.

Figure 4.9 shows the graphic chart for the results showed in Table 4.9, which clearly shows the word error rate obtained based on the acoustic model. In general, the results showed that less WER is achieved by using more data to train the acoustic model (meaning: more recognition accuracy achieved). Sriranjani et al. (2015) reported that the WER obtained using the Wall Street Journal (WSJ0) corpus and TI digits for control subjects was almost 36% when testing with Nemours corpus. The acoustic model enriched with speech data from the normal speakers and the adaptation data were used to improve the recognition accuracy of dysarthric speech (Al-Qatab et al., 2014; Mustafa et al., 2014)

Table 4.9: The overall WER of increasing data size for building the ASR system for dysarthric speakers

Acoustic Model	Size of data	WER (%)
WSJ1Model	77800 utterances (~73 hours of speech)	42.12
WSJ1TimitModel	82420 utterances (~76.14 hours of speech)	41.47
WSJ1TimitTorgoModel	86420 utterances (~100.14 hours of speech)	39.84

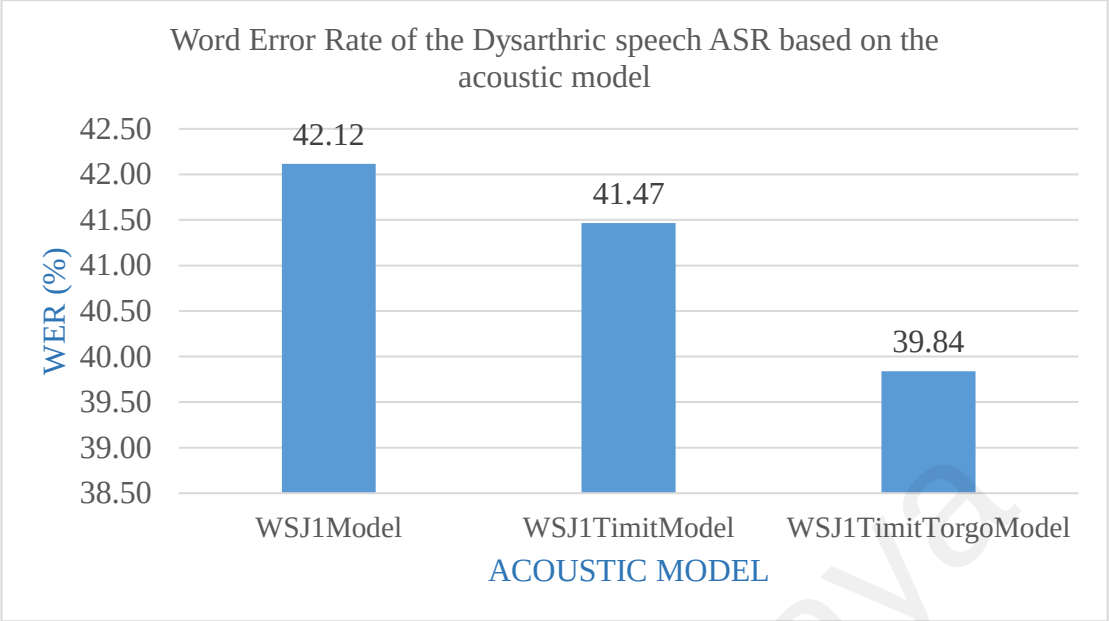

Figure 4.9: Graphic chart of overall WER for increasing data size for building the ASR system for dysarthric speakers

4.3.2 The Effectiveness of Using the Adaptation Data for Dysarthric Speech Recognition

In this section, the third acoustic model was used, which is the WSJ1-Timit-Torgo Model to obtain the results of the effectiveness of using adaptation data to improve the accuracy of the ASR for dysarthric speakers. In this section, there are four adaptation techniques that are used to perform the experiments. Two of them are the core adaptation techniques, which are Maximum Likelihood Linear Regression (MLLR), and Maximum A Posterior (MAP) adaptation techniques. The remaining two of them are the combination of the earlier two techniques in sequential order. They are the MLLR+MAP, which perform the MLLR adaptation, and the results model will be used to perform the MAP adaptation, and the MAP+MLLR, which start with MAP adaptation, and the resulting model will be used to perform the MLLR adaptation.

This section contains subsections that describe the results of the effectiveness of using the adaptation techniques with different severity levels as compared to the results obtained from the previous section with no adaptation techniques used.

4.3.2.1 The results using the MLLR adaptation techniques

Figure 4.10 illustrates the results in Table 4.10 as a graphic chart. The results reported in Table 4.10 show that the performance of the ADSR improved when using the MLLR adaptation techniques where WER dropped from 39.84% to 25.04% on overall severity level with a total of 14.80% of improvement of recognition accuracy. Table 4.10 reported that the more severe the severity level, the more the WER obtained as can be seen from WER of 16.28%, 26.29%, and 32.55% for mild, moderate, and severe severity level respectively.

Table 4.10: The WER for the dysarthric speech using adaptation technique MLLR.

Acoustic Model	Adaptation Technique	Adaptation Data	Mild (WER%)	Moderate (WER%)	Severe (WER%)	All (WER%)
WSJ1TimitTorgo Model	MLLR	NA	34.01	40.24	45.27	39.84
		A	16.28	26.29	32.55	25.04
NA is no adaptation used and A is adaptation used						

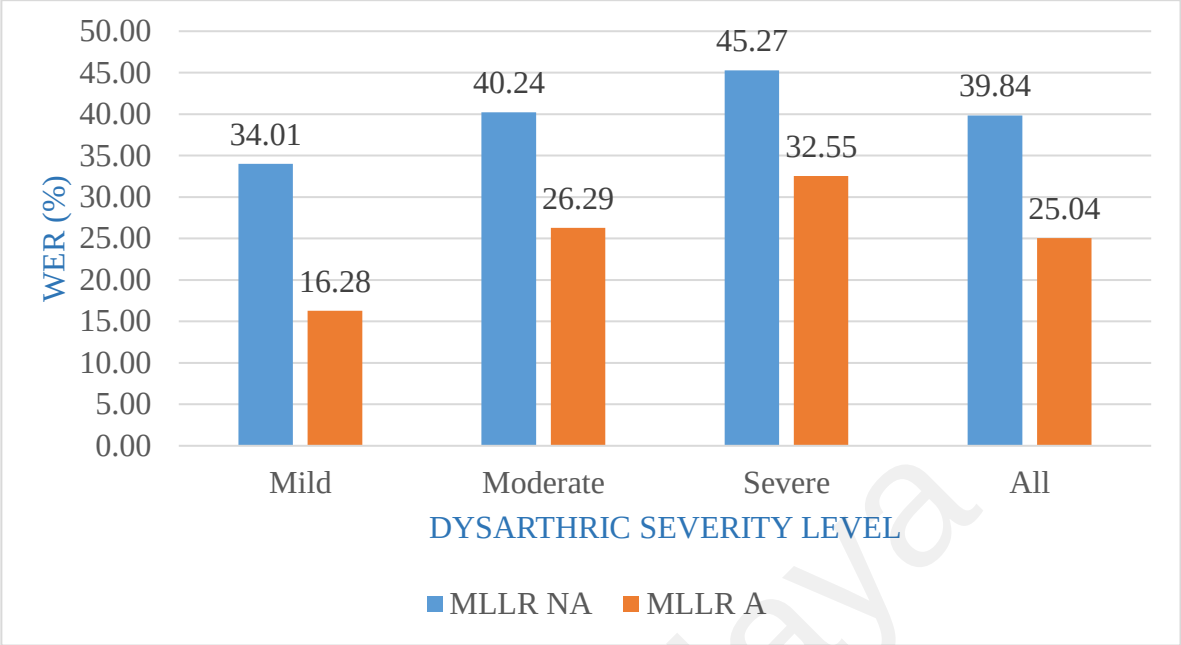

Figure 4.10: Graphic chart of WER for using the MLLR adaptation techniques

4.3.2.2 The results using the MAP adaptation techniques

Table 4.11 shows the results of the MAP adaptation techniques, which also include the results obtained when no adaptation data is used. Figure 4.11 depicts the results in Table 4.11 as a graphic chart that illustrates the WER with and without using the adaptation data. The results reported in Table 4.11 show that the performance of the ADSR improved when using the MAP adaptation techniques which showed the drop of WER from 39.84% to 23.36% on overall severity level with a total of 16.48% of improvement of recognition accuracy. The results showed that the MAP adaptation techniques outperform the MLLR adaptation techniques as it has the most improvement of recognition accuracy, at 16.48% compared to 14.80% recognition accuracy improvement of MLLR adaptation techniques.

Table 4.11: The WER for the dysarthric speech using the MAP adaptation technique

Acoustic Model	Adaptation Technique	Adaptation Data	Mild (WER%)	Moderate (WER%)	Severe (WER%)	All (WER%)
WSJ1TimitTorgo Model	MAP	NA	34.01	40.24	45.27	39.84
		A	14.55	23.60	31.92	23.36
NA is no adaptation used and A is adaptation used						

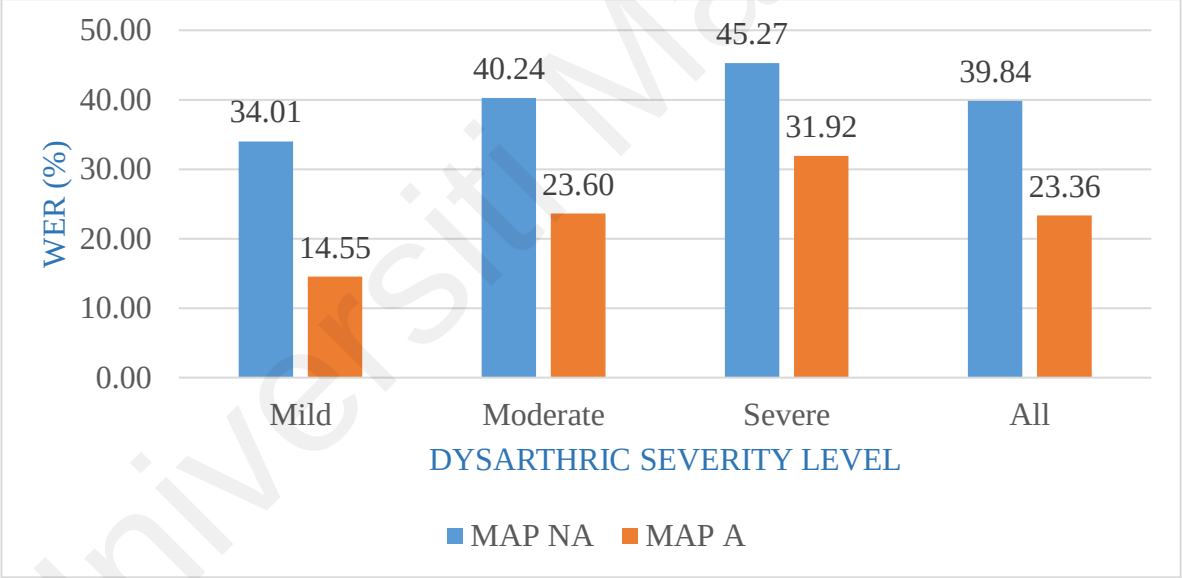

Figure 4.11: Graphic chart of WER for using the MAP adaptation technique

4.3.2.3 The results using the MLLR+MAP adaptation technique

Table 4.12 reports the results obtained from the ADSR system for each level of severity, which is mild, moderate and severe, as well as the overall severity level of dysarthric speakers. The WER decreased from 39.84% on overall severity level to

18.48%, with a total improvement of 21.36% of recognition accuracy. Figure 4.12 illustrates the graph chart for the results in Table 4.12 of WER before and after using the MLLR+MAP adaptation techniques.

Table 4.12: The WER for the dysarthric speech for the adaptation technique MLLR+MAP

Acoustic Model	Adaptation Technique	Adaptation Data	Mild (WER%)	Moderate (WER%)	Severe (WER%)	All (WER%)
WSJ1TimitTorgo Model	MLLR+MAP	NA	34.01	40.24	45.27	39.84
		A	11.46	19.17	24.80	18.48
NA is no adaptation used and A is adaptation used						

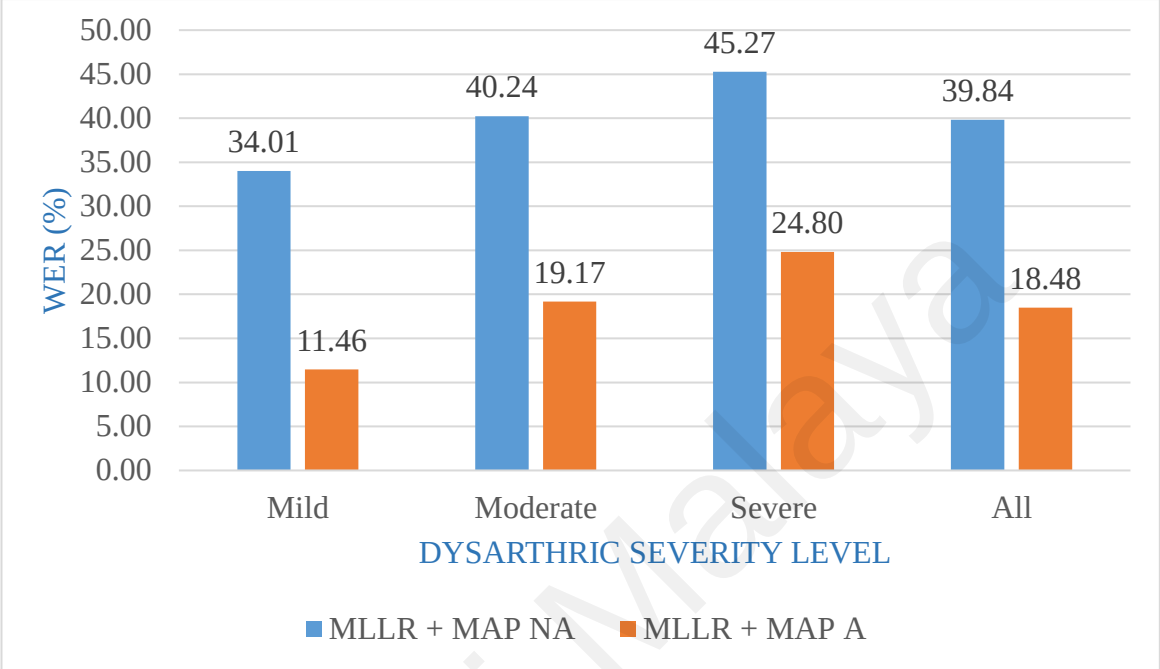

Figure 4.12: Graphic chart of WER for using the MLLR+MAP adaptation techniques

4.3.2.4 The results using the MAP+MLLR adaptation techniques

Table 4.13 shows the results of the combination of MAP+MLLR adaptation techniques as well as a result when adaptation is not performed. Figure 4.13 illustrates the results in Table 4.13 in the form of a graphic chart. The results reported in Table 4.13 show that the performance of the ADSR improved when using the MAP+MLLR adaptation techniques where WER dropped from 39.84% to 20.90% on the overall severity level with a total of 18.94% in the improvement of recognition accuracy.

Table 4.13: The WER for the dysarthric speech obtained using the adaptation techniques MAP+MLLR

Acoustic Model	Adaptation Technique	Adaptation Data	Mild (WER%)	Moderate (WER%)	Severe (WER%)	All (WER%)
WSJ1TimitTorgo Model	MAP + MLLR	NA	34.01	40.24	45.27	39.84
		A	12.98	21.08	28.63	20.90
NA is no adaptation used and A is adaptation used						

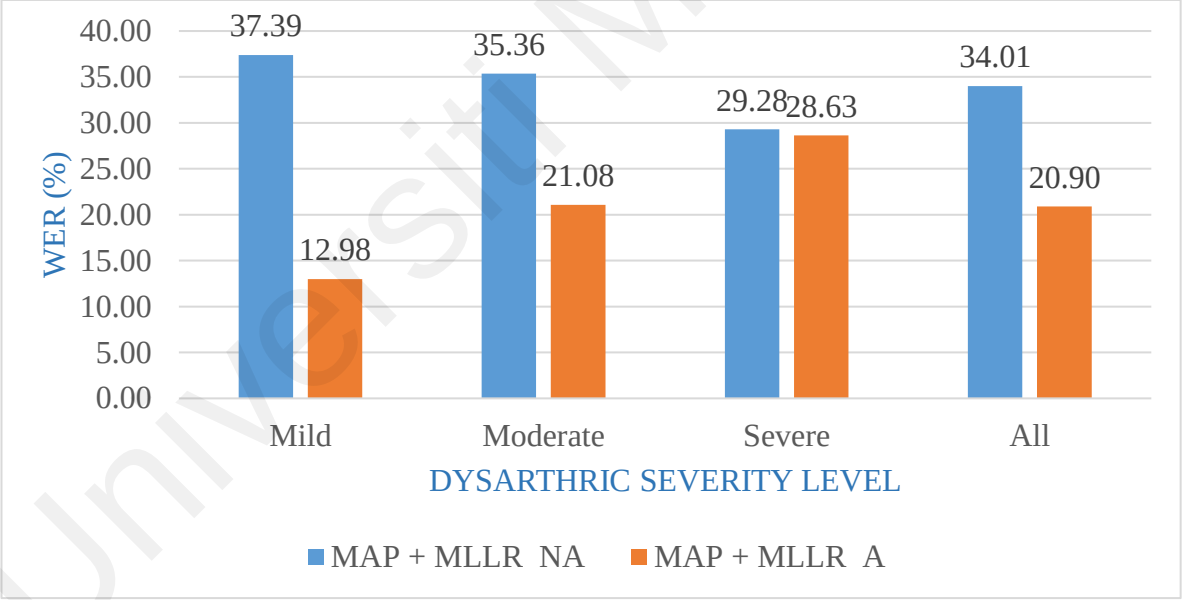

Figure 4.13: Graphic chart of WER using the MAP+MLLR adaptation technique

4.3.2.5 The results using the four adaptation techniques

Table 4.14 depicts the four adaptation techniques, to show the performance of each one independently, and in comparison with other adaptation techniques used in this

experiment. Comparing the WER of the four adaptation techniques showed that the MLLR+MAP outperforms the MLLR, MAP, and MAP+MLLR adaptation techniques. Figure 4.14 represents the graph chart for the results in Table 4.14 and it can be seen that the low WER is achieved by MLLR+MAP adaptation techniques with the WER at 11.46%, 19.17, and 24.80 for mild, moderate, and severe severity level respectively. Figure 4.15 shows the overall WER for the four adaptation techniques where the highest performance is achieved by the combination of different adaptation techniques of MLLR+MAP.

In general, the performance of the combination of the MLLR and MAP outperform standalone adaptation techniques. One possible explanation is that, as a transformation based approach, MLLR has no further improvement at a certain point although there is more adaptation data available (Shinoda, 2011). MLLR usually requires the recorded speech of a new speaker with the use of the same text or sentences recorded from the reference speaker, which is referred to as text-dependent (Digalakis & Neumeyer, 1996). On the other hand, MAP is more efficient as compared to the ML estimation technique when the data size is small. However, as the size of the data increases, the estimation of the parameter for MAP and ML is converging towards an equilibrium point (Kotler & Thomas-Stonell, 1997).

Table 4.14: The WER for the dysarthric speech recognition obtained using all the adaptation techniques used in this experiment

Acoustic Model	Adaptation Technique	Mild (WER%)	Moderate (WER%)	Severe (WER%)	All (WER%)
WSJ1TimitTorgoModel	**MLLR**	16.28	26.29	32.55	25.04
	MAP	14.55	23.60	31.92	23.36
	MLLR + MAP	11.46	19.17	24.80	18.48
	MAP + MLLR	12.98	21.08	28.63	20.90

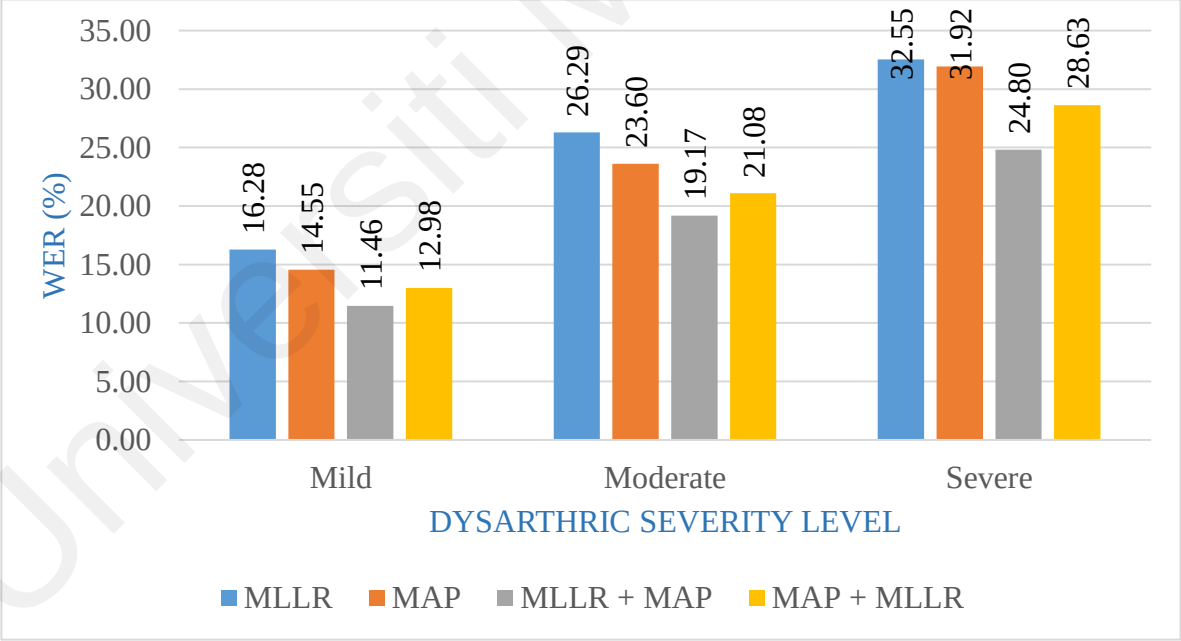

Figure 4.14: Graphic chart of WER for using the four adaptation techniques based on the severity level

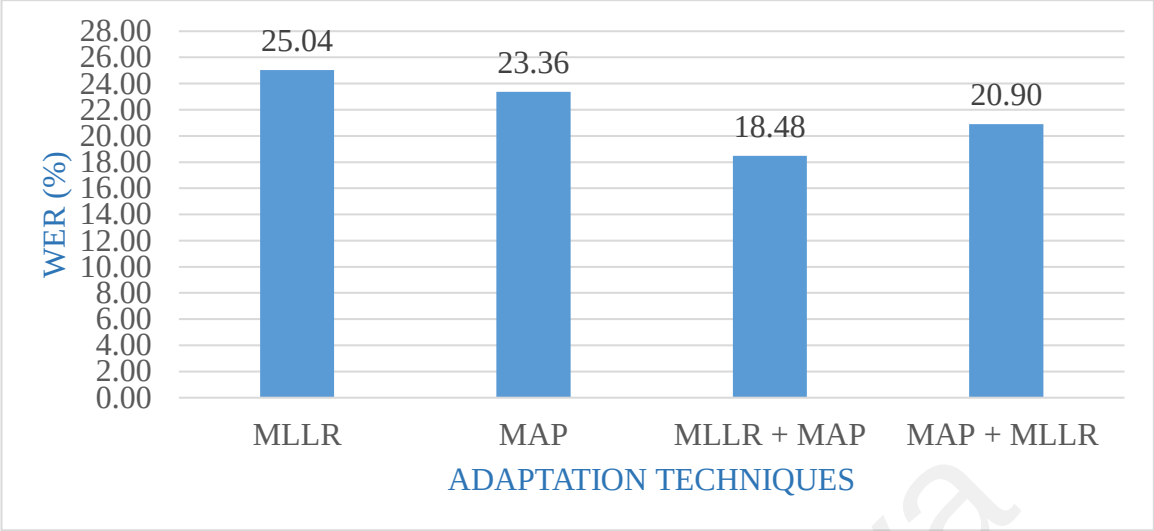

Figure 4.15: Graphic chart of overall WER based on using four adaptation techniques

4.3.2.6 The overall performance of the ADSR system using the adaptation techniques

The overall improvement of AWER will be described in this section according to the severity level of dysarthric speakers and the adaptation techniques used. The average improvement is calculated using the following equation:

$$AI = \frac{Orginal\ Value - New\ Value}{Orginal\ Value} \times 100\% \qquad (4.1)$$

Where AI is the Average Improvement of the performance based on WER, Original Value is the WER of the ADSR without adaptation techniques used, and New Value is the WER of ADSR when applied with the adaptation techniques.

Table 4.15 shows the Average Improvement in a percentage calculated according to equation 7 for all severity levels including all adaptation techniques applied. The results reported in Table 4.15 shows that the mild severity level has an average improvement of more than 52%, which has the highest average improvement, followed by the moderate

severity level with values varying between 34.66% and 52.35%. The severe severity level has the lowest average improvement with values varying from 28.10% to 45.20%. The overall average improvement showed that the MLLR+MAP adaptation techniques outperform the other adaptation techniques with an average improvement of 53.62%, followed by MAP+MLLR with an average improvement of 47.55%. The MLLR and MAP adaptation techniques achieved the lowest average improvement among the overall average improvement, with an average improvement of 37.15% and 41.37% respectively. Figure 4.16 depicts the graph chart for the overall average improvement of the severity levels according to the adaptation techniques used in this experiment.

Table 4.15: The overall average improvement of the ADSR system when using the adaptation techniques

Severity Level	Adaptation Technique	No Adaptation (WER%)	Use adaptation (WER%)	Average Improvement WER(%)
Mild	MLLR	34.01	16.28	52.12
	MAP	34.01	14.55	57.22
	MLLR + MAP	34.01	11.46	66.32
	MAP + MLLR	34.01	12.98	61.83
Moderate	MLLR	40.24	26.29	34.66
	MAP	40.24	23.60	41.34
	MLLR + MAP	40.24	19.17	52.35
	MAP + MLLR	40.24	21.08	47.61
Severe	MLLR	45.27	32.55	28.10
	MAP	45.27	31.92	29.48
	MLLR + MAP	45.27	24.80	45.20
	MAP + MLLR	45.27	28.63	36.76
All	MLLR	39.84	25.04	37.15
	MAP	39.84	23.36	41.37
	MLLR + MAP	39.84	18.48	53.62
	MAP + MLLR	39.84	20.90	47.55

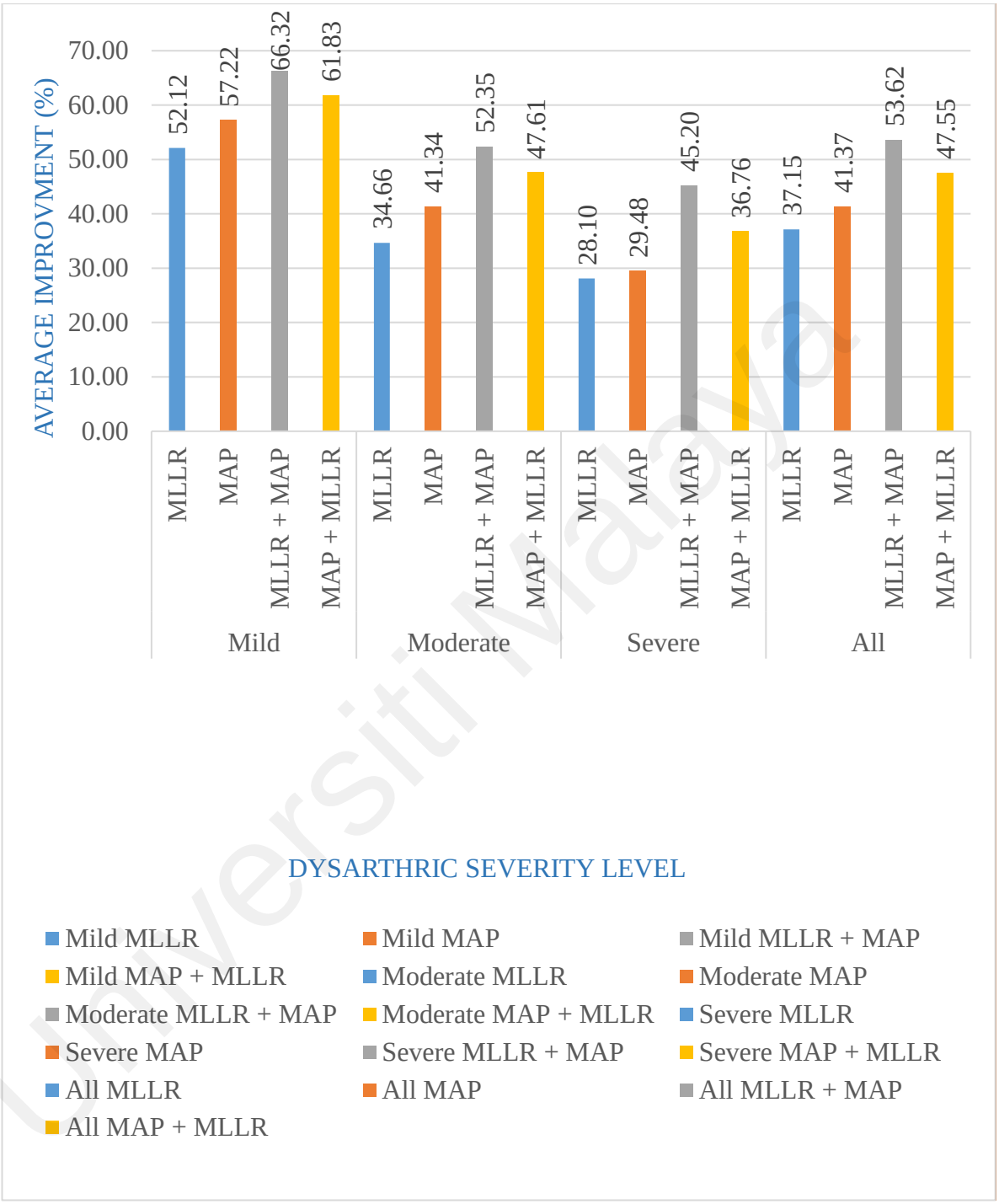

Figure 4.16: Graphic chart of overall improvement of the ADSR system when using the adaptation techniques

4.4 Comparing with other Related Work

This section provides a comparison of the current study with the related studies in regards to the Automatic Dysarthric Speech Recognition System for dysarthric speech. It includes the adaptation techniques used, type of speech, model type, data set used, and the word error rate as compared to the other studies.

Table 4.16 shows that many of the existing works adopted the adaptation techniques (Al-Qatab et al., 2014; Mustafa et al., 2014; Sriranjani et al., 2015). In the acoustic model type, the two types of acoustic models used are speaker-independent and speaker adaptive models. Unimpaired speech is used to enrich the acoustic model and is applied for all similar existing works as shown in Table 4.16. The WER of the current study is compared with the existing works to determine the effectiveness of the proposed technique. The WER of 18.48% by the proposed technique is considerably better than the existing works, where the WER can be reduced by as much as 50% (based on WER of 37.48% from (Al-Qatab et al., 2014)). This study used a large number of speech files to train the acoustic model which includes three corpora WSJ1, TIMIT, and TORGO to enrich the acoustic model that resulted in better performance of the adaptive model compared with other similar works. The results of this study conclude that enriching the acoustic model with the speech files from non-impaired speakers affected the recognition accuracy of dysarthric speech. Furthermore, the combination of the well-known adaptation techniques (MAP and MLLR) for dysarthric speech outperforms the standalone adaptation technique (Shinoda, 2011).

Table 4.16: Comparison of the current study with related studies of the ADSR System

Study	Adaptation Techniques	Type Of Speech	Acoustic Model Type	Testing Data Set	Recognition Accuracy WER (%)
Mustafa et al. (2014)	CMLLR	Continuous speech	Speaker independent and speaker adaptive model	NEMOURS Corpus	33.95%
Al-Qatab et al (2014)	MLLR	Continuous speech	Speaker Independent model	NEMOURS Corpus	37.48%
Sriranjani et al., (2015)	feature space MLLR (fMLLR)	Continuous speech	Speaker Independent model	NEMOURS Corpus	29.83 %
Present Study	MLLR+ MAP	Continuous speech	Speaker independent and speaker adaptive model	NEMOURS Corpus	18.48%

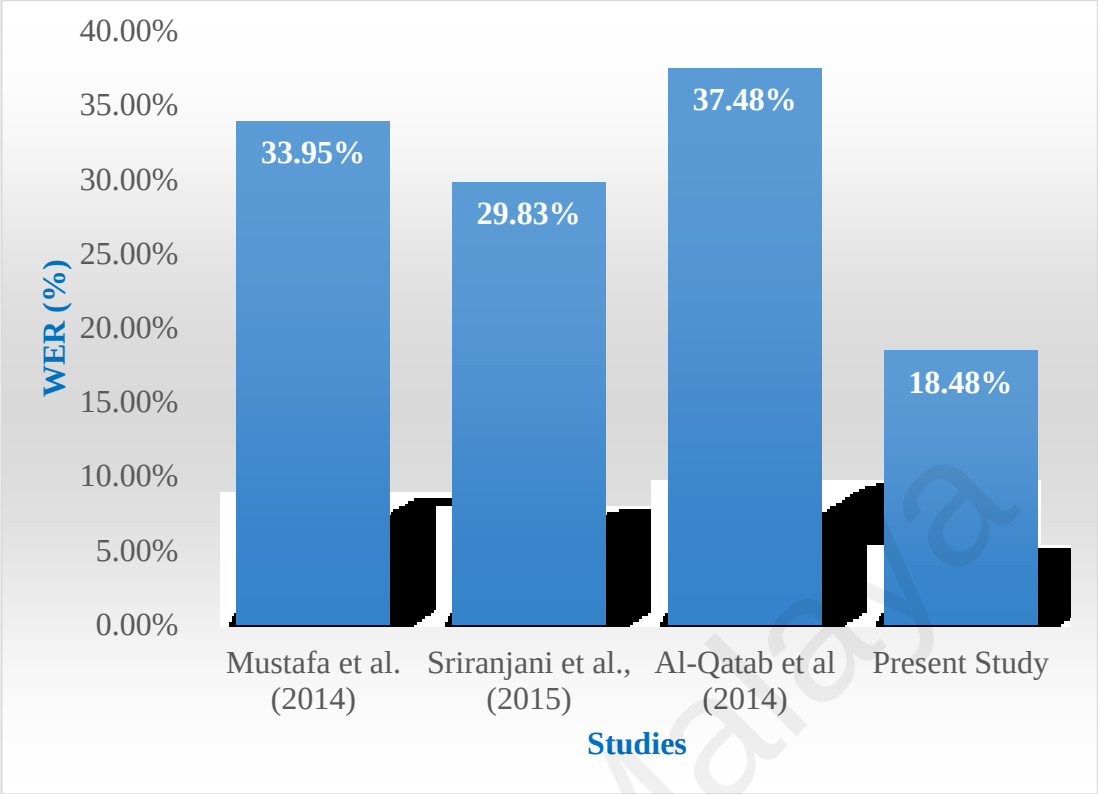

Figure 4.17: WER of the current study compared to the related studies

4.5 Summary

This chapter reports the results of the two processes of the system, the automatic classification of the dysarthric speakers, followed by the automatic dysarthric speech recognition system.

The analysis of the results for the first part, which is the classification of the dysarthric speakers based on severity level (mild, moderate, and severe) was reported in both the acoustic features level and classification algorithms level. By combining all the prosodic acoustic features, acoustic features' performance becomes the best, followed by the prosodic audispec(loudness) feature. In total, the combination of the feature for acoustic features had a competitive highest average accuracy. Furthermore, the Random Forest (RF) classification algorithms with the relief feature selection method have the highest average accuracy rate in classifying the severity level for dysarthric speakers.

The results of the dysarthric automatic speech recognition report that uses more data from non-impaired speakers are to enrich the acoustic model and enhance its performance. On the other hand, the performance of the dysarthric automatic speech recognition improved as the adaptation data based on the severity level is applied to the ADSR acoustic model. The average improvement in the average WER for the ADSR model shows that the adaptation data using the MLLR followed by the MAP (MLLR+MAP) have the highest average improvement for each severity level of dysarthric speakers.

Therefore, the automatic classification of severity level for dysarthric speakers is used in identifying the right adaptation model for each severity level of dysarthric speakers in the dysarthric automatic speech recognition systems. The combination of acoustic features with the Random Forest (RF) classification algorithms as well as using the adaptation techniques MLLR+MAP helps the system obtain the highest performance.

CHAPTER 5: CONCLUSION AND FUTURE WORKS

5.1 Overview

This research's main objective is to propose an automatic classification of dysarthric speech, using severity level based adaptation for automatic speech recognition of dysarthric speech. This chapter summarizes the work that was carried out in this research. The research objectives of this research listed in chapter one are revisited, research contributions, some limitations of this research, and some suggestions for future works are discussed in the following sections.

5.2 Fulfilment of Research Objectives

This section discusses the accomplishments of the research objectives defined for this research.

5.2.1 Research Objective 1

The first objective is to identify the suitable classification algorithms and acoustic features of dysarthria for automatic dysarthric severity level classification. This objective is achieved with the analysis of the findings in section 2.2 of chapter 2, reported in Table 2.4 as well as in chapter 2, sections 2.4 and 2.5 (subsection 2.5.1 and 2.5.2).

The findings describe the classification algorithms applied, the acoustic features, the features selection methods, and the performance of the system for each research. Some of the classification algorithms used are LDA, NB, Linear regression analysis, Gaussian mixture model, and ANN. The numerous features used in the classification algorithms are listed in Table 2.4 of chapter 2, which includes features like F0, MFCC, duration, prosody features, voice quality, and HNR features. For features selection, there are 5 out of 15 researches listed in Table 2.4. The reason for not using the feature selection methods is that this research applied a small number of features for classification algorithms.

The selection of the speech corpus, acoustic feature selection as well as the classification algorithms used to apply the intra-severity classification has been described in chapter 3 sections 3.5.1.1, 3.5.1.4, and 3.5.1.5 respectively.

The four research questions to be answered by this objective are:

RQ1: What is the importance of dysarthric speech severity level classification?

(Kim, Kent, & Weismer, 2011) stated that the classification accuracy using severity level and disease type outperforms the classification accuracy using the severity types. Also, the correlation between intelligibility and severity level of dysarthric speakers as reported in (Kayasith & Theeramunkong, 2009) encouraged the researcher to classify the dysarthric speakers based on severity levels.

RQ2: What are the acoustic features that affect the dysarthric speech severity level classification?

Table 2.4 summarizes the acoustic features used by researchers which shows that the acoustic features are used in the classification of dysarthric speakers. For example, the prosodic acoustic features and voice quality acoustic features, and the combination of both features were used by (Kim, Kumar, Tsiartas, Li, & Narayanan, 2015) for automatic intelligibility classification for dysarthric speakers and (Paja & Falk, 2012) for spastic severity disorder classification.

RQ3: What is the statistical function that can be used to determine the dimensional of features vector for each acoustic feature?

The statistical function is used by researchers to make the dimensional of the feature vector for acoustic features as in (Schlenck, Bettrich, & Willmes, 1993) is the prosodic

acoustic feature to distinguish between normal and dysarthric speakers. In (Eyben, 2015), large statistical features were extracted where this research has adopted several of them as listed in Table 2.4, chapter 2.

RQ4: What is the effect of the reduction of the statistical functional per acoustic feature?

A large number of statistical features extracted from each acoustic features required us to search for a method to reduce this large number, which can be achieved by using the feature selection algorithms that has the ability to rank the feature based on their relevance to classification type used in the training set (Kuncheva, 2007). Besides that, this research used the logarithmic base to total the number of features for better computation cost (Khoshgoftaar, Golawala, & Van Hulse, 2007; Samsudin, Shafri, Hamedianfar, & Mansor, 2015) described in section 3.5.1.3 of chapter 3.

5.2.2 Research Objective 2

The second objective is to identify the suitable adaptation techniques in relation to data size and level of severity of the dysarthric speech towards improvement in recognition accuracy of dysarthric speech recognition. This objective is achieved with the analysis of findings in section 2.3 of chapter 2 and reported in Table 2.5 as well as in chapter 2 section 2.4 and section 2.5.3. The findings described the automatic dysarthric speech recognition systems, and the adaptation techniques used. The combination of the different adaptation techniques were used in the previous researches like in (Dhanalakshmi & Vijayalakshmi, 2015), which are the CMLLR+MAP adaptation techniques for improving the intelligibility of dysarthric speakers as well as using the feature space MLLR adaptation as in (Bhat, Vachhani, & Kopparapu, 2016) to improve the recognition accuracy for

dysarthric speakers which provided an improvement of speaker-adapted systems of 55.81% and 63.67% for GMM-HMM-TA and DNN-HMM-TA respectively.

The used of enriching acoustic models to apply the adaptation techniques and the amount of adaptation data set used to improve the automatic dysarthric speech recognition system has been described in chapter 2 section 2.4, section 2.5.3 and chapter 3 section 3.5.2.4.

RQ1: What are the best adaptation techniques that obtain the highest recognition accuracy?

Bhat et al., (2016) and Sriranjani, Ramasubba Reddy, & Umesh (2015) reported that using the feature space MLLR based on speaker adaptive training improves the baseline acoustic model for dysarthric speakers which reaches up to 50% over the recognition of baseline acoustic model.

RQ2: What is the effect on ADSR's recognition accuracy by increasing the amount of adaptation data?

It can be seen from the studies that the standalone adaptation techniques have a limitation on improving the recognition accuracy of the ASR system. For example, in a transformation based approach, MLLR has no improvement up to a certain point, although more adaptation data is available (Shinoda, 2011). MLLR usually requires the recording of sentences for a new speaker with the same text recorded for the reference speakers, which is referred to as text-dependent (Digalakis & Neumeyer, 1996). However, for adaptation of the MAP, the size of the acoustic model can be adjusted to the data adaptation amount, with an update of each Gaussian component of the system. Moreover, it can function in a wide variability of pronunciation, like differences of

phonemes pronunciation standards, dialects, and foreign accents, since each phoneme is processed separately. One disadvantage of MAP is the poor estimation of the parameter of the correct acoustic transformation when there is limited or unsupervised adaptation data (Gales, 2001). MAP needs a large amount of adaptation data to fully update all the phonemes separately (Goronzy & Kompe, 1999).

5.2.3 Research Objective 3

The third objective is to design and develop the intra-severity automatic dysarthric speech recognition system using the identified classification and adaptation techniques in objectives 1 and 2. This objective is achieved by using a sequence of experiments as discussed in sections 3.5.1 and 3.5.2 of chapter 3. In Figure 3.2, the overall system development is depicted. The details of the first stage of the proposed system, which is the classification part, is described in detail in section 3.5.1. The acoustic features, statistical functions, features selection, and classification algorithms are described in detail in section 3.5.1. Section 3.5.2 describes the acoustic model and the adaptation techniques in detail. The experimental design and settings for both parts are shown in chapter 3.

> RQ1: What are the best classifier and adaptation techniques that can be used in tandem, to design and implement the proposed system for the improvement of the recognition accuracy of the automatic dysarthric speech recognition system (ADSR)?

The results analysis in Table 4.6 showed that the combination of prosodic acoustic features have the best performance as it has the lowest average ranking score of 4.48, followed by loudness plus prosodic acoustic features with an average ranking score of 4.83. The combination of all spectral acoustic features and the combination of all acoustic

features have the same average ranking score of 4.95. Also, the average ranking score depicted in Table 4.7 showed that the Random Forest (RF) algorithms with the "relief" feature selection method obtained the highest performance for classifying the severity level of dysarthric speakers with an average ranking score of 4.88. The second and third highest performance for classifying the severity level for dysarthric speakers were obtained by the RF algorithms with the "cmim" and "icap" features selection algorithms with an average ranking score of 5.29 and 6.41 respectively.

For the automatic dysarthric speech recognition (ADSR), the average improvement of the WER showed that the MLLR+MAP has the best performance due to its highest average improvement for each severity level of 66.32%, 52.35%, and 45.20% for mild, moderate, and severe severity level respectively, as well as the combination of all severity with an average improvement of 53.62%, as shown in Table 4.15.

5.2.4 Research Objective 4

The fourth objective is to evaluate the performance of the developed intra-severity automatic dysarthric speech recognition system by comparing it with the baseline acoustic model. This objective is achieved by applying the measurement techniques on each part of the system and by comparing the results with other related work.

RQ1: What are the measurements used to evaluate the classification accuracy and recognition accuracy of the severity level automatic dysarthric speech recognition system?

The measurements used in this research varied based on the part of the system. In the classification part, the K-Fold with (K=10) has been used for recognition accuracy of the severity level of dysarthric speech, depicted in Table 4.1 of chapter 4. The average

ranking score was obtained and reported in detail for a different analysis of the results as described in section 4.2.

For the ADSR, the Word Error Rate (WER) is used as the main measurement for the recognition accuracy. For evaluation of the system in comparison to baseline ADSR system, and to report the adaptation techniques' best performance, the average improvement of WER of the ADSR system was used, which showed the percentage improvement of each adaptation techniques compared to the percentage when no adaptation techniques were applied.

> RQ2: How are the results of the proposed system when compared to other baseline methods in terms of classification accuracy, recognition accuracy and the combination of both?

The proposed system is considered to be an effective solution as it obtained considerable performance among all the related studies. Section 4.4 showed a comparison of this study with related existing studies with regards to the two parts of the proposed system. It can be seen from the comparison that the proposed system performed well when compared with other related studies.

5.3 Research Contributions

The current research contributes to the field of ADSR by improving its recognition accuracy by applying the combined adaptation techniques based on severity level, as well as proposing the automatic classification of dysarthric speech based on severity level, which can be used to automatically assign the severity level to the corresponding adapting severity level model in the ADSR system. The contribution of this study can be listed as the following:

- Propose a design for intra-severity classification and adaptation techniques to help improve the recognition accuracy of the ADSR system for dysarthric speech.
- The proposed method helps to reduce the computation cost to the development of the ADSR system which mostly used in assistive technology to enhance dysarthric speaker's communication skills. The generalization of the severity level classification and adaptation proposed to reduce both the dimension of classification and the number of adaptive acoustic models of the ADSR system.
- Investigate a large acoustic feature space (around 6,000 features) and its affection on the classification of the severity level of dysarthric speech. Furthermore, ranking of the acoustic feature using the feature selection methods which help to identify the most suitable acoustic features for severity level classification.
- Introduce an automatic classification for the severity level of dysarthric speech which helps to the automatic election of the adaptive acoustic model for the ADSR system.
- The proposed method helps the ADSR system to use the available unimpaired corpus to develop the acoustic speech rather than used impaired corpus which is very limited.

5.4 Research Limitation

- This research focuses on the severity level of the dysarthric speech, which is mild, moderate, and severe, for both classification and adaptation techniques of the automatic dysarthric speech recognition system. In both the classification and adaptation for dysarthric speakers, the training and testing

stage the NEMOURS dysarthric speech corpus was used for. As for the results, it may be different when using other databases.

- The difficulties in obtaining data from dysarthric speakers in general, as well as finding the free corpus for dysarthric speakers, which include balancing the samples for severity level of dysarthric speakers, make our research limited to the available corpus, which is the NEMOURS corpus.
- All dysarthric speakers included in this research are young adult males.

5.5 Suggestions for Future Works

This section provides some suggestions as a result of carrying out this research, which can help to enhance the performance of the ADSR system.

- Include more corpora that are classified as severity level to enrich the speaker adaptation model with more speaker's variability, to help reduce the mismatch between the testing data and training data, so as to lead to more accurate results.
- Improve this proposed model to be included in assistive technology to help dysarthric speakers improve their communication skills for a better quality of life.

  Study the specific statistical function of acoustic features that is more effective in distinguishing between dysarthric speakers and normal speakers for better computational cost.
- Conducting more investigation of the acoustic features and its effectiveness on the dysarthric severity level which can help to identify the severity level of dysarthric speakers. This can help pathologies to automatically identify the level of severity to determine the suitable intervention plan.

- Apply more complicated classification algorithms like deep learning algorithms or ensemble classification for the classification part to enhance the classification accuracy.
- Enhance the proposed method to be applied in web-based applications. This will be required to apply the server-based acoustic model for dysarthric speech. Thus, the accessibility of the ADSR will be easy for both pathologists and parents to help improve the communication of dysarthric speakers.

REFERENCES

Al-Qatab, B. A., Mustafa, M. B., & Salim, S. S. (2014). *Severity Based Adaptation for ASR to Aid Dysarthric Speakers.* Paper presented at the 2014 8th Asia Modelling Symposium, Kuala Lumpur, Malaysia.

American Speech-Language-Hearing Association. (1993). Definitions of communication disorders and variations. Retrieved from http://www.asha.org/policy/RP1993-00208/

American Speech-Language-Hearing Association. (2007). Childhood apraxia of speech [Position Statement]. Retrieved from http://www.asha.org/policy/PS2007-00277/

Ansel, B. M., & Kent, R. D. (1992). Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech, Language, and Hearing Research, 35*(2), 296-308.

Auzou, P., Ozsancak, C., Morris, R. J., Jan, M., Eustache, F., & Hannequin, D. (2000). Voice onset time in aphasia, apraxia of speech and dysarthria: a review. *Clinical Linguistics & Phonetics, 14*(2), 131-150.

Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing, 18*, 1-8.

Bhat, C., Vachhani, B., & Kopparapu, S. (2016). Improving Recognition of Dysarthric Speech Using Severity Based Tempo Adaptation. In A. Ronzhin, R. Potapova, & G. Németh (Eds.), *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings* (pp. 370-377). Cham: Springer International Publishing.

Bowen, A., Hesketh, A., Patchick, E., Young, A., Davies, L., Vail, A., . . . Tyrrell, P. (2012). Effectiveness of enhanced communication therapy in the first four months after stroke for aphasia and dysarthria: a randomised controlled trial. *BMJ : British Medical Journal, 345*. doi:10.1136/bmj.e4407

Brazdil, P. B., & Soares, C. (2000). *A comparison of ranking methods for classification algorithm selection.* Paper presented at the European conference on machine learning.

Breiman, L., Friedman, J., & Olshen, R. (1984). Stone. In: CJ.

Brown, G., Pocock, A., Zhao, M.-J., & Luján, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research, 13*(Jan), 27-66.

Bunton, K., Kent, R. D., Kent, J. F., & Duffy, J. R. (2001). The effects of flattening fundamental frequency contours on sentence intelligibility in speakers with dysarthria. *Clinical Linguistics & Phonetics, 15*(3), 181-193.

Bunton, K., Kent, R. D., Kent, J. F., & Rosenbek, J. C. (2000). Perceptuo-acoustic assessment of prosodic impairment in dysarthria. *Clinical Linguistics & Phonetics, 14*(1), 13-24.

Center for Parent Information And Resources. (2011). Speech and Language Impairments(A legacy disability fact sheet from NICHCY). Retrieved from http://www.parentcenterhub.org/repository/speechlanguage/

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST), 2*(3), 27.

Chen, H., & Stevens, K. N. (2001). An acoustical study of the fricative/s/in the speech of individuals with dysarthria. *Journal of Speech, Language, and Hearing Research, 44*(6), 1300-1314.

Christensen, H., Casanueva, I., Cunningham, S., Green, P., & Hain, T. (2014). *Automatic selection of speakers for improved acoustic modelling: recognition of disordered speech with sparse data.* Paper presented at the Spoken Language Technology Workshop (SLT), 2014 IEEE.

Colorado (2017). Speech or Language Impairment (SLI), Colorado ECEA Rules [2.08(7)],. Retrieved from http://www.cde.state.co.us/cdesped/SD-SLI.asp

Constantinescu, G., Theodoros, D., Russell, T., Ward, E., Wilson, S., & Wootton, R. (2010). Assessing disordered speech and voice in Parkinson's disease: a telerehabilitation application. *International Journal of Language & Communication Disorders, 45*(6), 630-644.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969a). Clusters of Deviant Speech Dimensions in the Dysarthrias. *Journal of Speech, Language, and Hearing Research, 12*(3), 462-496. doi:10.1044/jshr.1203.462

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969b). Differential Diagnostic Patterns of Dysarthria. *Journal of Speech, Language, and Hearing Research, 12*(2), 246-269. doi:10.1044/jshr.1202.246

De Bodt, M. S., Hernández-Díaz Huici, M. E., & Van De Heyning, P. H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders, 35*(3), 283-292. doi:http://dx.doi.org/10.1016/S0021-9924(02)00065-5

DeGroot, M. H. (2005). *Optimal statistical decisions* (Vol. 82): John Wiley & Sons.

Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends® in Signal Processing, 7*(3–4), 197-387.

Depenau, J. (1995). Automated design of neural network architecture for classification. *DAIMI Report Series, 24*(500).

Despotovic, V., Walter, O., & Haeb-Umbach, R. (2018). Machine learning techniques for semantic analysis of dysarthric speech: An experimental study. *Speech Communication, 99*, 242-251.

Deterding, D. (2001). The measurement of rhythm: a comparison of Singapore and British English. *Journal of Phonetics, 29*(2), 217-230. doi:http://dx.doi.org/10.1006/jpho.2001.0138

Dhanalakshmi, M., & Vijayalakshmi, P. (2015). *Intelligibility modification of dysarthric speech using HMM-based adaptive synthesis system.* Paper presented at the 2015 2nd International Conference on Biomedical Engineering (ICoBE).

Dibazar, A. A., Berger, T. W., & Narayanan, S. S. (2006). *Pathological Voice Assessment.* Paper presented at the EMBS '06. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2006. .

Digalakis, V. V., & Neumeyer, L. G. (1996). Speaker adaptation using combined transformation and Bayesian methods. *Speech and Audio Processing, IEEE Transactions on, 4*(4), 294-300. doi:10.1109/89.506933

Doh-Suk, K. (2004). A cue for objective speech quality estimation in temporal envelope representations. *Signal Processing Letters, IEEE, 11*(10), 849-852. doi:10.1109/LSP.2004.835466

Doyle, P. C., Leeper, H. A., Kotler, A.-L., Thomas-Stonell, N., O'Neill, C., Dylke, M.-C., & Rolls, K. (1997). Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of Rehabilitation Research and Development, 34*, 309-316.

Duffy, J. R. (2006). History, current practice, and future trends and goals. *Motor speech disorders*, 7-56.

Enderby, P. (1980a). Frenchay dysarthria assessment. *International journal of language & communication disorders, 15*(3), 165-173.

Enderby, P. (1980b). Frenchay dysarthria assessment. *British Journal of Disorders of Communication, 15*(3), 165-173.

Eyben, F. (2015). *Real-time speech and music classification by large audio feature space extraction*: Springer.

Eyben, F., Weninger, F., Gross, F., Bj, #246, & Schuller, r. (2013). *Recent developments in openSMILE, the munich open-source multimedia feature extractor*. Paper presented at the Proceedings of the 21st ACM international conference on Multimedia, Barcelona, Spain.

Falk, T., Chan, W.-Y., & Shein, F. (2012). Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication, 54*(5), 622-631.

Falk, T. H., & Wai-Yip, C. (2010). Temporal Dynamics for Blind Measurement of Room Acoustical Parameters. *Instrumentation and Measurement, IEEE Transactions on, 59*(4), 978-989. doi:10.1109/TIM.2009.2024697

Ferrier, L. (1991). Clinical study of a dysarthric adult using a touch talker with words strategy. *Augmentative and Alternative Communication, 7*(4), 266-274. doi:10.1080/07434619112331276003

Ferrier, L., Shane, H., Ballard, H., Carpenter, T., & Benoit, A. (1995). Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication, 11*(3), 165-175. doi:10.1080/07434619512331277289

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of machine learning research, 5*(Nov), 1531-1555.

Fonville, S., Worp, H. B., Maat, P., Aldenhoven, M., Algra, A., & Gijn, J. (2008). Accuracy and inter-observer variation in the classification of dysarthria from speech recordings. *Journal of Neurology, 255*(10), 1545-1548. doi:10.1007/s00415-008-0978-4

Freed, D. B. (2012). *Motor Speech Disorders: Diagnosis & Treatment* (Second Edition ed.): Cengage Learning.

Fritzke, B. (1994). Growing cell structures—a self-organizing network for unsupervised and supervised learning. *Neural networks, 7*(9), 1441-1460.

Furui, S. (2005). 50 years of progress in speech and speaker recognition. *SPECOM 2005, Patras*, 1-9.

Gale, R., Chen, L., Dolata, J., van Santen, J., & Asgari, M. (2019). Improving ASR Systems for Children with Autism and Language Impairment Using Domain-Focused DNN Transfer Techniques. *Proc. Interspeech 2019*, 11-15.

Gales, M. J. F., & Woodland, P. C. (1996). Mean and variance adaptation within the MLLR framework. *Computer Speech & Language, 10*(4), 249-264. doi:http://dx.doi.org/10.1006/csla.1996.0013

Gao, W., Hu, L., & Zhang, P. (2018). Class-specific mutual information variation for feature selection. *Pattern Recognition, 79*, 328-339.

Garofolo, J. S., & Consortium, L. D. (1993). *TIMIT: acoustic-phonetic continuous speech corpus*: Linguistic Data Consortium.

Gauvain, J., & Lee, C. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *Speech and Audio Processing, IEEE Transactions on, 2*(2), 291-298. doi:10.1109/89.279278

Gauvain, J. L., & Lee, C. H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *Speech and Audio Processing, IEEE Transactions on, 2*(2), 291-298.

Goberman, A. M., Coelho, C. A., & Robb, M. P. (2005). Prosodic characteristics of Parkinsonian speech: The effect of levodopa-based medication. *Journal of Medical Speech-Language Pathology, 13*(1), 51-69.

Godino-Llorente, J. I., Gómez-Vilda, P., & Blanco-Velasc, M. (2006). Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters. *Biomedical Engineering, IEEE Transactions on, 53*(10), 1943-1953. doi:10.1109/TBME.2006.871883

Godino-Llorente, J. I., Osma-Ruiz, V., Sáenz-Lechón, N., Gómez-Vilda, P., Blanco-Velasco, M., & Cruz-Roldán, F. (2010). The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders. *Journal of Voice, 24*(1), 47-56. doi:http://dx.doi.org/10.1016/j.jvoice.2008.04.006

Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M. S., & Parker, M. (2003). *Automatic speech recognition with sparse training data for dysarthric speakers.* Paper presented at the INTERSPEECH.

Guerra, E. C. (2002). *A modern approach to dysarthria classification.* (Ph.D). University of New Brunswick, Canada.

Guerra, E. C., & Lovey, D. F. (2003). *A modern approach to dysarthria classification.* Paper presented at the Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2003.

Guerreiro, P. M. C. (2008). Linear Discriminant Analysis Algorithms. *Unpublished master's thesis). Technical University of Lisbon, Portugal*.

Haeb-Umbach, R., & Ney, H. (1992). *Linear discriminant analysis for improved large vocabulary continuous speech recognition.* Paper presented at the Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on.

Hamidi, F., Baljko, M., Livingston, N., & Spalteholz, L. (2010). CanSpeak: a customizable speech interface for people with dysarthric speech. *Computers Helping People with Special Needs*, 605-612.

Hartelius, L., Runmarker, B., & Andersen, O. (2000). Prevalence and characteristics of dysarthria in a multiple-sclerosis incidence cohort: relation to neurological data. *Folia Phoniatrica et Logopaedica, 52*(4), 160-177.

Hawley, M. S. (2002). Speech recognition as an input to electronic assistive technology. *The British Journal Of Occupational Therapy, 65*(1), 15-20.

Hawley, M. S., Enderby, P., Green, P., Cunningham, S., Brownsell, S., Carmichael, J., . . . Palmer, R. (2007). A speech-controlled environmental control system for

people with severe dysarthria. *Medical Engineering & Physics, 29*(5), 586-593. doi:http://dx.doi.org/10.1016/j.medengphy.2006.06.009

Hawley, M. S., Enderby, P., Green, P., Cunningham, S., & Palmer, R. (2006). Development of a Voice-Input Voice-Output Communication Aid (VIVOCA) for People with Severe Dysarthria. In K. Miesenberger, J. Klaus, W. L. Zagler, & A. I. Karshmer (Eds.), *Computers Helping People with Special Needs: 10th International Conference, ICCHP 2006, Linz, Austria, July 11-13, 2006. Proceedings* (pp. 882-885). Berlin, Heidelberg: Springer Berlin Heidelberg.

He, X., & Deng, L. (2008). Discriminative learning for speech recognition: theory and practice. *Synthesis Lectures on Speech and Audio Processing, 4*(1), 1-112.

Hill, A. J., Theodoros, D. G., Russell, T. G., Cahill, L. M., Ward, E. C., & Clark, K. M. (2006). An Internet-based telerehabilitation system for the assessment of motor speech disorders: A pilot study. *American Journal of Speech-Language Pathology, 15*(1), 45-56.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*: Prentice Hall PTR.

Hux, K., Rankin-Erickson, J., Manasse, N., & Lauritzen, E. (2000). Accuracy of three speech recognition systems: Case study of dysarthric speech. *Augmentative and Alternative Communication, 16*(3), 186-196. doi:10.1080/07434610012331279044

Ishibuchi, H., & Nojima, Y. (2013). Repeated double cross-validation for choosing a single solution in evolutionary multi-objective fuzzy classifier design. *Knowledge-Based Systems, 54*, 22-31.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31*(3), 264-323.

Joy, N. M., & Umesh, S. (2018). Improving acoustic models in torgo dysarthric speech database. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26*(3), 637-645.

Kayasith, P., & Theeramunkong, T. (2009). Speech confusion index : A confusion-based speech quality indicator and recognition rate prediction for dysarthria. *Computers & Mathematics with Applications, 58*(8), 1534-1549. doi:http://dx.doi.org/10.1016/j.camwa.2009.06.051

Kayasith, P., Theeramunkong, T., & Thubthong, N. (2006a). Recognition Rate Prediction for Dysarthric Speech Disorder Via Speech Consistency Score. In Q. Yang & G. Webb (Eds.), *PRICAI 2006: Trends in Artificial Intelligence: 9th Pacific Rim International Conference on Artificial Intelligence Guilin, China, August 7-11, 2006 Proceedings* (pp. 885-889). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kayasith, P., Theeramunkong, T., & Thubthong, N. (2006b). Speech Confusion Index (Ø): A Recognition Rate Indicator for Dysarthric Speakers. In T. Salakoski, F.

Ginter, S. Pyysalo, & T. Pahikkala (Eds.), *Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL 2006 Turku, Finland, August 23-25, 2006 Proceedings* (pp. 604-615). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kent, J. F., Kent, R. D., Rosenbek, J. C., Weismer, G., Martin, R., Sufit, R., & Brooks, B. R. (1992). Quantitative description of the dysarthria in women with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 35*(4), 723-733.

Kent, R., Kent, J., Weismer, G., & Duffy, J. (2000). What dysarthrias can tell us about the neural control of speech. *Journal of Phonetics, 28*(3), 273-302. doi:http://dx.doi.org/10.1006/jpho.2000.0122

Kent, R., & Kim, Y. (2008). Acoustic Analysis of Speech. *The Handbook of clinical linguistics*, 360.

Kent, R. D., Kent, J. F., Duffy, J. R., Thomas, J. E., Weismer, G., & Stuntebeck, S. (2000). Ataxic dysarthria. *Journal of Speech, Language, and Hearing Research, 43*(5), 1275-1289.

Kent, R. D., Kent, J. F., Weismer, G., Martin, R. E., Sufit, R. L., Brooks, B. R., & Rosenbek, J. C. (1989). Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. *Clinical Linguistics and Phonetics, 3*(4), 347-358. doi:10.3109/02699208908985295

Kent, R. D., & Kim, Y. J. (2003). Toward an acoustic typology of motor speech disorders. *Clinical Linguistics & Phonetics, 17*(6), 427-445.

Kent, R. D., Miolo, G., & Bloedel, S. (1994). The Intelligibility of Children's SpeechA Review of Evaluation Procedures. *American Journal of Speech-Language Pathology, 3*(2), 81-95.

Kent, R. D., Vorperian, H. K., Kent, J. F., & Duffy, J. R. (2003). Voice dysfunction in dysarthria: application of the Multi-Dimensional Voice Program™. *Journal of Communication Disorders, 36*(4), 281-306. doi:http://dx.doi.org/10.1016/S0021-9924(03)00016-9

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*(4), 482-499.

Kent, R. D., Weismer, G., Kent, J. F., Vorperian, H. K., & Duffy, J. R. (1999). Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of Communication Disorders, 32*(3), 141-186.

Khoshgoftaar, T. M., Golawala, M., & Van Hulse, J. (2007). *An empirical study of learning from imbalanced data using random forest.* Paper presented at the 19th IEEE international conference on Tools with Artificial Intelligence, 2007. ICTAI 2007. .

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., & Frame, S. (2008). *Dysarthric speech database for universal access research.* Paper presented at the Interspeech.

Kim, H., Martin, K., Hasegawa-Johnson, M., & Perlman, A. (2010). Frequency of consonant articulation errors in dysarthric speech. *Clinical Linguistics & Phonetics, 24*(10), 759-770.

Kim, J., Kumar, N., Tsiartas, A., Li, M., & Narayanan, S. (2012). Intelligibility classification of pathological speech using fusion of multiple subsystems. *Proc. of Interspeech, Portland, Oregon, USA*, 534-537.

Kim, J., Kumar, N., Tsiartas, A., Li, M., & Narayanan, S. S. (2015). Automatic intelligibility classification of sentence-level pathological speech. *Computer Speech & Language, 29*(1), 132-144. doi:http://dx.doi.org/10.1016/j.csl.2014.02.001

Kim, Y., Kent, R. D., & Weismer, G. (2011). An Acoustic Study of the Relationships Among Neurologic Disease, Dysarthria Type, and Severity of Dysarthria. *Journal of Speech, Language & Hearing Research, 54*(2), 417-429. doi:10.1044/1092-4388(2010/10-0020)

Kim, Y., Weismer, G., Kent, R. D., & Duffy, J. R. (2009). Statistical models of F2 slope in relation to severity of dysarthria. *Folia Phoniatrica et Logopaedica, 61*(6), 329-335.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine Learning Proceedings 1992* (pp. 249-256): Elsevier.

Kitzing, P., Maier, A., & Åhlander, V. L. (2009). Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phoniatrics Vocology, 34*(2), 91-96. doi:10.1080/14015430802657216

Klopfenstein, M. (2009). Interaction between prosody and intelligibility. *International Journal of Speech-Language Pathology, 11*(4), 326-331. doi:10.1080/17549500903003094

Kodrasi, I., & Bourlard, H. (2019). *Super-gaussianity of Speech Spectral Coefficients as a Potential Biomarker for Dysarthric Speech Detection.* Paper presented at the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*(1), 59-69. doi:10.1007/BF00337288

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*(9), 1464-1480. doi:10.1109/5.58325

Kotler, A.-L., & Thomas-Stonell, N. (1997). Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment. *Augmentative and Alternative Communication, 13*(2), 71-80. doi:doi:10.1080/07434619712331277858

Kuhn, R., Junqua, J. C., Nguyen, P., & Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing, 8*(6), 695-707.

Kuncheva, L. I. (2007). *A stability index for feature selection.* Paper presented at the Artificial intelligence and applications.

Lebedev, A., Westman, E., Van Westen, G., Kramberger, M., Lundervold, A., Aarsland, D., . . . Tsolaki, M. (2014). Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical, 6*, 115-125.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436.

LeGendre, S. J., Liss, J. M., & Lotto, A. J. (2009). Discriminating dysarthria type and predicting intelligibility from amplitude modulation spectra. *The Journal of the Acoustical Society of America, 125*(4), 2530-2530. doi:10.1121/1.4783544

Leggetter, C., & Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer speech and language, 9*(2), 171.

Leggetter, C. J., & Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language, 9*(2), 171-185.

Lin, D., & Tang, X. (2006). *Conditional infomax learning: an integrated framework for feature extraction and fusion.* Paper presented at the European Conference on Computer Vision.

Linguistic Data Consortium. (1994). CSR-II (WSJ1) Complete LDC94S13A. DVD. In. Philadelphia.

Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., & Caviness, J. N. (2009). Quantifying Speech Rhythm Abnormalities in the Dysarthrias. *Journal of speech, language, and hearing research : JSLHR, 52*(5), 1334-1352. doi:10.1044/1092-4388(2009/08-0208)

Liu, H.-M., Tsao, F.-M., & Kuhl, P. K. (2005). The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *The Journal of the Acoustical Society of America, 117*(6), 3879-3889.

Liu, H.-M., Tseng, C.-H., & Tsao, F.-M. (2000). Perceptual and acoustic analysis of speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *Clinical Linguistics & Phonetics, 14*(6), 447-464.

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., & Nöth, E. (2009). PEAKS – A system for the automatic evaluation of voice and speech disorders. *Speech Communication, 51*(5), 425-437. doi:http://dx.doi.org/10.1016/j.specom.2009.01.004

Maier, A., Haderlein, T., Stelzle, F., Nöth, E., Nkenke, E., Rosanowski, F., . . . Schuster, M. (2009). Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. *EURASIP Journal on Audio, Speech, and Music Processing, 2010*(1), 1.

Manochiopinig, S., Thubthong, N., & Kayasith, P. (2007). *Dysarthric speech characteristics of Thai stroke patients assessed by the computerized articulation test*. Paper presented at the Proceedings of the 1st international convention on Rehabilitation engineering & assistive technology: in conjunction with 1st Tan Tock Seng Hospital Neurorehabilitation Meeting, Singapore.

Manochiopinig, S., Thubthong, N., & Kayasith, P. (2008). Dysarthric speech characteristics of Thai stroke patients. *Disability and Rehabilitation: Assistive Technology, 3*(6), 332-338.

Mathew, J. B., Jacob, J., Sajeev, K., Joy, J., & Rajan, R. (2018). *Significance of Feature Selection for Acoustic Modeling in Dysarthric Speech Recognition.* Paper presented at the 2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET).

McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition* (Vol. 544): John Wiley & Sons.

McLachlan, G., Do, K.-A., & Ambroise, C. (2005). *Analyzing microarray gene expression data* (Vol. 422): John Wiley & Sons.

McRae, P. A., Tjaden, K., & Schoonings, B. (2002). Acoustic and perceptual consequences of articulatory rate change in Parkinson disease. *Journal of Speech, Language, and Hearing Research, 45*(1), 35-50.

Menendez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., & Bunnell, H. T. (1996). *The Nemours database of dysarthric speech.* Paper presented at the ICSLP 96 Proceedings., Fourth International Conference on Spoken Language, 1996. .

Mengistu, K., & Rudzicz, F. (2011). Comparing Humans and Automatic Speech Recognition Systems in Recognizing Dysarthric Speech. In C. Butz & P. Lingras (Eds.), *Advances in Artificial Intelligence* (Vol. 6657, pp. 291-300): Springer Berlin Heidelberg.

Mengistu, K. T., & Rudzicz, F. (2011). *Adapting acoustic and lexical models to dysarthric speech.* Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011

Meyer, P. E., & Bontempi, G. (2006). *On the use of variable complementarity for feature selection in cancer classification.* Paper presented at the Workshops on Applications of Evolutionary Computation.

Middag, C., Bocklet, T., Martens, J.-P., & Nöth, E. (2011). *Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment.* Paper presented at the 12th Annual conference of the International Speech Communication Association (Interspeech 2011).

Middag, C., Martens, J.-P., Van Nuffelen, G., & De Bodt, M. (2009). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing, 2009*, 3.

Mokbel, C., Mauuary, L., Jouvet, D., Monne, J., Sorin, C., Simonin, J., & Bartkova, K. (1996). *Towards improving ASR robustness for PSN & GSM telephone applications.* Paper presented at the Interactive Voice Technology for Telecommunications Applications, 1996. Proceedings., Third IEEE Workshop on.

Morales, S. O. C., & Cox, S. J. (2009). Modelling errors in automatic speech recognition for dysarthric speakers. *EURASIP Journal on Advances in Signal Processing, 2009*, 2.

Mustafa, M. B., Salim, S. S., Mohamed, N., Al-Qatab, B., & Siong, C. E. (2014). Severity-based adaptation with limited data for ASR to aid dysarthric speakers. *PLoS One, 9*(1), e86285. doi:10.1371/journal.pone.0086285

Narendra, N., & Alku, P. (2018). *Dysarthric Speech Classification Using Glottal Features Computed from Non-words, Words and Sentences.* Paper presented at the Interspeech.

Narendra, N., & Alku, P. (2019). Dysarthric speech classification from coded telephone speech using glottal features. *Speech Communication, 110*, 47-55.

Neave, H. R., & Worthington, P. L. (1988). *Distribution-free tests*: Unwin Hyman London.

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of Speech, Language, and Hearing Research, 51*(3), 574-585.

Niimi, M. N. S. (2001). Speaking rate and its components in dysarthric speakers. *Clinical Linguistics & Phonetics, 15*(4), 309-317. doi:10.1080/02699200010024456

Ozawa, Y., Shiromoto, O., Ishizaki, F., & Watamori, T. (2001). Symptomatic Differences in Decreased Alternating Motion Rates between Individuals with Spastic and with Ataxic Dysarthria: An Acoustic Analysis. *Folia Phoniatrica et Logopaedica, 53*(2), 67-72. Retrieved from http://www.karger.com/DOI/10.1159/000052656

Özsancak, C., Auzou, P., Jan, M., & Hannequin, D. (2001). Measurement of Voice Onset Time in Dysarthric Patients: Methodological Considerations. *Folia Phoniatrica*

et Logopaedica, 53(1), 48-57. Retrieved from http://www.karger.com/DOI/10.1159/000052653

Paja, M. O. S., & Falk, T. H. (2012). *Automated Dysarthria Severity Classification for Improved Objective Intelligibility Assessment of Spastic Dysarthric Speech.* Paper presented at the INTERSPEECH.

Parker, M., Cunningham, S., Enderby, P., Hawley, M., & Green, P. (2006). Automatic speech recognition and training for severely dysarthric users of assistive technology: The STARDUST project. *Clinical Linguistics & Phonetics, 20*(2-3), 149-156. doi:10.1080/02699200400026884

Parmar, C., Grossmann, P., Rietveld, D., Rietbergen, M. M., Lambin, P., & Aerts, H. J. (2015). Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Frontiers in oncology, 5*, 272.

Paul, D. B., & Baker, J. M. (1992). *The design for the wall street journal-based CSR corpus*. Paper presented at the Proceedings of the workshop on Speech and Natural Language, Harriman, New York.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems, 24*(3), 45-77. doi:10.2753/MIS0742-1222240302

Prelock, P. A., Hutchins, T., & Glascoe, F. P. (2008). Speech-language impairment: how to identify the most common and least diagnosed disability of childhood. *The Medscape Journal of Medicine, 10*(6), 136. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491683/

Preminger, J. E., & Van Tasell, D. J. (1995). Quantifying the relation between speech quality and speech intelligibility. *Journal of Speech, Language, and Hearing Research, 38*(3), 714-725.

Qian, Y., Soong, F., Chen, Y., & Chu, M. (2006). An HMM-Based Mandarin Chinese Text-To-Speech System. In Q. Huo, B. Ma, E.-S. Chng, & H. Li (Eds.), *Chinese Spoken Language Processing: 5th International Symposium, ISCSLP 2006, Singapore, December 13-16, 2006. Proceedings* (pp. 223-232). Berlin, Heidelberg: Springer Berlin Heidelberg.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing, 10*(1), 19-41. doi:http://dx.doi.org/10.1006/dspr.1999.0361

Roger, J., & Lewis, M. (2000). *An introduction to classification and regression tree (CART) analysis.* Paper presented at the Annual meeting of the society for academic emergency medicine.

Roth, C. (2011). Dysarthria. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (pp. 905-908). New York, NY: Springer New York.

Rudzicz, F. (2007). *Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech*. Paper presented at the Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility, Tempe, Arizona, USA.

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2011). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 1-19.

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation, 46*(4), 523-541.

Rueda, A., Vásquez-Correa, J. C., Rios-Urrego, C. D., Orozco-Arroyave, J. R., Krishnan, S., & Nöth, E. (2019). Feature Representation of Pathophysiology of Parkinsonian Dysarthria. *Proc. Interspeech 2019*, 3048-3052.

Saeys, Y., Abeel, T., & Van de Peer, Y. (2008). *Robust feature selection using ensemble feature selection techniques.* Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Samsudin, S. H., Shafri, H. Z., Hamedianfar, A., & Mansor, S. (2015). Spectral feature selection and classification of roofing materials using field spectroscopy data. *Journal of Applied Remote Sensing, 9*(1), 095079.

Santana, L. E. A. d. S., & de Paula Canuto, A. M. (2014). Filter-based optimization techniques for selection of feature subsets in ensemble systems. *Expert Systems with Applications, 41*(4), 1622-1631.

Schlenck, K. J., Bettrich, R., & Willmes, K. (1993). Aspects of disturbed prosody in dysarthria. *Clinical Linguistics & Phonetics*, *7*(2), 119-128. doi:10.3109/02699209308985549

Schuller, B. W. (2013). *Intelligent audio analysis*: Springer.

Sharma, H. V., & Hasegawa-Johnson, M. (2013). Acoustic model adaptation using in-domain background models for dysarthric speech recognition. *Computer Speech & Language, 27*(6), 1147-1162. doi:http://dx.doi.org/10.1016/j.csl.2012.10.002

Sharma, H. V., Hasegawa-Johnson, M., Gunderson, J., & Perlman, A. (2009). *Universal access: Preliminary experiments in dysarthric speech recognition.* Paper presented at the Proc. 10th Annual Conf. of the Internat. Speech Communication Association.

Shinoda, K. (2011). Speaker Adaptation Techniques for Automatic Speech Recognition. *Proc. APSIPA ASC 2011 Xi'an*.

Solomon, N. P., & Hixon, T. J. (1993). Speech breathing in Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 36*(2), 294-310.

Sriranjani, R., Ramasubba Reddy, M., & Umesh, S. (2015). *Improved acoustic modeling for automatic dysarthric speech recognition.* Paper presented at the Communications (NCC), 2015 Twenty First National Conference on.

Stern, R., & Lasry, M. (1987). Dynamic speaker adaptation for feature-based isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 35*(6), 751-763. doi:10.1109/TASSP.1987.1165203

Strand, E. A., & McCauley, R. J. (2008). Differential Diagnosis of Severe Speech Impairment in Young Children. *The ASHA Leader, 13*(10), 10-13. doi:10.1044/leader.FTR1.13102008.10

Stuntebeck, S. (2002). *Acoustic Analysis of the Prosodic Properties of Ataxic Speech*: University of Wisconsin--Madison.

Takashima, Y., Takiguchi, T., & Ariki, Y. (2019). *End-to-end Dysarthric Speech Recognition Using Multiple Databases.* Paper presented at the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Teixeira, J. P., & Fernandes, P. O. (2014). Jitter, Shimmer and HNR Classification within Gender, Tones and Vowels in Healthy Voices. *Procedia Technology, 16*, 1228-1237. doi:http://dx.doi.org/10.1016/j.protcy.2014.10.138

Theodoridis, S., & Koutroumbas, K. (2006). Clustering: basic concepts. *Pattern Recognition*, 483-516.

Thomas-Stonell, N., Kotler, A.-L., Leeper, H., & Doyle, P. (1998). Computerized speech recognition: influence of intelligibility and perceptual consistency on recognition accuracy. *Augmentative and Alternative Communication, 14*(1), 51-56. doi:10.1080/07434619812331278196

Thubthong, N., Kayasith, P., Manochiopinig, S., Leelasiriwong, W., & Rukkharangsarit, O. (2005). Articulation Analysis of Thai Cerebral Palsy Children with Dysarthric Speech.

Tjaden, K., & Wilding, G. E. (2004). Rate and Loudness Manipulations in DysarthriaAcoustic and Perceptual Findings. *Journal of Speech, Language, and Hearing Research, 47*(4), 766-783.

Tuv, E., Borisov, A., Runger, G., & Torkkola, K. (2009). Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of machine learning research, 10*(Jul), 1341-1366.

Van der Graaff, M., Kuiper, T., Zwinderman, A., Van de Warrenburg, B., Poels, P., Offeringa, A., . . . De Visser, M. (2009). Clinical Identification of Dysarthria Types among Neurologists, Residents in Neurology and Speech Therapists.

European Neurology, 61(5), 295-300. Retrieved from http://www.karger.com/DOI/10.1159/000206855

Van Nuffelen, G., Middag, C., De Bodt, M., & Martens, J. P. (2009). Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language & Communication Disorders, 44*(5), 716-730. doi:10.1080/13682820802342062

Vapnik, V. (1998). *Statistical learning theory. 1998*: Wiley, New York.

Wang, Y.-T., Kent, R. D., Duffy, J. R., & Thomas, J. E. (2005). Dysarthria associated with traumatic brain injury: speaking rate and emphatic stress. *Journal of Communication Disorders, 38*(3), 231-260. doi:http://dx.doi.org/10.1016/j.jcomdis.2004.12.001

Weismer, G., Jeng, J.-Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica, 53*(1), 1-18.

Weismer, G., Kim, Y., Maassen, B., & van Lieshout, P. (2010). Classification and taxonomy of motor speech disorders: What are the issues. *Speech motor control: New developments in basic and applied research*, 229-241.

Weismer, G., Martin, R., & Kent, R. (1992). Acoustic and perceptual approaches to the study of intelligibility. *Intelligibility in speech disorders*, 67-118.

Whitehill, T. L., & Ciocca, V. (2000). Speech errors in Cantonese speaking adults with cerebral palsy. *Clinical Linguistics & Phonetics, 14*(2), 111-130.

Wilson, E. M., Abbeduto, L., Camarata, S. M., & Shriberg, L. D. (2019). Speech and motor speech disorders and intelligibility in adolescents with Down syndrome. *Clinical Linguistics & Phonetics, 33*(8), 790-814. doi:10.1080/02699206.2019.1595736

Witt, S. M. (1999). *Use of speech recognition in computer-assisted language learning*: University of Cambridge.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Woodland, P. C. (2001). *Speaker adaptation for continuous density HMMs: A review.* Paper presented at the ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition.

Xiong, F., Barker, J., & Christensen, H. (2018). *Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition.* Paper presented at the Speech Communication; 13th ITG-Symposium.

Yorkston, K. M., Beukelman, D. R., & Traynor, C. (1984). *Computerized assessment of intelligibility of dysarthric speech*. Tigard, Ore. :: C.C. Publications.

Young, V., & Mihailidis, A. (2010). Difficulties in Automatic Speech Recognition of Dysarthric Speakers and Implications for Speech-Based Applications Used by the Elderly: A Literature Review. *Assistive Technology, 22*(2), 99-112. doi:10.1080/10400435.2010.483646

Zaidi, B.-F., Boudraa, M., Selouani, S.-A., Addou, D., & Yakoub, M. S. (2019). *Automatic Recognition System for Dysarthric Speech Based on MFCC's, PNCC's, JITTER and SHIMMER Coefficients.* Paper presented at the Science and Information Conference.

Zhang, H. (2004). The optimality of naive Bayes. *AA*, *1*(2), 3.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Articles

- Mustafa, M. B., Salim, S. S., Mohamed, N., Al-Qatab, B., & Siong, C. E. (2014). Severity-based adaptation with limited data for ASR to aid dysarthric speakers. *PLoS One, 9*(1), e86285. doi:10.1371/journal.pone.0086285.
- Al-Qatab, B. A., Mustafa, M. B., & Salim, S. S. (2015). Maximum Likelihood Linear Regression (MLLR) for ASR Severity Based Adaptation to Help Dysarthric Speakers. International Journal of Simulation--Systems, Science & Technology, 15(6).
- Al-Qatab, B. A., Mustafa, M. B., & Salim, S. S. (2014, 23-25 Sept. 2014). Severity Based Adaptation for ASR to Aid Dysarthric Speakers. Paper presented at the 2014 8th Asia Modelling Symposium.