The University Of Sheffield.

Dysarthric speech analysis and automatic recognition using phase based representations

By:

Siddharth Sehgal

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield

Faculty of Medicine, Dentistry and Health

Department of Human Communication Sciences

April 2018

I dedicate this thesis to my mother. Her love and support is eternal and unconditional.

Table of Contents

Table of Contents **iv**

List of Tables **viii**

List of Figures **x**

Abbreviations **xiv**

Abstract **xvi**

Acknowledgements **xvii**

1 Introduction **2**
1.1 Motivations 2
1.2 Scope of the thesis 3
1.3 Structure of the thesis 4

2 Background on Dysarthria **6**
2.1 Dysarthria and its causes 7
2.1.1 Neurological basis of dysarthria 8
2.2 Types of dysarthria 11
2.2.1 Flaccid dysarthria 11
2.2.2 Spastic dysarthria 11
2.2.3 Ataxic dysarthria 12
2.2.4 Hypokinetic dysarthria 12
2.2.5 Hyperkinetic dysarthria 12
2.2.6 Mixed dysarthria 13
2.3 Statistics of dysarthric etiologies 13
2.4 Effects of dysarthria 14
2.5 Severity and impact on intelligibility 19
2.6 Treatment and management of dysarthria 22
2.6.1 Directions for management of speech disorders 23
2.6.2 Approaches for management of speech disorders 23

2.6.2.1 Medical intervention . . . 24
2.6.2.2 Prosthetic management . . . 25
2.6.2.3 Behavioural management . . . 26
2.6.2.4 Augmentative and alternative communication (AAC) . . . 27

3 ASR and its Applications in Dysarthric Speech 30
3.1 Generic architecture of an ASR System . . . 31
3.2 Feature extraction . . . 33
3.3 Acoustic modelling . . . 37
3.3.1 Pre-Statistical approaches . . . 37
3.3.2 Statistical approaches . . . 38
3.3.2.1 Generative learning . . . 39
3.3.2.2 Conditional learning . . . 45
3.3.2.3 Discriminative learning . . . 46
3.3.3 Advanced learning architectures . . . 47
3.4 Language and pronunciation modelling . . . 49
3.5 Adaptation and adaptive training . . . 52
3.5.1 Feature based adaptation . . . 53
3.5.2 Maximum a Posteriori (MAP) adaptation . . . 55
3.5.3 Linear transformation . . . 56
3.5.3.1 Maximum Likelihood Linear Regression (MLLR) . . . 56
3.5.3.2 Constrained MLLR (CMLLR) . . . 59
3.5.4 Adaptive training . . . 59
3.5.4.1 Speaker Adaptive Training (SAT) . . . 60
3.5.4.2 Cluster adaptive training and eigenvoices . . . 61
3.5.5 Discriminative adaptation . . . 62
3.6 Automatic recognition of dysarthric speech . . . 63
3.6.1 Commercial speech recognition . . . 64
3.6.2 Modelling approaches . . . 66
3.6.3 Acoustic features and enhancement . . . 70
3.6.4 Adaptation of dysarthric speech . . . 72
3.6.5 Other approaches . . . 73

4 Recognition and Analysis of Dysarthric Speech 78
4.1 Part-A: Baseline recognition results of dysarthric speech . . . 80
4.1.1 Experimental setup . . . 80
4.1.1.1 Data preparation . . . 80
4.1.1.2 Acoustic modelling . . . 81
4.1.1.3 Methodology . . . 82
4.1.2 Experimental results . . . 84
4.1.2.1 SI systems . . . 84
4.1.2.2 SI adapted systems . . . 85
4.1.2.3 SAT-adapted vs other systems . . . 86

4.1.2.4 MLLR-MAP for severity groups . . . 88
4.1.3 Discussion of baseline results . . . 91
4.2 Part-B: Acoustic analysis of the UASPEECH database . . . 94
4.2.1 Temporal analysis . . . 96
4.2.1.1 Speech rate . . . 96
4.2.1.2 Voice onset time . . . 100
4.2.2 Spectral analysis . . . 107
4.2.2.1 F1-F2 space . . . 107
4.2.2.2 F1-F2 quantification . . . 111
4.2.3 Relationship of acoustic analysis with the ASR accuracy . . . 118
4.2.4 Zeros of the z-Transform (ZZT) analysis for the vowel segments . . . 122
4.2.4.1 ZZT Analysis of a basic signal: An example . . . 122
4.2.4.2 Relationship between ZZT, phase and articulation . . . 125
4.2.4.3 ZZT analysis of a typical vowel segment . . . 128
4.2.4.4 ZZT analysis of a dysarthric vowel segment . . . 129
4.2.5 Summary of the acoustic analysis . . . 135

5 Phase-based Analysis of Dysarthric Speech 140
5.1 Phase-slope deviation . . . 141
5.2 PSD analysis of dysarthric vowels . . . 141
5.2.1 PSD and dysarthric intelligibility . . . 142
5.2.2 The behaviour of PSD on a secondary data source (VIVOCA) . . . 144
5.2.3 An operational understanding of PSD . . . 146
5.2.4 Correcting PSD in dysarthric speech . . . 150
5.3 PSD effect on dysarthric ASR performance . . . 152
5.3.1 Experimental setup . . . 152
5.3.2 Dysarthric ASR results . . . 153
5.3.2.1 Supervised correction . . . 153
5.3.2.2 Semi-supervised correction . . . 155
5.3.2.3 Unsupervised correction . . . 157
5.4 Conclusion . . . 164

6 Feature Representations based on Phase Spectrum 168
6.1 Phase-based feature representations for speech recognition . . . 169
6.1.1 Group Delay Function . . . 170
6.1.1.1 Properties of Group Delay Function . . . 171
6.1.1.2 Problem with Group Delay Function . . . 174
6.1.2 Modified Group Delay Function (MODGDF) . . . 177
6.1.3 Product Spectrum (PS) . . . 177
6.1.4 Cepstral coefficients based on MODGDF and PS representations . . . 178
6.2 Phase based features for dysarthric speech . . . 179
6.2.1 Frequency representation using phase spectrum . . . 180
6.2.2 Better class separability . . . 181

6.3 Experiments on phase vs magnitude based features of dysarthric speech . . 190
6.4 PSD enhanced phase based feature representation for dysarthric ASR . . . 192
6.5 Conclusion 194

7 Discussions and Future Work **198**

A Acoustic Analysis of UASPEECH **210**
A.1 F1-F2 vowel quadrilaterals for UASPEECH dysarthric speakers 210

B The VIVOCA Data Source **215**
B.1 Missing vowel tokens 215

C Phase Alignment for Control and Dysarthric Speakers **217**
C.1 Unwrapped phase alignment for control and dysarthric speakers 217

D Standard Deviational Ellipses **219**
D.1 Standard Deviational Ellipses for control and dysarthric intelligibility groups 219

Bibliography **224**

List of Tables

2.1 Different modes of AAC strategies and components 28

4.1 A summary of each training corpus in the system 81
4.2 Summary of baseline systems and the corpus used for its preparation 82
4.3 Absolute word accuracy for SD and SI/SAT baseline systems adapted using MLLR-MAP 89
4.4 Cochran's Q analysis for various intelligibility groups 91
4.5 Configuration used for measuring VOT of voiceless and voiced stops 100
4.6 Expected correlation trend between the ASR score and acoustic variables . 119

5.1 Vowel categories with the list of phonetic tokens examined 142
5.2 Summary of the VIVOCA users 144
5.3 A hypothesised relationship between the PSD metric and intelligibility . . . 148
5.4 PSD and intelligibility prediction using regression 149
5.5 ASR systems re-tested after PSD corrections 152
5.6 Absolute ASR word accuracy after the PSD corrections 155
5.7 Absolute ASR word accuracy after the global PSD corrections 157
5.8 Phonetic and corresponding vowel-consonant alignment 158
5.9 Operational duration range for each intelligibility group 161
5.10 Absolute ASR word accuracy after the unsupervised PSD corrections 163

6.1 Phonetic tokens examined for certain speech production errors 182
6.2 Absolute ASR word accuracy for phase based features 191
6.3 Absolute ASR word accuracy for PSD correction with phase features 194

B.1 The availability of data for the vowel tokens of VIVOCA speakers. The red blocks indicate that there was no speech utterance for the specific speaker-vowel pair. 216

List of Figures

2.1 Schematic diagram for a human speech production 7
2.2 Incidence and prevalence of some major dysarthric etiologies 14
2.3 Spectrogram comparison for typical and dysarthric speakers with varying severity 17
2.4 F2 trajectory comparison for dysarthric and typical speaker 18
2.5 A holistic listening model for perceptual judgement of intelligibility 20
2.6 Management approaches for motor speech disorder 24
2.7 Dysarthric speakers who could benefit from AAC devices 29

3.1 Generic architecture for the ASR systems 32
3.2 Schematic diagram for the MFCC generation process 36
3.3 Statistical learning paradigms 38
3.4 An example of a five state HMM 42
3.5 Illustration of a DBN architecture 48
3.6 Word pronunciation for typical and dysarthric speaker 51
3.7 An overview of the SAT framework 61

4.1 Average word accuracy for the baseline SI systems 85
4.2 Adaptation scores for the baseline SI systems 86
4.3 Comparison of SD and MLLR-MAP based SI & SAT systems 87
4.4 Accuracy for the baseline SI systems for various intelligibility groups 90
4.5 MLLR-MAP scores for the SAT & SI systems for various intelligibility groups 90
4.6 Temporal analysis experiments 96
4.7 sypse for the control and dysarthric speakers 97

4.8 sypse for various intelligibility groups 97
4.9 Scatter plot between sypse and intelligibility 98
4.10 Speaker-wise sypse analysis 99
4.11 VOT timing spectrogram for control and dysarthric speakers 101
4.12 Voice Onset Times for the voiceless stops /p/, /t/, /k/ 102
4.13 Voice Onset Times for the voiced stops /b/, /d/, /g/ 102
4.14 Voice Onset Times for the voiceless stops /p/, /t/, /k/ across various intelligibility groups 103
4.15 Voice Onset Times for the voiced stops /b/, /d/, /g/ across various intelligibility groups 103
4.16 VOT of UASPEECH Speakers 104
4.17 Spectral analysis experiments 107
4.18 F1-F2 plot for vowels of control and dysarthric speakers 108
4.19 F1-F2 plot for the diphthongs of control and dysarthric speakers 108
4.20 Standard deviational ellipses for the control and dysarthric speakers 109
4.21 F1-F2 plot for the vowels of dysarthric intelligibility groups 110
4.22 F1-F2 plot for the diphthongs of dysarthric intelligibility groups 111
4.23 F1-F2 area computation 112
4.24 F1-F2 area compression factor for dysarthric speakers 114
4.25 F1-F2 vowel quadrilateral shape for various intelligibility groups 115
4.26 An exhibit of the distance measure between centres of vowel quadrilaterals . 117
4.27 Vowel quadrilateral distance between dysarthric and control speaker 117
4.28 Correlation analysis of five acoustic parameters against the SAT system . . 119
4.29 Correlation analysis of five acoustic parameters against the SAT performance for each intelligibility group. 120
4.30 ZZT patterns for the exponential function 124
4.31 Effect of phase group delay filter on a time-domain signal 125
4.32 ZZT pattern for a typical vowel segment 128
4.33 ZZT patterns for the dysarthric vowel segment /iy/ 130
4.34 ZZT patterns for the dysarthric vowel segment /uw/ 132
4.35 Unwrapped phase plot for the complex roots of dysarthric vowel segments . 133

4.36 Effect of the incorrect glottal closure alignment on the unwrapped phase . . 135

5.1 PSD analysis for speakers with dysarthria in UASPEECH database 143
5.2 PSD analysis for the speakers with dysarthria in UASPEECH and VIVOCA databases 145
5.3 Relationship between PSD metric and Intelligibility for UASPEECH database 147
5.4 Predicted values for PSD metric and Intelligibility for VIVOCA database . 149
5.5 Correction measure to reduce the PSD effect 150
5.6 Corrected PSD alignment for various intelligibility groups 151
5.7 Relative ASR improvement after the PSD correction 154
5.8 Comparison between specific and global vowel PSD transforms 156
5.9 Recognition output of a vowel-consonant classifier 159
5.10 Histogram comparison between forced-aligned and vowel-consonant classifier prediction 160
5.11 Comparison between specific, global and unsupervised vowel PSD transform 162
5.12 Comparison between Original and PSD corrected file - I 165
5.13 Comparison between Original and PSD corrected file - II 165

6.1 High resolution property of the Group Delay Spectrum 173
6.2 Reason for spikes in the Group Delay Spectrum 175
6.3 Spikes in the Group Delay Spectrum 176
6.4 Main steps for the generation of phase based cepstral coefficients 178
6.5 GDF Resolution for dysarthric speech vowel 180
6.6 Two dimensional projection for the MFCC & PSCC features for very-low and low intelligibility speakers 185
6.7 Two dimensional projection for the MFCC & PSCC features for mid and high intelligibility speakers 186
6.8 Two dimensional projection for the MFCC & PSCC features for very-low and low intelligibility speakers 187
6.9 Two dimensional projection for the MFCC & PSCC features for mid and high intelligibility speakers 188
6.10 Relative ASR gains for the phase based feature representations 190

6.11 ASR gains for the PSD enhanced phase features 193
6.12 Pole locations of the magnitude and phase spectra 195

7.1 Proposed framework for the current thesis 199
7.2 Overall results for each proposed method in the study 203
7.3 Per word recognition improvement example 206

A.1 F1-F2 Vowel Quadrilateral for speakers with very-low intelligibility 211
A.2 F1-F2 Vowel Quadrilateral for speakers with low intelligibility 212
A.3 F1-F2 Vowel Quadrilateral for speakers with mid intelligibility 213
A.4 F1-F2 Vowel Quadrilateral for speakers with high intelligibility 214

C.1 Unwrapped phase alignment for control and dysarthric speakers 218

D.1 Standard deviational ellipse for the very-low intelligibility group 220
D.2 Standard deviational ellipse for the low intelligibility group 221
D.3 Standard deviational ellipse for the mid intelligibility group 222
D.4 Standard deviational ellipse for the high intelligibility group 223

Abbreviations

AAC Augmentative and Alternative Communication
ALS Amyotrophic Lateral Sclerosis
ANN Artificial Neural Network
ASR Automatic Speech Recognition
CMLLR Constrained Maximum Likelihood Linear Regression
CN Cranial Nerves
CNS Central Nervous System
CP Cerebral Palsy
DBN Deep Belief Network
DCT Discrete Cosine Transform
DFT Discrete Fourier Transform
DNN Deep Neural Network
DTW Dynamic Time Warping
FFT Fast Fourier Transform
GDF Group Delay Function
GMM Gaussian Mixture Models
HMM Hidden Markov Model
LMN Lower Motor Neuron
LPC Linear Predictive Coding
LSVT Lee Silverman Voice Treatment
MAP Maximum a Posteriori
MFCC Mel Frequency Cepstral Coefficients
ML Maximum Likelihood
MLLR Maximum Likelihood Linear Regression
MND Motor Neuron Disease

MODGDFCC Modified Group Delay Function Cepstral Coefficients

MSD Motor Speech Disorder

PCA Principal Component Analysis

PD Parkinson's Disease

PNS Peripheral Nervous System

PSCC Product Spectrum Cepstral Coefficients

PSD Phase Slope Deviation

SVM Support Vector Machine

SYPSE Syllables Per Second

TBI Traumatic Brain Injury

UMN Upper Motor Neuron

VIVOCA Voice Input Voice Output Communication Aid

VOCA Voice Output Communication Aid

VOT Voice Onset Time

ZZT Zeros Of The Z-Transform

Abstract

Dysarthria is a neurological speech impairment which usually results in the loss of motor speech control due to muscular atrophy and poor coordination of articulators. Dysarthric speech is more difficult to model with machine learning algorithms, due to inconsistencies in the acoustic signal and to limited amounts of training data. This study reports a new approach for the analysis and representation of dysarthric speech, and applies it to improve ASR performance.

The Zeros of Z-Transform (ZZT) are investigated for dysarthric vowel segments. It shows evidence of a phase-based acoustic phenomenon that is responsible for the way the distribution of zero patterns relate to speech intelligibility. It is investigated whether such phase-based artefacts can be systematically exploited to understand their association with intelligibility.

A metric based on the phase slope deviation (PSD) is introduced that are observed in the unwrapped phase spectrum of dysarthric vowel segments. The metric compares the differences between the slopes of dysarthric vowels and typical vowels. The PSD shows a strong and nearly linear correspondence with the intelligibility of the speaker, and it is shown to hold for two separate databases of dysarthric speakers. A systematic procedure for correcting the underlying phase deviations results in a significant improvement in ASR performance for speakers with severe and moderate dysarthria.

In addition, information encoded in the phase component of the Fourier transform of dysarthric speech is exploited in the group delay spectrum. Its properties are found to represent disordered speech more effectively than the magnitude spectrum. Dysarthric ASR performance was significantly improved using phase-based cepstral features in comparison to the conventional MFCCs. A combined approach utilising the benefits of PSD corrections and phase-based features was found to surpass all the previous performance on the UASPEECH database of dysarthric speech.

Acknowledgements

After completing my thesis I realise that there are too many people to thank in too compressed a space, so I owe my gratitude to whom my debt is the greatest hoping that others will realise my predicament.

I would like to extend my utmost gratitude to my supervisor Dr. Stuart Cunningham for all his support and guidance throughout the course of this study. His invaluable and insightful suggestions played a significant role in shaping the course of this thesis and helped me to envisage new ideas. He has been an incredible guide and manager all throughout who has helped me to evolve as a researcher and mentored me in my personal and professional development.

I express my appreciation and special thanks to Prof. Phil Green for his indispensable advice and constructive suggestions during the most important stages of my thesis. His feedback and meticulous approach to problem solving has greatly assisted me to ponder on many critical aspects of the thesis and helped me to pull my work together in a coherent way.

I would also like to extend my sincere appreciation to Prof. Mark Hawley and my supervisor Dr. Stuart Cunningham for keeping me employed on speech based projects like VIVOCA, STAR etc. during the course of my study. These projects and my other day-to-day work has in some way constantly helped me to develop on my research pursuits and I am indebted to their continual support.

I would like to thank all my past and present colleagues at work for all the encouragement and tea time discussions. I want to particularly thank Richard Simmonds for all the interesting pub talks and for kindly proofreading my thesis and providing me with his valuable comments.

Lastly, I thank my wonderful family for their unending love, support and encouragement and always standing by me in all thick and thin. I will especially thank my mother for her constant tenacity in keeping me motivated all throughout.

Chapter 1

Introduction

Dysarthria is the collective name for a group of neurological speech disorders which result from damage to the central or peripheral nervous system. Dysarthric speech is usually characterised by changes such as reduced stress, slow speech rate, hypernasality, muscular rigidity, spasticity, monopitch and limited range of speech movements (Darley, Aronson, and Brown, 1969a; Duffy, 2005). It can have debilitating effects on speech production and can affect the subglottal, laryngeal and articulatory systems. The most prevalent causes of such motor speech disorders in the UK are stroke, cerebral palsy and Parkinson's disease (RCSLT, 2006). Reports suggest that there is an ever growing need to improve human-to-machine interaction for people with dysarthria in order to promote overall wellbeing and independence (Enderby et al., 2013). People with dysarthria can often have physical impairment making usual input methods (typing, touchscreen, etc.) difficult to use, so speech could provide an attractive interface for a natural and faster mode of interaction (Hawley, 2002).

1.1 Motivations

The automatic recognition of dysarthric speech has been pursued as a research problem for more than three decades (Coleman and Meyers, 1991; Fried-Oken, 1985; Roberts, 1985), but performance is still far behind that for typical speech, which has potentially reached human performance, especially under controlled conditions (Xiong et al., 2016). To date there is no commercially available system that can reliably recognise dysarthric speech. In addition, there is also a huge gap between listener and machine recognition of such material. The carers and close family members of a person with dysarthria can be regarded as the

oracle in comprehending their speech with highest accuracy, as listener familiarisation to dysarthric speech is shown to significantly increase performance (Tjaden and Liss, 1995).

The prime objective of this thesis is an attempt to bridge the gap between human and machine performance by having a wider understanding of such material and consequently increasing the performance of ASR systems for dysarthric speech. This will improve the overall acceptability of speech based communication or control devices for people with dysarthria, that can potentially increase their participation. Some of the difficulties in recognising dysarthric speech are often associated with the high degree of inter and intra speaker variations, data sparsity issues and malformed phonetic space (Blaney and Wilson, 2000; Kent et al., 2000; Morris, 1989). Hence, it is imperative to ask questions to find the gaps in our knowledge of dysarthric speech and to maximise the usage of the available data by searching for additional discriminatory information within the available acoustic space. This thesis attempts to answer some basic questions as stated below:

- Is there any additional information in the acoustics of dysarthric speech that can give cues about the underlying nature of the disorder?
- Can functional links be formed between such information and the underlying severity of impairment?
- Is there any alternate feature encodings that can characterise dysarthric speech more effectively than standard magnitude based spectral representations?
- Can such representations of dysarthric speech be utilised by machine learning algorithms to improve ASR performance?

1.2 Scope of the thesis

This thesis aims to systematically address each of the above questions in a quest to improve feature representations and ASR performance on dysarthric speech. It is generally a misconception that the success of any speech recognition system relies only on designing efficient classifiers that require a lot of data for optimal performance. This leads researchers to focus on the design of the classifier rather than understanding the data that must be modelled. The task is even more difficult for dysarthric speech recognition, where data sparsity is an issue and design of an optimal classifier is further restricted by the limited

understanding of "what makes the acoustics of dysarthric speech different?". To the best of our knowledge, to date any attempt to understand the nuances of the dysarthric acoustics is limited to the search for information encoded within the magnitude spectrum. This thesis introduces a novel approach to the analysis and characterisation of dysarthric speech that looks beyond the magnitude spectrum. We explore the phase spectrum for the analysis and better representation of dysarthric speech. It shows how the information encapsulated in the phase of a signal can be extremely useful for functionally discerning the underlying dysarthric intelligibility and improve the automatic recognition of various speech systems. Unlike the magnitude domain, where the spectral structure has been studied extensively to exhibit a more direct relation to the understanding of speech, the phase spectrum is much more difficult to interpret, especially for dysarthric speech. However, as will be seen that despite such impediments, the phase spectrum is extremely beneficial for the better representation and recognition of dysarthric speech signals. This thesis will not attempt to improve the quality of dysarthric speech from a perceptual standpoint. It will provide a series of coherent phase based approaches that will utilise the same amount of given data for statistically improving the machine performance of various dysarthric speech recognition systems. It will further give an alternate perspective to the analysis of such speech for quantitatively interpreting the underlying intelligibility of a signal.

1.3 Structure of the thesis

The thesis begins in Chapter 2 giving a physiological description about the types and causes of dysarthria. Chapter 3 provides the background on the basic fundamentals of an ASR system and how research has progressed for the automatic recognition of dysarthric speech. Chapter 4 is divided into two parts. The first part extends the work of earlier researchers and exploits advanced adaptation methods for giving the best results for our baseline system. The second part gives an account on the acoustic analysis of dysarthric speech. The last sections of Chapter 4 and Chapters 5 and 6 contain the novel contribution of the thesis and Chapter 7 summarises the main outcomes and gives a purview for future research. An overview of the content of each chapter is given below.

Chapter 2

The chapter provides a background of dysarthria from an anatomical perspective. It will discuss the causes and types of dysarthria and how such conditions can be managed or

treated using various medical and non-medical interventions. The chapter also gives a brief overview on the effects that dysarthria can have on the underlying acoustics or intelligibility.

Chapter 3

The chapter gives a literature survey on the ASR architecture in general. It will cover some important components like front-end processing, acoustic modelling and adaptation techniques and conclude by discussing its impact on dysarthric ASR research.

Chapter 4

The chapter is covered in two parts. The first part will extend the work of earlier researchers and explore various adaptation approaches to model dysarthric variabilities. It will also explore adaptive training techniques and test its efficacy. The collective results will form our baseline systems for further research. The second part of the chapter will deal with the acoustic analysis of UASPEECH database and explore a novel investigation strategy that is based on the analysis of the zeros of the Z-transform (ZZT) of dysarthric vowel segments.

Chapter 5

This chapter will extend the idea of ZZT analysis of dysarthric vowel segments and introduce a new metric based on phase slope deviations (PSD) that are observed in the unwrapped phase spectrum. It will explore the possibility of a functional associations between the PSD of dysarthric vowel segments and the underlying intelligibility. The later part of the chapter will develop a systematic approach for correcting the effect of such PSD aberrations that eventually results in significant ASR gains across varied dysarthric speech systems.

Chapter 6

This chapter will study the phase based feature representations of dysarthric signals. It will explore the properties of such phase representations from a theoretical and practical viewpoint. The efficacy of the features will be tested by measuring the ASR performance. The chapter will also explore if such phase based speech features can be supplemented with the benefits of PSD corrections to improve ASR results.

Chapter 7

It will give a collective discussion of the main contributions in the thesis and outline some of the potential areas for future research.

Chapter 2

Background on Dysarthria

This chapter will give a background summary of dysarthria from a physiological perspective. It will discuss the prime causes of dysarthria and enumerate categorical descriptions under which various types of dysarthria are medically classified. The chapter also discusses broad level effects that such neurological speech impairments can have on the acoustics. Towards the end of the chapter, a short summary is given on how the severity factors in dysarthria affects speech intelligibility followed by an overview of the common treatment and management approaches for motor speech disorders.

2.1 Dysarthria and its causes

Human speech results from the highly coordinated functioning of several components that make up the human vocal apparatus. The apparatus can be broadly conceptualised as being composed of three major components (Stevens, 2000): (1) parts below the larynx forming the subglottal structure, (2) the larynx and its surrounding structure which comprises the vocal folds, and (3) the parts above the larynx forming the supraglottal structure.

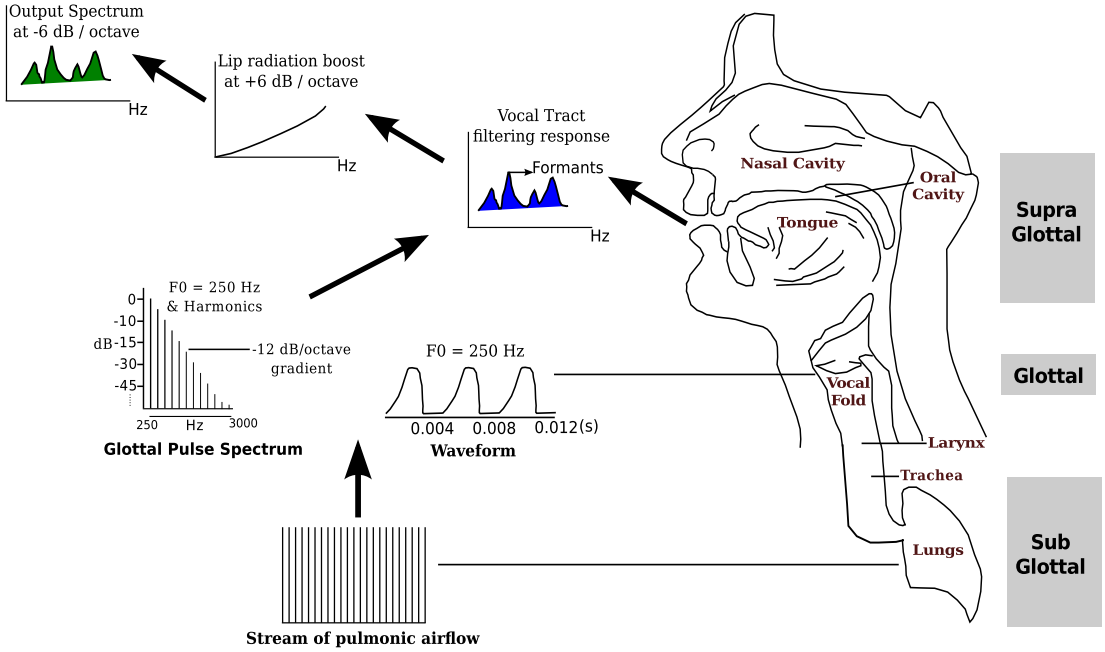

Figure 2.1: Schematic diagram for human speech production at the sub-glottal, glottal and supra-glottal levels. The process is exhibited for a fundamental frequency of 250 Hz.

Speech production is initiated by a stream of pulmonic air flowing from the subglottal structure, which acts as the source of sound energy that moves up to the larynx where it passes through the glottis. The intended speech can make the vocal folds vibrate to produce the fundamental frequency and its corresponding harmonics. This is the glottal pulse whose spectrum has a slope of approximately -12 dB per octave (Ladefoged, 1996; Stevens, 2000). The pulse train moves to the supraglottal structure in the vocal tract which comprises of the oral cavity, nasal cavity, hard palate, soft palate, velum, lips and tongue (Ladefoged, 1993; Stevens, 2000). The vocal cavity acts as a resonator which filters the waveform of the glottal pulses, and passes some frequencies better than others depending on the configuration of the

articulators at a given instant. This resonant function shapes the spectrum of the speech waveform. Figure 2.1 shows a schematic diagram for the three stages of speech production.

The initiation of speech production originates in various centres of the brain, where the coordinated process of *programming*, *planning* and *sequencing* for the implementation of the anticipated speech takes place (Kent et al., 2000). These processes are carried out by a neural activation that directs a timed and coordinated message to the musculoskeletal structure responsible for speech production. However, the musculoskeletal structure can work in an uncoordinated fashion if there is a damage to the central nervous system (CNS) or peripheral nervous system (PNS). Such a damage can result in a group of motor speech disorders (MSD), caused by weakness and incoordination in the speech musculature (Darley, Aronson, and Brown, 1969b; Duffy, 2005). MSD's resulting from such neurological impairments often leads to the condition of dysarthria. The condition of dysarthria is not localised to only the impairment in the musculoskeletal structure resulting from neural damage, but it can also have a wider effect on multiple parts of the supraglottal, laryngeal and the subglottal system depending on the extent and severity of dysarthria (Kent et al., 2000).

More formally dysarthria can be defined as:

"a collective name for a group of neurologic speech disorders resulting from abnormalities in the strength, speed, range, steadiness, tone, or accuracy of movements required for control of the respiratory, phonatory, resonatory, articulatory, and prosodic aspects of speech production." (Duffy, 2005, p. 5)

2.1.1 Neurological basis of dysarthria

Dysarthria is a result of single or multiple lesions inside the structures of the central or peripheral nervous system which might result in the loss of motor speech control due to muscular atrophy and lack of coordination. In order to understand the causes and types of dysarthria it is necessary to demarcate the regions of the central and peripheral nervous system at an anatomical level that are broadly divided in four categories:

- **Anterior and Middle brain:** The anterior and middle brain constitutes the largest

part of the brain, the cerebrum, initial cranial nerves (CN), designated CN-I and CN-II, important speech motor control structures[1] of the cerebral cortex and the basal ganglia residing inside the anterior brain and connected to the cerebral cortex.

- **Posterior Brain:** The brainstem and ten pairs of cranial nerves emerging from it designated as CN III - CN XII and the cerebellum. Out of ten pairs of cranial nerves, six are involved in controlling and innervating the motor speech system in some way. For example, trigeminal (CN V) and facial (CN VII) nerves control all the facial, mouth and jaw movements. glassopharyngeal (CN IX) and vagus (CN X) nerves contribute to control the movements of the pharynx, larynx and palatal regions. The accessory (CN XI) nerve control the shoulder and neck movements and the hypoglossal (CN XII) nerve is responsible for tongue movements (Duffy, 2005). All the above structures act as important articulators in the speech production process and damage or weakness in any of these can result in one or more forms of dysarthria.

- **Spinal Nerves:** A collection of 31 pairs of nerves connected to the spinal cord via posterior and anterior nerve roots. It is responsible for carrying motor and sensory messages to the various organs (Duffy, 2005).

- **Peripheral Nervous System:** The cranial nerves emerging from the brainstem and the spinal nerves, which innervates various muscle groups.

The anterior, middle and posterior parts of the brain contain specialised nerve fibers known as tracts or pathways. These pathways extend from the pyramidal system (Direct Motor System), which is composed of the corticobulbar tract that connects the cortex to the brainstem and the corticospinal tract. The pathways are also found in the extrapyramidal system (Indirect Motor System), which is mainly composed of the basal ganglia, red nucleus, substantia nigra and the cerebellum. The pyramidal and extrapyramidal system together is known as the Upper Motor Neuron (UMN) system[2]. The pairs of cranial nerves emerging from the brainstem along with the spinal nerves are collectively referred to as the Lower Motor Neuron (LMN) system.

[1] The cerebral cortex houses these structures in each of its hemisphere, which is responsible for its role in speech production. The components include *Primary Motor Cortex*, *Brocas Area*, *Insula* and *Supplementary Motor Area*. (Duffy, 2005; Kent et al., 2000)

[2] It should be noted that the UMN does not include the basal ganglia and the cerebellum parts of the extrapyramidal system (Duffy, 2005)

The UMN is responsible for direct voluntary and skilled movements and also controls posture and tone. The LMN, on the other hand, is responsible for producing muscle reflex, tone, and carrying out UMN commands for voluntary movements (Duffy, 2005). However, the coordinated and skilled muscle movements directed by the upper and lower motor neurons are a result of a pre-processing stage which comprises of programming, planning and sequencing of internal gestures of speech for carrying out respective motor actions. The intention to speak activates the core structures inside the central nervous system, which initiates the building of these internal models. For example, the cerebellum is responsible for creating internal models that simulate the dynamics of the musculoskeletal system, which is then used by the cortex directly to carry out the actions, rather than using the musculoskeletal system directly and the insular cortex plays a crucial role of sequencing speech segments (Kent et al., 2000). Hence, the critical steps of carrying out any linguistic plans, setting up prosodic constituents and timing of various syllables to produce phonetic segments is used implicitly by the upper and lower motor neuron systems to produce the desired output.

Motor speech disorders can be characterised by damage inside the central and peripheral nervous systems. The damage can either be acute or chronic and the extent of damage can be on a single structure, multiple structures or can be symmetrically spread across the central and peripheral nervous system (Duffy, 2005). Dysarthria can either be congenital in nature, such as cerebral palsy (CP), or acquired, where it develops preceding a phase of typical speech, as it can occur in stroke. The principal manifestations of dysarthria can be a result of the following disorders:

Degenerative Disorder is a motor neuron disease often characterised by a gradual decline and death of neuronal activity, as in amyotrophic lateral sclerosis (ALS).

Traumatic Disorder results from a head injury, as in traumatic brain injury (TBI).

Vascular Disorder is a cerebrovascular disease often characterised by obstruction of blood supply to the neurons resulting in oxygen and nutrient starvation, as in stroke.

Neurochemical Disorder often results from deficiencies and imbalance in the neurochemical system. Examples include Parkinson's Disease (PD) due to death in dopamine cells, myasthenia gravis due to death in acetylcholine cells.

Inflammatory and Neoplastic Disorders, which often results from attack to the nervous system due to toxins, micro-organisms and tumours.

Irrespective of the cause of a specific disorder, the prognosis for dysarthria is assessed by monitoring its developmental pattern over a period of time. For example, *Progressive* symptoms of dysarthria can exacerbate by presenting new symptoms over a course of time, hence making it difficult for treatment by failing to localise it. Parkinson's disease is an example of a progressive disorder. *Recovering* symptoms refer to a reduced effect of severity over time, as in stroke, but fails to recover completely. *Stable* symptoms are those that persist and do not show any further signs of deterioration over prolonged periods of time, as in cerebral palsy. *Recurring* symptoms are those that show initial phases of improvement followed by a sudden deterioration, and lastly *Transient* symptoms are less common but, usually heal completely over due course (Duffy, 2005).

2.2 Types of dysarthria

The types of dysarthria described below are more commonly referred to for diagnosis and treatment, and were originally formalised by Darley, Aronson, and Brown (1969a,b).

2.2.1 Flaccid dysarthria

Flaccid dysarthria is caused by damage to the lower motor neurons. It is sometimes referred to as bulbar palsy and it is characterised by hypotonia, weakness in muscle movement and poor reflexes. It can be caused by either unilateral or bilateral damage to the cranial or spinal nerves. It results in speech that is characterised by hypernasality, breathy voice, monopitch and imprecise consonant production. The most common causes of flaccid dysarthria result from stroke, surgical trauma, degenerative disease and muscular dystrophy. (Darley, Aronson, and Brown, 1969b; Duffy, 2005)

2.2.2 Spastic dysarthria

Spastic dysarthria is caused by damage to the upper motor neurons. It is sometimes referred to as pseudobulbar palsy and it is characterised by spasticity, hyperreflexia and

Babinski sign[3]. It results in speech that is characterised by imprecise consonant production, monopitch, reduced stress, slow rate, harsh and strained voice and hypernasality. The most common causes of spastic dysarthria result from degenerative disorders, with the most common being due to ALS. Multiple stroke and TBI are other common causes. (Darley, Aronson, and Brown, 1969b; Duffy, 2005)

2.2.3 Ataxic dysarthria

Ataxic dysarthria is caused by damage to the cerebellum. Ataxia (*"lack of order"*) induces errors in range, force, timing and direction of the speech muscle movements. It is characterised by slowness, hypotonia, jerky & dysrythmic movements, incoordination and lack of smoothness. It affects the respiratory, phonatory and articulatory aspects of speech production. The most common causes of ataxic dysarthria results from cerebellar degeneration as evident in Friedreich's ataxia (Eigentler et al., 2011), which is hereditary in nature (Delatycki, Williamson, and Forrest, 2000), and by damage to the myelin sheath of the neurons, as seen in multiple sclerosis (Darley, Aronson, and Brown, 1969b; Duffy, 2005).

2.2.4 Hypokinetic dysarthria

Hypokinetic dysarthria is caused by damage to the extrapyramidal system's basal ganglia circuitry. It occurs when the neurons in the substantia nigra component of the basal ganglia are destroyed that leads to the death of dopamine cells. It is characterised by rigidity in muscles, slowness and limited range in speech movements with low frequency termors of around 3-8 Hz. It results in speech with reduced stress, imprecise consonant production, monopitch, monoloudness and phases of inappropriate silences. The most common causes of hypokinetic dysarthria are an outcome of hypokinesia, which are a result of mostly Parkinson's disease and parkinsonism (Darley, Aronson, and Brown, 1969b; Duffy, 2005).

2.2.5 Hyperkinetic dysarthria

Hyperkinetic dysarthria is caused by damage to the extrapyramidal system's basal ganglia component or portions of the cerebellar circuitry. It is characterised by hyperkinesia, where there is a presence of abnormal and usually unexpected involuntary movements. The

[3]It is a reflex which is characterised when the big toe moves toward the top surface of the foot and the other toes fan out after the sole of the foot has been firmly stroked.

movement disorders evident in hyperkinetic dysarthria are exhibited in dyskinesia (abnormal involuntary movements), dystonia (slow hyperkinesia), chorea (fast hyperkinesia) and tremors (Duffy, 2005). Some of the common causes of hyperkinetic dysarthria are seen in degenerative disorders such as Huntington's disease, where the neurons in the caudate and putamen structures of the basal ganglia component are destroyed (Walker, 2007).

2.2.6 Mixed dysarthria

Mixed dysarthria is usually a combination of any of the above forms. For example Spastic-Flaccid dysarthria is caused by damage to both the upper and lower motor neurons. Ataxic-Spastic dysarthria is another common type of mixed dysarthria. The most common causes of mixed dysarthria results from degenerative (ALS, PD) and vascular disorders (stroke). By far ALS is seen as one of the major contributors to the mixed dysarthric types (Darley, Aronson, and Brown, 1969b).

2.3 Statistics of dysarthric etiologies

There are no official figures that gives an estimate of incidence and prevalence of various types and etiologies of dysarthria in the UK. It is roughly estimated that around 1% of UK population is diagnosed with a neurological disorder each year, which includes both progressive and non-progressive disorders, and not necessarily all the conditions lead to dysarthria (RCSLT, 2006, 2009).

Figure 2.2 gives an approximation of the incidence and prevalence of various dysarthric etiologies. The incidence is the newly diagnosed cases of an etiology within a period of time and prevalence refers to the actual number of cases that lead to dysarthria during a period of time or a particular date in time when the data was recorded. Stroke has the highest prevalence of dysarthria in the UK with around 416 per 100,000 individuals affected by the condition. It is followed by cerebral palsy with a prevalence of 200-300 per 100,000. The etiologies from motor neuron disease down to multiple system atrophy in figure 2.2 can be collectively termed under progressive neurological disorders. They are the least affected groups of dysarthria in UK.

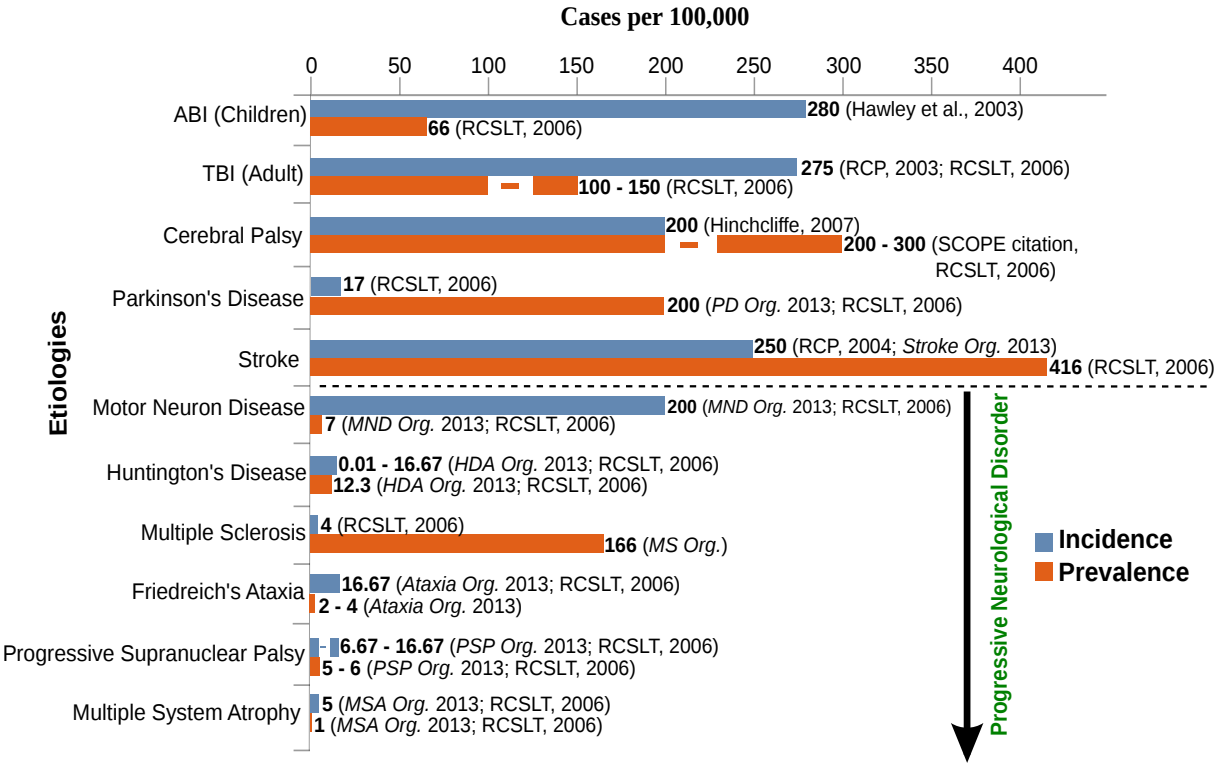

Figure 2.2: Incidence and prevalence of some major dysarthric etiologies. All estimates are based on the current UK population of approximately 60 million.

2.4 Effects of dysarthria

Dysarthria can simultaneously affect one or more components of the speech production system where an individual can have an impaired respiratory system, laryngeal or velopharyngeal dysfunction or imprecise articulation due to weakness, abnormal muscle tone and incoordination (Kent and Kim, 2003a). An individual can show laryngeal deficiency as evident in flaccid dysarthria due to the affected vagus nerve (CN-X) or exhibit multilevel system failure as its more evident in hypokinetic dysarthria associated with PD or stroke (Kent et al., 1999a).

The complexity that is associated with the classification of neurological speech disorders based on etiology, type or severity is a challenging task (Kent and Kim, 2003a; Kim, Kent, and Weismer, 2011a). Since the introduction of the first classification system by Darley, Aronson, and Brown (1969a,b), researchers have used this classification system for both

research and clinical purposes. However, there has been uncertainty in regard to the degree of accuracy, validity and reliability of the perceptually motivated classification system (Van Der Graaff et al., 2009; Zyski and Weisiger, 1987). This can give sub-optimal classification accuracy that might not be clinically acceptable for a robust management and treatment plan for dysarthria. The failure of perceptual methods alone to give accurate dysarthric classification results can be attributed to the following measures:

- **Judgement Inconsistency** resulting from inter and intra judgement scoring variations, which can be further affected by lack of skilled and experienced judges. For example, one way to increase the robustness of the perceptual assessment process is to get an affirmation that the judges will describe similar type and degree of errors, which is not often the case. One way to achieve sensitivity and increase inter-judge reliability could be to use a broad scoring system (Enderby, 1988).

- **Inaccurate Detection** of concurrent neurological impairments occurring in two or more components of the speech production system is another shortcoming of perceptual analysis. This is likely to happen because different judges might not have sufficient training (Kent and Kim, 2003a; Kent et al., 1999a).

- **Lack of Quantification** of perceptual features might result in inconsistency of speech dimensions. For example, manifestation of critical frequency bands depicting long-term phonatory instability are termed as wow (1-2 Hz), tremors (2-10 Hz) and flutters (10-20 Hz) of dysarthric speech (Hartelius, Buder, and Strand, 1997). These could act as one of the unique distinguishing feature to identify the neurological disorder. Ataxic dysarthria associated with cerebellar lesions, usually exhibit a low frequency tremor of around 3 Hz during a sustained vowel phonation task (Ackermann and Ziegler, 1991). A similar experiment showed a higher frequency tremor on individuals with amyotrophic lateral sclerosis (Aronson et al., 1992). The detection of such a delineating feature is very difficult to be assessed by perceptual judgement alone and the exact frequency level of these disturbances need a closer inspection.

- **Overlapping Features** is another persistent problem in perceptual analysis, which makes it extremely difficult to classify various neurologic disorders. In the original classification system by Darley, Aronson, and Brown (1969b), imprecise consonant is one of the deviant speech dimension, which is common across all dysarthric types.

Hence, it cannot act as an indicator for discerning any particular dysarthria, and is rendered ineffective for this purpose. One way to make imprecise consonant as a useful indicator could be to map its acoustic correlates which might give a clear and quantifiable description of its worthiness in distinguishing across dysarthria types.

In order to overcome the limitations of the perceptual assessment process and to increase the reliability of identifying dysarthria types, and assessing intelligibility, an instrumental approach can be followed. As an example, the objective and quantitative process of acoustic analysis can act as one such tool, either in addition to the perceptual assessment or as a standalone process.

As an example, figure 2.3 shows a spectrogram comparison for typical and dysarthric speakers with varying severity. The spectrogram is plotted for the word *backspace*, which is one of the command words in the UASPEECH database (Kim et al., 2008). It can be seen in part-(a) that the typical production of the utterance is around 650 ms with well defined temporal and spectral structures for the expected vowels, stops and the fricatives. The part-(b), which exhibits a dysarthric speaker with high intelligibility is seen to be very similar to the typical speech. However, it is about 20% slower than a typical utterance of the same word and the temporal delay is noticeable at the intra syllabic level. Lastly, part-(c), which shows the same utterance for a dysarthric speaker with very low intelligibility shows both temporal and spectral disfluencies. It is around 2.5 times slower than the typical speech production. There is also marked difference between the production of the vowels. The second vowel is about twice as long as the first one, possibly due to fatigue. Lastly, the sibilant at the end of the word *backspace* is generally characterised by concentration of high frequency energies (Ladefoged, 1993), which seems to be less prominent in low intelligibility speaker in comparison to the typical. This might possibly hint towards respiratory insufficiency and lack of muscle coordination. Although, figure 2.3 presents information only for a single typical and dysarthric speaker with very-low and high intelligibility, it gives some informative cues about the underlying acoustic realisation of the observed phonetic tokens.

In addition, instead of a broad acoustic examination (as in spectrogram), a more fine tune approach can also be followed, where a single or a small group of acoustic variables are examined. For example, it has been shown that a reduced F2 slope is usually found to be proportional to intelligibility (Kent et al., 1989; Rong et al., 2012a). As a demonstration, figure 2.4 shows the F2 plot comparison of multiple utterances of the word *"alpha"* for a typical and dysarthric speaker with cerebral palsy. It is evident that the F2 trajectories

tend to overlap each other for the typical speech and shows a more consistent pattern across utterances. On the other hand, dysarthric speech shows a flat and skewed F2 pattern across the utterances.

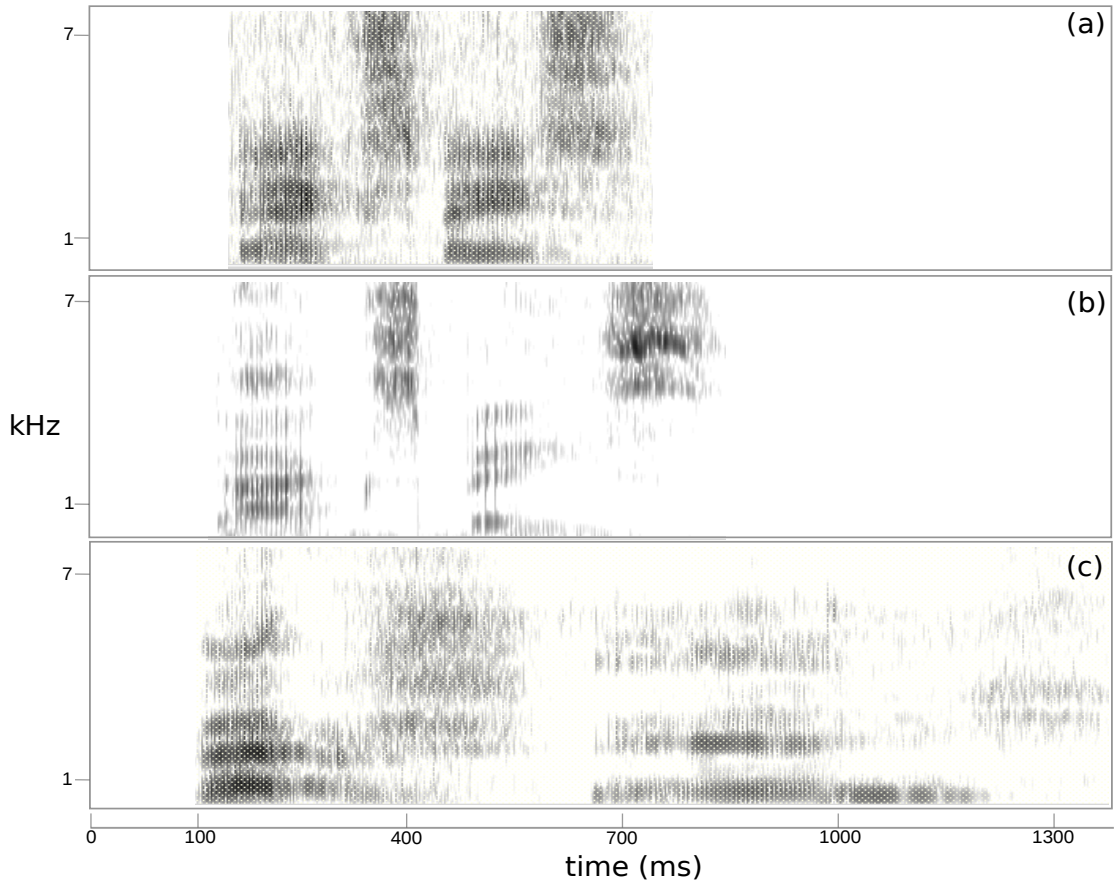

Figure 2.3: Spectrogram comparison of the word *backspace* for a (a) typical speaker, (b) dysarthric speaker with high intelligibility and (c) dysarthric speaker with very low intelligibility.

Acoustic analysis gives a promising aspect to systematically quantify certain perceptual aspects of speech. For example, the demonstration in figure 2.3 can tempt us to conclude that dysarthric speech tends to get slower with decreasing intelligibility or it gives an insight into understanding the relation that intelligibility might have with the distribution of formants. Also, observations of the F2 trajectory as shown in figure 2.4 gives a convincing qualitative perspective, but it does not have a simple quantitative measure that can be

useful from a classification or ASR perspective. Hence, the inter and intra speaker variabilities in the dysarthric speech can make the temporal and spectral analysis a very difficult process.

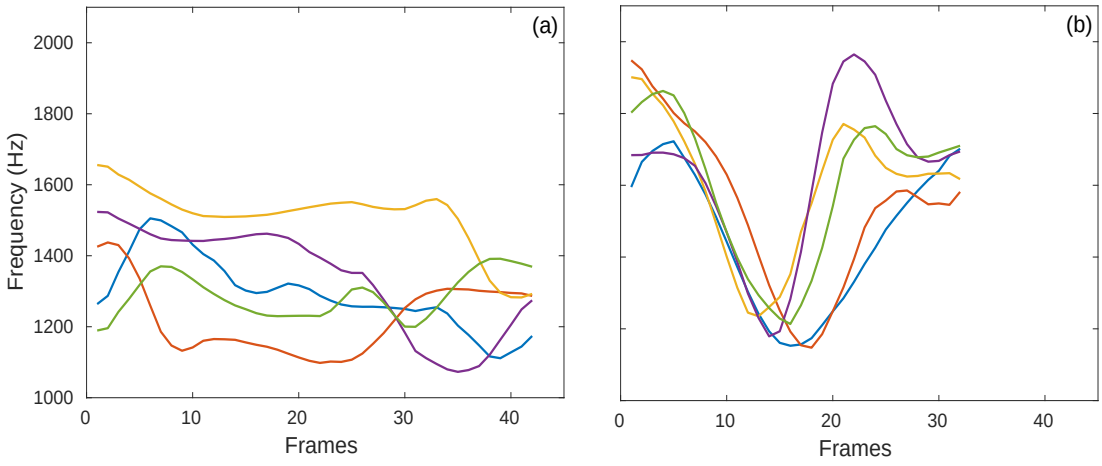

Figure 2.4: F2 trajectory comparisons for a (a) speaker with cerebral palsy and (b) typical speaker for five utterances of the word *alpha*. Moving average smoothing was applied on the F2-slopes with a span of 5.

Despite the challenges in conducting acoustic analysis, it is used as an informative tool for automatic prediction of dysarthric intelligibility and type. The original Mayo System (Darley, Aronson, and Brown, 1969a,b) hypothesised the classification process according to dysarthria type as a better alternative, which has been followed in lot of classification studies. However, the perceptually driven approaches have been proved otherwise by more recent acoustically motivated studies. One such comprehensive work was conducted by Kim, Kent, and Weismer (2011a) on 107 speakers with dysarthria. It covered four broad etiological types (Parkinson's disease, stroke, traumatic brain injury and multiple system atrophy) and seven different dysarthric types with varying degree of severity. Eight acoustic variables were used in the study and they achieved an overall score of 68.6% for etiological classification and 54.9% for severity classification, which was significantly better than the type classification scheme, which gave a score of 31.7% across all the speakers. In addition to classification tasks, acoustic enhancement approaches have been exploited for increasing the overall intelligibility of dysarthric speech (Kain and Santen, 2009; Kain et al., 2007; Lalitha, Prema, and Mathew, 2010).

Research still has a long way to go in fully quantifying the differences present in the dysarthric speech signals. Such variabilities could be exploited to improve the ASR framework for designing better dysarthric speech systems. However, in order to achieve this, a more detailed understanding of the relationship between the acoustic measures of disordered speech with its underlying perceptual correlates and the articulatory features is needed.

2.5 Severity and impact on intelligibility

Severity and Intelligibility of dysarthric speech are terms often used as an analogue (Kim et al., 2008). Severity is broadly operationalised by the speech intelligibility index as ***Mild***, ***Moderate*** or ***Severe*** or any condition within, such as, ***Mild-Moderate*** etc. An intelligibility rating itself is generated by perceptual judgement, which is based on an objective scoring scheme. There is no common consensus of how to group speakers with dysarthria into different intelligibility groups. For example, Kim, Kent, and Weismer (2011a) used a direct magnitude estimation approach (Gescheider, 1976) conducted by non-expert listeners to give intelligibility scores as scaled ratios, which were divided into broad severity groups of *mild, moderate and severe* based on some operationally selected scoring range. In another study, a multiple regression modelling approach was followed by De Bodt, Hernndez-Daz Huici, and Van De Heyning (2002) to quantify intelligibility as a linear combination of four perceptual global dimensions of speech: voice quality, articulation, prosody and nasality. The analyses was conducted by expert listeners and articulation was found to be the strongest predictor of intelligibility.

However, there is often a problem in such perceptual judgements to predict intelligibility. It is usually a costly and time-consuming process to recruit listeners and conduct tests. Also, inter-judge and intra-judge variations can often give variable results. An expert can assign higher intelligibility scores to an otherwise less intelligible person, because of their better comprehension skills of dysarthric speech (Bunton et al., 2007; Tjaden and Liss, 1995). Since intelligibility is partially related to the proficiency of the listener to understand the acoustic signal with accuracy, one of the ways to increase the effectiveness of the perceptual judgement methods is by presenting supplementary information to the listener. Figure 2.5 shows an example framework that can be adapted by the listeners to increase the efficacy of perceptual judgements. The block model shows three levels of information (Hustad, 2008) that can be exposed to the listener:

- Level 1 (Surface Code): Exact syntax and morphology of the spoken text.
- Level 2 (Propositional Model): This is the intermediate level, which contains the textbase information, which refers to the propositions or meanings that are extracted from Level 1 (the semantics of the message).
- Level 3 (Situational Model): This is the highest level of information representation which contains the assimilation of Level 2 information with the world knowledge. The external knowledge that can be integrated is usually in the form of comprehension and contextual cues viz. audio-visual, gesture etc.

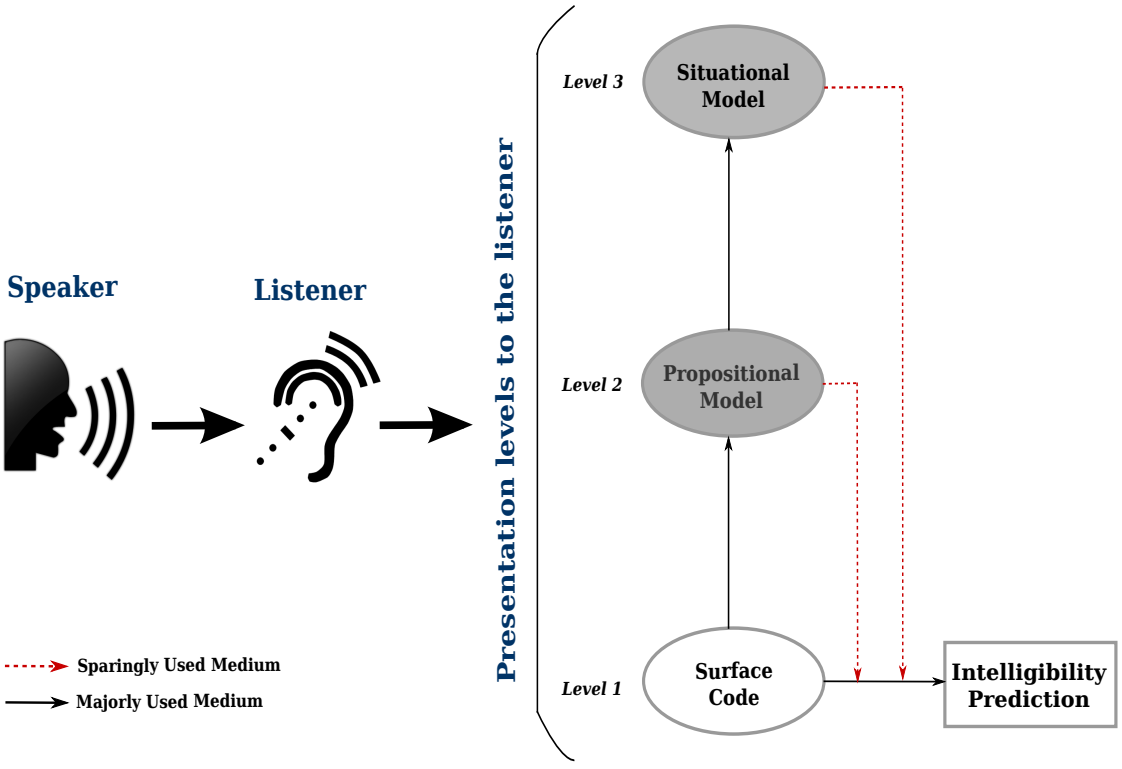

Figure 2.5: A holistic listening model for perceptual judgement of intelligibility.

In practice, for the purpose of quantifying intelligibility in a laboratory setup, only Level 1 (Surface Code) information is ever used by the listener to predict the scores, where the focus is primarily on phonetic identification accuracy and other higher level information is either not utilised or ignored. However, in real life, anyone interacting with a person with

dysarthria will usually use all three levels of information to communicate. If the listener is able to comprehend other contextual information in addition to the acoustic signal, it can greatly assist in a meaningful communication exchange. One such study by Hustad (2008) tried to map a relationship between the listeners comprehension abilities and the intelligibility scores. Although, in the study they did not find any significant relationship, but it was observed that listeners were better able to comprehend dysarthric speech rather than just predict intelligibility scores based purely on orthographic transcriptions. The effect of using higher levels of contextual information, such as, listener familiarisation with the dysarthric speech (Hustad and Cahill, 2003; Tjaden and Liss, 1995), audio-visual cues (Hunter, Pring, and Martin, 1991; Hustad and Cahill, 2003) or letter cueing (Hunter, Pring, and Martin, 1991) has proved to be beneficial for improved intelligibility ratings.

Although, the listener model shown in figure 2.5 forms a more accurate framework to get improved judgement of intelligibility, it will however come with functional limitations. There is a need to explore other methodologies which are more suited to give realistic estimates of intelligibility. One of the ways to address this problem is to explore the acoustic domain to find a set of variables that could assist in predicting intelligibility by direct signal scanning. If such a tool can be devised, it can help to minimise the negative effects of perceptual ratings discussed earlier.

For example, in a study by Kim, Kent, and Weismer (2011a) they predicted scaled intelligibility scores based on perceptual ratings and used regression analysis for selected acoustic variables against the scaled perceptual scores. They found out that both F2 slope and articulation rate exhibited a strong relationship to each other. If articulation rate can be shown to have a strong association with intelligibility then such acoustic measures can be utilised to quantify speech intelligibility automatically. In another study, a reduced F2 slope (typically long and flat in nature), which is indicative of minimum or no articulatory movement, amongst a group of ALS patients was found to be highly correlated with the speech intelligibility (Kent et al., 1989). Similar findings of reduced F2 slopes amongst speakers with cerebral palsy was found to have a relationship with reduced intelligibility (Rong et al., 2012a). The studies mentioned do give a strong indication that F2 slope transitions could be one of the acoustic measures that can prove to be a strong indicator to quantify intelligibility. However, there is no exhaustive study to ascertain the efficacy of F2 slope transitions on other neurological speech disorders or large vocabularies.

Despite the difficulties, researchers have attempted to measure the effectiveness of some

common acoustic variables for intelligibility estimation. For example, the PEAKS system (Maier et al., 2009) was designed to analyse voice and speech disorders and automatically predict intelligibility. It uses a forward acoustic feature selection approach based on multiple linear regression that selects the best weighted combination of acoustic features. The experimental setup of PEAKS system predicted intelligibility scores of 41 laryngectomees and 31 children with cleft lip/palate in agreement to five expert listeners rating within 95% confidence interval. A similar sequential forward feature selection approach was adapted to sift the best acoustic features by Paja and Falk (2012). They managed to select the 9 most salient acoustic features out of the extracted 50. It showed a 13% improved intelligibility prediction using discriminant analysis on 10 speakers with spastic dysarthria from the UASPEECH database (Kim et al., 2008). Another novel approach for predicting intelligibility was conducted by Kim and Kim (2012) that used an iterative feature selection approach. It minimised prediction errors and kept low mutual dependency amongst the selected features at each incremental step to select the best feature sets.

Despite some of the recent advances, it can be said that it is generally difficult for a single acoustic measure to give a reasonable estimate on intelligibility and if more than one acoustic variable is used, we can run into data sparsity issues manifest in dysarthric speech. Hence, there is a need to explore such variable(s) that are independent of speech material and are more robust to data sparseness problems. The study conducted in this thesis will attempt to explore one such acoustic measure in the later chapters that will be effective at predicting intelligibility on different datasets.

2.6 Treatment and management of dysarthria

A complete discussion of the treatment and management approaches is outside the scope of the thesis, but it is important to conclude the chapter by giving a brief summary. It will not only shed light into the wide range of interventions that are applied for alleviating the symptoms of dysarthria, but it will also show the importance of using computer aided devices for long-term management of severe cases of dysarthria. Since, speech can be a natural interface for such computer aided devices, it will re-emphasise the importance of improving dysarthric speech recognition in the future.

The goal of managing motor speech disorder is to effectively increase the efficiency and naturalness of the intended communication. However, this is not easy, because a single

etiology can be responsible for various kinds of dysarthric conditions and the severity of dysarthria can further make the management process difficult. Other factors such as societal limitations and specific communication needs add to the difficulty in reaching a consensus on management and treatment of various dysarthrias (Duffy, 2005). Despite such challenges, constant effort has been put in the last five decades to formalise the foundations for an effective management and treatment plan for various motor speech disorders. This section will give a brief overview of various management and treatment approaches.

2.6.1 Directions for management of speech disorders

The approaches for the management of disordered speech is mainly intended to work in three main directions as shown in Duffy (2005):

- *Restoration:* Aims to reduce the effect of impairment and is dependent on the underlying etiology and severity of the speech disorder.
- *Compensation:* Aims to promote the usage of prosthetic and alternative communication devices to increase the overall intelligibility and communication efficacy. Compensation is usually applied when full recovery of speech is not possible.
- *Adjustment:* This approach is followed when it becomes known that the underlying speech will be difficult to manage through restoration and compensation. For example, a degenerative dysarthria will have a regular deterioration of speech over a period of time and can often be adjusted by gradually changing the lifestyle and environment.

2.6.2 Approaches for management of speech disorders

Management approaches can be categorised into five major categories: *Medical Intervention, Prosthetic Management, Behavioral Management, Augmentative and Alternative Communication and Counselling & Support* (Duffy, 2005). Since there is no single approach for treating specific motor speech disorder, the management and treatment approaches tend to overlap with each other. It facilitates a multidisciplinary approach that can be effective in treating a variety of dysarthric etiologies like PD (Marck et al., 2009; Skelly, Lindop, and Johnson, 2012), stroke (Langhorne, Bernhardt, and Kwakkel, 2011), ALS (Oliveira and Pereira, 2009) etc. Figure 2.6 shows an overlapping view that any possible combination of approaches are used for the management and treatment of motor speech disorder.

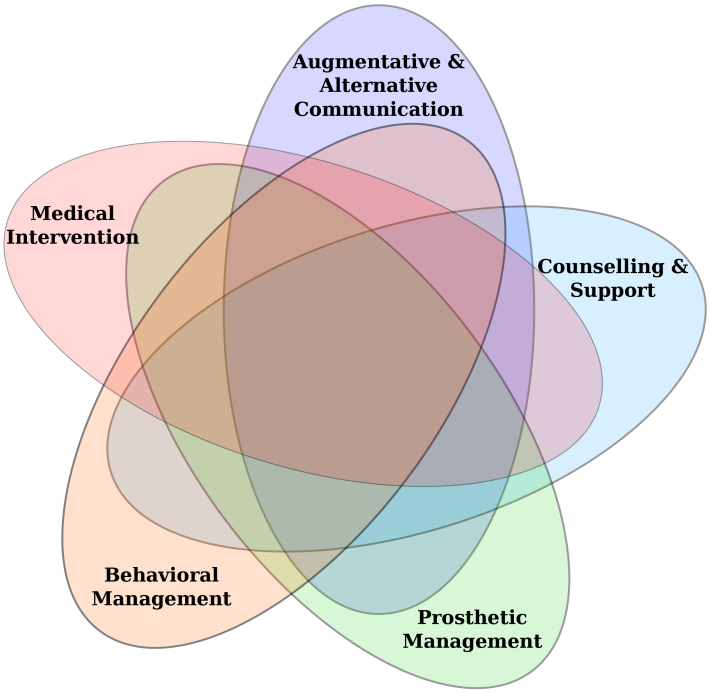

Figure 2.6: Management approaches for motor speech disorder.

2.6.2.1 Medical intervention

Medical intervention can follow one of the following approaches:-

Pharmacological Approach is a drug based intervention aimed majorly at treating the underlying cause of the neurological disorder. For example, to treat Parkinson's disease, the drug levodopa remains the main approach of treatment (Katzenschlager and Lees, 2002), and its effectiveness can be enhanced if used in conjunction with inhibitors like carbidopa (Hussain and Manyam, 1997; Tourtellotte et al., 1982) for improving rigidity and speed response of upper body extremities. Some other inhibitors like Mestinon have proved beneficial in maintaining normal phonation activity in flaccid dysarthria (Neiman, Mountjoy, and Allen, 1975). To minimise the long term side effect of drugs like levodopa (Fedorova and Chigir, 2007; Lieu et al., 2010; Skodda, Visser, and Schlegel, 2010), alternative medicine approaches are also explored that showed the neuroprotective and therapeutic utility of natural herbs (Mythri, Harish, and Bharath, 2012). For example, extracts of

Mucuna Prurien[4] seeds have proven to be more effective than the synthetically available l-DOPA in animals with a reduced risk of drug-induced dyskinesia (Hussain and Manyam, 1997; Lieu et al., 2010) and the herb also showed significant improvement in 23 PD patients with mild or infrequent side effects (Vaidya et al., 1978).

Surgical Approach involves treating the underlying neurological speech impairment directly using surgical intervention. Neurosurgery is performed at the part of the brain where a possible neural defect is detected. For example, Deep Brain Stimulation, which involves implanting electrodes for sending high-frequency electrical impulses to affected areas of the brain has proved as an essential surgical treatment in overcoming detrimental effects of drug-induced dyskinesia and motor fluctuations as evident in parkinsonism, essential tremors (Halpern et al., 2007) and episodic genetic movement disorders associated with hyperkinetic movements (Kaufman, Mink, and Schwalb, 2010). Other surgical approaches like Thyroplasty and Arytenoid adduction are successfully applied to correct speech disorders due to glottal insufficiency and vocal fold paralysis. Another beneficial approach is found in the usage of Botox (botulinum toxin) injections that has shown to improve spasticity, hyperkinetic movements and tremors (Evidente and Adler, 2010) and damaged muscles in patients with cerebral palsy (Kalinina et al., 2000).

2.6.2.2 Prosthetic management

Prosthetic management incorporates a number of mechanical and electronic devices to improve specific motor speech functions. For example, dysarthria that results from velopharyngeal insufficiencies often lead to the problem of hypernasality. The design of an appropriate Palatal Lift Prostheses can be implemented for soft palate elevation to enhance voice quality and articulation (Lang, 1967; Witt et al., 1995) in people with dysarthria (Esposito, Mitsumoto, and Shanks, 2000; Marshall and Jones, 1971; Ono et al., 2005). Other devices that has been helpful to increase the intelligibility of speakers with dysarthria (Shimura and Kakehi, 2011; Van Nuffelen et al., 2009, 2010) include Pacing Boards (Lang and Fishbein, 1983).

[4] A legume extensively used in Ayurvedic medicine, which contains high concentration of levodopa.

2.6.2.3 Behavioural management

Behavioural management approaches aim to maximise communication between the speaker and the listener. It can be **Speech-oriented** or **Communication-oriented**. The former uses physiologic support system or compensation strategies to improve overall speech intelligibility, efficiency and naturalness and the latter uses modification strategies targeted to maximise the impact of communication even when the speech itself is not improved (Duffy, 2005). Before any approach can be adopted, it is the role of the clinical team to identify any specific need for the routine to be followed. For example, to use behavioural management approach for the treatment of deteriorating speech in ALS, the patients must score a 6 or 5 on the speech subscale[5] of the ALS severity scale (Yorkston et al., 2012). Behavioural management approaches encompass a number of strategies to improve inadequacies pertaining to the respiratory, phonatory, resonatory & articulatory aspects of speech.

Respiration: The techniques focus on improving the subglottal system where the breathing and respiratory system is compromised and is more common in flaccid dysarthria (Duffy, 2005). It employs non-speech techniques like biofeedback therapy (Thompson-Ward, Murdoch, and Stokes, 1997) and pushing-pulling techniques, which are suited for people unable to generate sufficient subglottal air pressure (Yorkston, Spencer, and Duffy, 2003). Another approach that is effective to increase the pulmonary insufficiency and vital capacity involves imitating "frog breathing (glossopharyngeal breathing)" (Harries and Lawes, 1957; Johansson, Nygren-Bonnier, and Schalling, 2012; McKeever and Miller, 2002).

Phonation: The techniques focus on improving vocal fold adduction problems that are either incomplete (hypo) or excessive (hyper). Phonation techniques are more commonly applied on patients with flaccid, hypokinetic dysarthria for hypoadduction, hyperkinetic dysarthria for hyperadduction of the vocal folds (Duffy, 2005; Yorkston, Spencer, and Duffy, 2003). Lee Silverman Voice Treatment (LSVT) is one such approach, which, focuses on increasing the vocal loudness and reducing breathiness through a structured exercise programs. The therapeutic efficacy of LSVT has been reported in improving intelligibility, loudness and phonation, in PD (Ramig et al., 1995, 1996, 2001; Sapir et al., 2002, 2007; Theodoros et al., 1999; Whitehill et al., 2011), TBI & stroke (Mahler and Ramig, 2012; Wenke, Theodoros, and Cornwell, 2008, 2011), CP (Fox and Boliek, 2012), ataxia (Sapir et al., 2003) and multiple sclerosis (Sapir et al., 2001).

[5] The Speech Subscale assigns a numerical score for speech, swallowing, lower extremity & upper extremity functions within a range of 10 - 1. (Yorkston et al., 2012)

Resonance: It is generally believed that resonance effects are not usually benefited from behavioural management approaches, and prosthetic or surgical methods should be sought. However, some approaches that are followed in the literature guidelines include modification of speaking patterns to reduce the effects of velopharyngeal inadequacies, exaggerate jaw movement for wider oral opening and supine position speaking (Duffy, 2005).
Articulation: The techniques focus on improving the place and manner of articulation for optimal vowel and consonant production. Strength Training is beneficial for improving motor control (Smith and Kurian, 2012) and is used to increase the strength of specific articulatory muscles, e.g. tongue-strengthening program (Dworkin and Hartman, 1979), orofacial myofunctional therapy to improve the strength and mobility of the buccal, facial, labial and lingual musculature to improve speech intelligibility (Ray, 2002).

For some speakers, behavioural speech-oriented approaches may be inappropriate and ineffective due to the severity of the condition or the progressive nature of dysarthria. In these situations, clinician may focus on communication-oriented approaches to improve speaker intelligibility for effective message exchange. These are supplemental augmentative and alternative communication strategies that require an effort on part of the speaker, listener or both to increase comprehension. For example, a speaker can usually prompt and alert the listener about their intended mode of communication. The presentation of prompts like *Alphabet & Topic Cues* and *Iconic Hand Gestures* reported a significant increase in the intelligibility and listener's attitude towards effective communication (Hustad, 2005; Hustad and Garcia, 2005; Hustad and Gearhart, 2004; Toy and Joubert, 2008).

2.6.2.4 Augmentative and alternative communication (AAC)

AAC can be defined as:

"Any method of communicating that supplements (augments) or replaces (provides an alternative to) the usual methods of speech and/or writing where these are impaired or insufficient to meet the individual's needs." (Murray and Goldbart, 2009)

Mode	Example
Unaided no-tech	Facial Expressions, Gestures, Sign Languages, Symbols
Aided low/light-tech (mechanical/electronic)	Symbols, Communication Boards, Alphabet Supplementation, Portable Writing System
Aided high-tech (mostly electronic/computer-based)	Speech generating devices, Voice output and Voice input/output communication devices

Table 2.1: Different modes of AAC strategies and components.

AAC is a mixture of a wide variety of tools ranging from unaided low-cost strategies to aided high-tech dedicated computer systems. Since technology is changing rapidly with more robust and effective AAC devices, it is not the intent of the current section to cover them in detail. We give a very brief summary of some generic tools mentioned in table 2.1 and its potential usage for certain types of dysarthria.

Around 0.05% of the UK population could benefit from power aided communication devices (Enderby et al., 2013). For example, Yorkston et al. (2012) asserts the usage of AAC in progressive diseases such as ALS, when the patient's speech subscale rating drops to 4 or 3 (see section 2.6.2.3). The level of AAC intervention is highly dependent on the degree of speech deterioration. A collective review (Beukelman et al., 2007) showed the increased acceptance and usage of AAC for people with ALS and in a study by Ball, Beukelman, and Pattee (2004), nearly all the ALS patients continued with the usage of AAC devices.

AAC has proved as an effective self-repairing system for potential communication breakdown which is common in acquired progressive dysarthria (Bloch and Wilkinson, 2004). A review study based on assessment and interventions defined on WHO's ICF-CY[6] framework for measuring health and disability showed an effective categorisation of the best suited AAC systems for children with CP (Clarke and Price, 2012). Also, the usage of both low and high tech AAC devices like Alphabet Supplementation and VOCA has shown limited, though significant success for PD speakers (Armstrong, Jans, and MacDonald, 2000; Yorkston et al., 2012). The percentage of people from various dysarthric etiologies who could benefit from AAC devices is shown in figure 2.7. The percentages are taken as a gross estimation of the prevalence estimates in figure 2.2.

[6]International classification of functioning, disability and health for children and youth.

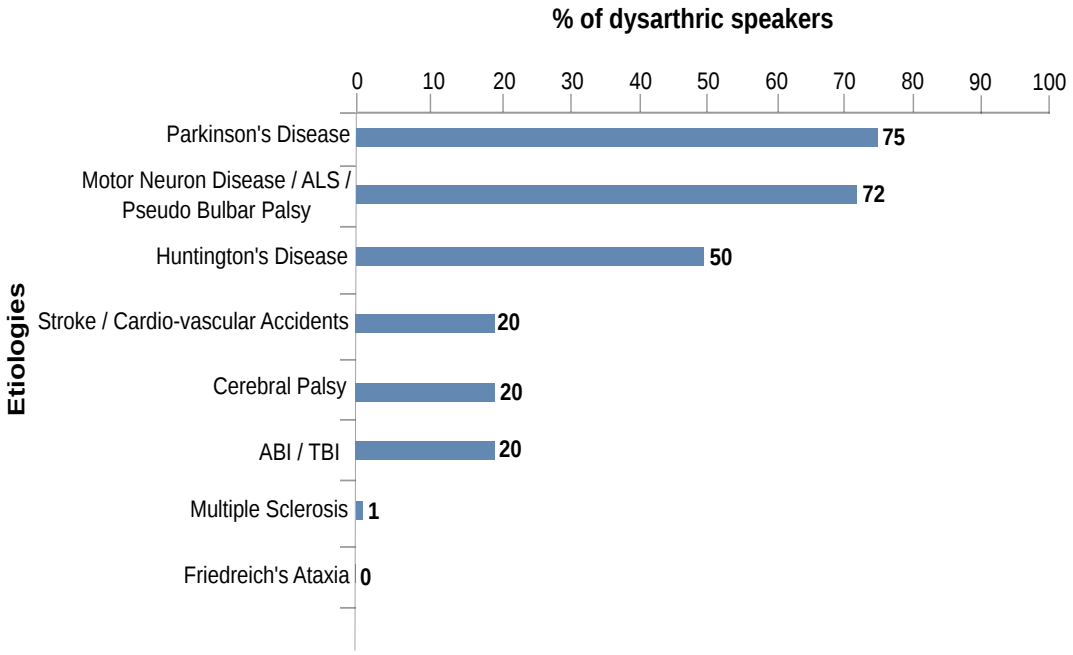

Figure 2.7: Dysarthric speakers who could benefit from AAC devices (Enderby et al., 2013).

Despite the advent of AAC strategies, there is still a greater need to understand individual AAC needs and address possible limitations in the current systems. Studies like Dowden (1997) and Light et al. (2007) elaborate on the functional knowledge needed to re design better AAC systems in the future, especially for young children, which will be more effective. Another problem with high-tech AAC devices such as VOCA is its over reliance on key or switch based interaction, which is slower and does not give a real-time communication experience. A possible solution for this problem can be envisaged in incorporating a new form of AAC device, viz., Voice Input Voice Output Communication Aid (VIVOCA) that recognises disordered speech, builds appropriate messages and coverts it into synthetic speech, thus facilitating real-time and natural communication (Hawley et al., 2012).

Chapter 3

ASR and its Applications in Dysarthric Speech

Speech offers the potential as an effective and natural interface of communication for people with dysarthria. The design of any dysarthric speech system generally requires careful customisation and optimisation to various components of an ASR framework. It is thus imperative to give a general overview of these components and understand its internal working. The chapter will give a brief explanation of the main component of an ASR framework with emphasis on the speech components that will be more relevant in context of the thesis. The chapter will conclude by expanding the discussion of ASR in general to how research has progressed in the field of dysarthric ASR.

3.1 Generic architecture of an ASR System

An ASR system generally comprises of components with pre-defined tasks that combine to translate the input speech into output text. The components either work individually or in mutual correspondence with each other to produce the output of the system. The accuracy of an ASR system is related to the optimal performance of each of these components. Figure 3.1 shows the architecture of a generic ASR system. It is not mandatory for an ASR system to have all the components and it can have fewer or even more components catering to specific tasks. A brief description of each ASR component is given below:

Feature Extraction converts the incoming speech into feature vectors that are used by machine learning algorithms for modelling the data.

Vocabulary is the domain of words that will be used. It is one of the constraints under which the ASR system produces its output. For example, isolated digit task will usually have a vocabulary of less than 10 words and a large vocabulary speech system can have its vocabulary exceeding 60k words.

Acoustic Models are responsible for representing a relationship between the incoming acoustic features to its respective speech units (words, tri-phones etc.) as constrained by the vocabulary.

Language Models are responsible for assigning estimates to the occurrence of word or sub-word units in a particular language, which puts a syntactic and semantic constraint on the overall recognition task.

Pronunciation Models refine the search hypothesis of ASR systems by feeding knowledge about the way particular words are spoken.

Adaptation Module is a predecessor layer which adjusts the various components in the modelling layer. It attempts to reduce the mismatch between the training and test condition (particular speaker, task, etc.).

Hypothesis Search produces the output hypothesis that best matches the incoming speech signal, and is constrained by the knowledge of language and pronunciation modelling components.

Corpus is a collection of audio, video and text files that are used for training and adaptation of the ASR systems.

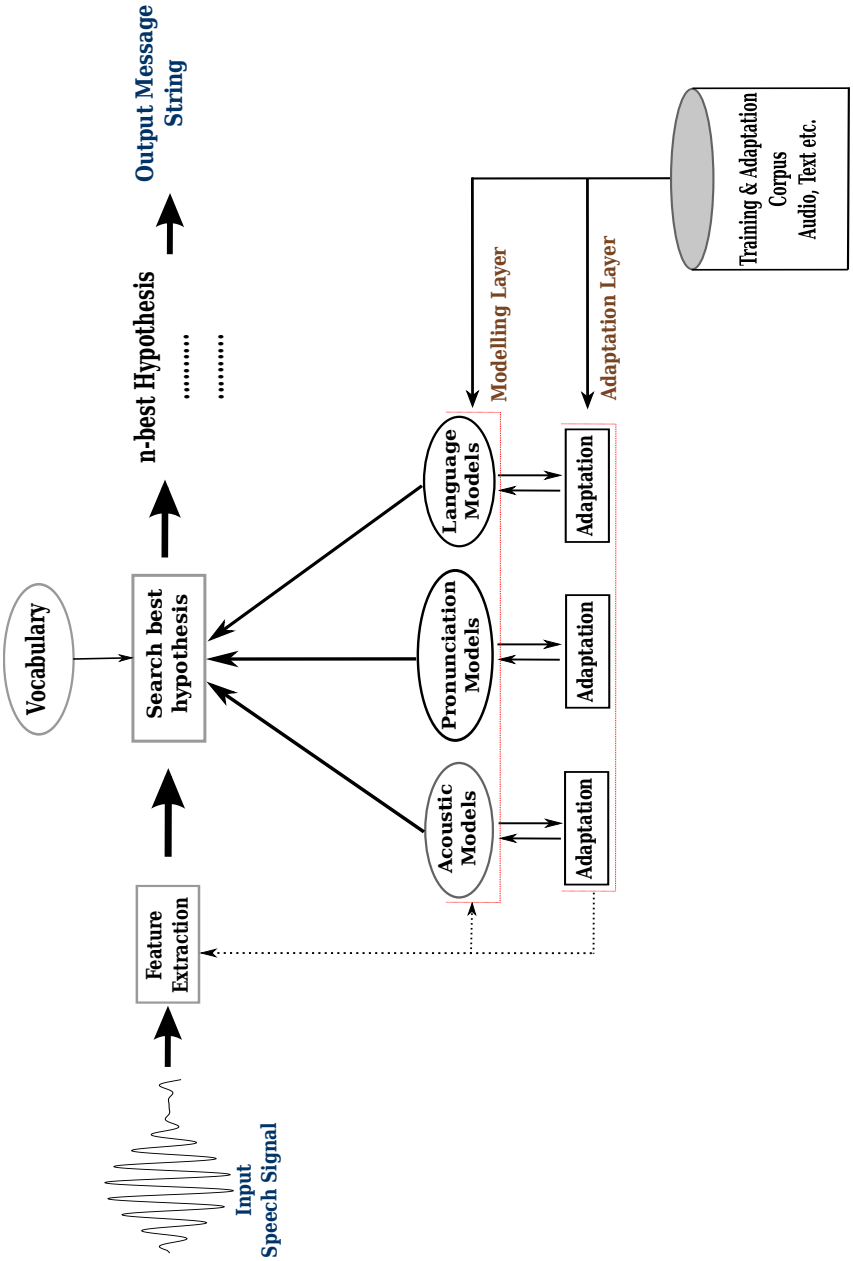

Figure 3.1: Generic architecture for the ASR systems.

3.2 Feature extraction

Feature extraction converts the incoming acoustic signals into a particular representation of feature vectors, which are more suited for the speech recognition process. It starts by sampling the incoming continuous sinusoidal wave into a set of discrete sample points. It is important to ensure that the sampling rate is high enough to capture the rapidly fluctuating speech patterns. Selection of the sampling rate for any system is given by the rule:

$$f_s > 2f_n \tag{3.1}$$

f_s is the sampling rate and f_n is commonly known as the *Nyquist frequency*, which is the highest frequency component in the wave that is being sampled (Oppenheim and Schafer, 1989). Any frequency component above the Nyquist frequency will be incorrectly represented by the same amount below the Nyquist frequency. This anomaly is often referred to as *aliasing* that can be avoided by removing the high frequency components using a low-pass filter before any further processing. The inequality 3.1 ensures that the sampling frequency will be high enough to represent the frequencies below f_n. The amplitude information of these sample points are represented by numeric values using a process called **quantisation**. The values are determined by the number of bits used for each sample, also called bit-depth. The signal-to-noise (SNR) ratio is represented in decibels as:

$$SNR_{DB} = 20\log(2^B) \tag{3.2}$$

where B is the number of quantisation bits used. For example, it can be computed from equation 3.2 that a 4-bit representation gives a SNR of around 24 dB and a 16-bit representation results in 96 dB SNR, which is about the full range of intensities that the human ear can tolerate without pain (Ladefoged, 1996). Hence, bit-depth affects the SNR and the dynamic range of the signal and lower bit-depth will result in lowered SNR values.

The output spectrum of any speech falls off at a rate of -6 dB/octave. This approximation is a result of -12 dB/octave fall in the spectrum of the glottal pulse and a boost of +6 dB/octave due to radiation at the lips (Ladefoged, 1996). In order to compensate for this slope descent in the output spectrum, it is common to apply **pre-emphasis** on the input speech signal $s[n]$ as a first order filter (Rabiner and Schafer, 2007):

$$\hat{s}[n] = s[n] - \lambda s[n-1] \tag{3.3}$$

where λ is an empirically determined value in the range of $0.9 \leq \lambda \leq 1.0$.

Due to the non-stationary nature of speech, the signal is generally divided into a series of consecutive overlapping "frames" before applying the above pre-emphasis filter. The speech within each frame can have discontinuities at the start and end that can give rise to pseudo high frequency components in the signal. This is avoided by **windowing** each speech frame to smooth out the edges. A variety of window functions can be used ranging from rectangular, triangular to more smoother bell shaped family of curves. For example, Hamming window is one of the popular choices represented by the cosine sum (Oppenheim and Schafer, 1989):

$$w[n] = \begin{cases} 0.54 - 0.46\cos(2\pi n/M) & 0 \leq n \leq M \\ 0 & elsewise \end{cases} \tag{3.4}$$

where M is the number of discrete samples in the window. Due to the non-stationary nature of speech, it is a common practice to take short overlapping windows, during which we can assume the signal to be stationary. The size of the window taken is also of relative importance. Too short windows give poor frequency resolution and long windows give poor time resolution. A usual trade-off is to set the window size between 20-25 ms with an overlap window size of around 10 ms (Holmes and Holmes, 2001). In some special cases, like dysarthric speech, the window size can be greater than 25 ms with an overlap extended to 15 ms. This is done, because of slow articulation the dynamics of the speech are not varying rapidly and a reduced frame rate can yield better results (Selouani et al., 2012; Yakcoub, Selouani, and O'Shaughnessy, 2008).

Any speech signal is generally characterised by its spectral envelope that results from the vocal tract filter response, which gives rise to dominant frequency component in the signal and the harmonic structure due to the excitation source. The feature extraction process aims to encode the spectral shape by conducting short-time spectral analysis of the windowed frames using Discrete Fourier transform (DFT). For a windowed signal $w[n]$ the DFT is given by:

$$X(k) = \sum_{n=0}^{N-1} w[n] e^{\frac{-j2\pi kn}{N}} \qquad 0 \leq k \leq (N-1) \tag{3.5}$$

where k is the k^{th} frequency bin of N uniformly spaced frequencies and $e^{j\theta}$ is the Euler constant. DFT is a computationally expensive procedure with $O(n^2)$ complexity and a **Fast**

Fourier Transform (FFT) with $O(n \log n)$ complexity is usually applied that operates on window sizes that are in multiples of two. FFT is the basis for some common forms of feature analysis in ASR, like *Cepstrum Analysis* (Bogert, Healy, and Tukey, 1963) & *Linear Predictive Coding (LPC) Analysis* (Atal and Schroeder, 1970; Itakura and Saito, 1970).

The result of cepstral analysis is a **cepstrum**, which is the inverse Fourier transform of the log magnitude spectrum (Bogert, Healy, and Tukey, 1963). The logarithm of the first stage FFT ensures that the source and filter frequency components transform from multiplicative domain into an additive domain. The filterbank energies are generally correlated due to the overlapping filters. The energies are decorrelated using a simplified version of a frequency transform, usually a **Discrete Cosine Transform (DCT)**[1], to get the actual components of the source and filter. The DCT equation is given by:

$$c_j = \sqrt{2/N} \sum_{i=1}^{N} A_i \cos\left(\pi j (i - 0.5)/N\right) \qquad 0 \leq j \leq N \tag{3.6}$$

where c_j is the j^{th} cepstral coefficient and A_i is the log magnitude of the i^{th} channel (Holmes and Holmes, 2001). The low order cepstrum represents the response of the vocal tract filter and the high order cepstrum represents the excitation source. As cepstrum features are orthogonal, all the off-diagonal entries in a covariance matrix of features are zero. This greatly simplifies the computational load. Since most relevant information is found in the low order cepstrum features, a truncation procedure known as liftering [2] is applied to remove the unwanted high order features.

In contrast, LPC analysis finds the speech parameters by considering the excitation source and the vocal tract response as a combined system and then applying an inverse filter to the speech spectrum to obtain a near zero output. It works on the idea that any sample of speech at time n, say $x[n]$, can be considered as an approximated linear combination of past p samples (known as p^{th} order LPC analysis). An estimate of $x[n]$ is given as:

$$\hat{x}[n] = \sum_{k=1}^{p} a_k x[n-k] \tag{3.7}$$

where a_k is the predictor coefficient. Both cepstral and LPC analysis gives high resolution frequency information. However, they do not take into account the way frequencies

[1] DCT is a simplified version of the DFT, since the first stage Fourier transform is a symmetric function.
[2] An anagram for filtering in the frequency-frequency domain.

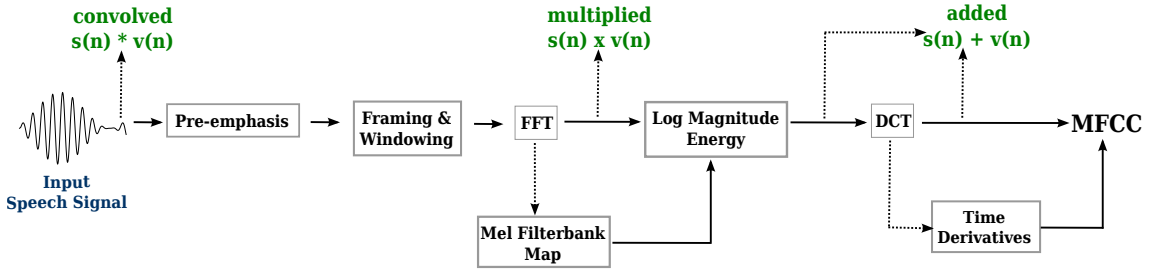

Figure 3.2: Schematic diagram for the MFCC generation process.

are perceived by human ear. The ear is more sensitive to changes in absolute pitch at lower frequencies. The perceived pitch of a tone increases approximately linearly with its frequency up to 1000 Hz and logarithmically thereafter up to 10000 Hz (Ladefoged, 1996). This relationship was determined by numerous psychophysical experiments that led to the development of non-linear scales, which showed the mapping between the actual frequency and the corresponding pitch percept of the human auditory system. Two commonly applied scales in ASR are the *Mel* and *Bark* scales represented as:

$$f = \begin{cases} 1127 \ln(1 + f/700) & , Mel \\ 13 \tan^{-1}(0.00076f) + 3.5tan^{-1}((f/7500)^2) & , Bark \end{cases} \tag{3.8}$$

ASR systems usually implement perceptually motivated outputs of the cepstral and LPC analysis. The most widely used representations of speech include the **Mel Frequency Cepstral Coefficients (MFCC)** (Davis and Mermelstein, 1980) and the **Perceptual Linear Prediction Coefficients (PLPC)** (Hermansky, 1990). MFCC feature representation is used in all the experiments in the thesis. A schematic diagram for MFCC generation is shown in figure 3.2.

The MFCC representation discussed so far generates the set of static features that describe the power spectral envelope. In order to boost the performance of ASR systems, MFCC trajectories are also computed over time. This involves finding additional speech information of a feature with respect to its neighbouring features. These dynamic feature sets are the first order regression coefficients (delta) that are appended to the original MFCC representation. They are computed using the regression formula:

$$delta_t = \frac{\sum_{w=1}^{\theta} w(f_{t+w} - f_{t-w})}{2\sum_{w=1}^{\theta} w^2} \tag{3.9}$$

where f_t is the frame at time t and θ is the total number of adjacent frames (usually 2). In a similar way second order regression coefficients (delta-delta) are also computed from the delta coefficients.

3.3 Acoustic modelling

3.3.1 Pre-Statistical approaches

Pattern classification for ASR has changed a lot over last six decades. The earliest methods included rule based acoustic-phonetic approaches which exploited the properties of speech signals directly to perform recognition. These were used for isolated unit recognition (Davis, Biddulph, and Balashek, 1952; Forgie and Forgie, 1959) and in conjunction with other knowledge based systems (Juneja, Deshmukh, and Espy-Wilson, 2002; Olivier et al., 1996). However, these approaches were not very successful due to the limited and partial knowledge the system had about the acoustic properties of various phonetic units, which lead to sub-optimal selection of features during recognition (Rabiner and Juang, 1993).

The acoustic-phonetic approaches were replaced by pattern recognition methods that implement the categorical learning of speech classes by direct observation of speech patterns without incorporating any explicit feature determination knowledge. Such an approach usually requires sufficient data to learn the acoustic properties associated with the observed patterns. This process of observation and learning is iterated over all the speech classes that are relevant for a given task. The pattern learning algorithms deployed are used to generate acoustic models of speech that are either deterministic or stochastic in their structure. The former accumulates an average snapshot of all the observed patterns for a particular speech class, called reference templates, and the latter represents the patterns using an appropriate probability density function (pdf). In both cases, the classification is made by selecting a particular template or probability density function that has the best fit with the unknown speech pattern.

The reference template approach, also known as template matching is a non-statistical method that measures the distance between an unknown speech unit and a finite set of

pre-stored speech templates for classification. Due to temporal differences between different speech utterances dynamic programming techniques are generally employed (Bellman, 1953). Dynamic Time Warping (DTW) is one such algorithm that models the timescale differences, as it warps the time-axis between the spoken speech pattern and the underlying template by non-linear modelling. DTW attempts to find the best cumulative distance at each time instant and iterates it over the entire speech utterance. Euclidean measure is the most commonly used distance metric, however other distance estimates based on smoothed LPC group delay spectrum (Itakura and Umezaki, 1987) and weighted cepstrals (Tohkura, 1986) have been effective in DTW applications. DTW has been used for isolated word recognition (Brown and Rabiner, 1982; Sajjan and Vijaya, 2012; Xu and Ke, 2012; Yang et al., 2011), speaker verification (Andrei, Paleologu, and Burileanu, 2011; Geppener, Simonchik, and Haidar, 2007; Wen, Liu, and Liu, 2003) and connected word recognition (Myers and Rabiner, 1981b; Nakagawa, 1984; Ney, 1984; Rabiner and Schmidt, 1980; Sakoe, 1979). Details of DTW and its optimisation strategies are discussed in a comparative study report by Godin and Lockwood (1989) and Myers and Rabiner (1981a).

Despite progress in template matching approaches, it lacked the ability to effectively model the variations and inconsistencies in the speech patterns. The limitations can be attributed to factors such as speaking speed, co-articulation, pitch variations etc. In order to overcome these impediments, template based approaches has been overtaken by more powerful statistical approaches that use probability distributions to model the variations.

3.3.2 Statistical approaches

The statistical learning paradigm can be divided into three broad categories, viz. *Generative, Conditional and Discriminative* modelling approaches.

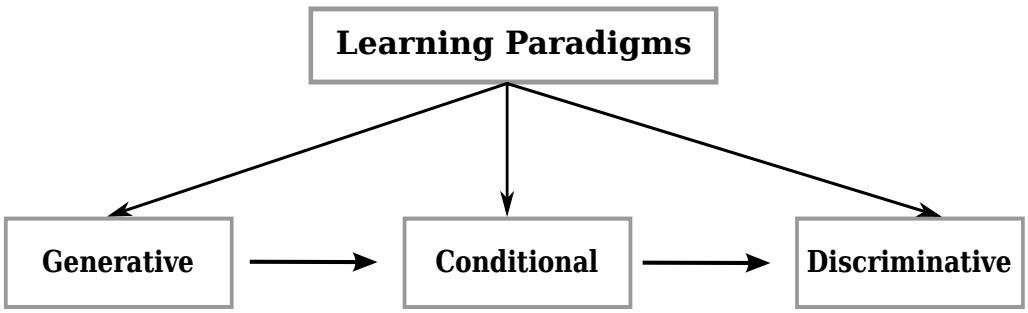

Figure 3.3: Statistical learning paradigms.

For example, the generative approach attempts to model all the input, output and other unobserved variables in the system by using joint probability distributions. Its goal is to optimise the model parameters to maximise the likelihood of the observed data and does not focus on the end classification task. In scenarios where it is possible to explicitly establish the end task, a conditional distribution can be directly optimised that will be used for final classificiation or regression instead of optimising the entire generative model. Discriminative learning follow a more minimalistic approach and do not focus on modelling any underlying distribution. Its prime goal is to have a robust input, output mapping and focuses on minimising the classification errors.

3.3.2.1 Generative learning

Generative learning aims to learn about the system or the phenomenon to be observed by modelling all input, output, observed and unobserved variables through a single joint probability distribution function. The joint distribution can then be used for classification by marginalisation and conditioning of the learnt distribution. The generative learning formalism is usually approached in light of the classical Bayesian Inference theory. It relies on the principle of Bayes' rule that states, if ϕ is a hypothesis and o is an observation, then the posterior estimate is given by:

$$\underbrace{P(\phi|o)}_{\text{Posterior}} = \frac{\overbrace{P(o|\phi)}^{\text{likelihood}}\overbrace{P(\phi)}^{\text{prior}}}{\underbrace{P(o)}_{\text{marginal likelihood}}} \tag{3.10}$$

$$P(\phi|o) \propto P(o|\phi)P(\phi)$$

where $P(\phi)$ is the prior probability of the hypothesis ϕ before any data is observed and $P(o|\phi)$ is the likelihood function which gives the power of estimating an observation o, given the hypothesis ϕ. Since, $P(o)$ is independent of ϕ, it can be treated as the marginal estimate and taken as a normalising factor in Bayes computation (Holmes and Holmes, 2001).

Now, the joint Bayesian estimate for a set of training examples $d \in \mathbb{D}$ and all permitted hypothesis/models $\phi \in \Phi$ is given as (Bishop, 1995):

$$P(d|\mathbb{D}) = \int_{\phi \in \Phi} P(d|\phi, \mathbb{D}) P(\phi|\mathbb{D}) d\phi \tag{3.11}$$

The above estimation depends on the family of pdf's $p(\phi|\lambda)$, where λ represents the vector of hyperparameters and the training examples in $\mathbb{D}$ are a 2-Tuple sequence of input vectors and output class labels represented as (x,y). The above integral is often not easy to compute and a scaled down approach is adopted where the mode of the posterior distribution is computed as an approximation. Instead of summing over all possible hypothesis ϕ, the most probable one is considered. It gives the maximum a posteriori (MAP) and maximum likelihood (ML) estimates as:

$$\begin{aligned} P(d|\mathbb{D}) &\approx P(d|\tilde{\phi}, \mathbb{D}) \\ \tilde{\phi} &= \arg\max_{\phi \in \Phi} P(\phi|\mathbb{D}) \\ &= \begin{cases} \arg\max_{\phi \in \Phi} P(\mathbb{D}|\phi) P(\phi) & MAP \\ \arg\max_{\phi \in \Phi} P(\mathbb{D}|\phi) & ML \end{cases} \end{aligned} \tag{3.12}$$

MAP takes advantage of incorporating prior belief about ϕ into the system and ML assumes uniform priors. Logarithms are generally taken in equation 3.12 as it makes the optimisation process easier due to monotonous nature of the log function and is well suited for exponential family of distributions.

$$\tilde{\phi} = \begin{cases} \arg\max_{\phi \in \Phi} \Big(\sum_{\mathbb{D}} \ln P(d|\phi) + \ln P(\phi) \Big) & MAP \\ \arg\max_{\phi \in \Phi} \Big(\sum_{\mathbb{D}} \ln P(d|\phi) \Big) & ML \end{cases} \tag{3.13}$$

In case sufficient data is available for modelling, both MAP and ML can efficiently find the optimal parameters of ϕ using an exponential class of distribution and show good convergence. However, the training data is almost always insufficient for producing the optimal model set. In addition, acoustic models that implement the Gaussian mixture hidden Markov model (HMM-GMM) framework has additional problems of unknown and hidden variables. These are the state sequences that correspond to a particular sequence of observations. The estimates under these constraints are found using an iterative procedure, known as expectation-maximisation (EM) algorithm (Dempster, Laird, and Rubin, 1977).

If $\mathbb{H}$ is a vector of such hidden variables and $L(\phi)$ be the likelihood function that needs to be optimised, then EM tries to simultaneously formulate the conditional expected value of the model likelihood with all integrated variables (*E-Step*) and then maximises this formulation (*M-Step*) as:

$$\Theta(\bullet) = L(\phi) - L(\phi_n) = \ln\Big(\sum_{h\in\mathbb{H}} P(\mathbb{D}|h,\phi)P(h|\phi)\Big) - \ln P(\mathbb{D}|\phi_n) \tag{3.14}$$

In the above equation, ϕ is the updated estimate that needs to optimised and ϕ_n is the current estimate at the n^{th} iteration. Equation 3.14 has a logarithmic sum and the EM algorithm takes advantage of the concavity of the log function and applies Jensen's inequality to create a lower bound for the log-likelihood function which ensures that every successive iteration will produce at least as good an estimate as the previous one. Both MAP and ML have been extensively used in the acoustic modelling of ASR systems. A theoretical and implementation framework of MAP Bayesian learning has been succinctly provided for HMM model estimation, speaker adaptation and language modelling (Gauvain and Lee, 1992, 1994).

The generative models using HMM-GMM has been the most popular framework implemented in ASR systems till date. Since the empirical work for the generation of acoustic models in the current thesis will use the HMM-GMM setup, the next section will briefly discuss its theory and implementation.

Hidden Markov Models

Any stochastic process that satisfies the Markov property is called a Markov random process and its model is known as a Markov model. The Markov property asserts that the state of the system at any time $t+1$ only depends on the state of the system at time t. This is also called the *"Property of Forgetfulness"*, since any predictions for the future state of the system is only dependent on the present state of the system and not the past. This is usually called the first-order Markov property. In theory we can have k-order assumptions that will take into account the past k states of the system to predict the current state. However, it makes the computations intractable and usually only first-order assumptions are taken into account.

The simplest Markov model is a Markov chain in which the state sequences are observable. In case of non-observable or hidden state sequences, it is known as a **Hidden**

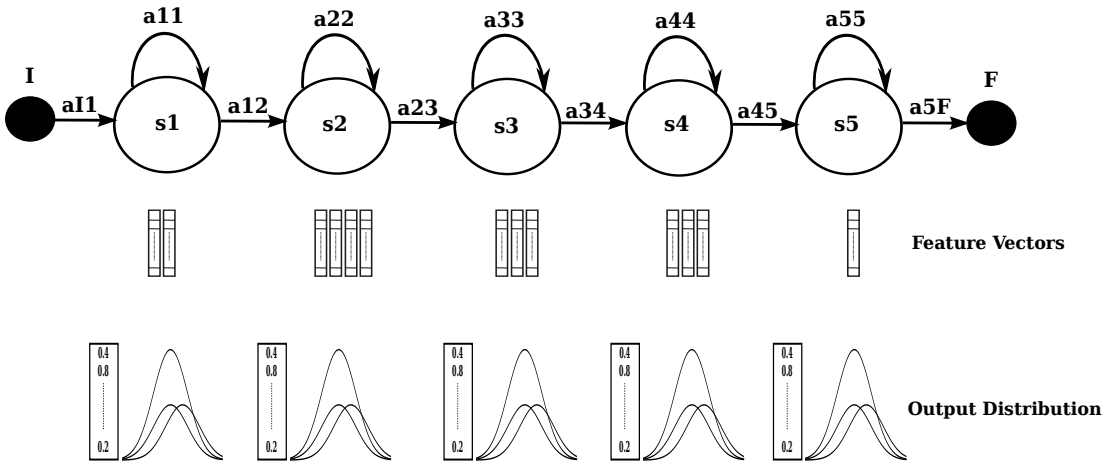

Figure 3.4: An example of a five state HMM. It is a strictly left-to-right topology due to the transition probabilities and the output distributions are either discrete or estimated using a mixture of Gaussian densities.

Markov Model (HMM) which is represented as a stochastic finite state machine with a fixed number of states. Although the stochastic process of state sequence generation is not directly observable in HMM, it can be observed through another set of stochastic process that produce the sequence of observations (Rabiner, 1989).

Figure 3.4 shows a five state HMM with a strict left-to-right topology. The HMM is fully characterised by the number of states, transition probabilities between the states and likelihood of a state observing a feature vector. The output probability distribution of states can either be discrete or continuous depending on the data being modelled. HMM topology usually also have special non-emitting initial (I) and final state (F) which acts as place holders for expanding the HMM network by concatenating several HMM's. This is usually the basis for continuous HMM based speech recognition systems. The formation of any HMM system asserts the following fundamental assumptions:

First Order Markov Assumption: If $s_1 s_2 ..... s_{t-1}$ is the history of system states in the past $(t-1)$ time units, then the transition probability of the system at time t is given by

$$P(s_t | s_1 s_2 ..... s_{t-1}) = P(s_t | s_{t-1}) \tag{3.15}$$

Stationarity: Let a_{ij} represent the transition probability of going form state i to state j. Then a_{ij} is independent of the time instant at which the actual transition takes place.

$$P(s_{t'} = j|s_{t'-1} = i) = P(s_{t''} = j|s_{t''-1} = i) \qquad t', t'' \in \overset{\text{random time}}{\text{instants}} \tag{3.16}$$

Observation Independence: This asserts that the probability of the observation at time t, say o_t, is only dependent on the state of the system at time t and is statistically independent of all past observations and states.

$$P(o_t|o_1o_2.....o_{t-1}, s_1s_2.....s_t) = P(o_t|s_t) \tag{3.17}$$

HMM is widely used in acoustic modelling of speech recognition systems (Jelinek, 1976) due to its inherent capability to capture temporal variations in a statistical framework. To formulate an HMM system for speech classification we consider $\Phi = W_1, W_2, ..., W_n$ as the set of all possible hypotheses or speech units and $Y = y_1, y_2, ..., y_T$ as an unknown observation of length T. According to Bayes' rule in equation 3.10, the posterior probability $P(W_i)$ is given by:

$$P(W_i|Y) = \frac{\overbrace{P(Y|W_i)}^{\text{acoustic model}} \quad \overbrace{P(W_i)}^{\text{language model}}}{P(Y)} \tag{3.18}$$

The posterior probability for each W_i is estimated and the hypothesis that is most likely to generate Y is the most probable output. If the HMM passes through a state sequence $s_1, s_2, ..., s_T$ in observing Y, then the joint probability estimate is given as (Holmes and Holmes, 2001):

$$\begin{aligned} P(y_1, y_2, ..., y_T) &= \sum_{\substack{\text{state} \\ \text{sequences} \\ \text{of length T}}} P(y_1, y_2, ..., y_T, s_1, s_2, ..., s_T) \\ &\approx \underbrace{\max_{\substack{\text{state} \\ \text{sequences} \\ \text{of length T}}} P(y_1, y_2, ..., y_T, s_1, s_2, ..., s_T)}_{ViterbiApproximation} \end{aligned} \tag{3.19}$$

The intractable summation is usually equated by Viterbi approximation for the best state sequence. If each state output distribution, also called **emission pdf's**, are optimally trained, then the best path should give a reasonable probability estimate that will

be very close to the summed score. For practical setups, the implementation of Viterbi approximations for a number of competing hypothesis is generally formulated using the token passing algorithm (Young, Russell, and Thornton, 1989). The success of an HMM-GMM based system largely depends on generating optimal estimates for the acoustic model sets. One of the principal approaches for maximum likelihood estimations is the Baum-Welch algorithm (Baum et al., 1970), which is the motivational basis for the generalised EM algorithm. The method corresponds to the Bayesian generative learning explained earlier and estimates the optimal parameter set by jointly observing data from multiple dimensions, both directly observable and hidden.

The Baum-Welch algorithm relies on the computation of a *forward-backward* procedure, which recursively computes the probability of partially observing an example and being in a particular state at a given time instant. The forward and backward scores are then combined to give the total likelihood of emitting the entire set of feature vectors and occupying a particular state at any time t. This is called the **state occupation probability** and is the most important component in the re-estimation procedure. In its minimal form it is written as $\gamma_j(t,e) = P(s_t = j|y_1, y_2, ..., y_T, \phi)$ for being in state j at time t and emitting the entire set of feature vectors of length T. This is estimated for a single example e with the current model parameters ϕ. The forward, backward and state-occupation scores are used in the re-estimation of transition probabilities of the HMM and the means and variances of the emission pdf. In speech recognition the emission pdf's are usually represented by a mixture of multivariate Gaussian distribution and the HMM is known as a Continuous Density Hidden Markov Model (CDHMM). For a state j to emit a feature vector y of dimension D at time t, the pdf is represented as:

$$b_j(y_t) = \sum_{m=1}^{M} \frac{w_{jm}}{(2\pi)^{D/2}|\sum_{jm}|^{1/2}} \exp\left[-\frac{1}{2} \sum_{d=1}^{D} \frac{(y_t[d] - \mu_{jm}[\mathrm{d}])^2}{\sigma_{jm}^2[\mathrm{d}]} \right] \tag{3.20}$$

where M is the total number of mixture components and w_{jm} is the weight associated with the m^{th} Gaussian mixture of state j such that $\sum_{m=1}^{M} w_{jm} = 1$ and $|\sum_{jm}|^{1/2}$ is the determinant of the full covariance matrix. Speech representations like MFCC's that output an uncorrelated set of feature vectors allows for the reduction of full-covariance matrix into a diagonal-covariance matrix ($|\sum_{jm}|^{1/2} \approx |\Pi_{d=1}^{D} \sigma_{jm}^2[d]|^{1/2}$). This greatly improves real time efficacy of ASR systems by reducing the computational complexity. The means and variances are re-estimated using the following relationship:

$$\widehat{\mu_{jm}} = \frac{\sum_{\forall e \in E} \sum_{\forall t \in T_e} \gamma_{jm}(t,e)\, y_{te}}{\sum_{\forall e \in E} \sum_{\forall t \in T_e} \gamma_{jm}(t,e)} \tag{3.21}$$

$$\widehat{\sum_{jm}} = \frac{\sum_{\forall e \in E} \sum_{\forall t \in T_e} \gamma_{jm}(t,e)\, (y_{te} - \widehat{\mu_{jm}})(y_{te} - \widehat{\mu_{jm}})^T}{\sum_{\forall e \in E} \sum_{\forall t \in T_e} \gamma_{jm}(t,e)} \tag{3.22}$$

where y_{te} is the observation vector and $\gamma_{jm}(t,e)$ is the state occupation probability of being in state j at time t and emitting the entire set of feature vectors using the mixture component m for the example word e. In the above Baum-Welch re-estimations, $\gamma_{jm}(t,e)$ imposes soft decision boundaries by allowing every state to have some weighted contribution. If $\gamma_{jm}(t,e)$ is bound to only have values of 0 (not occupied) or 1 (occupied), it leads to Viterbi training (Holmes and Holmes, 2001; Rabiner and Juang, 1993).

3.3.2.2 Conditional learning

Conditional Bayesian learning is an intermediate approach between the generative and discriminative modelling. Unlike the generative approach where a joint probability distribution models all the input, output, observed and hidden variables, conditional approach only models the output provided one knows what inputs to condition it on. In simple symbolic terms it means:

$$\underbrace{P(input, output, observed, hidden, ...)}_{\text{Generative Learning}} \Longrightarrow \overbrace{P(output|input)}^{\text{Conditional Learning}} \tag{3.23}$$

Recalling from section 3.3.2.1, where $\mathbb{D}$ represented both input and output variables of a system, one can divide the set separately into input($\mathbb{X}$) and output($\mathbb{Y}$). In conditional domain, instead of jointly learning $(\mathbb{X}, \mathbb{Y})$, equation 3.11 modifies to:

$$P(y|x, \mathbb{X}, \mathbb{Y}) = \int_{\phi \in \Phi^c} P(y|x, \phi^c, \mathbb{X}, \mathbb{Y}) P(\phi^c|\mathbb{X}, \mathbb{Y}) d\phi^c \tag{3.24}$$

This is an important simplification if we know the specific task to be learnt and now modelling of the distribution only focuses on learning the population of $\mathbb{Y}$ conditioned over $\mathbb{X}$. The parameters of the learnt system ϕ^c are effectively independent of the density of $\mathbb{X}$.

Despite the constraints imposed by the conditional learning, the integral in equation 3.24 is intractable and the mode can be computed to get the most probable model as:

$$\begin{aligned} P(y|x) &\approx P(y|x,\tilde{\phi}) \\ \tilde{\phi} &= \underset{\phi\in\Phi^c}{\arg\max} P(\phi|\mathbb{Y},\mathbb{X}) \\ &= \begin{cases} \underset{\phi\in\Phi^c}{\arg\max} P(\mathbb{Y}|\mathbb{X},\phi)P(\phi|\mathbb{X}) & \textit{C-MAP} \\ \underset{\phi\in\Phi^c}{\arg\max} P(\mathbb{Y}|\mathbb{X},\phi) & \textit{C-ML} \end{cases} \end{aligned} \tag{3.25}$$

Conditional Maximum a Posteriori (C-MAP) and Conditional Maximum Likelihood (C-ML) has been used in various branches of speech recognition. For example, variants of C-MAP are used in voice activity detection (Choi and Chang, 2012; Kim et al., 2010b; Shin et al., 2008) and C-ML has shown some performance gains over standard ML optimisation in HMM and Neural Network based hybrid systems (Krogh and Riis, 1999).

3.3.2.3 Discriminative learning

Generative and conditional learning approaches offer an appealing framework to model the system and the inter-dependencies between its variables. The former attempts to model the joint distribution $P(\mathbb{X},\mathbb{Y})$ and the latter models the conditional distribution $P(\mathbb{Y}|\mathbb{X})$. In both cases, the task of optimising a probability distribution exists as an intermediate step before any classification can be done. This can be difficult, especially when sufficient data is not available to model all the parameters of the system being observed or if the distributions representing the system fail to classify an unseen data resulting in a poor classifier.

In discriminative learning the focus shifts from modelling towards classification only. Unlike generative and conditional approaches, which are oblivious to the end classification goal, discriminative approach only focuses on adjusting the classification boundaries to get an optimal input to output mapping. Although, discriminative approach might lack the elegance of probabilistic modelling of variables and their inter-dependencies, it compensates by only focusing on producing an optimal classifier. Since the task of discriminative learning is not on modelling the underlying system, the Bayesian integral optimisation is no longer the target as seen in sections 3.3.2.1 and 3.3.2.2.

Some of the commonly used discriminative approaches include *Maximum Mutual Information (MMI)* (Bahl et al., 1986; Chow, 1990; Nadas, Nahamoo, and Picheny, 1988) and *Minimum Classification Error (MCE)* (Juang and Katagiri, 1992). The former is based on principles of information theory that attempts to probabilistically maximise the mutual information between the training data and its corresponding transcripts. The later optimises the performance of the system by minimising the classification errors through a smoothed, continuous and differentiable function. The MCE discriminant function defined for speech recognition tasks is represented using joint log likelihoods (He, Deng, and Chou, 2008; Juang, Hou, and Lee, 1997; Reichl and Ruske, 1995) that are suited for optimsation tasks such as gradient descent. Another discriminative criteria known as Minimum Phone Error/ Minimum Word Error (MPE/MWE) was presented by Povey and Woodland (2002). Unlike MCE, which considers the entire transcript of the utterances for optimisation objective, MPE/MWE only takes weighted contributions of correctly classified phones/words between the competing and true hypothesis strings. This flexibility allows the inclusion of not only the true hypothesis, but also partially true and completely incorrect hypothesis strings in optimising the discriminative criteria. Another important discriminative criteria is the Large Margin Estimation (LME), which aims to maximise minimum margins of training data for a generalised approach to classifier design. The margins are basically the distance between each data in the training set and the decision boundary. The margins are used to bound generalisation errors which are important to various machine learning tasks (Mitchell, 1997). LME has been successfully applied for estimation of CDHMM by maximising the minimum multi-class margins and had been shown to produce better results than the MMI and MCE approaches for isolated word task (Li, Jiang, and Liu, 2005; Sha and Saul, 2007). The same concept has been extended to continuous speech recognition which showed gains over standard MCE approaches (Liu, Jiang, and Rigazio, 2005). Another novel approach for continuous speech recognition was presented by Kaiser, Horvat, and Kacic (2002), which minimised the overall risk of misclassification on the training database using the Levenshtein distance metric between the correct and n-best competing hypothesis.

3.3.3 Advanced learning architectures

The learning paradigms discussed so far can be broadly classified as shallow architectures that contain a single layer of non-linear feature transformations. However, a family of more

complex algorithms that fall under the domain of deep learning architecture are gradually gaining equal prominence. It inherently deploys multiple layers of non-linear interactions that aim to increase the modelling and representational power of the underlying information. There is more than a decade of extensive research conducted in the domain of deep learning from both machine learning and mainstream ASR viewpoint. The reader can refer to the overview articles by Deng (2014) and Deng and Li (2013) for an exhaustive summary of the developments. The current section will give a general overview of the deep learning framework.

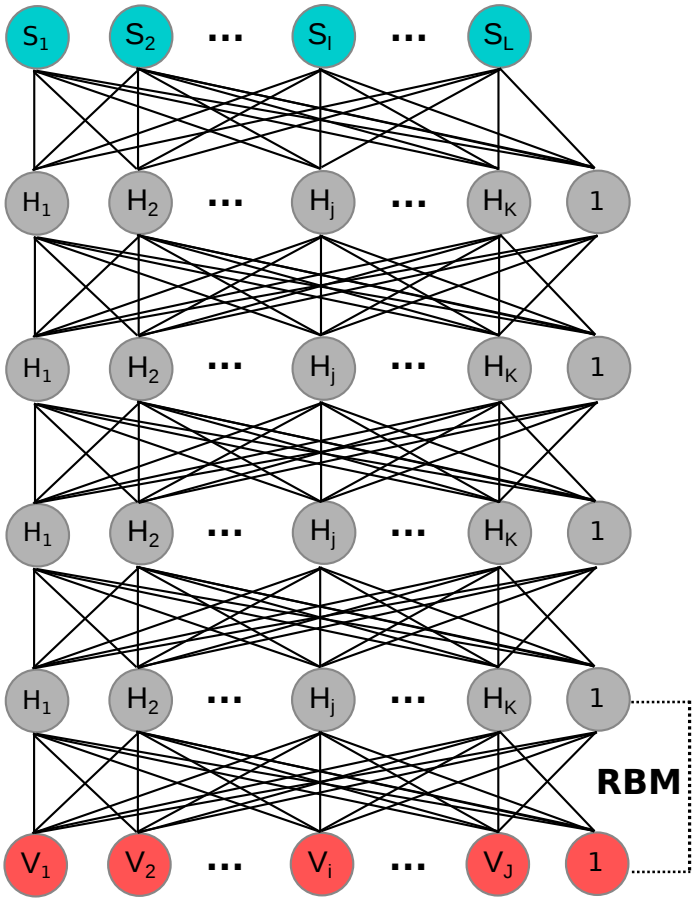

Figure 3.5: Illustration of a DBN architecture.

Deep learning is based on the principles of artificial neural networks and it owes its recent developments and success to the seminal work by Hinton, Osindero, and Teh (2006) and Hinton and Salakhutdinov (2006) in which an optimised class of networks were introduced,

known as the deep belief networks (DBN) that forms the foundation for deep learning.

Figure 3.5 shows a schematic illustration of a DBN network. It is constructed by stacking a series of restricted Boltzmann machines (RBM), where each RBM is a specialised Markov random field that has a single layer of stochastic hidden units and a single layer of stochastic visible units. In an RBM construction all the visible and hidden units are connected to each other with no visible-visible and hidden-hidden connection. The activation units of each RBM layer are used as the input data for training the visible units of the above layer RBM. This iterative and efficient layer-by-layer greedy learning process runs bottom up across the entire network (Hinton, Osindero, and Teh, 2006).

Just like learning paradigms discussed earlier, deep learning also falls into generative, discriminative and hybrid architectures. Out of these, the hybrid deep architecture is of prime interest in ASR research. In this, the DBN is subject to the process of generative learning for "pre-training" the parameters of the network, which is followed by the discriminative learning process to "fine-tune" the weights for optimising the entire network performance. In context of ASR, the discriminative fine-tuning is generally performed by adding a final layer of expected speech unit labels denoted by $S_1, S_2, ..., S_l, ...S_L$ in figure 3.5 and using the backpropagation algorithm to adjust the weights in the network. When such a generative-discriminative DBN setup is modelled, it is also called the deep neural network (DNN). Since DBN/DNN represent a static architecture with fixed dimensional input/output, it is natural to extend the DBN/DNN framework with dynamic models that can better capture the temporal and co-articulatory properties of variable length input/output of speech utterances. The extended architecture of DNN-CRF (Mohamed, Yu, and Deng, 2010) and DNN-HMM (Dahl et al., 2011; Mohamed, Dahl, and Hinton, 2009) have been successfully used in large vocabulary ASR tasks.

3.4 Language and pronunciation modelling

Language Models

The language model (LM) is responsible for assigning probabilistic estimates for the occurence of word sequence. It imposes syntactic and semantic constraint on the overall recognition task. The LM is represented as a prior probability (equations 3.10) in the computation of the posterior estimates. The probability estimate for a hypothesis consisting of n independently occurring words as $\mathcal{H} = \{w_1, w_2, w_3, ..., w_n\}$ is given as:

$$P(\mathcal{H}) = P(w_1, w_2, w_3, ..., w_n) = \Pi_{m=1}^{n} P(w_m | w_{m-1}, ..., w_1) \tag{3.26}$$

Due to intractable numbers of word sequence combinations, the above equation is usually approximated by only considering the past $(n-1)$ tokens. This is the **n-gram language model** that is approximated as $P(\mathcal{H}) \approx \Pi_{m=1}^{M} P(w_m | w_{m-1}, ..., w_{m-n+1})$. In trivial cases, a maximum likelihood estimate is used for the n-gram estimate, which counts the frequency of a particular sequence as:

$$P(w_m | w_{m-1}, ..., w_{m-n+1}) = \frac{Count(w_{m-1}, ..., w_{m-n+1}, w_m)}{\sum_w Count(w_{m-1}, ..., w_{m-n+1}, w)} \tag{3.27}$$

Due to data sparseness, the above equation can fail to assign any meaningful estimates for missing n-gram sequences. Such issues are handled by applying discounting and backoff techniques where the former shifts the probability mass from the non-zero count n-grams to the zero/low count n-grams, and the later assigns a zero-count n-gram with a scaled factor of its corresponding lower order n-gram counts (Jurafsky and Martin, 2000).

Some of the common discounting and backoff approaches include the Witten-Bell discounting (Witten and Bell, 1991), Good-Turning discounting (Good, 1953), Katz backoff smoothing (Katz, 1987) and smoothing via linear interpolation for different n-gram orders (Jelinek and Mercer, 1980). Class based LM models are also commonly applied (Brown et al., 1992; Shuanghu et al., 1998; Yamamoto, Isogai, and Sagisaka, 2001) for overcoming the problem of data sparseness by introducing syntactic and semantic matching. For example, $P(Albert|Hi), P(Peter|Hi) \approx P(< proper\_noun > |Hi)$ since proper name occurrences can be rare even for a large corpus. LM's are usually evaluated using its perplexity score [3] over a test corpus where a lower score indicates a better model (Jurafsky and Martin, 2000).

Pronunciation Models

The Pronunciation Model (PM) assists in the search for predicting the best hypothesis by capturing alternate phonetic realisations for the words in the vocabulary. It can be an important module in the design of dysarthric speech systems, where the phonetic realisation of a word can vary considerably from that of a typical pronunciation. An example of this variation is depicted in figure 3.6, which shows how the pronunciation of a dysarthric speaker with high degree of severity deviates from a typical pronunciation pattern by addition,

[3] For a sequence of words of length N, Perplexity$(w_1, w_2, ..., w_N) = 2^{Entropy(w_1, w_2, ..., w_N)}$

substitution or deletion of phones. Limited research has been conducted in the domain of dysarthric PM (Seong, Park, and Kim, 2012a) and there is no model to predict the most reliable phonetic target for dysarthric speakers of varying etiologies and severities. Although this thesis will not use PM in building dysarthric speech systems, it will give informative cues that further directs towards the potential for future research.

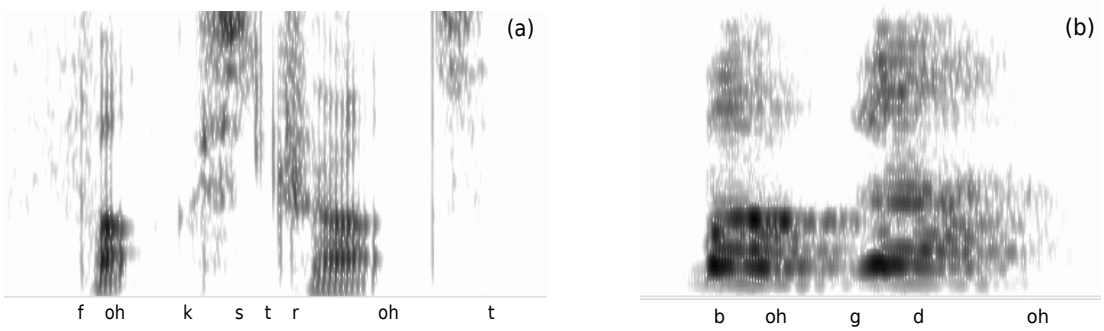

Figure 3.6: The word *foxtrot* spoken by (a) typical and (b) dysarthric speaker with high degree of severity. The dysarthric pronunciation [b oh g d oh] shows a greater degree of variation from the typical pronunciation [f oh k s t r oh t].

The pronunciation models are largely categorised as *Knowledge-base* or *Data-driven*. The knowledge-base models (Bartkova and Jouvet, 2006; Tjalve and Huckvale, 2005) use phonological rules to generate pronunciation variants and the lexicon can be handcrafted for specific ASR task. Data-driven models (Kessens, Cucchiarini, and Strik, 2003; Kim, Oh, and Kim, 2007) attempts to derive pronunciation variations directly from the acoustic signals. It uses phone level recognitions and smoothing techniques such as decision trees to get reliable estimates (Humphries, Woodland, and Pearce, 1996; Humphries and Woodland, 1998; Strik and Cucchiarini, 1999; Wester, 2003). The kind of model to use is generally dependent on factors such as time, linguistic expertise, quality of training data and vocabulary. Although knowledge-based models are refined through years of research and expert knowledge, it might miss some unexpected variations that are more likely to be captured by data-driven approaches. Data-driven models however come with additional problems of encoding noise within the pronunciations (Fosler-Lussier, 2003). The posterior probability estimates for a speech system can be easily expanded to include the pronunciation models. For a set of model parameters ϕ, observation vector O and a pronunciation model P_w, equation 3.12 can be modified as:

$$\begin{aligned}\arg\max_{\phi} P(\phi|O,P_w) &= \arg\max_{\phi} \sum_{P_w} P(O|\phi,P_w)P(\phi,P_w) \\ &= \arg\max_{\phi} \sum_{P_w} P(O|\phi,P_w) \underbrace{P(P_w|\phi)}_{\substack{\text{Pronunciation}\\ \text{Model}}} P(\phi)\end{aligned} \tag{3.28}$$

Researchers have explored various techniques to generate optimal pronunciation variants and reduce lexical confusions. These include the use of Bayesian framework (Sakti, Markov, and Nakamura, 2008; Sakti et al., 2010), discriminative techniques like MCE (Adde and Svendsen, 2011; Adde et al., 2010), maximum likelihood estimation (Holter and Svendsen, 1999), HMM's (Magimai-Doss and Bourlard, 2005; Saralar, Nock, and Khudanpur, 2000; Seong-Jin, Yung-Hwan, and Gyung, 1997) and the use of Finite State Transducers (Fosler-Lussier, Amdal, and Kuo, 2005; Hazen et al., 2005). In addition, pronunciation models have also been useful for speaker verification tasks (Leung et al., 2005), phone-to-articulatory mapping (Bowman and Livescu, 2010) and searching pronunciation alternatives at syllable level (Ng and Hirose, 2012).

3.5 Adaptation and adaptive training

The training of ASR systems aim to generate acoustic models that can be speaker dependent (SD), which is modelled to recognise only a particular speaker or speaker independent (SI), which is a generic model to recognise a range of speakers including the ones who have not engaged during the training process. SD systems on average produce a word error rate (WER), which is two to three times lower than an SI system trained using same amount data (Woodland, 2001). It is desirable for an ASR system to reduce the gap between SI and SD models, however, due to highly variable and dynamic nature of speech, the modelling of SI systems is a very challenging task. The speech variations can be either a direct result of speakers age, gender, health, intonation, speaking rate, accent, etc. or have an indirect impact through environmental conditions and effects of microphone and transmission channel. Moreover, desirable SD systems are also difficult to build, due to lack of sufficient training data available for individual speakers. This problem is further exacerbated for dysarthric speech, where sufficient data is usually unavailable due to various physical and social constraints.

The aim of speaker adapted (SA) methods is to compensate for these shortcomings by reducing the mismatch between a generic baseline SI acoustic model and a target test speaker. Adaptation process can be carried out in the signal domain to reduce the speaker or environmental induced variations and is referred to as *feature-based adaptation*. The other technique called *model-based adaptation* transforms the acoustic models in the SI system into a set of adapted models which has more SD-like characteristics using the adaptation data. The process of speaker specific adaptation can be either *supervised* or *unsupervised*, based on whether the adaptation transcripts are available or not. Further, the adaptation data used to prepare SA systems can either be used all at once (*static*) or iteratively (*dynamic*) over a period of time. The remainder of this section will present a brief overview of some of the commonly used state-of-the-art adaptation techniques.

3.5.1 Feature based adaptation

Feature based adaptation, also known as feature normalisation, aims to reduce speaker or environmental induced variations by directly adjusting the speech features in the acoustic domain. The most common techniques used in ASR systems are given below.

Vocal Tract Length Normalisation (VTLN)

Robust generation of SI acoustic models require steps to compensate for the anatomical differences between the vocal tract lengths of various speakers. For example, the females have shorter vocal tract producing higher formant frequencies as compared to the longer vocal tracts of males producing lower formant frequencies (Lin and Che, 1995). One of the methods to overcome this was suggested by Lee and Rose (1996), in which linear frequency warping was applied by modifying the filter-banks in the MFCC analysis as:

$$f_w = \alpha f \tag{3.29}$$

where f_w and f are the warped and original frequencies and α is the scaling factor. The study conclusively suggested formant frequency compression for shorter vocal tracts and expansion for longer vocal tracts. The speaker specific warping factor α for the frequency-axis normalisation is determined empirically by applying a grid search over a range of warp factors which maximised the likelihood of the warped observations and it reflects the

approximate 25% range in the vocal tract length differences of humans. A value of $\alpha < 1.00$ indicates frequency compression and $\alpha \geq 1.00$ indicates frequency expansion.

Other VTLN methods include selective non-warping schemes which directly attempts to manipulate the upper frequency limits in the FFT spectrum before feature extraction (Lin and Che, 1995), bilinear transforms (McDonough et al., 1998; Wang, Bing-xi, and Qi, 2004), applying linear transformation in the cepstral space (Pitz et al., 2001), or using a linear warping matrix (Claes et al., 1998) to determine the optimal warping factor.

Cepstral Mean and Variance Normalisation

For the MFCC generation process, the convolution of the signal in time-domain is additive in the log-magnitude domain (figure 3.2). Hence, the effect of any convolved noise in the signal due to channel distortion or microphone characteristics can be reduced by simply subtracting it from the cepstral features. Such distortions are generally uniform over the entire length of the speech utterance. *Cepstral Mean Normalisation* (CMN) or *Cepstral Mean Subtraction* (Atal, 1974) is one of the technique used to remove such channel noise from the signal without removing any useful speech information. It simply subtracts the mean of the observation feature vectors from each frame in the observation so that the transformed cepstral feature vectors have zero mean. It is applied as:

$$\hat{o}_t[d] = o_t[d] - \frac{1}{T}\sum_{t=1}^{T} o_t[d] \tag{3.30}$$

where d is the dimension of the feature vector and T is the total number of frames in the observation. Another form of CMN, called the *Tied Mixture Normalization* (TMN) was implemented by Anastasakos et al. (1994) to reduce the microphone mismatch by mapping vector quantised (VQ) codebook entries between the adaptation data collected over the training and test microphones respectively and using the VQ maps to compute the altered means and covariances of the Gaussian distributions.

Similarly, *Cepstral Variance Normalisation* (CVN) is applied in conjunction with CMN to reduce the variance induce by the channel state. CVN transformation ensures that each dimension of the observation vector has a variance of unity. It is estimated as:

$$\hat{o}_t[d] = \frac{o_t[d]}{\sqrt{\sigma_d^2}}$$
$$\sigma_d^2 = \frac{1}{T}\sum_{t=1}^{T} o_t^2[d] \tag{3.31}$$

Both CMN and CVN are computationally inexpensive operations which are effective in reducing the environmentally induced variations.

3.5.2 Maximum a Posteriori (MAP) adaptation

Maximum a posteriori (MAP) estimation was discussed in section 3.3.2.1 and it was defined at the mode of the posterior distribution (equation 3.12). Thus, for a given adaptation data $\mathcal{A}$ and its related transcription $\mathcal{A}^{\mathcal{T}}$, MAP maximises the posterior probability of the model parameters ϕ, as:

$$\phi_{MAP} = \arg\max_{\phi} P(\phi|\mathcal{A}, \mathcal{A}^{\mathcal{T}}) = \arg\max_{\phi} P(\mathcal{A}|\phi, \mathcal{A}^{\mathcal{T}})P(\phi|\lambda) \tag{3.32}$$

where λ is the vector of hyperparameters for the prior distribution $P(\phi|\lambda)$. The inclusion of the prior distribution in the training process aids towards robust model estimation in the presence of limited data, which is often the case in speaker adaptation. It should be noted that model estimation can be formulated in a convenient way if the prior and posterior distributions are same, i.e., the prior forms a conjugate prior to the corresponding likelihood function. In the MAP setup the likelihood function is usually represented by an HMM with mixtures of Gaussians which has no well defined conjugate prior. An alternate approach suggested that the choice of prior densities for the given HMM parameters was adequately represented as the product of the Direchlet and normal-Wishart densities (Gauvain and Lee, 1994). It was shown that SA models using MAP outperforms the SD models for small amounts of available data (Gauvain and Lee, 1994). As the data increases it was shown that the MAP adapted models converged to the maximum likelihood estimated SD models.

One of the disadvantages of MAP adaptation is that it only updates parameters that are observed in the adaptation data, which makes the MAP estimates very slow and cumbersome, especially if large number of Gaussians are left unadapted. One of the technique,

known as regression-based model prediction (Ahadi and Woodland, 1997) uses linear regression of the adapted parameters to update the poorly adapted or unobserved parameters. Another approach known as Structural MAP was implemented by Shinoda and Chin-Hui (1997) to increase the speed of standard MAP adaptation. It uses bias and scaling vectors to update the means and variances of the Gaussian pdf's, while arranging the Gaussians in a hierarchical tree structure.

3.5.3 Linear transformation

This is an alternative approach that applies a linear transform to adapt the means and variances of the Gaussian mixture components that represent the output distributions of say an HMM system. The linear transformation allows Gaussian parameters to be adjusted for the target speaker with minimal data. It also has the advantage of sharing the same linear transform across multiple Gaussians or even all (global transform) the Gaussian. Some commonly used linear adaptation approaches are discussed below.

3.5.3.1 Maximum Likelihood Linear Regression (MLLR)

MLLR (Leggetter and Woodland, 1995b) is the most successful linear transformation approach. The adaptation of the m^{th} Gaussian component of dimension d is given by:

$$\begin{aligned} \hat{\mu}_m &= A_m \mu_m + b \\ &= \begin{bmatrix} A_m & b \end{bmatrix} \begin{bmatrix} \mu_m \\ 1 \end{bmatrix} \\ &= \hat{\boldsymbol{W}}_{\boldsymbol{m}} \xi_m \end{aligned} \tag{3.33}$$

where $\hat{\mu}_m$ is the adapted mean, $\hat{\boldsymbol{W}}_{\boldsymbol{m}}$ is the $d \times (d+1)$ regression matrix which performs the linear affine transform and ξ_m represents the extended mean vector. The affine transform ensures that the point-space ratio and collinearity of the mean vector is preserved in the affine space after the transformation. The transformation matrices $\hat{W}_m$ are calculated to maximise the likelihood of the adaptation data using the expectation-maximisation framework. If $\mathcal{A}$ is the adaptation data of length T, then it maximises the objective function $P(\mathcal{A}|\phi)$ by defining an auxiliary function of the form (Gales and Woodland, 1996; Leggetter and Woodland, 1995b):

$$\mathcal{Q}(\phi,\hat{\phi}) = \widehat{K} - \frac{1}{2}P(\mathcal{A}|\phi)\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)\Big[K_m + log(|C_m|) + (a_t - \hat{W_m}\xi_m){C_m}^{-1}(a_t - \hat{W_m}\xi_m)'\Big] \tag{3.34}$$

where $\widehat{K}$ is the constant term that is independent of state sequences and time, K_m is the normalisation constant associated with mixture m, $\gamma_m(t)$ is the occupation probability for the component m at time t and $\mathcal{A} = (a_1, a_2, ...a_T)$ is the adaptation data. Equation 3.34 is expanded and differentiated to maximise the auxiliary function and get the general form for $\hat{W_m}$ as:

$$\sum_{t=1}^{T}\gamma_m(t){C_m}^{-1}a_t\xi_m' = \sum_{t=1}^{T}\gamma_m(t){C_m}^{-1}\hat{W}_m\xi_m\xi_m' \tag{3.35}$$

The transformation $\hat{W_m}$ can be either applied to a single Gaussian component or it can globally transform all the Gaussian components in the system. In practice though, a single transformation is usually shared with multiple Gaussian components. This allows for the generation of regression classes, which contains a cluster of associated or similar Gaussian components updated by a particular regression matrix. This not only allows precise adaptation of specific clusters of Gaussians, but also overcomes the problem of adapting with less data by tying regression matrices across clustered mixture components.

If the covariance matrix in equation 3.35 is restricted to the diagonal, then the k^{th} row of the transform $\hat{w}_k$ is given as (Leggetter and Woodland, 1995b):

$$\hat{w}_k = (G_k)^{-1}z_k \tag{3.36}$$

$$G_k = \sum_{t=1}^{T}\sum_{\substack{m\in \\ \text{shared} \\ \text{compo} \\ \text{-nents}}}\frac{\gamma_m(t)}{\sigma_{mk}^2}\xi_m\xi_m' \tag{3.37}$$

$$z_k = \sum_{t=1}^{T}\sum_{\substack{m\in \\ \text{shared} \\ \text{compo} \\ \text{-nents}}}\frac{\gamma_m(t)a_{tk}}{\sigma_{mk}^2}\xi_m \tag{3.38}$$

In the above equations a_{tk} is the k^{th} element of the observation vector a_t and σ_{mk}^2 is the k^{th} diagonal element of C_m.

The construction of regression classes mentioned earlier is achieved by putting together mixture components that belong to similar phonetic classes (e.g. stops, fricatives etc.) (Leggetter and Woodland, 1995b) or by using more robust hierarchical sorting technique, such as *Regression Class Trees* (Leggetter and Woodland, 1995a,b). In regression trees, the mixture components are arranged in a top-down tree structure, with root node containing all the mixture components in the ASR system and the individual leaf nodes correspond to specific Gaussian mixtures. The mixtures at the leaf nodes are then clustered together based on their likelihood similarity to generate intermediate higher level regression classes, also known as base classes. Depending on the quality and quantity of adaptation data available, any number of intermediate levels can be constructed to generate higher level regression classes. The rationale is to apply similar transformation to the components clustered in these higher level regression classes. If any regression class in the tree is inadequate for adaptation purpose, the tree allows a flexible structure of bubbling up to higher levels, all the way to the root node, for optimal transform application.

Although, the most important speaker specific information is believed to be characterised by the Gaussian means (Leggetter and Woodland, 1995b; Woodland, 2001), the linear Gaussian variance transform can also be applied. The transforms can be applied to the matrix in full, diagonal or block-diagonal modes, depending on the adaptation data (Gales, Pye, and Woodland, 1996; Gales and Woodland, 1996). The transform for adapting the covariance matrix of the k^{th} Gaussian component is given by

$$\hat{\Sigma}_k = L_k^T \hat{H}_k L_k \tag{3.39}$$

where $\hat{H}_k$ is the linear transform to be estimated and L_k is the inverse of the Cholesky factor of ${\Sigma_k}^{-1}$, such that ${\Sigma_k}^{-1} = C_k C_k^T$ and $L_k = C_k^{-1}$. Cholesky decomposition involves high computational complexity of the order $O(n^3)$ (Hammerlin and Hoffmann, 1991). An alternative method is proposed by Gales (1998b) that modifies the mean and observation vector for the transformation in the following form.

$$\hat{\Sigma}_k = H_k \Sigma_k H_k^T \tag{3.40}$$

For the diagonal covariance case, this simple mean and observation vector modification reduces to matrix-vector multiplication and an addition operation.

3.5.3.2 Constrained MLLR (CMLLR)

In the previous section, the linear transforms were applied independently on the means and variances. However, in a constrained case the same transform is applied to both means and variances. It is depicted as

$$\hat{\mu}_p = A^c \mu_p + b^c \tag{3.41}$$

$$\hat{\Sigma}_p = A^c \Sigma_p A^{cT} \tag{3.42}$$

where A^c is the constrained linear transform and b^c is the bias for the mean vector. The constrained MLLR was first introduced for the diagonal covariances (Digalakis, Rtischev, and Neumeyer, 1995) and was later extended to full covariance (Gales, 1998b). The constrained MLLR leads to a simplification which is primarily equivalent to transforming the observation vector as

$$\hat{o}_t = A^{c-1} o_t + A^{c-1} o_t b^c \tag{3.43}$$

Both variants of MLLR described above give reliable estimates for the adaptation, however if sufficient statistic is not available the MLLR transformations can be even poorer than the SI models (Woodland, 2001). In order to overcome this problem extended combined approaches has been timely suggested to increase the practical viability of MLLR. A MAP-like weighting scheme was suggested (Goronzy and Kompe, 1999) for rapid, unsupervised MLLR speaker adaptation which transforms the means as weighted linear combinations of the MLLR predicted means. The weighted scheme was applied both in static and dynamic modes. Another approach known commonly as MAP-Linear Regression (MAPLR) (Chesta, Siohan, and Lee, 1999) uses the prior distribution of the mean transformation matrix, which acts like a constraint to avoid getting poor estimates in the underlying structure of the acoustic space.

3.5.4 Adaptive training

SI modelling is intrinsically difficult due to inter and intra speaker variabilities which are accentuated by the presence of varied acoustic environments in which the data was collected.

The speaker variations is largely attributed to the differences in the vocal tract, accents, dialects etc. In case of dysarthric speech the problem is further aggravated due to the underlying neurological impairment causing atrophy of the musculoskeletal structure.

As detailed in section 3.5.1 some feature normalisation techniques like VTLN, CMN, CVN are used to reduce the speaker and channel induced variations to some extent, but it often fails to completely remove the underlying effects. Moreover, it also lacks the modelling framework to deal with a wide variety of speaker variations. *Adaptive Training* (Anastasakos et al., 1996) provides such a framework in which it attempts to model the speaker and non-speech variations separately.

3.5.4.1 Speaker Adaptive Training (SAT)

One of the most commonly used adaptive training method is the *Speaker Adaptive Training* (SAT) (Anastasakos, McDonough, and Makhoul, 1997; Anastasakos et al., 1996). It provides a common framework to explicitly model the speaker induced variations and parameter estimation of the canonical HMM models in a single unified training regime. The speaker specific characteristics are modelled using a linear transformation of the Gaussian means (Leggetter and Woodland, 1995b) and the canonical HMM models are updated from these speaker specific transforms. The integrated training phase is iteratively estimated using the EM algorithm (Baum et al., 1970; Dempster, Laird, and Rubin, 1977), which jointly estimates the parameters of the canonical HMM and the speaker specific transforms. SAT provides a natural framework for implicitly annihilating the inter-speaker variations in a corpus. This might be beneficial for testing the efficacy of dysarthric speech systems, where such variabilities are more prominent.

An overview of the SAT framework is shown in figure 3.7. The speaker induced variations are modelled by using linear regression transformations that generates a set of speaker specific transforms as $\mathcal{W} = \{W^{(1)}, W^{(2)}, ..., W^{(\mathcal{S})}\}$ for the $\mathcal{S}$ speakers in the system. The regression matrix for the transformation can either be limited to a single transform for a speaker or it can have multiple transforms through regression classes. Finally, given the current set of transforms $\mathcal{W}$, an auxiliary function can be defined for estimating the canonical model parameters as:

$$\mathcal{Q}(\phi, \phi^{SAT}) = \sum_{\Theta} P(\mathcal{O}, \Theta|\mathcal{W}, \phi)\ logP(\mathcal{O}, \Theta|\mathcal{W}, \phi^{SAT}) \tag{3.44}$$

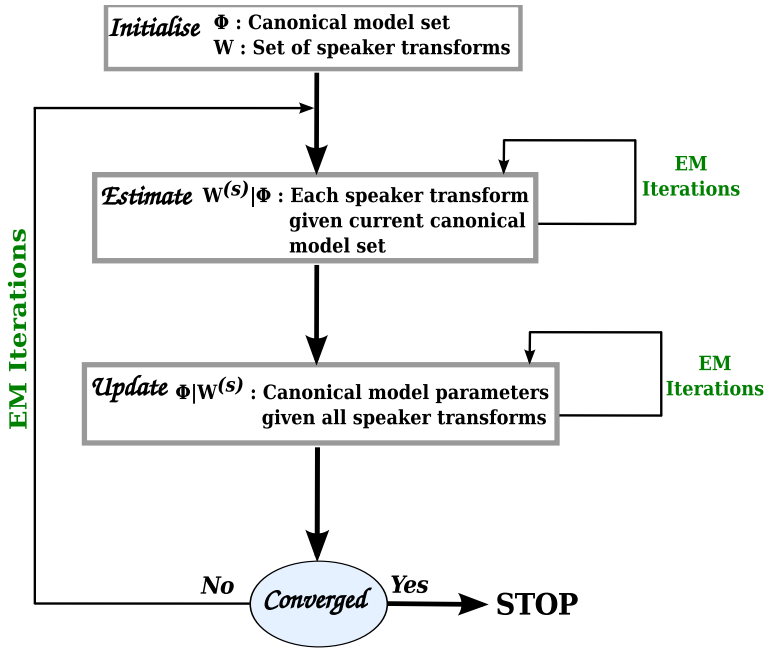

Figure 3.7: An overview of the SAT framework.

where $\mathcal{O} = (O^{(1)}, O^{(2)}, ..., O^{(S)})$, is the set of transcribed speech data from each speaker in the system and Θ is the set of all possible state sequences. The auxiliary function is expanded to maximise the likelihood of the data from all the speakers in the system in an iterative fashion. The details of the estimation process are given in Anastasakos et al. (1996) and Leggetter and Woodland (1995b).

SAT can be applied using the MLLR and CMLLR transforms. SAT based on MLLR transforms generates robust canonical model estimates, but it comes with computational and memory overheads (Spyros et al., 1997), making it less practical for implementation. This can be avoided by applying constrained MLLR (CMLLR) (Digalakis, Rtischev, and Neumeyer, 1995; Gales, 1998b) that applies a common transform for both means and variances. SAT with CMLLR results in a kind of feature normalisation during model training and has the same computational overhead as any other standard HMM update.

3.5.4.2 Cluster adaptive training and eigenvoices

Other important adaptive approaches are the *Cluster Adaptive Training (CAT)* (Gales, 1998a, 2000) and *Eigenvoices Method* (Kuhn et al., 1998). Both the approaches aim to

estimate speaker specific parameters as a weighted combination of different speaker clusters that form the canonical model sets. The Gaussian variances, weights and the transition matrices are assumed similar across all the clusters and only the Gaussian mean components are adapted. Unlike SAT, where only a single set of canonical model is used, CAT and eigenvoices use multiple canonical models derived from each cluster. It should be noted that the eigenvoices method finds the canonical models or clusters (eigenvoices) by applying *Principal Component Analysis (PCA)* (Jolliffe, 2002) on a set of supervectors that is constructed from all mean values in a set of speaker-dependent HMM systems (Kuhn et al., 1998, 2000). These principal components are representative of important variations between the training speakers.

To symbolically define the parameter update process, let us consider $\mathcal{C}$ distinct clusters. Then in context of the CAT approach we can define:

$$\Upsilon_m = \left[\mu_m^{(1)}, \mu_m^{(2)}, ..., \mu_m^{(\mathcal{C})}\right] \qquad \lambda_{(s)} = \left[\lambda_{(s)}^{(1)}, \lambda_{(s)}^{(2)}, ..., \lambda_{(s)}^{(\mathcal{C})}\right]^T \tag{3.45}$$

where $\mu_m^{(c)}$ is the mean associated with the component m of the cluster c and $\lambda_{(s)}^{(c)}$ is the interpolation weight associated with speaker s for the cluster c. The adapted mean for speaker s is then given by

$$\mu_m^{(s)} = \Upsilon_m \lambda_{(s)} \tag{3.46}$$

The CAT approach is generally described as "model-based", where there is a separate mean for each component of every distinct cluster or "transform-based" that represents clusters as cluster-specific MLLR transforms of a common set of canonical model means. Also, extensions of eigenvoice based methods have been suggested, which include MLLR transformations (EMLLR) to form supervector for eigenvoice based adaptation (Chen et al., 2000), Segmental Eigenvoice Method (SEV) (Tsao, Lee, and Lee, 2005), and a hybrid approach between eigenvoices and SEV called Hierarchical Eigenvoice Method (HEV) (Onishi and Iso, 2003) that works through Gaussian component clustering to control the complexity of the adaptation process.

3.5.5 Discriminative adaptation

The first category of discriminative adaptation techniques is know as the *Discriminative Linear Transform (DLT)*. It is similar to the MLLR, but instead of optimising a likelihood

function, it aims to optimise one of the discriminative criteria like MCE (He and Chou, 2003), MMI (Gunawardana and Byrne, 2001) or MPE (Wang and Woodland, 2004). DLT has proven to be effective under sparse data conditions and has been formulated for both supervised and unsupervised adaptations. In addition, hybrid DLT approaches have also been presented (Uebel and Woodland, 2001) that uses a linear interpolation of the maximum likelihood criteria and the MMI objective function for linear transform estimation.

It should be noted that all the above discriminative adaptation methods showed superior performance over the standard MLLR techniques under supervised adaptation modes. Unsupervised adaptation is usually modelled more effectively using the ML estimation procedure. It is because discriminative optimisation criteria are more sensitive to errors in the hypothesis, as they are based on phone or word errors (Yu, Gales, and Woodland, 2008). In order to overcome this issue a *Discriminative Mapping Transform (DMT)* (Yu, Gales, and Woodland, 2008) method was proposed. It obtains a discriminative speaker-independent mapping transform from a speaker-specific ML transform, which ensures that the sensitivity to the underlying hypothesis will be reduced. The basic idea behind DMT is to define a speaker-independent criteria mapping function to obtain an indirect estimate of the final speaker-specific discriminative transform. The details of estimation procedure are given in Yu, Gales, and Woodland (2009).

The above discriminative adaptation approaches have been further extended to accommodate the adaptive training schemes. For example, discriminative-SAT estimates ML based speaker transforms and updates the canonical models in a discriminative fashion using criteria like MPE (Wang and Woodland, 2003) and MMI (Mcdonough, Schaaf, and Waibel, 2002; Tsakalidis, Doumpiotis, and Byrne, 2003). A similar approach has also been extended to the case of canonical models derived from multiple clusters known as discriminative-CAT (Yu and Gales, 2006).

3.6 Automatic recognition of dysarthric speech

The use of ASR to control assistive technology has been discussed for more than 30 years (Cohen and Graupe, 1980; Noyes, Haigh, and Starr, 1989). Speech is a potentially attractive

input medium for people with physical impairments who find keyboard, mice and touch-screens difficult to use. However for a significant minority of people with physical impairments, especially with motor speech disorders resulting in dysarthria, speech recognition technology has been able to enhance human-human & human-computer interaction and provides an effective medium for controlling environmental systems (Noyes and Frankish, 1992). For example, the integration of ASR with assistive technology can enable speakers with dysarthria to control standard devices at home through speech commands (Hawley et al., 2007) or engage in communication via voice-input-voice-output communicators (Hawley et al., 2012). Such use of speech driven assistive devices could enable users to participate in social situations where they can interact with non-familiar communication partners.

Speech as an interface can also be quicker for people with dysarthria than other inputs. For example, it was found in a study that a 100 word delivery task takes around 44 minutes to be conveyed using a combination of ASR and switch scanning, in comparison to 100 minutes taken by switch-scanning alone (Hawley, 2002). Another study showed that when ASR was integrated in an environmental control system, it resulted in a mean task completion time of 7.7 seconds for ASR only versus 16.9 seconds for switch-scanning alone (Hawley et al., 2007). Although the studies showed the fact that ASR accuracy was generally lower than switch-scanning, but the final message transfer was much faster even with mis-recognitions followed by corrections.

Instead of giving an exhaustive summary, the remainder of this section will review some key findings on the current state of ASR for speakers with dysarthria under some broad application domains. A good historical overview of speech recognition of mild, moderate and severely dysarthric speech is detailed in Patel (2000).

3.6.1 Commercial speech recognition

Commercial speech recognition systems are intended for typical speakers and may successfully be applied to large vocabulary with continuous speech. Commercial applications sometimes require an enrolment phase, where a user adapts with some pre-defined text that enables the system to become more representative of that particular speaker. The usage of such commercial systems has been applied within the research domain for dysarthric speech with some moderate levels of success.

One of the earliest reports of such systems was reported by Roberts (1985), where the IBM and Dragon Naturally Speaking v5 commercial packages were evaluated by 11

speakers with dysarthria. Only 6 speakers with mild and mild-moderate dysarthria were able to use the system with an accuracy ranging from 30%-80%. The remaining 5 speakers with moderate to severe dysarthria failed to complete the enrolment phase. Another study conducted on 10 speakers with cerebral palsy using the Shadow VET/2 system (Coleman and Meyers, 1991) also reported lower recognition performance for speakers with dysarthria.

The use of commercial packages was also employed to understand the effect of a speech training program on the user performance. A study by Kotler and Stonell (1997) with a single user who had cerebral palsy and mild dysarthria using the IBM VoiceType version 1.0 system had a 57% reduction in errors after an effective speech training programme was implemented. The speech training consisted of practice sessions for the speaker to improve upon initial consonant production (*e.g. fair vs air*) and the final consonant production (*e.g. steak vs state*). In a similar study on 3 speakers with CP and 3 speakers with TBI using the same IBM system, Stonell et al. (1998) reported an absolute mean increase of recognition score from 27% to 80% after five training sessions. The training sessions were also found to be effective for mild/moderate dysarthric users using other commercial systems like the DragonDictate (Ferrier et al., 1992). The studies showed that the acquisition learning curves for both typical and dysarthric speakers had similar slopes despite low recognition accuracy for the speakers with dysarthria. A common consensus from these studies is that the ASR system performance was directly associated with the underlying severity of dysarthria.

As commercial speech systems improved, researchers have continued to investigate the utility of them for people with dysarthria. A comparative study of three ASR systems (Microsoft Dictation, Dragon Naturally Speaking (3.0) and Voicepad Platinum) was reported for a speaker with mild dysarthria. The Dragon system was around 13% more accurate when averaged across five sessions of read and spontaneous speech (Hux et al., 2000). Despite this, the accuracy reported using the Dragon system for the speaker with dysarthria was around 26% less than the control speaker.

Another study conducted on four speakers with dysarthria of varying severity compared a speaker-adaptive and a speaker-dependent system (Raghavendra, Rosengren, and Hunnicutt, 2001). The former was a phoneme based Prototype Swedish Dragon Dictate (PSDD) system that was adapted from the English variant of Dragon Dictate (Bamberg, 1990) and the later was the Infovox RA unit (Elenius and Blomberg, 1986). The PSDD was better at adapting and recognising the mild/moderate dysarthric speech. Despite the baseline accuracy for the most severely affected speaker being less than 30% in both the systems,

the PSDD was able to adapt better than the speaker-dependent Infovox system after only three sessions that increased the accuracy to 75% for the most severe user.

The commercial systems seem to perform reasonably for speakers with mild to moderate dysarthria, but the performance of the these systems tend to degrade substantially with increasing severity, possibly due to high degree of variabilities. Also, the results reported using the commercial systems are mostly conducted in control conditions, and its usage in more practical setups is unknown. Another difficulty in using such systems might be the inadequacy of the user to successfully complete the enrolment sessions due to physical constraints. Although these commercial systems are constantly evolving with the most up-to-date ASR technology, it still lacks the sophistication to cater to the variabilities manifest in dysarthric speech, especially with increasing severity.

3.6.2 Modelling approaches

Dysarthric speech recognition has also been pursued as a research problem for more than three decades. Instead of using commercial speech software, researchers have attempted to improve the modelling techniques to overcome the problems associated with variabilities in dysarthric speech. One of the earliest known attempts used the traditional template-matching approach for the isolated word task on two speakers with spinal cord injury, who acquired dysarthria as a secondary symptom (Fried-Oken, 1985). They got an average accuracy of 80% for the two speakers in controlling a computer-based educational software package for rehabilitation.

Template based approaches were soon replaced by more robust statistical methods. In an isolated word recognition study by Deller, Hsu, and Ferrier (1991), discrete HMM's were used to explore the effects of phonemic transitions in dysarthric speech. A transition clipping scheme was implemented in an ergodic HMM topology to remove the inconsistent transitions in the linear prediction feature space. When the setup was tested for 3 speakers with CP with intelligibility ranging from 22%-65%, an absolute accuracy of 88% was reported. This was found to be significantly better than other HMM topologies (ergodic (72%), bakis (50%)) with no transition clipping. Discrete HMM's have also been utilised to evaluate the effect of multiple training sessions on the recognition performance (Chen and Kostov, 1997).

Discrete HMM's were replaced by continuous density HMM approaches in nearly all

of the more recent studies of dysarthric ASR. Rudzicz (2007) used a CDHMM modelling framework for adaptation of speakers with mild and moderate dysarthria. They showed a relative error reduction of 23.1%-4.9% on the Nemours database of dysarthric speech (Menendez-Pidal et al., 1996). The finding were similar to those reported by Raghavendra, Rosengren, and Hunnicutt (2001) that also showed the efficacy of adaptive approaches for speakers with mild to moderate dysarthria. They also showed that speakers with severe dysarthria were better recognised using SD modelling. SD systems can however prove as a viable option for mild dysarthric users in high perplexity tasks with increased lexicon size (Sanders et al., 2005). Another finding reported in several studies showed some benefits in using complex HMM topologies like ergodic over standard left-to-right models (Deller, Hsu, and Ferrier, 1991; Polur and Miller, 2005a,b) for handling dysarthric variabilities.

Discriminative modelling has also been used as an alternative to the standard HMM technique. One of the earliest study was reported by Jayaram and Abdelhamied (1995) that used Artificial Neural networks (ANN) for isolated word recognition. They tested the neural network on a single participant with CP. The network was trained using FFT and formant features as two separate input vectors and the results were compared against the perceptual scores of five expert listeners and a commercial Intervoice speech system. FFT feature vectors were reported to give the best recognition score of 76.3%, and was significantly better than formant features (42.5%), listener test (42.4%) and the commercial system (37.5%). Other discriminative approaches for isolated word recognition include the use of Support Vector Machines (SVM), which performed better (Wan and Carmichael, 2005) or comparable (Hasegawa-Johnson et al., 2006) to the standard DTW and HMM based recognition.

In a more comprehensive study, a comparative report of the ANN and SVM techniques was reported (Rudzicz, 2009). It tested four separate discriminative classifiers using both acoustic and articulatory features (manner and place of articulation, tongue height, voicing etc.) on the Nemours database. The discriminative classifiers were based on feed-forward and recurrent Elman neural networks, and the two SVM variants were based on the radial basis and the polynomial-DTW sequence kernels (Wan and Carmichael, 2005). In general, SVM methods were reported to outperform the neural network based methods in classifying the articulatory features from MFCC acoustic vectors on an average between 4.9% to 9.3% with a relative error reduction of 19.8% across all the speakers. On the phone-classification task, SVM reduced the frame-level error relatively by 6.9% to 8.8% when only MFCC's

were used, and between 1.5% to 10.9% when MFCC's were augmented with articulatory features.

In addition to the individual generative and discriminative methods, hybrid approaches have been used for dysarthric ASR. An isolated word recognition task that uses the HMM-ANN system, implemented a multi-layered feed-forward neural network linked to an ergodic HMM framework. The hybrid approach gave a relative 5% improvement over the standard HMM system for three speakers with CP (Polur and Miller, 2006). Another hybrid system proposed by Selouani et al. (2012) suggested the usage of a new activation function based on class posterior distributions for the hierarchical multi-layer perceptron network. This connectionist approach was used to classify severity of the dysarthria prior to any recognition. The paper also suggested the usage of rhythm metric in addition to the standard MFCC's[4]. For the assessment of severity, the hybrid system was more than 3% better than the baseline GMM system, and the inclusion of rhythm features further boosted the performance from 3% to 6% absolute. The speech recognition task was conducted on three speakers of varying severity from the Neumors database using either (i) speaker-specific SD-system or (ii) severity-specific SD-system prepared from clusters of speakers under same severity levels. The former system gave an average accuracy of 64.5% and the later system gave an accuracy of 60% after the prior classification of severity was performed using the proposed connectionist approach. Although the proposed hybrid system gave lower recognition accuracy, it had the advantage of sharing data across speakers to build the models. More recently, specialised forms of HMM, based on Kullback-Leibler divergence (KL-HMM) (Rasipuram and Magimai.-Doss, 2013) have been successfully implemented within a DNN-HMM framework for dysarthric ASR. In one such study the KL-HMM framework was evaluated for 30 native Korean speakers with mild and moderate dysarthria (Kim et al., 2017). In order to minimise the effect of imprecise articulation in dysarthric speech, KL-HMM provides a natural framework where the HMM emission probabilities were modelled as categorical distributions using phoneme posteriors from a DNN system. The speaker-specific phonetic variations were further adapted using a combination of L2 and lexical confusion-reducing regularisation methods. The proposed methods in the study concluded with significant gains over the conventional GMM-HMM and DNN-HMM setups.

[4]Rhythm metric components include features such as vocalic and consonantal durations, voiced and unvoiced regions etc. For details of all the rhythm measures used in the study please refer to the paper (Selouani et al., 2012).

Researchers have also attempted to improve the dysarthric ASR performance by employing methods targeted to maximise the usage of available data. One such novel approach was presented in a study by Ons, Gemmeke, and Van hamme (2014) that learnt the recurrent acoustic patterns in spoken command and control type of sentences by using binary semantic frame descriptors. The method was based on the decomposition of the non-negative matrix factorisation (NMF) technique and it showed high accuracy gains even after training was performed using a single utterance. When the study was evaluated against a word based speaker dependent HMM-GMM system (Gemmeke et al., 2014), it showed absolute gains in the range of 5-40% for the isolated word recognition task when relatively less data was used. The ASR performance for both the systems was however comparable as the amount of training data increased and the HMM-GMM systems were better for handling complex grammar network even with lesser training data than the NMF approach. Recently DNN based modelling approaches have also gained prominence over the classical HMM-GMM framework in mainstream ASR for typical speech. These approaches are extended to the dysarthric domain also and it is believed that the optimisation of the complex network structure might be able to handle the dysarthric variabilities and sparseness problems (Yilmaz et al., 2016, 2017) more robustly.

More recently, DNN framework has been applied to some of the common open databases of dysarthric speech. In one such study (Espaa-Bonet and Fonollosa, 2016) the hybrid DNN-HMM models outperformed the HMM-GMM systems on the TORGO database (Rudzicz, Namasivayam, and Wolff, 2012) of dysarthric speech. The authors employed a wide variety of feature and model parametrisations for the experiments. For example, the generation of MFCC's was refined by applying Maximum Likelihood Linear Transform (MLLT) for a unique transformation of each speaker, and speaker-specific feature space normalisation was derived by using feature-space MLLR (fMLLR) (Gales, 1998b). The standard monophone/triphone GMM-HMM systems were further tested using subspace GMM (Povey et al., 2011) that makes it possible to optimally train under sparse data conditions. The DNN networks were prepared using cross-entropy measures and discriminative minimum Bayes risk was applied at state-level. In addition time delayed and Long-term Short-Term memory networks were also tested. In summary, the DNN-HMM systems improved the word error rate (WER) by 3% for control and 13% for dysarthric speakers relative to the best classical GMM-HMM architecture. The study was further extended by other researchers to improve the baseline accuracy on the TORGO database (Joy and Umesh, 2018). They suggested

further refinement and optimisation to improve the relative WER by 17.6% compared to the earlier study by Espaa-Bonet and Fonollosa (2016). In addition to refining basic parameters like frame rate, number of context-dependent states, the authors also suggested the usage of dropout and sequential discriminative strategies and generalised distillation framework.

The performance of DNN have also been reported on the UASPEECH database. In a study by Bhat, Vachhani, and Kopparapu (2016b) DNN-HMM performed slightly better (2% absolute) than the GMM-HMM systems on task specific recognition of a smaller vocabulary of digits and computer commands in the UASPEECH corpus. In another recent study on hybrid paradigms, representational learning framework has been explored that proposes an HMM-SVM modelling schema for classification (Chandrakala and Rajeswari, 2017). It builds generative sub-models for each phonetic class, which are termed as example specific HMMs (ESHMM). The log-probability of an utterance being generated by each of the ESHMM gives a score vector that is fed as an input to the SVM discriminative classifier. The authors reported that the proposed method outperformed the conventional HMM and DNN-HMM systems by approximately 66% and 21% respectively for small vocabulary (digits and computer commands from UASPEECH) ASR tasks with a $3/4 - 1/4$ split for the training and test data. Although, it is worth highlighting that some of the earlier studies (Christensen et al., 2012; Sehgal and Cunningham, 2015) that used a $2/3 - 1/3$ split for the training and test data had reported better results using conventional HMM systems on a bigger UASPEECH test vocabulary of 255 words instead of 19.

3.6.3 Acoustic features and enhancement

Acoustic features and their properties have been exploited for increasing recognition performance and perceptual intelligibility of speakers with dysarthria. In a study by Polur and Miller (2005b), three different representations of speech signals, viz. FFT, LP and MFCC coefficients were compared for three speakers with CP that used an ergodic HMM topology for an isolated word recognition task. MFCC coefficients gave the best recognition output at 92%, which was around 3%-12% higher for the models trained with LP and FFT coefficients. The effect of window size for generating speech frames has also been investigated for speakers with dysarthria, where it was found that a window greater than 25 ms worked on an average 8%-10% better than the smaller windows (Selouani et al., 2012; Yakcoub, Selouani, and O'Shaughnessy, 2008). In another acoustic study, the effective limit in the high frequency spectral regions were investigated. It was reported that no significant

information was available for dysarthric speech above 5.5 kHz (Polur and Miller, 2005a).

Some studies have focussed on the modification to the dynamics of the incoming speech signal with a view of making the disordered speech signal more intelligible and more amenable for ASR. For example, a study by Rudzicz (2013) applied signal transformation routines to dysarthric speech and were able to increase the perceptual intelligibility by around 20% and ASR scores by 15% absolute on the TORGO database (Rudzicz, Namasivayam, and Wolff, 2012). The signal transformation included: (i) devoicing of improperly voiced regions by using a high-pass filter, (ii) correcting pronunciations by the insertion of deleted sounds and deletion of repeated sounds using Levenshtein metric between the actual and expected phone sequence, (iii) tempo contractions for spuriously longer phoneme sequences identified as sonorants and (iv) adjusting the formant trajectories by using an anchor based frequency morphing of the spectrum. It was also found in the study that the intelligibility ratings were largely influenced by the apparently inappropriate insertion and deletion of phonemes and intelligibility improved after the corrections were applied. This highlighted the importance of lexically correct phoneme sequences from human comprehension viewpoint. Another study that aimed to increase the intelligibility and recognition of dysarthric speech was conducted by Tolba and El Torgoman (2009). The first two formants of speech from 11 Arabic speakers with dysarthria were transformed to approximate values more closely associated with typical speakers. The average intelligibility assessed by 12 naive listeners was increased from 7% on the original speech to 84% on the formant modified. The re-synthesis further increased the quality of dysarthric speech and improved the recognition scores from 28.5% to 71.4% for a group of seven words extracted from short Arabic sentences.

In more recent studies researchers have attempted to adjust particular dysarthric idiosyncrasies for signal and feature enhancement. For example, severity based tempo adaption of the sonorants in dysarthric speech (Bhat, Vachhani, and Kopparapu, 2016a), and generating multi-tapered MFCC's that are appended with important pathological voice parameters (jitter, shimmer, F0 etc.) (Bhat, Vachhani, and Kopparapu, 2016b) have been used for the evaluation of dysarthric ASR systems. Both the studies reported improvements to the GMM-HMM and DNN-HMM based speech systems on small vocabulary tasks of digits and commands from the UASPEECH database. In addition, the enhancement of the MFCC's using belief networks like deep autoencoders have proved effective to model speech with mild to severe dysarthria for very small vocabulary tasks (Vachhani et al., 2017).

3.6.4 Adaptation of dysarthric speech

The fundamental approach of any adaptation process is very appealing when modelling a process with sparse data. This is especially the case for speakers with dysarthria who are unable to record large amounts of data due to physical constraints, fatigue and muscular atrophy. Although, little work has been done in the field of dysarthric speech adaptation, some novel attempts have paved the path for further research and investigation.

Some of the earlier research seems to have reached a common consensus that adaptation techniques were more suited to model mild/moderate dysarthric speech and speakers with severe dysarthria were better represented using SD models (Raghavendra, Rosengren, and Hunnicutt, 2001; Rudzicz, 2007). However, more recent studies reported contrary results, which suggests severity as not a good indicator for an optimal selection of modelling approaches. This was established in a study by Sharma and Hasegawa-Johnson (2010), where they tested a SA system based on MAP, SD left-right HMM system and an ergodic HMM model that was prepared using linear interpolation of transition probabilities between fully ergodic and left-right topologies. In this study the SI HMM systems for MAP adaptation were prepared from typical speech in the TIMIT database (Garofolo et al., 1993) and seven speakers with spastic dysarthria were taken from the UASPEECH database (Kim et al., 2008). The results showed that the SA systems with or without transition matrix interpolation were better than SD systems for 5 out of 7 tested speakers. The other two speakers who benefitted from SD models had an average absolute gain of around 1.6% over the best SA systems. A relative accuracy gain of around 25.1% and 71.2% was reported for the two least and two most-severe speakers. MAP adaptation has been investigated further with all 15 speakers with dysarthria from UASPEECH and it was found that for most of the speakers MAP based adaptation outperformed SD systems (Christensen et al., 2012). Their reported results were on average 34.5% relatively better than the earlier published result on a similar kind of study (Sharma and Hasegawa-Johnson, 2010).

More recently, MAP adaptation has been explored in a completely novel framework, developed especially for dysarthric speech, known as Background Interpolation MAP (BI-MAP) (Sharma and Hasegawa-Johnson, 2013). The idea behind BI-MAP is to obtain an intermediate prior acoustic model which will narrow the gap between two disparate SI systems, viz. one trained on dysarthric speech and the other trained on typical speech. This intermediate model then provides as an optimal starting point for a base model on which adaptation techniques like MAP can be applied. This interpolation is described as:

$$\underbrace{\Phi_{BIM}}_{\substack{\text{Background}\\ \text{Interpolated Model}}} = \Delta \overbrace{\Phi_{NSI}}^{\substack{\text{Unimpaired}\\ \text{SI Model}}} + (I - \Delta) \underbrace{\Phi_{DSI}}_{\substack{\text{Dysarthric}\\ \text{SI Model}}} \tag{3.47}$$

where $\Delta = diag(\delta_i)_i$ is a $P \times P$ diagonal matrix such that $0 \leq \delta_i \leq 1 \ \forall i$, P is the dimensionality of the acoustic model parameter space and I is a P-dimensional identity matrix. The Φ_{NSI} models were prepared from the TIMIT corpus and the Φ_{DSI} models were prepared from UASPEECH. Once the Φ_{BIM} is determined, a MAP adaptation is applied to it as a second stage process. It was found that the interpolation technique improved the recognition accuracy of the BI-MAP system by an absolute of 8% (40% relative) over the standard SI (typical speech data) adapted MAP system for all the dysarthric speakers in the UASPEECH database.

The application of adaptation techniques was further explored by Mengistu and Rudzicz (2011), who in addition to constructing a robust baseline system also presented a comparison between the MLLR and MAP adaptation methods. The baseline system used both dysarthric and typical speech for the SI models, which gave an absolute gain of around 13% and 18.5% in comparison to typical-only and dysarthric-only SI models. It was one of the first studies conducted on a relatively large vocabulary task of around 1500 words of the TORGO database. The MLLR transformed both means and variances and gave an absolute 16.24% WER reduction. A pass of MAP adaptation on the MLLR transformed models further reduced the absolute WER by 3.9%. In addition to acoustic adaptation, the study also investigated the effects of a speaker-dependent lexical adaptation by adding alternate word pronunciations to the dictionary. The speaker specific pronunciation lexicon boosted the recognition accuracy in both SD and SA systems. It was also reported that when acoustic and lexical adaptation was used together, they managed to get an overall 22.87% absolute (42.11% relative) WER reduction.

3.6.5 Other approaches

The acoustics of dysarthric speech are highly variable and often lacks the presence of robust cues for a particular phonetic token. In lieu of this, attempts have been made to harness alternative source of knowledge in the speech production process. One such additional source of information is present in the estimates of articulatory parameters, which might compensate for any partial or missing acoustic data or provide additional information that

can be utilised for building robust dysarthric speech systems. The feasibility of articulatory knowledge for dysarthric speech was demonstrated by Rudzicz (2010), where the author managed to reduce the ambiguities in the acoustics of dysarthric speech by reducing the average relative differential entropy by 18.3% for three speakers with cerebral palsy. The paper suggested that dysarthric speech should not be considered as a deviation from typical speech, but rather a noisy channel distortion of an abstract representation of articulatory goals. An elaborate practical demonstration for the usage of articulatory knowledge for improving dysarthric speech recognition was given in a study by Rudzicz (2011), where five different modelling techniques, viz. HMM, DBN, Latent-Dynamic Conditional Random Fields (LDCRF), ANN and SVM were explored. The methods were used to classify articulatory features (AF) from the given acoustic signals and for phone level classification using *"acoustic/articulatory only"* and *"acoustic/articulatory augmented"* feature vectors. LDCRF gave the best average AF classification output which had an approximate mean of around 64% across seven separate dimensions of articulatory features[5]. When the LDCRF predicted articulatory features were augmented to the acoustic MFCC feature vector, it reduced the relative error between 0.5% and 7.1%, when compared to acoustic-only predictions. LDCRF and SVM techniques were the most successful for phone classification accuracy and the study showed advantages of using articulatory features in addition to acoustics only.

This work was extended by integrating it within a framework of a dynamical system, known as *task-dynamics*, which represents a combined model of skilled articulator motion and the planning of abstract vocal tract configurations. Task-dynamics tends to encapsulate the dynamic patterns of speech as overlapping gestures, which are treated as high level abstractions of reconfigurations of the vocal tract. Each gesture is represented as tract variables, such as lip aperture, tongue tip constriction etc. that are modelled using non-homogeneous second-order differential equations (Rudzicz, 2012). In this work, the acoustic-to-articulatory inversion was based on the theory of task-dynamics and used the MOCHA (Wrench, 1999) and TORGO (Rudzicz, Namasivayam, and Wolff, 2012) corpora. It used the kernel canonical correlation analysis (KCCA) and mixture density network (MDN) to estimate the task-dynamic features of the vocal tract from the given acoustics. Once the acoustic-articulatory inversion was established, it was integrated within an ASR

[5]The seven articulatory features used in the study included manner and place of articulation, high/low & front/back positions of the tongue, voicing, lip-rounding and static/dynamic nature of the articulators (Rudzicz, 2011)

framework, which estimated the best output hypothesis as a weighted combination of the acoustic likelihoods and the corresponding articulatory realisations in task-dynamics. Both KCCA and MDN gave an average WER of 34.1% for the dysarthric data.

The articulatory data in Rudzicz (2011, 2012) was collected using a procedure known as electromagnetic articulography (EMA). Other signal forms such as surface electromyographic (sEMG) signals have also been used in dysarthric speech recognition. A study by Deng et al. (2009) used MFCC and sEMG signals to develop a multi-modal framework for isolated word recognition. The experiments were conducted on five dysarthric speakers with an intelligibility range of 60% to 92%. Their study showed that when acoustic and sEMG signals were used together in a speaker-dependent system, it gave the best average recognition score of around 96% in comparison to acoustic-only score of 94%. Although, when only sEMG signals were used, it gave a low average recognition score of around 62%, albeit it highlighted some benefits of using sEMG signals which are more immune to background noise.

Researchers have also focussed on techniques that attempt to model the differences in pronunciation of dysarthric speech. One such novel approach incorporated a metamodel of speaker's phonetic confusion matrix into the ASR framework (Morales and Cox, 2007; Morales and Cox, 2009). The basic idea of such a framework is to resolve the confusion between the decoded and the postulated phone sequences. For example, if the decoded sequence of phones is given by $\mathcal{S}_D$ and the postulated sequence is given by $\mathcal{S}_P$, then the aim is to jointly estimate the probability:

$$P(\mathcal{S}_D, \mathcal{S}_P) = P(\mathcal{S}_D|\mathcal{S}_P)P(\mathcal{S}_P) \tag{3.48}$$

The LHS of equation 3.48 is usually estimated by combining a confusion-matrix model that estimates $P(\mathcal{S}_D|\mathcal{S}_P)$ with a language model. However, using only confusion-matrix can be too restrictive in practice as it is unable to optimally resolve the phone insertion errors. Instead, more flexible discrete density HMM models are used for every phoneme in the system and are termed metamodels. Results reported using the metamodels showed significant gains (5% absolute) over the standard MLLR adaptation for the dysarthric speakers with low intelligibility, especially under limited adaptation data. Sometimes metamodels derived from confusion matrices can prove to be insufficient at modelling some specific phone sequences. In order to overcome these limitation refined approaches based on weighted finite state transducers (WFST) have been developed. In one such study (Morales and Cox, 2009),

a cascade of WFST at the confusion matrix, word and language levels were generated. The WFST results were significantly better than both metamodel and MLLR approaches for speakers within all intelligibility groups. Another study (Seong, Park, and Kim, 2012b) corrected the errors in the dysarthric speech by using an interpolated context-dependent confusion matrix to build a WFST framework and integrating it with a dictionary and language model. The test results on Nemours database reduced the relative WER by 13.7% when compared to the MLLR adapted baseline system, and by 5.9% when compared to the error correction system based on the context-independent pronunciation variation model.

Chapter 4

Recognition and Analysis of Dysarthric Speech

The previous chapter gave an overview of various components of a speech recognition system and its implementation for the recognition of dysarthric speech. Some of the difficulties in recognising dysarthric speech are due to high degree of inter and intra speaker variations, data sparsity issues and malformed phonetic space (Blaney and Wilson, 2000; Kent et al., 2000; Morris, 1989). Broadly categorising, researchers have tried to address these problems in three ways: *(i) Acoustic modelling using both generative and discriminative techniques, (ii) Speaker adaptation approaches and (iii) Signal transformation and enhancement tehcniques*. Despite being pursued as a research problem for more than three decades, the performance is still far behind that for typical speech, which has potentially reached human like performance, especially under controlled conditions (Coleman and Meyers, 1991; Fried-Oken, 1985; Xiong et al., 2016). The effectiveness of any commercial speech system in recognising dysarthric speech has been limited due to reliability and setup constraints in modelling such material. The implementation of any of the existing techniques have shown varying levels of success and there still remains a large gap between the human and machine performance of dysarthric speech. One of the reasons for this difference may be the inadequacy to explicitly address any of the underlying variabilities in dysarthric speech.

This chapter is divided into two parts. The first part (**Part-A**) will extend the work of earlier researchers and use adaptation techniques that might be more suited to implicitly handle variabilities of dysarthric speech. The efficacy of such techniques will be evaluated

on the UASPEECH database of speakers with dysarthria (Kim et al., 2008). It is the largest database of speech of people with cerebral palsy with a vocabulary of 455 distinct words. The aim is to systematically review the alternative existing adaptation techniques and compare with the best available results of similar studies done in the past. This will set a performance benchmark for further research to be measured against. The second part of this chapter (**Part-B**) will give a summary of an acoustic analysis performed on the UASPEECH database. It will cover some key areas like dysarthric phoneme/word timings, F1-F2 vowel plots, frequency phase response and ZZT (Zeros of the Z-Transform) plots of dysarthric vowels. The outcome of the analysis will be useful to illustrate specific differences between typical and dysarthric speech manifest in the acoustic signal.

4.1 Part-A: Baseline recognition results of dysarthric speech

Acoustic modelling and adaptation techniques primarily aim at producing speaker independent (SI), speaker dependent (SD) or speaker adapted (SA) systems. Each of these system comes with its own strengths to handle variabilities in dysarthric speech and might be better suited in certain practical setups than others. For example a good SD system might be more apt for handling small vocabulary tasks, provided a sufficiently large amount of data is available for each intended item in the vocabulary. Data sparsity is a persistent problem for dysarthric ASR. The issue is not likely to be resolved due to physical and ethical constraints. There is a growing need to investigate speaker adapted (SA) systems, which can be trained with less data and aim to achieve SD like performance on small vocabulary tasks. SA systems also have the potential to be extended to larger vocabulary tasks using the same limited data. The success of a good SA system usually depends on robust SI models, which are not too distinct from the target speakers. The focus of the current section will be to investigate the following questions:

- What is the optimal SI system that can be used as a base to adapt to a dysarthric speaker?
- Which is the best technique for adapting a set of recognition models to a speaker with dysarthria?
- Can methods that seek to minimise inter-speaker variability at training time be used to model a dysarthric speaker?

4.1.1 Experimental setup

4.1.1.1 Data preparation

All the experiments presented in this section used two corpora of typical speech, viz., WSJ0 SI-84 (Paul and Baker, 1992) that consists of read speech from 84 North American English speakers with texts drawn from a machine-readable corpus of Wall Street Journal news, and, WSJCAM0 (Robinson et al., 1995) , which is a British English version of WSJ database

The work presented in the current section has already been published in the 6^{th} workshop on Speech and Language Processing for Assistive Technologies, *Model adaptation and adaptive training for the recognition of dysarthric speech (Sehgal and Cunningham, 2015).*

that consists of data from 92 training speakers. For WSJCAM0, data was also included for speakers from the development and two evaluation test sets.

The UASPEECH (Kim et al., 2008) corpus was used, which consists of data from 15 speakers with dysarthria and 13 control speakers. The corpus consists of 765 isolated words (455 distinct words) per speaker collected in three separate blocks, where each block consists of 10 digits, 26 international radio alphabets, 19 computer commands, 100 common words and 100 distinct uncommon words, which were not repeated across blocks. In addition, the corpus also provides an estimate of the intelligibility for each dysarthric speaker by five naive listeners. The ratings will be used in all the experiments for grouping the speakers by the severity of the dysarthria.

Corpus	**Speakers**	**Training Files**
WSJ SI-84	84	14377
WSJCAM0 †	136	18537
UA-CTL	13	41819
UA-DYS	15	44277

Table 4.1: A summary of each training corpus in the system. UA-CTL and UA-DYS codes are used for UASPEECH control and dysarthric speakers. (†) Four evaluation speakers with no secondary microphone data were excluded from WSJCAM0.

All the block one (B1) and block three (B3) data from UASPEECH was used for training or adaptation purposes and block two (B2) was used solely for testing. Because speakers with dysarthria can take longer to utter words, the UASPEECH training data had to be logically resegmented to remove extra silences around approximate word boundaries. 200 ms of silence was left to either side of the word for training. However, test data block B2 was left untouched to maintain the natural speaking conditions. Data from all the microphones was used for each corpus for training and adaptation purpose and a summary is given in Table 4.1.

4.1.1.2 Acoustic modelling

For acoustic modelling, data from all the corpora was processed as 12 dimensional MFCC features with c_0 and cepstral mean normalisation. First and second order time derivatives

were also appended giving a 39 dimensional feature vector per frame. Speech was analysed with a 25 ms window with a 10 ms target shift rate. Continuous density HMMs used in all the experiments are word-internal tied-state triphone models. They use a strict left-to-right topology with 3 emitting states and 16 Gaussian components used per state. Silence models used 32 Gaussian components per state. A phonetic decision tree was constructed for state clustering.

4.1.1.3 Methodology

An objective of these experiments is to determine the best baseline SI models that can be used as the basis of a SA system. There can usually be a large number of combinations in which multiple data sources can be mixed together to build SI models. Hence, in context of the databases that are used in this study, the most appropriate logical combinations of data sources are used. Table 4.2 summarises the SI systems that were built for future adaptation experiments.

System Code	**Training Dataset Used**
SI-00	WSJ SI-84 + WSJCAM0
SI-01	UA-DYS excluding target test speaker
SI-02	UA-DYS
SI-03	UA-CTL
SAT	UA-DYS

Table 4.2: Summary of baseline systems and the corpus used for its preparation.

The SI systems generate a set of models for a particular kind of speaker (e.g. British English, typical speech, American English, dysarthric speech etc.). It assumes that the acoustic realisations of such a speaker group is constant throughout the database. During typical speaker adaptation, the optimal model set $\tilde{\Phi}$, given a set of S speakers in the system is generally represented as:

$$\tilde{\Phi} = \arg\max_{\phi} \mathcal{L}(O;\phi) = \arg\max_{\phi} \prod_{s=1}^{S} \mathcal{L}(O^{(s)};\phi) \tag{4.1}$$

where $\mathcal{L}(O^{(s)};\phi)$ is the likelihood of the observation sequences from speaker s, given the current set of model estimates ϕ.

A justification for choosing the systems in table 4.2 is given below:

- **SI-00:** Systems prepared from such a combination of typical speech data represent an extremum model set, which will be acoustically distant to dysarthric speech. It will enable our understanding to study the gap between typical and dysarthric speech from a modelling perspective. It further rules out any peculiarities about regional accents (*which are not strongly exhibited in dysarthric speech*) due to data merging and is independent of any particular dysarthric database vocabulary.

- **SI-01, SI-02:** The two systems are trained using the entire UASPEECH dysarthric data. The SI-02 system includes the training data of the target test speaker and SI-01 system is unbiased towards any particular target test speaker. Since dysarthric speech has high inter and intra speaker variability, testing the two systems in parallel will give a good understanding on the importance of speaker specific data in building models for speaker adaptation.

- **SI-03:** This system is trained with typical speech data that has the same specific vocabulary of the target test speakers. The system is prepared from much less data than the analogous SI-00 system. It will help understand if better modelling for dysarthric speech adaptation will benefit from large quantities of data or sparse data with a complementary vocabulary.

- **SAT:** Systems that use SAT based training routines have the inherent capability to implicitly reduce the inter speaker variabilities. Thus, SAT is better suited to produce robust baseline models for speaker adaptation. SAT training splits the data into blocks, where each block assumes homogeneity of the underlying acoustics, e.g. data pertaining to a particular speaker is regarded as a homogeneous block for incorporating particular speaker induced variations. It uses two sets of parameters, a canonical model ϕ_c, usually hypothesised to represent phonetically relevant speech variabilities, and the set of transforms $\mathcal{T}^{(s)}$ to represent the speaker variabilities. This is given as:

$$(\tilde{\Phi}_c, \tilde{\mathcal{T}}) = \underset{(\phi_c, \mathcal{T})}{\arg\max} \prod_{s=1}^{S} \mathcal{L}(O^{(s)}; \mathcal{T}^{(s)}(\phi_c)) \tag{4.2}$$

  In the above equation speaker variations are modelled by $\mathcal{T}$ and the canonical model is updated, given each transform. The entire SAT paradigm works iteratively in an

interleaved fashion. Refer to section 3.5.4.1 for details on the SAT framework.

The baseline adaptation results presented in section 4.1.2 for the SI and SAT models will use the standard MLLR and MAP techniques. In addition a hybrid MLLR-MAP approach is also presented. SAT canonical models are intentionally trained using only UA-DYS speakers to minimise the effect of inter-speaker variabilities associated with speakers with varying degree of intelligibility. The MLLR implemented uses a two-pass static adaptation procedure. The first pass performs a global transformation and the second pass uses the global transforms to produce more accurate transforms using a regression class tree with 32 terminal leaf nodes. It should be noted that SAT based on MLLR transforms should be able generate robust canonical models, however, it comes with computational and memory overheads (Spyros et al., 1997), making it impractical for implementation. Such issues can be avoided by applying constrained MLLR (CMLLR) (Digalakis, Rtischev, and Neumeyer, 1995; Gales, 1998b), which uses the same transform for both means and variances. The transforms are computed for each homogeneous block of data. SAT with CMLLR results in a kind of feature normalisation during model training and has the same computational load as any other standard HMM update process. Unlike SI models which can be directly used for recognition, SAT canonical model sets are not suitable for direct decoding. Both systems are usually adapted to some target condition.

4.1.2 Experimental results

All the test results presented here were obtained using test set B2 of the UASPEECH corpus (see 4.1.1.1). Since the database comprises of single word utterances, the decoding grammar was strictly restricted to only one of the possible test words, preceded and succeeded by silences. There are 255 distinct competing words in the test block with a total of 22281 files from all speakers and microphones.

4.1.2.1 SI systems

The first set of experiments involved obtaining recognition scores of all the baseline SI systems explained in table 4.1. These were then compared alongside the speaker dependent (SD) performance. Figure 4.1 shows the average baseline accuracy of all the speaker independent systems. SI-00 has the lowest baseline result, which would be expected as it was

trained only on typical speech. The highest accuracy was obtained using the SI-02 system, which was trained on the largest amount of dysarthric speech data.

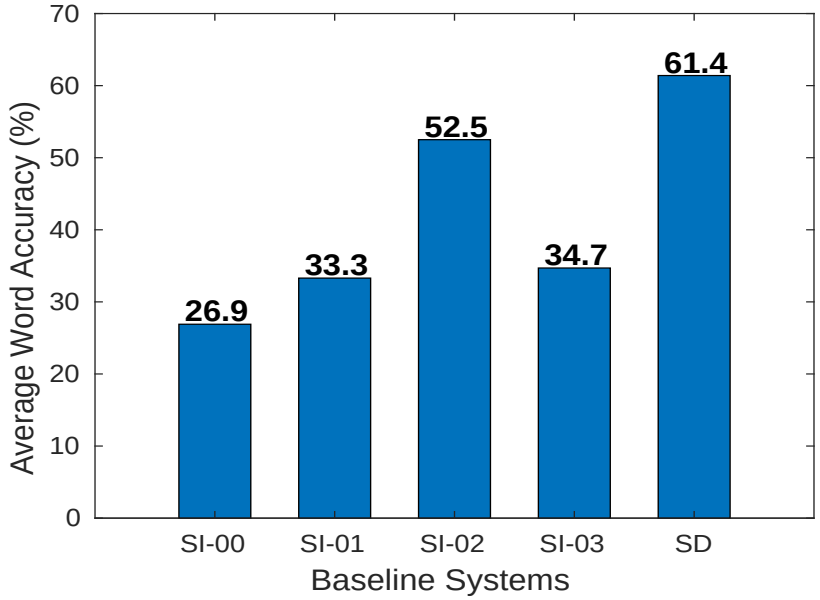

Figure 4.1: *Average word accuracy for the baseline SI systems along with the SD result.*

4.1.2.2 SI adapted systems

All of the baseline SI systems were adapted for each test speaker. Standard techniques were used and the results are shown in figure 4.2 and table 4.3. MAP outperforms the MLLR based adaptation in all cases except SI-00 models that are trained from WSJ0 + WSJCAM0 datasets. Since SI-00 models use only typical speech data, it might not present much useful information about the parameter distributions of the adaptation and test datasets. This may be a classic example of non-informative priors that does not assist in reducing the training and test mismatch.

Following on from this observation a combined approach (MLLR-MAP) was implemented that involves generating MLLR transforms for the target speaker followed by MAP adaptation. By doing this, MLLR adapted parameters can act as informative priors for the MAP process. For all the SI systems, the MLLR-MAP combination outperformed all other adaptation approaches. Intuitively, it may be thought that SI-01 or SI-02 should form an

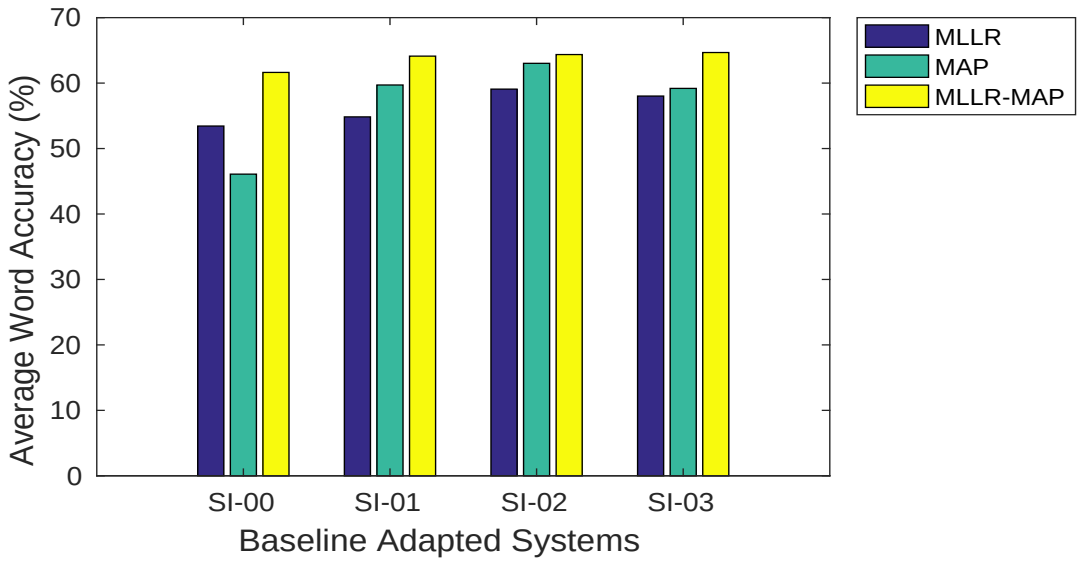

Figure 4.2: *Adaptation scores for the baseline SI systems.*

optimal set of baseline models for adaptation, since they exhibit less difference between the training, adapted and test conditions. Overall, the best MLLR-MAP scores for dysarthria and typical speech based SI systems was found to be for SI-02 and SI-03.

The remainder of the thesis will only present results obtained using the MLLR-MAP adaptation approach.

4.1.2.3 SAT-adapted vs other systems

One of the investigative goals of section 4.1 is to study the effect of SAT based modelling, which has the inherent capability to minimise the effect of inter-speaker variations during training time. It is known that such variabilities are present in the acoustics of dysarthric speech. Figure 4.3 and table 4.3 shows a comparison of the MLLR-MAP based SI and SAT systems. SAT-adapted model sets outperform all the other tested systems

It should be noted that the SD system performs more poorly than all the adapted systems. Indeed, it can be seen in table 4.3 that SD system does not perform better

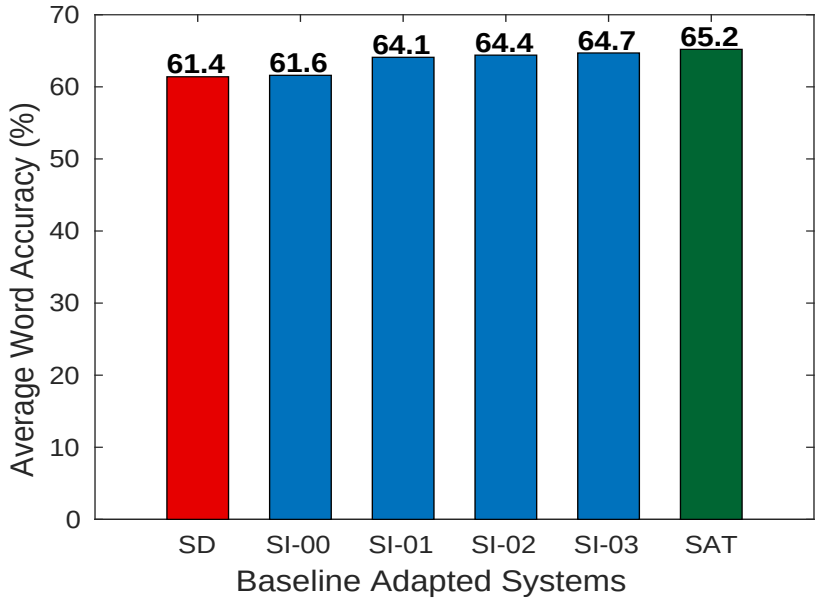

Figure 4.3: *Comparison of SD and MLLR-MAP based SI & SAT systems.*

than any of the SA systems (*except one speaker*) in each of the intelligibility sub-groups. This gives us a basis to assuming that adaptation can be an effective approach to model dysarthric speech of varying intelligibilities. A similar finding about the efficacy of SA systems was also reported in a study by Sharma and Hasegawa-Johnson (2010). However, the findings are contrary to some of the earlier published results (Raghavendra, Rosengren, and Hunnicutt, 2001; Rudzicz, 2007), which show better performance with SD systems for decreasing intelligibility. Christensen et al. (2012) found that SI systems trained with only dysarthric speech produced better baseline models for adaptation that was beneficial for most of the speakers.

To the best of our knowledge, all the systems we tested gave significantly better results than the earlier similar published results in the literature. Please refer to table 4.4 and section 4.1.3 for a comparative discussion. The findings further suggest that SI systems like SI-03, prepared from typical speech can also adapt as well as a dysarthric speech based SI system, especially for speech with increasing intelligibility. In order to test this interpretation, the effectiveness of all the MLLR-MAP based SAT and SI systems along with SD system were statistically analysed using Cochran's Q test (Cochran, 1950; Gillick and Cox, 1989).

Cochran's Q is a non-parametric test to verify whether a group of s different treatments (speech systems in our case) have identical or different effect in the recognition process. In this study the null hypothesis for the Cochran's Q test defines that *"different speech systems are equally effective to model the data"*. The outcome of each system is recorded as a binary response indicating if the recognition of an utterance was successful or not. The test statistic TS is given by:

$$TS = \frac{(s-1)\left[s\sum_{j=1}^{s}\hat{x}_j^2 - \sum_{i=1}^{n}\sum_{j=1}^{s}x_{ij}^2\right]}{s\sum_{i=1}^{n}\sum_{j=1}^{s}x_{ij} - \sum_{i=1}^{n}\hat{x}_i^2} \tag{4.3}$$

where

s is the number of speech systems tested

$\hat{x}_j$ is the sum of correctly recognised files for the j^{th} speech system

n is the total number of speech files recognised (also called blocks)

$\hat{x}_i$ is the total number of speech systems that correctly recognised the i^{th} file

The null hypothesis is rejected if the test statistic is in the critical region ($TS > \chi^2_{critical}$) of the chi-squared distribution with $s-1$ degrees of freedom. For $s = 2$, the Cochran's Q test reduces to a pairwise examination that is equivalent to McNemar's test.

All the six speech systems were tested for differences across all the test speakers. The null hypothesis was rejected at $\alpha = 0.01$ (degrees of freedom = 5), which meant that all the systems were not equally effective for modelling dysarthric speech in general. Later a pairwise Cochran's Q test was conducted between the system with the best absolute average score (SAT) and all others. The test showed that SAT was significantly different to all other systems at $p < 0.01$, except for the SI-03 system in the higher intelligibility group.

4.1.2.4 MLLR-MAP for severity groups

So far the reported findings are averaged across all the test speakers irrespective of intelligibility. However, to have a more informed approach for preparing systems for specific speakers it is important to study the effect of SD and SA based systems in each of the intelligibility groups. Figure 4.4 shows an overall picture of how the baseline SI systems performed for the different intelligibility sub-groups as defined for the UASPEECH corpus. Figure 4.5 shows the effect of adapting the respective baseline systems along with SAT estimates. Systems trained or adapted with some dysarthric data or the SAT based system

Intelligibility	Speaker (Intelligibility%)	SD	MLLR-MAP				
			SI-00	SI-01	SI-02	SI-03	SAT
Very Low	M04 (2%)	6.54	8.98	9.5	8.54	8.11	9.68
	F03 (6%)	32	27.61	37.49	36.01	36.81	38.36
	M12 (7%)	32.24	17.76	35.08	32.31	30.71	32.9
	M01 (17%)	16.76	27.03	28.32	28.22	27.46	29.22
Sub Acc.		23.52	20.61	28.82	27.36	26.95	28.71
Low	M07 (28%)	62.33	69.7	69.26	68.89	61.91	66.06
	F02 (29%)	61.08	37.62	50.12	54.02	50.93	56.93
	M16 (43%)	64.29	68.08	62.76	66.47	65.23	66.55
Sub Acc.		62.48	57.89	60.56	62.92	59.03	62.98
Mid	M05 (58%)	70.48	64.27	69.93	70.6	67.47	71.83
	M11 (62%)	58.18	56.57	63.8	66.06	68.1	65.62
	F04 (62%)	62.66	76.06	70.57	68.48	74.52	70.57
Sub Acc.		64.44	66.12	68.34	68.51	70.13	69.54
High	M09 (86%)	80.96	83.11	84.43	85.62	87.82	86
	M14 (90%)	77.76	80.4	80.09	79.2	85.71	80.84
	M10 (93%)	84.28	91.77	86.28	87.21	91.33	88.08
	M08 (95%)	85.86	87.96	87.21	86.47	87.4	87.34
	F05 (95%)	86.46	92.14	92.01	92.33	90.58	92.08
Sub Acc.		83.07	87.08	86.01	86.17	88.57	86.87
Overall Acc.		61.44	61.63	64.12	64.36	64.67	65.15

Table 4.3: Absolute word accuracy for SD and SI/SAT baseline systems adapted using MLLR-MAP. The table also shows sub accuracy scores under various intelligibility groups. The best scores are highlighted in grey for each row.

were the most effective for recognising speakers with lowest intelligibility, while systems prepared from typical speech data resulted in improved recognition for the high intelligibility group of speakers. Table 4.3 gives a detailed report for all the dysarthric test speakers.

In order to test the differences between the systems, a Cochran's Q test was again applied for various intelligibility groups. The summary of the results of this test are shown in Table 4.4. It shows that SAT system is statistically equivalent to some other systems in the *very-low, low and mid* intelligibility group of speakers.

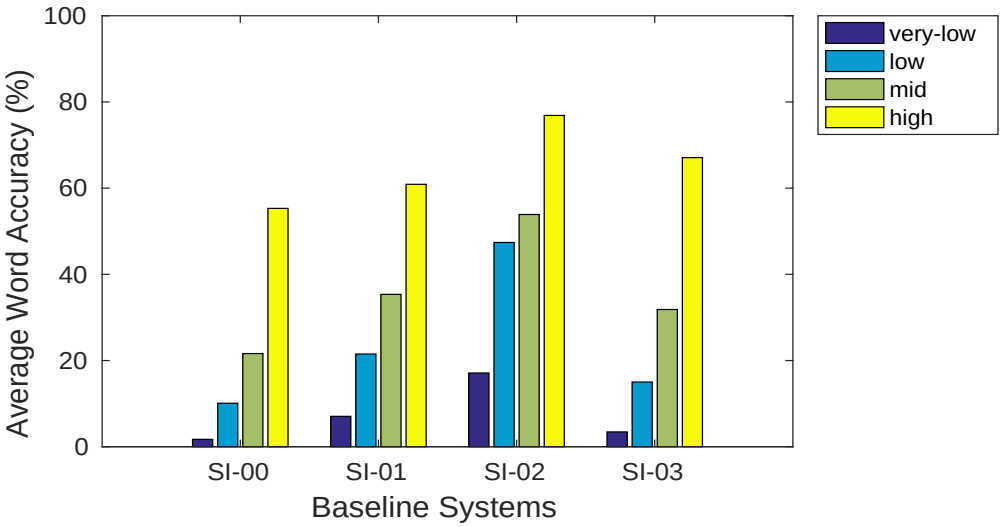

Figure 4.4: Accuracy for the baseline SI systems for various intelligibility groups.

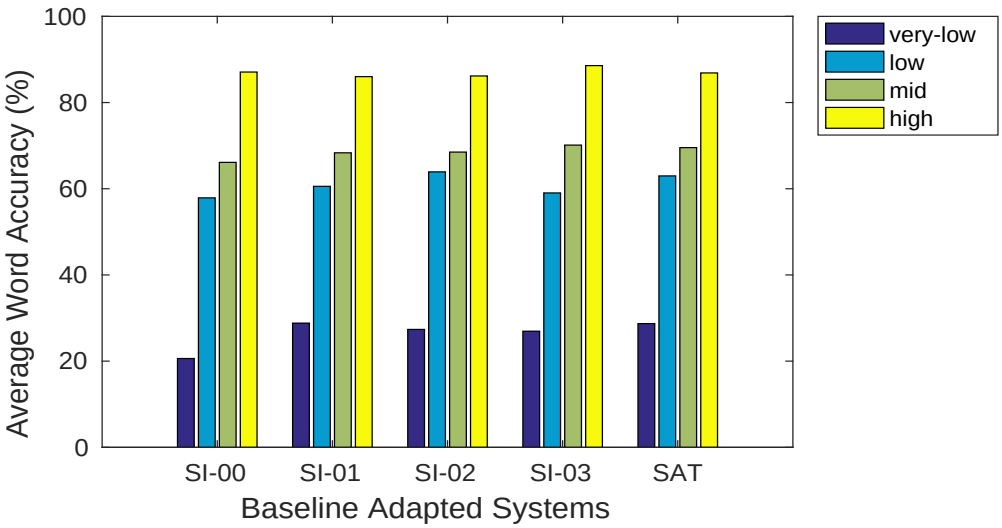

Figure 4.5: MLLR-MAP scores for the SAT & SI systems for various intelligibility groups.

For the *high* intelligibility sub-group, system trained from typical speech data with the same recording and vocabulary as the test conditions was significantly better than all the other systems.

Intelligibility	**Best performing systems** ($p < 0.05$)
Very Low	SAT, SI-01
Low	SAT, SD, SI-02
Mid	SAT, SI-03
High	SI-03

Table 4.4: Cochran's Q analysis on the tested speech systems for various intelligibility groups. The best performing systems for each intelligibility group are statistically equivalent.

4.1.3 Discussion of baseline results

The results reported in section 4.1.2 show that it is difficult to train a single system to model the variabilities in dysarthric speech and to also generalise to speakers of different severities. For example, when studying the performance of various baseline systems in section 4.1.2.1, it was interesting to note that SI-03 had similar performance to SI-01 system, despite being trained from typical speech data. We think that SI-03 models benefit from homogeneous vocabulary and the same recording conditions in the test data.

The findings also show that SD system were not the most effective to model dysarthric speech except in the low intelligibility group. This can be partially attributed to the relatively small amount of data per speaker in UASPEECH, especially when compared to previous studies in the literature (Raghavendra, Rosengren, and Hunnicutt, 2001; Rudzicz, 2007). The test block B2 also comes with many unseen words in the form of 100 unique "uncommon words" and an SD system is usually only trained to maximise the model fit for the seen data during training. In contrast, a SA system might overcome this problem to some extent by using acoustic information present from other speakers in the baseline SI systems. This might be a contributing factor for all the adapted systems to be significantly better than majority of SD system.

Another point of interest was reported in section 4.1.2.3. It indicated that to model dysarthric speech with higher intelligibility, SAT and SI-03 systems were not significantly different. Hence, it is more flexible to prepare the baseline model sets from disparate data sources to adapt dysarthric speech with reduced severity.

The results suggest that the variabilities in dysarthric speech can be better accommodated from modelling both typical and dysarthric domains. One such attempt was reported by Sharma and Hasegawa-Johnson (2013), where background interpolation MAP was implemented to obtain an intermediate prior acoustic model to narrow the gap between two disparate SI systems (typical & dysarthric), albeit, the reported results were worse than those reported by Christensen et al. (2012). Our best overall results were obtained on the MLLR-MAP adapted SAT systems. It gives an absolute gain of 23% (54% relative) over results of Sharma and Hasegawa-Johnson (2013) and an absolute gain of 11% (21% relative) over results of Christensen et al. (2012). In a more recent study, the results for 6 speakers of UASPEECH was reported using the DNN-HMM framework (Tejaswi and Umesh, 2017). The author used various methodologies like knowledge distillation, multitask learning and model adaptation alongside the conventional GMM-HMM systems. Although, the study reported DNN-HMM to be 13% relatively better than the GMM-HMM systems, the results did not outperform our SAT based GMM-HMM systems that deployed hybrid adaptation procedures.

The choice of a particular system for a given target speaker is not completely clear, even when analysis is carried out at specific intelligibility groups. Table 4.4 shows there are possible choices in the lower intelligibility group of speakers. Since dysarthric speech is likely to be more variable in the lower intelligibility group, the presence of SI-01 and SI-02 speech systems does not come in as a surprise as they will be inherently capable of modelling some of the common differences. The statistically equivalent performance of SD system in the *low* intelligibility sub-group was rather unexpected. Upon closer examination it was found that group mean was getting biased for the speaker **F02** who exhibits a huge variation in the reported accuracy.

Despite the fact that several alternatives appear to be equivalent for different groups of speakers, it is noticeable that SAT-based systems are among the best performing for the very low to mid intelligibility groups. This may be due to the implicit capability of SAT to remove the speaker induced variations during training. This speaker normalising might be having a nullifying effect on some complex variabilities present across all the speakers.

Among systems trained with typical speech, both SI-00 and SI-03 were found to the most effective for higher intelligibility group of speakers. SI-03 is a significantly better base model for adaptation than SI-00. This is despite being trained with a smaller dataset and it may suggest that large quantities of typical speech data might not be necessary for

training models meant for people with dysarthria. In addition to acoustic similarities, SI-03 system also has an additional benefit of homogeneous vocabulary and recording conditions. It was also observed that SI-03 was the best performing system for speakers with a high intelligibility. This was expected because high intelligibility dysarthric speech is more similar to typical speech and table 4.3 clearly shows the inclination of typical speech baseline systems (*SI-00, SI-03*) to model *high* intelligibility sub-group of speakers. Finally it was also noted that SI-00 marginally scored better for two speakers in the low intelligibility group despite its dissimilarity to the target test speakers relative to the other systems. However, the difference was noted to be statistically insignificant. Indeed it can be seen from table 4.4 that SI-00 did not show up as a statistically equivalent system in any of the intelligibility groups.

4.2 Part-B: Acoustic analysis of the UASPEECH database

Acoustic studies of dysarthric speech can be very informative in delineating certain aspects of the underlying inconsistencies manifest in specific phonetic realisations. However, widespread variabilities present across various types of dysarthria makes it an extremely challenging task. There has been a lot of progress made in the past three decades with a greater understanding of the relationship between specific acoustic variables and their underlying perceptual correlate.

Broadly speaking, any acoustic analysis carried out on dysarthric speech is intended for three types of tasks:- **(i) Prediction of intelligibility** (Feenaughty, Tjaden, and Sussman, 2014; Kent and Kim, 2003b; Kim, Hasegawa-Johnson, and Perlman, 2011; Magnuson and Blomberg, 2000; Rong et al., 2012b; Weismer et al., 2001), **(ii) Classification of various dysarthric types or etiologies** (Kim, Kent, and Weismer, 2011b; Skodda, Visser, and Schlegel, 2011; Weismer et al., 2001) and **(iii) Assessing disordered speech by comparing it against the acoustic measures of typical speech** (Blaney and Wilson, 2000; Kent et al., 2000; Morris, 1989). A systematic coverage of acoustic analysis methods and its applications is detailed in the comprehensive paper by Kent et al. (1999b).

There haven't been any studies that have looked specifically at the acoustic-phonetics of the speech with respect to how differences might affect the dysarthric ASR performance. This might be due to the lack of modelling or correcting any explicit dysarthric anomaly that is manifest in the acoustic signal. This section will attempt to search for variability in the acoustic domain, which can help to build a hypothesis for improved signal parametrisation or modelling. The investigative work can be summarised as:

- Investigate acoustic variables postulated by earlier studies that might show quantifiable differences for dysarthric speech.
- Examine the differences between dysarthric and typical speech in the Z-domain. This will be carried out by conducting a ZZT (*zeros of the z-Transform*) analysis of the vowel segments.
- Develop a hypothesis based on the ZZT analysis of dysarthric vowel segments, which can help to formulate a novel approach for understanding the relationship between such material and intelligibility.

The acoustic analysis in this section will use data from the dysarthric and control speakers of the UASPEECH corpus.

4.2.1 Temporal analysis

The temporal analysis will look at the durational patterns of phoneme and word productions of control and dysarthric speakers. It will be studied at both the speaker and intelligibility levels. The aim is to investigate if patterns in the speaking rate can have any potential associations with intelligibility or other aspects of speaker's dysarthria. A summary of the time based analysis, which will be studied in this subsection is shown in figure 4.6.

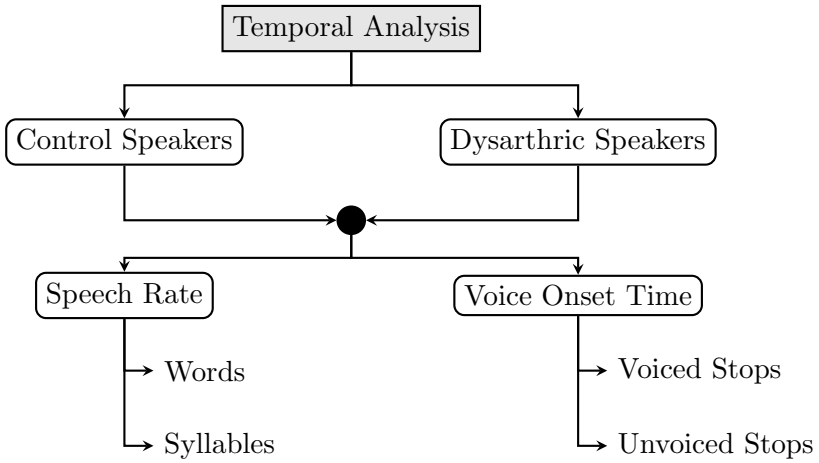

Figure 4.6: Temporal analysis experiments conducted for the UASPEECH database.

4.2.1.1 Speech rate

Since UASPEECH database only consists of word based utterances, the term speech rate and articulation rate are used interchangeably and hold similar connotations. The word durations were extracted from forced-aligned label files, which included times for any pauses and non-speech events within words. The search engine uses the Viterbi process to align the spoken utterance with its exact transcription. In addition, since UASPEECH database consists of both mono and poly syllabic words ranging from $1-6$ syllables, the articulation rate was measured in **syllables per second (sypse)** to remove any bias for the word length.

Figure 4.7 shows that dysarthric speakers have a much slower speaking rate at 1.86 **sypse** in comparison to 3.0 **sypse** for the control group. Dysarthric speakers also exhibit more than twice the degree of variation ($\sigma = 0.80$) than the control group ($\sigma = 0.33$). When the analysis was conducted across various intelligibility groups, it showed an association between

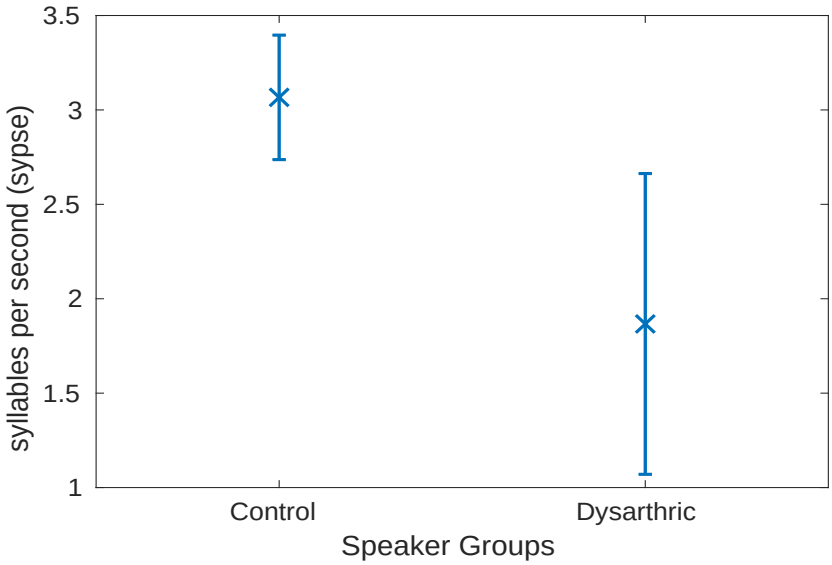

Figure 4.7: **sypse** for the control and dysarthric speakers averaged across all the words.

slow speaking rate and intelligibility of speaker groups. This is depicted in figure 4.8, where three out of four intelligibility groups fall below the average dysarthric **sypse** rate and the high intelligibility group tends to be similar to the control speaking.

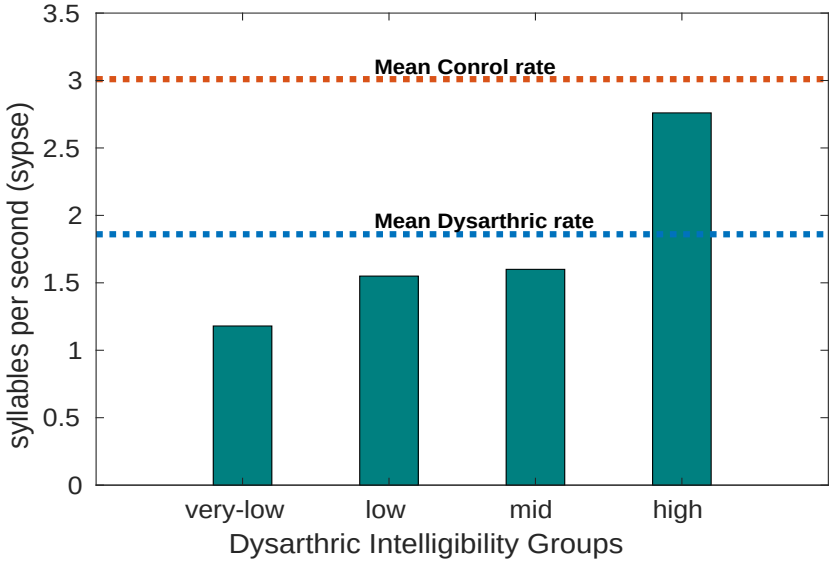

Figure 4.8: **sypse** for various intelligibility groups averaged across all the words.

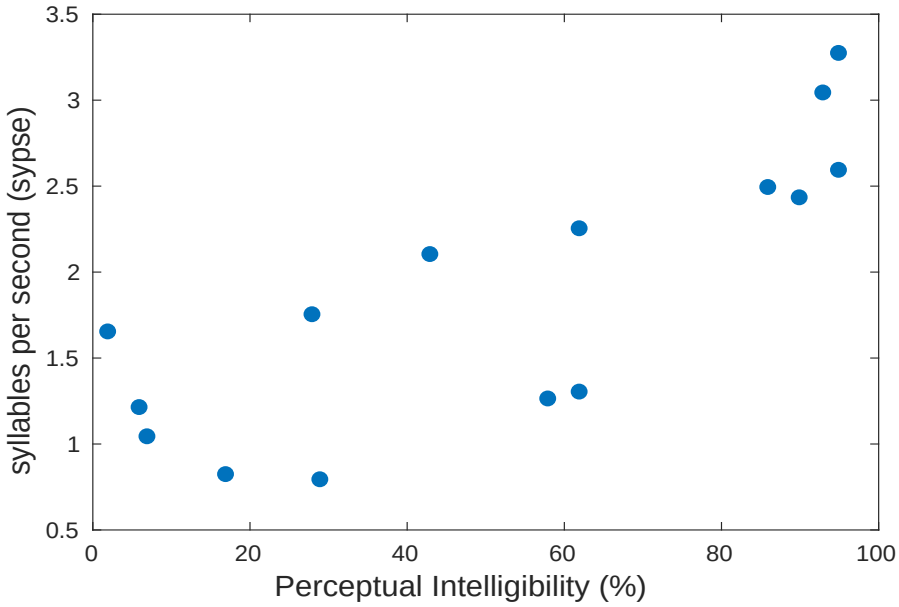

Figure 4.9: Scatter plot between sypse and intelligibility.

So far we have seen a clear association of **sypse** with various dysarthric intelligibility groups. Figure 4.9 further corroborates this observation where the scatter plot shows a strong positive correlation of $r = 0.81 (p < 0.01)$ between **sypse** and the perceptual intelligibility of various dysarthric speakers.

Despite the strong association of **sypse** with intelligibility, it does not show a predictable pattern of speaking rates when it was analysed across speakers within each intelligibility group as shown in figure 4.10. The ideal expected trend should have followed an upward staircase pattern for the speakers arranged from left (least intelligible) to right (highest intelligible). For example, the measured **sypse** rate for the very-low and low speaker groups is contrary to the predicted intelligibility and the **sypse** trend only seems to fall within an ideally expected range as we move higher up in the intelligibility spectrum. As a specific example, if we observe **sypse** readings for the least intelligible speaker (M04), it is falling within the average range of mid-intelligibility group of speakers. It indicates that the speaker is not speaking at a substantially slower rate, which could be detrimental from speech recognition perspective. This can have two possible interpretations:- (i) Intelligibility might not be a strong indicator of speaking rate patterns in dysarthric speech with reduced intelligibility and (ii) It is unfeasible to construct a generic framework for modelling the

underlying temporal disfluencies, and it would be better to handle timing inconsistencies on a speaker specific basis.

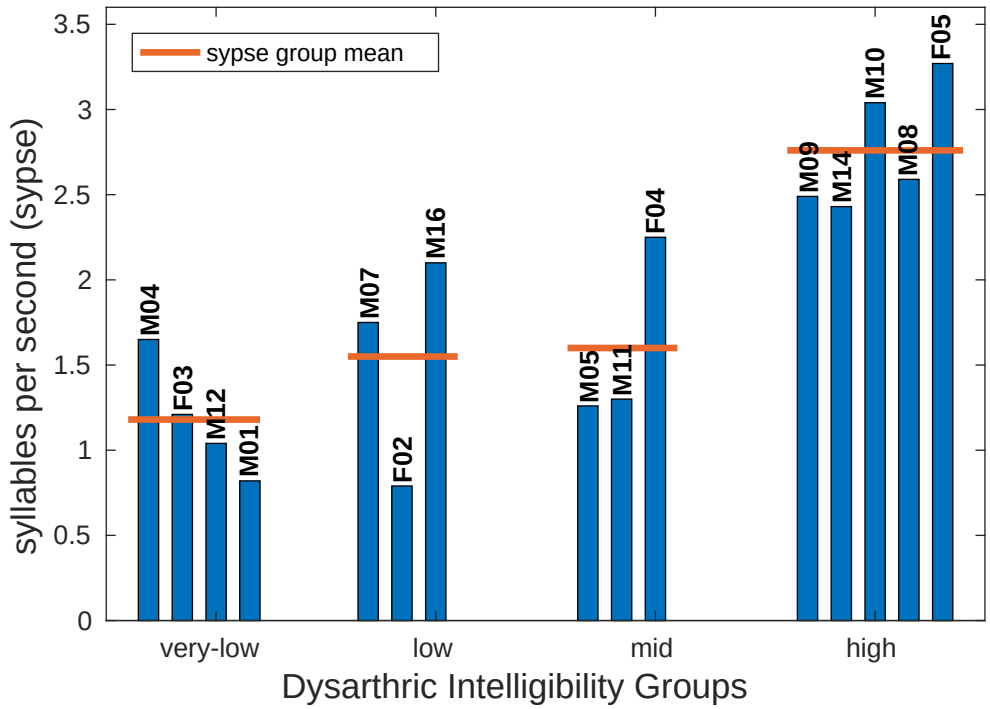

Figure 4.10: Speaker-wise **sypse** analysis averaged across all the words. Speakers are ordered according to their increasing intelligibility from left to right.

In a HMM-GMM based acoustic model, the transition matrix implicitly models the durational information for phones through a prior geometric distribution. However, modelling such information is mostly avoided in ASR since the standard use of HMM's cannot optimally model phonetic durations. Multi variate distributions have shown to better model the durational information at both the phonetic level (Pylkknen and Kurimo, 2004) and speaking rate of individuals (Samudravijaya, Singh, and Rao, 1998).

Since incorporating any such explicit distributions into the HMM topology to model durations violates the first-order Markov property, various duration modelling approaches have been proposed. Some well known techniques involve modifying the HMM topology by either introducing explicit state duration distributions (hidden semi-Markov models, HSMM) or splitting specific states into sub-HMM systems (expanded state HMM, ESHMM)

that implement an optimal state duration distribution with specific topology and transition probability (Russell and Cook, 1987). These techniques have further been supplemented with optimisation techniques for improving the real time performance (Bonafonte, Ros, and Marifio, 1993). More recently, feed-forward neural networks (Rao and Yegnanarayana, 2004) and DNN based systems (Shreekanth, Udayashankara, and Chandrika, 2015) have been used for explicit duration modelling, albeit, most of the work has been carried out in the speech synthesis domain.

4.2.1.2 Voice onset time

Another time based feature that has been investigated in the acoustic study of speech is the voice onset time (VOT). It gives an insight into the production of stop consonants. It is often used as a quantitative metric for intelligibility preiction and discrimination of various dysarthric etiologies and types (Lisker and Abramson, 1964; Morris, 1989). For example in a study of twenty dysarthric speakers, it was found that the VOT measure for flaccid and ataxic speakers showed significantly greater variability than the VOT measure for spastic and hypokinetic speakers, with spastic speakers exhibiting the shortest VOT value (Morris, 1989). A complete VOT study of dysarthric speech is beyond the scope of this thesis and this section will only focus on some broad level aspects. It will primarily report on a comparison between dysarthric and control speakers, association of VOT with the intelligibility of dysarthria and some general qualitative observations.

Voiceless-Voiced Pairs	**Vowel Context**	**Word Production**
/p/ - /b/	/iy/ (Front-High)	people, be
/t/ - /d/	/uw/ (Back-High)	two, do
/k/ - /g/	/aa/ (Back-Low)	copy, golf

Table 4.5: The configuration parameters used for the VOT measurements of the voiceless and voiced stop consonants. The examples in the "word production" column was used to extract the VOT values. The symbolic notation of stops and vowels are taken from the ARPABET phonetic transcription codes.

For VOT measurements, both the voiceless (/p/, /t/, /k/) and voiced stops (/b/, /d/,

/g/) are examined. VOT measurements are usually extracted by recording repeated productions of the type /*stop*∧/ (Morris, 1989). Since UASPEECH does not have such recordings, the VOT is measured by studying the voiceless/voiced stops in context of one of the front or back vowels. Table 4.5 shows a summary of the vowels and words that were examined in context of various stop consonants for VOT measurements.

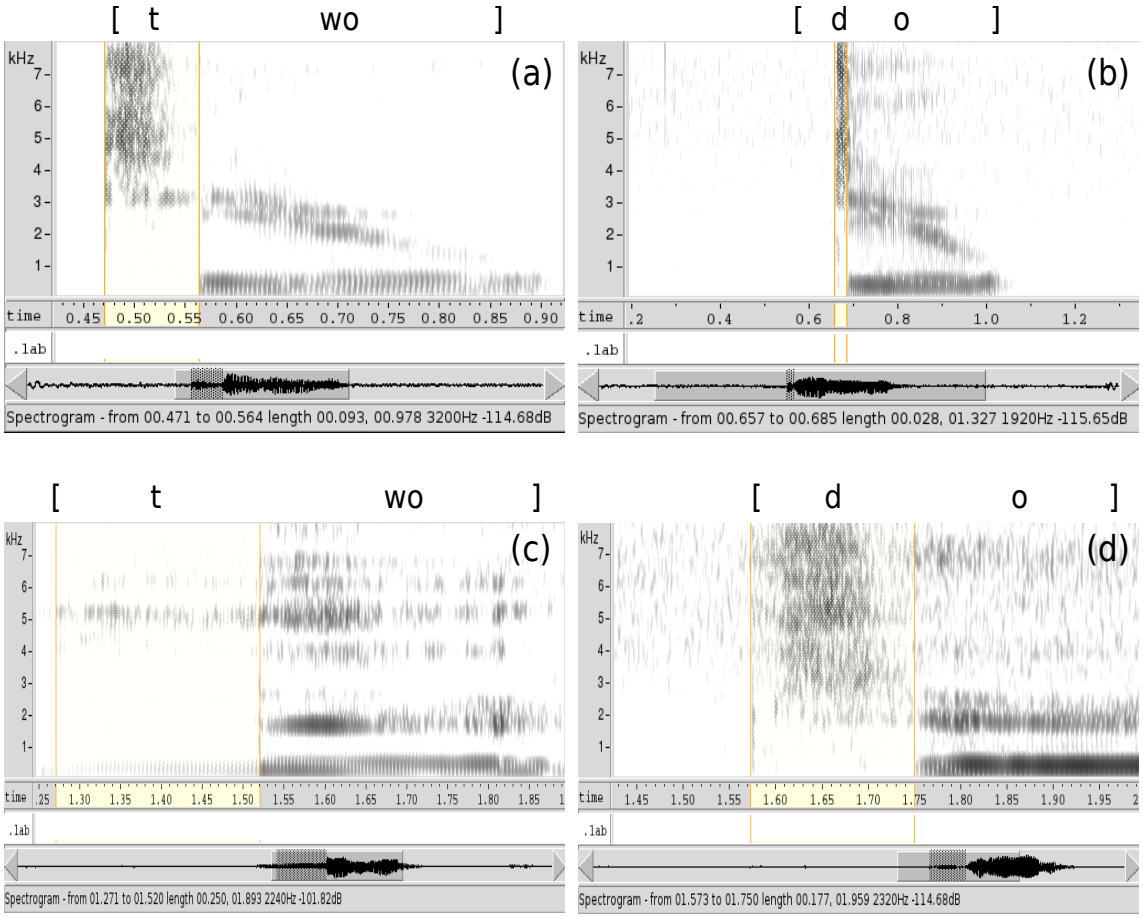

Figure 4.11: VOT timing spectrograms for English words "two" and "do". The VOT values are shown by the highlighted area for the voiceless and voiced alveolar stops /t/ and /d/. The figure shows the (a) VOT value of **93 ms** for the phoneme /t/ for a control speaker, (b) VOT value of **28 ms** for the phoneme /d/ for a control speaker, (c) VOT value of **250 ms** for the phoneme /t/ for a severe dysarthric speaker, (d) VOT value of **177 ms** for the phoneme /d/ for a severe dysarthric speaker.

The VOT measurements were manually extracted by visual inspection. No uncommon words were taken from UASPEECH corpus as it does not have repeated utterances across the three blocks of data. Since there are 3 utterances from all the blocks for each word, in all 18 utterances were inspected for each speaker to get the VOT scores for all the stops. This was repeated for both the control and dysarthric speakers.

VOT was measured by hand using the standard procedure described in Lisker and Abramson (1964). It involves examining the wideband spectrograms and evaluating the segment between the stop-release and the onset of glottal pulse vibrations. The point of voicing onset was marked by examining the first instance of the regularly spaced vertical striations. Figure 4.11 shows an example of how the VOT measurements were taken for the control and dysarthric group of speakers.

The exact data is examined by repeating the process over all the speakers across all the stops. It should be noted that during the VOT marking process, there was no subtle distinction made between voiced and voiceless aspirated and unaspirated plosives. It, thus rules out the possibility of reporting negative VOT values. It can be seen in figure 4.11 that dysarthric speakers with lowest intelligibility usually manifests an inflated VOT score.

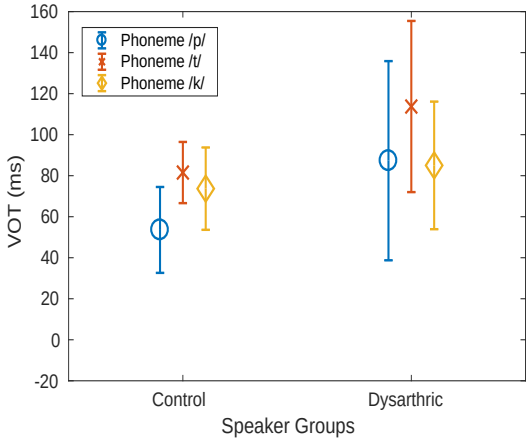

Figure 4.12: Voice Onset Times for the voiceless stops /p/, /t/, /k/

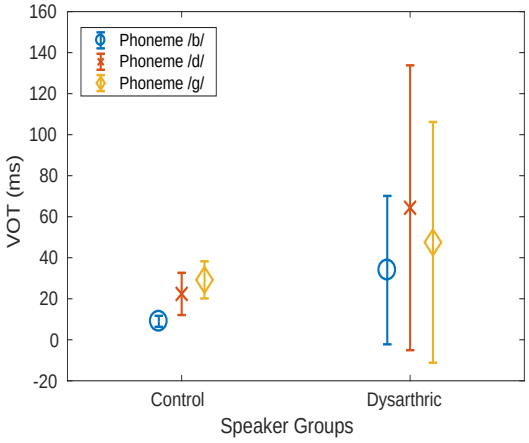

Figure 4.13: Voice Onset Times for the voiced stops /b/, /d/, /g/

It is evident from figure 4.12 and 4.13 that speakers with dysarthria have escalated mean VOT values for both voiceless and voiced stop consonants. The average VOT in ms for both group of speakers are /p/$\rightarrow$ $(54_C, 87_D)$, /t/$\rightarrow$ $(82_C, 114_D)$ and /k/$\rightarrow$ $(74_C, 85_D)$

for voiceless stops and /b/→ $(9_C, 34_D)$, /d/→ $(22_C, 64_D)$ and /g/→ $(29_C, 48_D)$ for voiced stops. Dysarthric VOT exhibits longer lag times, which is accentuated by greater standard deviations of /p/→ $(21_C, 49_D)$, /t/→ $(15_C, 50_D)$ and /k/→ $(20_C, 31_D)$ for voiceless stops and /b/→ $(3_C, 36_D)$, /d/→ $(10_C, 69_D)$ and /g/→ $(9_C, 58_D)$ for voiced stops. On average, the standard deviation of dysarthric speech across the VOT values of all the stop consonants is more than 2.5 times greater than control group of speakers.

For the control speakers, a pattern of increasing VOT is observed as the point of occlusion moves posteriorly inside the oral cavity for the voiced stops. A slight dip was however observed for the voiceless stops. However, an increasing-decreasing pattern is observed for both voiced and voiceless stops of dysarthric speech with a higher degree of variation across the mean. It is hard to make any committed judgement about the relationship between the point of occlusion and the VOT times, since the stop consonants are not observed within the same vowel context. However, a larger standard deviation with an increasing-decreasing pattern for dysarthric speakers might indicate lack of muscular control as the occlusion moves towards the velum.

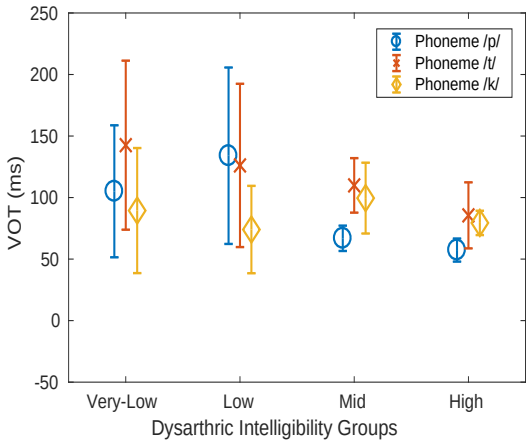

Figure 4.14: Voice Onset Times for the voiceless stops /p/, /t/, /k/ across various intelligibility groups

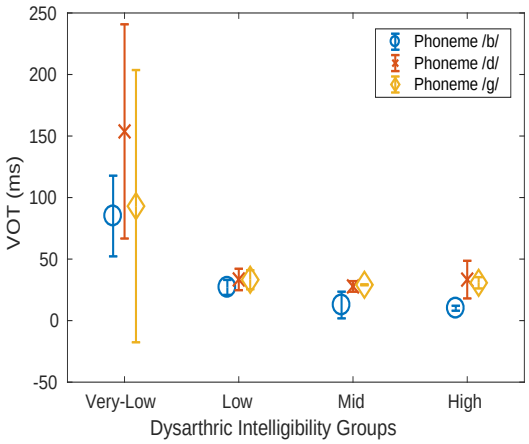

Figure 4.15: Voice Onset Times for the voiced stops /b/, /d/, /g/ across various intelligibility groups

Indeed a similar trend is also evident for the least intelligible (very-low) group of speakers for all the stop consonants as shown in figure 4.14 and 4.15. It is interesting to note that the low intelligibility group manifests a decreasing VOT with escalated deviations for the

voiceless stops. It might be indicative of a lack of muscular coordination as the occlusion moves towards the lips. In both voiceless and voiced cases, the VOT follows a control like pattern as it moves high up in the intelligibility spectrum.

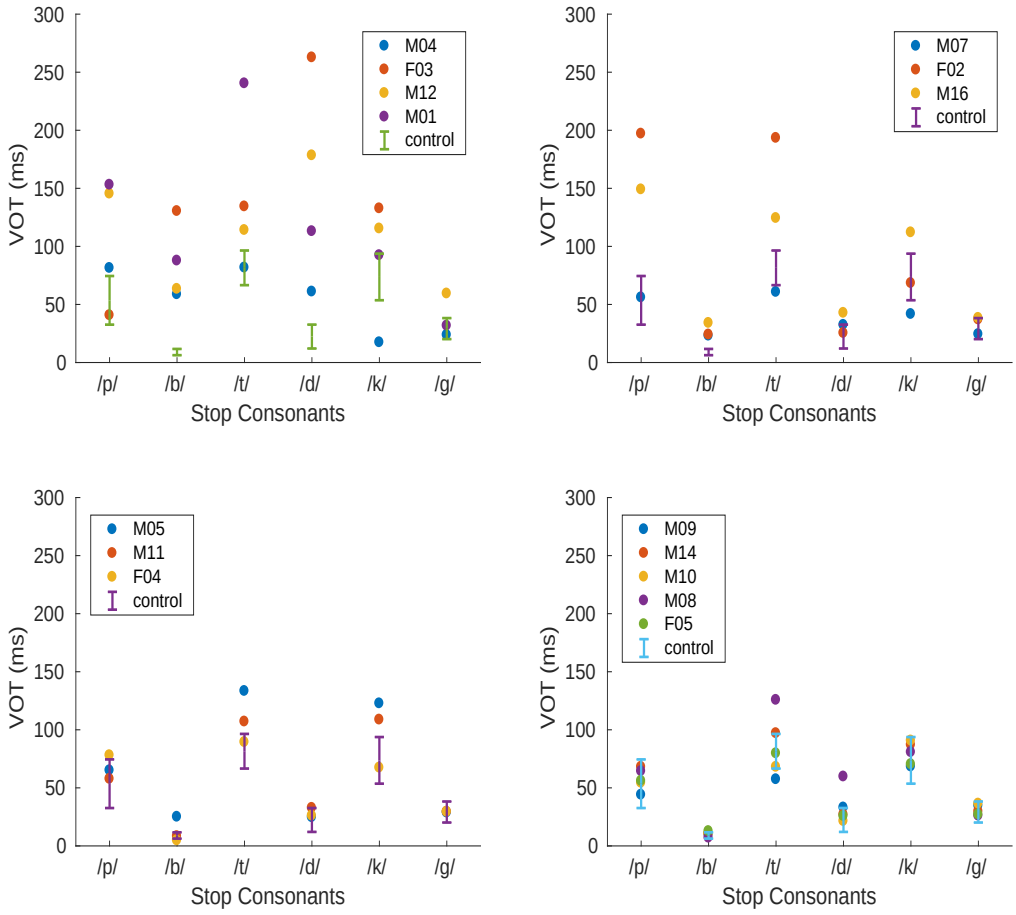

Figure 4.16: Voice Onset Times for various speakers in the UASPEECH database marked against the average control rating. It shows speakers who fall in *(i) very-low (top-left), (ii) low (top-right), (iii) mid (bottom-left) and (iv) high (bottom-right)* intelligibility groups.

In order to be more precise in understanding the relationship between VOT and articulatory movements, it is ideal to analyse VOT's on a per-speaker basis instead of any specific intelligibility groups. Figure 4.16 gives the VOT of individual speakers grouped within

their intelligibility domain. The charts also show the average control VOT values with their standard deviation. Any VOT value that falls outside the control range can be treated as a deviation from normal stop production. Such anomalies can be predicted to be associated with certain articulatory insufficiencies.

It is observed that nearly all the VOT timing differences occur for the very-low and low intelligibility groups for both the voiceless and voiced stops. Most of the mid and high intelligibility group of speakers fall within the standard deviation of the control range except a few outliers.

For example, all the speakers in the very-low category have escalated VOT values for the voiced plosive /b/ and /d/, which might be indicative of an incomplete closure of the lips or impartial alveolar closure of the tongue. As another example, three out of four very-low and two out three low intelligibility speakers exhibit velo-pharyngeal insufficiency by exhibiting deflection on either side of the normal range for the voiceless stop /k/.

It can be concluded that slow speaking rate was found to be a characteristic of speakers with lower intelligibility. They exhibited reduced sypse and thus increased VOT values relative to speakers in the control or higher intelligibility group. The VOT also showed a greater degree of variation for dysarthric speakers for all the observed stop consonants. For speakers with lower intelligibility, the VOT values were also indicative of some underlying physiological insufficiency as the production of the stop consonants moved from lips towards the velum. VOT measurements also exhibited a noteworthy case of phonemic errors, where VOT values of voiceless and voiced stop tokens are reversed in their expected range. For example, when the expected VOT values of a voiceless consonant /p/ falls within the expected VOT range of its voiced counterpart /b/ or vice versa (Morris, 1989). This was noticed for the laryngeal stop pair of /k/-/g/ for the speaker M04 and M07.

The majority of research into VOT of dysarthric speech mentioned in the literature is around qualitative assessment of dysarthric speech, intelligibility prediction and dysarthria classification tasks. However, if such specific timing trends related to VOT can be utilised as an explicit domain of information, it could inform a new approach to durational or pronunciation modelling.

Despite the evidence of temporal dysfluencies in dysarthric speech, no explicit duration modelling will be explored in context of the current thesis. Other areas of acoustic analysis involving frequency and phase information will be studied to introduce methods for improving ASR performance of dysarthric systems.

4.2.2 Spectral analysis

The frequency analysis will focus on the formant analysis for the vowels and diphthongs of control and dysarthric speakers. In particular the first two formants are studied across various speakers and intelligibility groups. A summary of the frequency based analysis, which will be studied in this subsection is shown in figure 4.17.

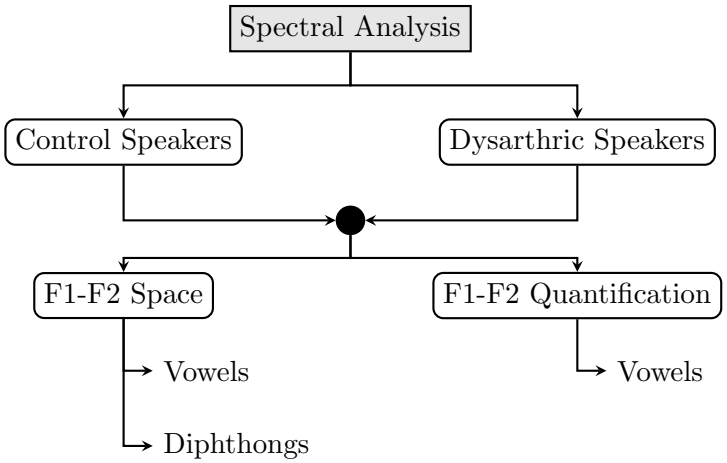

Figure 4.17: Spectral analysis experiments conducted for the UASPEECH database.

4.2.2.1 F1-F2 space

The first analysis was conducted for each group of speakers to study the F1-F2 space for all the vowels and diphthongs. The formant estimation uses all the training data (Blocks B1 + B3) of the UASPEECH corpus for the control and dysarthric speakers. The data is pre-segmented using forced alignment to given an approximate location of the phonetic tokens. The formant trajectories were extracted using the Snack sound toolkit (Sjlander, 2004) that selects the formant frequencies by solving the roots of the linear predictor polynomial. First six formants were extracted for each of the vowels and diphthongs using a 25 ms window and a 10 ms overlap. The order of the LPC analysis was set to 16 for formant extraction. All the other setting were left to default as provided by the `formant` command of the toolkit.

Figure 4.18 and 4.19 shows the visual representation of the F1-F2 space for the vowels and diphthongs. Each data point[1] is averaged across all the control and dysarthric speakers.

[1]Each point in figures 4.18 - 4.22 is averaged across the entire UASPEECH database covering every phonetic context as constrained by the vocabulary.

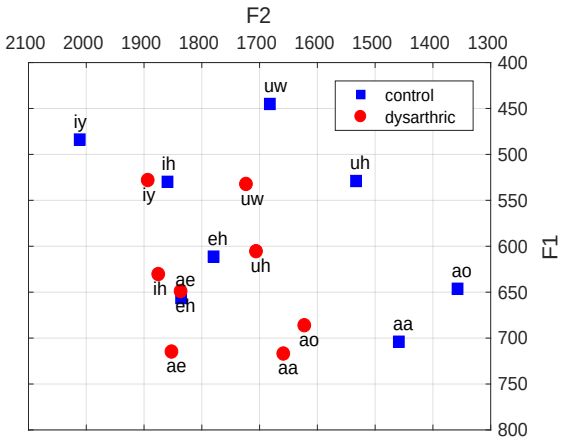

Figure 4.18: F1-F2 plot for vowels of control and dysarthric speakers.

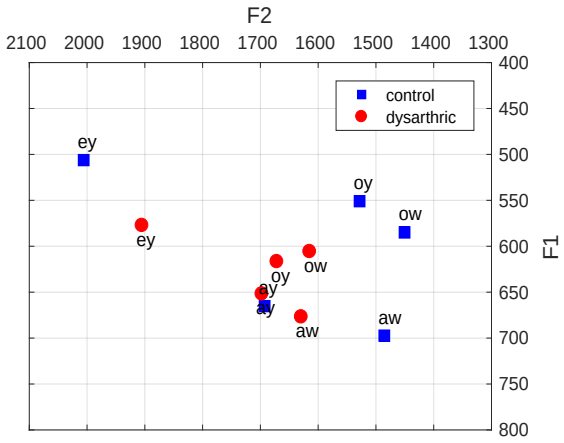

Figure 4.19: F1-F2 plot for diphthongs of control and dysarthric speakers.

It is clearly evident that dysarthric speakers have a reduced F1-F2 space. For dysarthric speech on average, this compression results in an increased F1 for all the vowels and diphthongs (except /ay/ and /aw/) and an increased F2 for all the vowels and diphthongs (except /iy/ and /ey/). As a general rule in the acoustic-phonetic studies, there is an inverse relationship of the first and second formants with tongue height and advancement (Kent et al., 1999b), i.e.,

$$F1 \propto \frac{1}{Tongue\,Height}$$
$$F2 \propto \frac{1}{Tongue\,Advancement}$$

In the F1, F2 findings for UASPEECH, it is observed that on average for dysarthric speakers, there seems to be reduced horizontal movement of the tongue. This can be detrimental for speech recognition tasks as it will reduce the phonatory discrimination between front and back vowels/diphthongs. In order to see the natural variation in the data, figure 4.20 extends the previous plots to show the deviation of the data represented as ellipses. The axes of each ellipse define the variation of the data in the F1 and F2 directions.

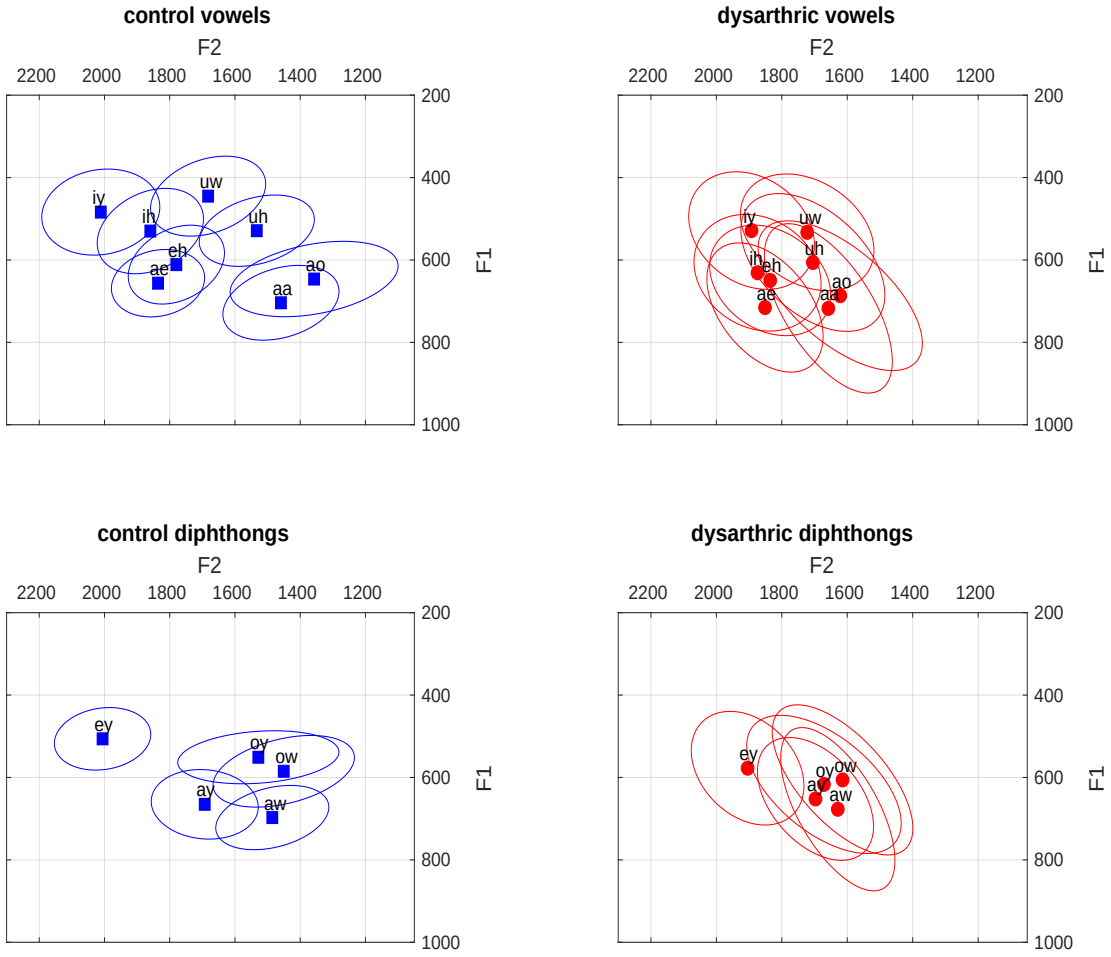

Figure 4.20: Standard deviational ellipses for the control and dysarthric speakers. The top and bottom graphs show the variations for vowels and diphthongs respectively.

It can be seen that the ellipses for the control group are more dense and tend to form segregated clusters. On the other hand, the ellipses for the dysarthric group show greater overlap of sounds. These differences may be due to the higher degree of inter- and intra-speaker variations in dysarthric speech and could be affected by the phonetic context. Further investigation, not attempted here, would be required to determine the true significance of the variations shown in these graphs.

The F1-F2 space analysis was extended across each of the dysarthric intelligibility groups and the results are shown in figure 4.21 for vowels and 4.22 for diphthongs. The high

intelligibility group of speakers show a greater degree of similarity to the control group and the very-low intelligibility group of speakers show the contrary that exhibit a highly skewed mapping of vowels and diphthongs across the F1-F2 plane.

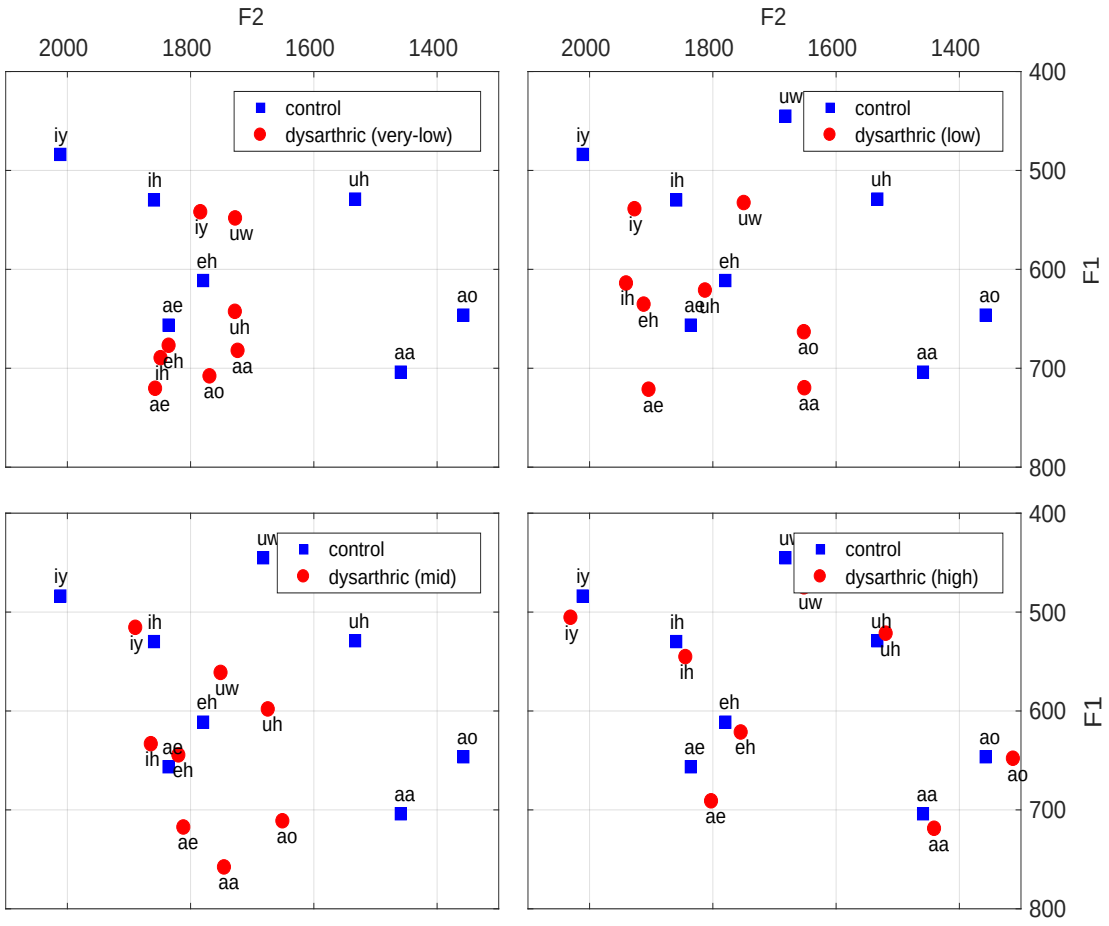

Figure 4.21: F1-F2 plot for the vowels of dysarthric intelligibility groups.

It is indicative that majority of very-low intelligibility speakers appear to have limited range in their tongue movements. This reduces the differences between the phonetic tokens as they tend to converge towards a densely packed cluster of vowels with restricted and overlapping formants. For the low and mid intelligibility group, some discernible patterns for distinguishing amongst various vowels/diphthongs is still visible, albeit, the positioning of vowel/diphthong tokens for the low intelligibility group of speakers is more similar to the control group than the mid group of speakers.

It is emphasised that the F1-F2 space was not analysed under any specific phonetic context and the main aim was to highlight the differences that are observed for various dysarthric intelligibility groups regardless of context.

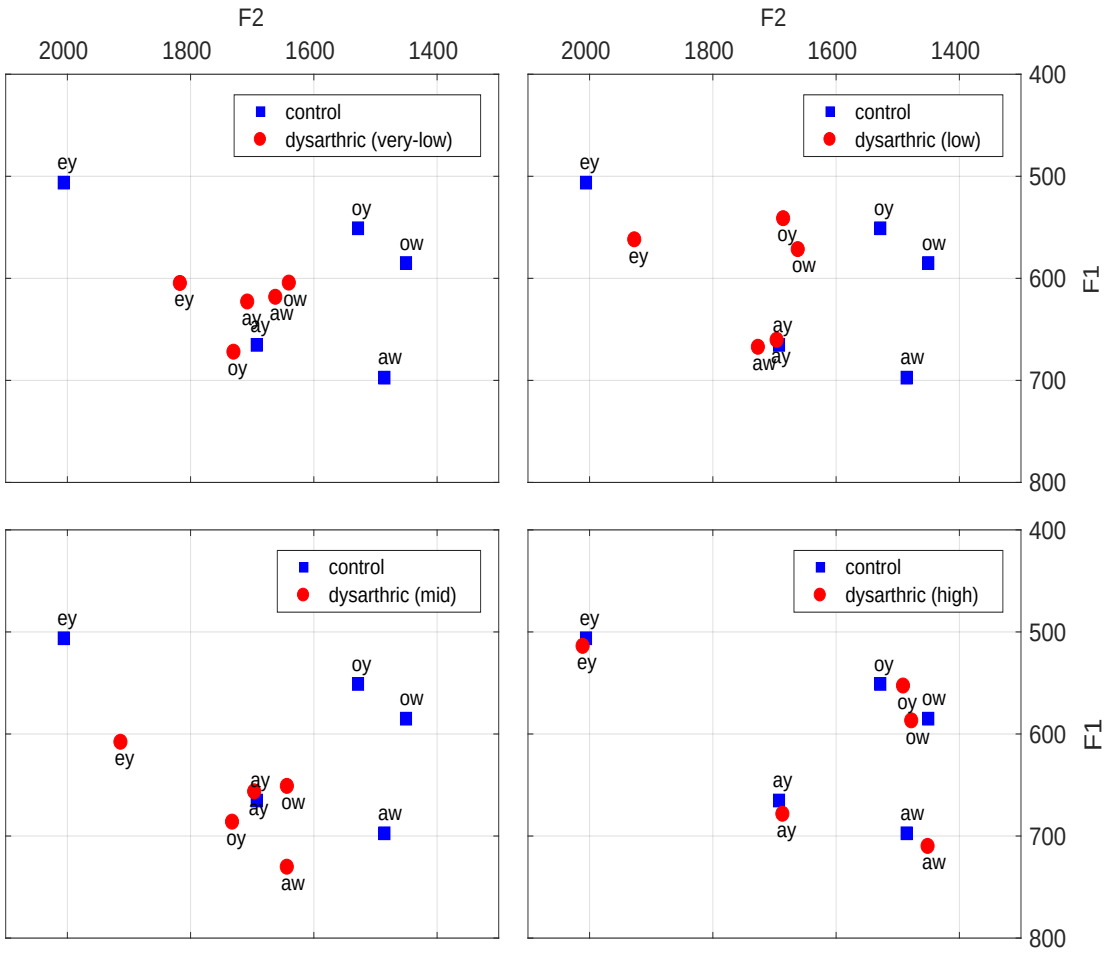

Figure 4.22: F1-F2 plot for the diphthongs of dysarthric intelligibility groups.

The standard devaition ellipses for the various intelligibility groups are shown in appendix D.

4.2.2.2 F1-F2 quantification

This section will look at quantification of some of the vital components of the F1-F2 space. It will study the **Area** and introduce two new measures of **Shape and Displacement** for

analysis of the F1-F2 vowel quadrilateral.

The notion of computing **Area** of the F1-F2 space has already been explored in research on acoustic analysis. It has been a useful quantitative metric in studies on speech intelligibility and to examine acoustic changes due to physiological factors. For example, a reduced F1-F2 space was found to be a dominant characteristic of speakers with dysarthria (Turner, Tjaden, and Weismer, 1995; Weismer et al., 1995), and in another study it was found to be related to the process of ageing (Vorperian and Kent, 2007). The computation of the F1-F2 area in this section is based on the technique described by Turner, Tjaden, and Weismer (1995), which is roughly explained in the following text.

The total **Area** can be quantified by measuring the quadrilateral space bounded by the corner vowels /iy/, /ae/, /aa/ and /uw/. It is computed by splitting the quadrilateral into two triangles (say $\Delta 1$ and $\Delta 2$), whose area is computed individually and summed to get the total area expressed in Hz^2.

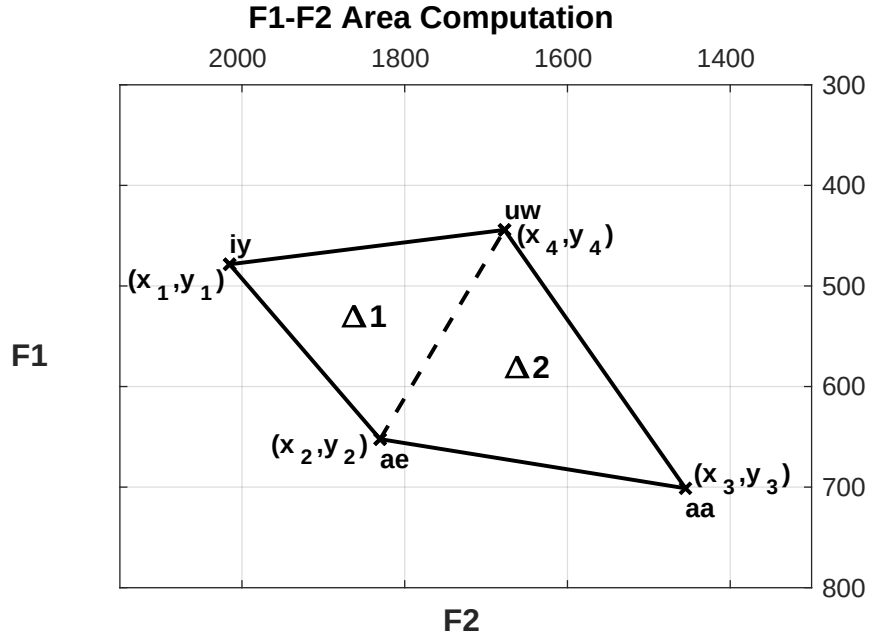

Figure 4.23: F1-F2 area computation by splitting the quadrilateral into two triangles.

Figure 4.23 shows a diagramatic representation of the quadrilateral split into two triangles. Each corner vowel is marked with a coordinate point and the total area is computed using the determinant rule as:

$$Area(Quad.) = Area(\Delta_1) + Area(\Delta_2)$$
$$\frac{1}{2} \cdot \left\{ \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_4 & y_4 & 1 \end{vmatrix} + \begin{vmatrix} x_4 & y_4 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} \right\} \tag{4.4}$$

where $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$ represent the vertices of the corner vowels /iy/, /ae/, /aa/ and /uw/.

In order to have an understanding of F1-F2 area of dysarthric speech, it is referenced with the quadrilateral space for the average control speakers. To measure this effect we introduce the log compression factor (CF), which will be defined as:

$$CF = ln\left(\frac{(F1F2\ \ Area)_{control}}{(F1F2\ \ Area)_{dysarthric}} \right) \tag{4.5}$$

The CF will give a quantitative estimate of the extent to which the vowel discrimination has been reduced. Greater CF indicates higher area of compression. Since the vowel area for dysarthric speech generally tends to be less than that for typical speech, CF is expected to be greater than 0. The analysis in this section is conducted in terms of speaker and intelligibility groups. Figure 4.24 shows the average CF value for various speakers and the intelligibility groups. Any compression factor above zero might be an indication of an atypical vowel space. It can be seen that there is no expected linear relationship between the intelligibility and CF values. For example, speakers can manifest similar CF scores even if they fall at different ends of the intelligibility spectrum or ASR performance (see table 4.3), e.g., M04(very-low)/M11(mid), M12(very-low)/M09(high).

Although CF scores for the expected intelligibility groups (very-low, low, mid) fall above the control CF threshold of zero, it has not come in as a strong indicator to draw any firm conclusion about its association with speech intelligibility and ASR scores. There is also an unexpected CF trend observed within the very-low, low and mid intelligibility groups, which indicates the presence of inter-speaker variations manifest in dysarthric speech. Increased CF scores can be seen as altering the dynamics of the F1-F2 space, but the degree to which this affects the speech from a perceptual or machine processing point of view is still vague.

In addition to computing the CF ratio, the newly introduced notion of **Shape** of F1-F2 space was also examined. This requires to check if the quadrilateral under consideration is

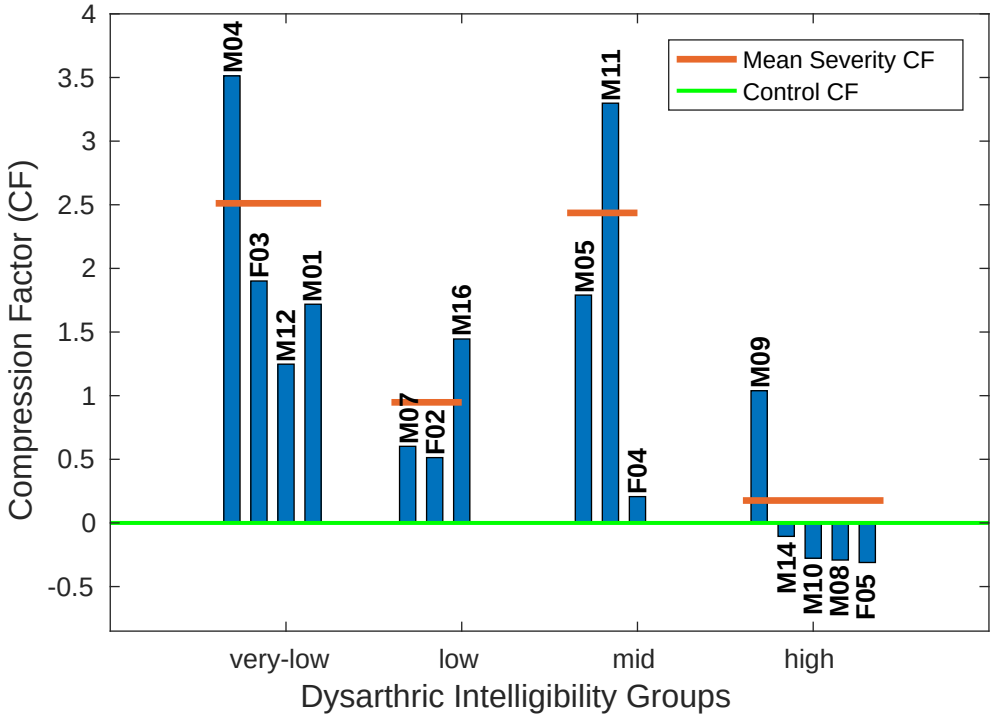

Figure 4.24: F1-F2 area compression factor (CF) for dysarthric speakers under different intelligibility groups. Speakers are ordered according to increasing intelligibility from left to right. The red line shows the average CF for each intelligibility group and the green line is the reference CF for a typical speech area.

convex, concave or flipped in presentation. The way quadrilateral presents itself can give an insight into the placement of vowels in the F1-F2 plane. It can be a useful tool to understand the range of vowel tokens that are easily discernible for an individual dysarthric speaker or intelligibility group, and can be helpful to distinguish confusing or overlapping vowel productions. Figure 4.25 shows the F1-F2 quadrilateral area for an example speaker from each of the intelligibility groups. The speakers are selected in a way so that it shows every possible presentation (*concave, convex, flipped*) of the quadrilateral. For the F1-F2 plot of all the speakers please refer to Appendix D.1.

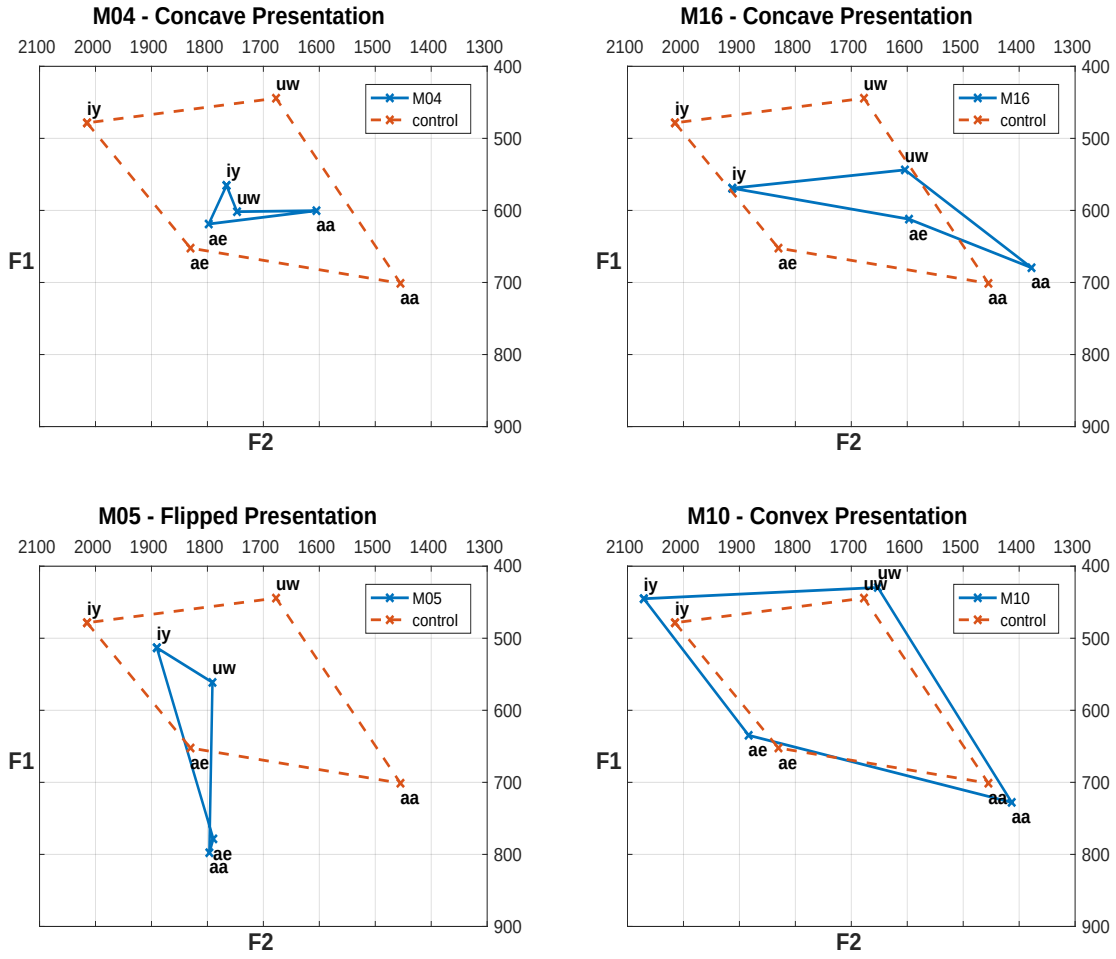

Figure 4.25: F1-F2 vowel quadrilateral for speakers with **very-low** (M04), **low** (M16), **mid** (M05) and **high** (M10) intelligibility. The red polygon represents the average vowel quadrilateral for the control speakers in UASPEECH database.

It can be seen from the above figure that except for high intelligibility speaker M10, the speakers in other intelligibility groups show a reduced F1-F2 space, which generally projects as a skewed map of the vowels, especially for high CF values. This pattern is evident for the very-low, low and mid intelligibility groups, which can be easily inferred from the CF values shown in figure 4.24 and the per speaker F1-F2 vowel space shown in appendix D.1.

The vowel quadrilateral can also give useful cues for understanding the lower number formant distributions. For example, a convex quadrilateral with low CF value as shown for

user M10 has a greater degree of similarity to the F1-F2 space for average typical speech. It indicates that most of the methods for signal processing and modelling for typical speech can prove equally effective for such low severity dysarthric speech since most of the vowel tokens have a well quantified presentation in the F1-F2 plane. Such speech can more or less be regarded as similar to typical speech. However, the other two presentations (*concave, flipped*) do not have a straight forward interpretation. A concave (M04, M16) or a flipped (M05) quadrilateral might indicate unexpected formant frequency for specific vowels. While a concave presentation will usually result in a group of vowels populating the F1-F2 space meant to be occupied by other vowels, the flipped presentation might result in exchanged frequency regions between a pair of vowels. In addition, if the concave and flipped presentation is accompanied by higher CF value, the vowels will tend to overlap with other vowels much quicker than a convex presentation. All this eventually results in reduced phonetic discrimination. The study of CF values along with the shape of vowel quadrilateral can be a useful tool for (i) better design of user dictionaries, (ii) better design of phonetic decision trees for acoustic model clustering and (iii) reducing data sparsity issues by merging vowel tokens with overlapping tendencies.

To complete the investigation of the F1-F2 vowel space, we lastly introduce the measure of **Displacement** for the quadrilateral. This will be defined as the distance between the centres of the dysarthric and control vowel quadrilaterals. The following methodology is used for computing the centre of different quadrilateral presentations:

- **Convex / Flipped:** Centroid of the quadrilateral (C_x, C_y) as

$$(C_x, C_y) = (\sum_{i=1}^{4} \frac{x_i}{4} \ , \ \sum_{i=1}^{4} \frac{y_i}{4}) \tag{4.6}$$

  where (x_i, y_i) are the coordinates of the quadrilateral vertices.

- **Concave** Mid-point between the points where concavity is present.

The distance of each quadrilateral centre is measured against the average control quadrilateral centre. An exhibit for measuring the centre distance for a convex and concave case is shown in figure 4.26.

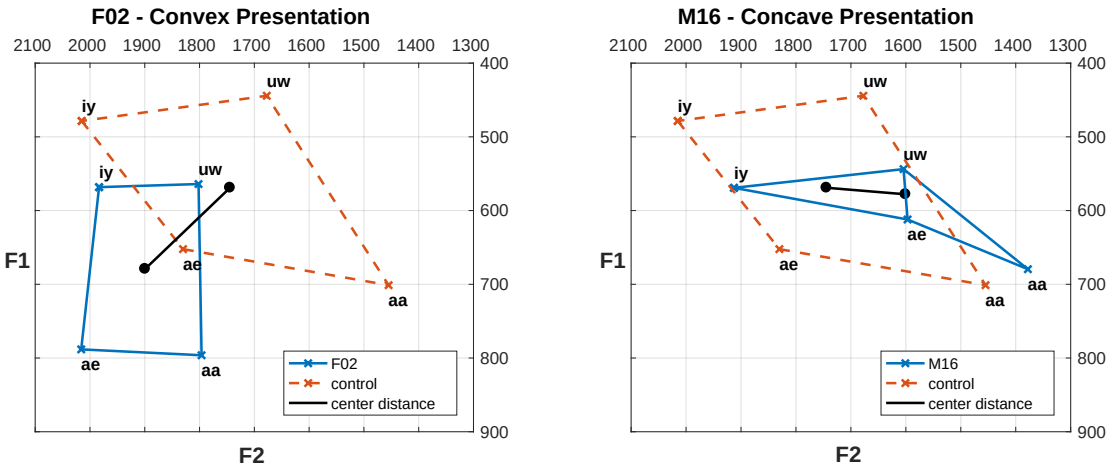

Figure 4.26: An exhibit of the distance measure between two different presentations (*convex, concave*) of the vowel quadrilateral.

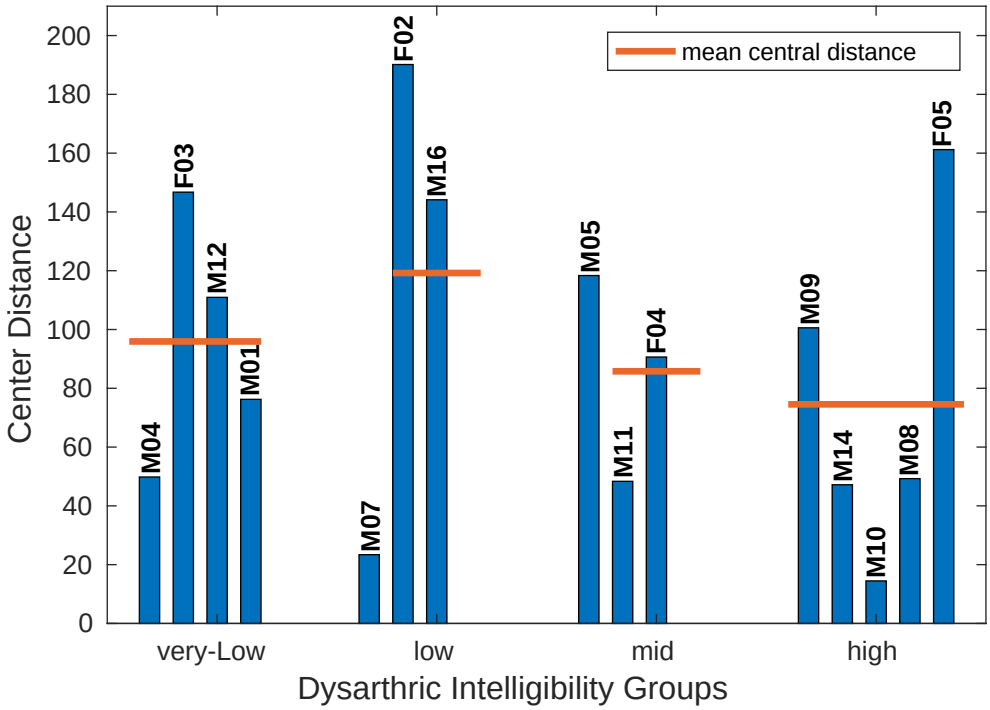

Figure 4.27: Distance between the centres of vowel quadrilaterals of dysarthric speakers and the average control speaker.

The displacement measure gives a quantitative estimate for the positioning of vowel space in the F1-F2 plane. In addition to CF scores and the shape of the quadrilateral space, this added variable can provide useful information about the functional limits for the lower formants. It can be derived from figure 4.27 that there is no predictable pattern emerging from the analysis of UASPEECH speakers. To some extent, the very-low and low intelligibility groups show an increased average quadrilateral displacement than the mid and high intelligibility groups. The F1-F2 quadrilateral displacement can be regarded more as a speaker-wise phenomenon rather than something predictable at an intelligibility level.

4.2.3 Relationship of acoustic analysis with the ASR accuracy

One of the purposes of any acoustic analysis of dysarthric speech is to understand the association between the signal properties and the characteristics of dysarthria. This is already a widely researched topic in the literature as referenced at the start of section 4.2. However, such properties are hardly explored with an aim to enhance ASR performance. Since improvement of dysarthric ASR is a goal of the thesis, this section will attempt to find some functional understanding between all the acoustic parameters examined so far against the best baseline ASR system. For this, the SAT recognition results of table 4.3 are correlated against five quantified acoustic variables, viz, *sypse, voiceless-VOT, voiced-VOT, CF & Displacement*. It should be made clear that all the acoustic variables observed so far are only studied with the intention to quantitatively comprehend some of the underlying dysarthric artefacts and will not be explored any further for improving dysarthric classification or ASR performance in the current thesis.

In this section we will examine the linear relationship between the five acoustic variables and the ASR performance. For this the $Pearson(r)$ coefficient will be computed all throughout. For the correlation analysis, the ideal trend expected between the r and the ASR results is summarised in table 4.6. The analysis was conducted for the ASR scores of each dysarthric speaker and intelligibility group and correlated against the various acoustic parameters.

Figure 4.28 shows the corresponding correlations. The bars marked as red were noted to be signficant at $p < 0.05$. The correlation can be observed both for directionality and strength. In the current study of analysis, directionality plays a much more important role than strength as the expected correlation trend dictates the behaviour of an acoustic variable in relation to the ASR performance.

Variables	Expected Correlation
ASR, sypse	+
ASR, voiceless-VOT	-
ASR, voiced-VOT	-
ASR, CF	-
ASR, Displacement	-

Table 4.6: Expected correlation trend between the ASR score and acoustic variables.

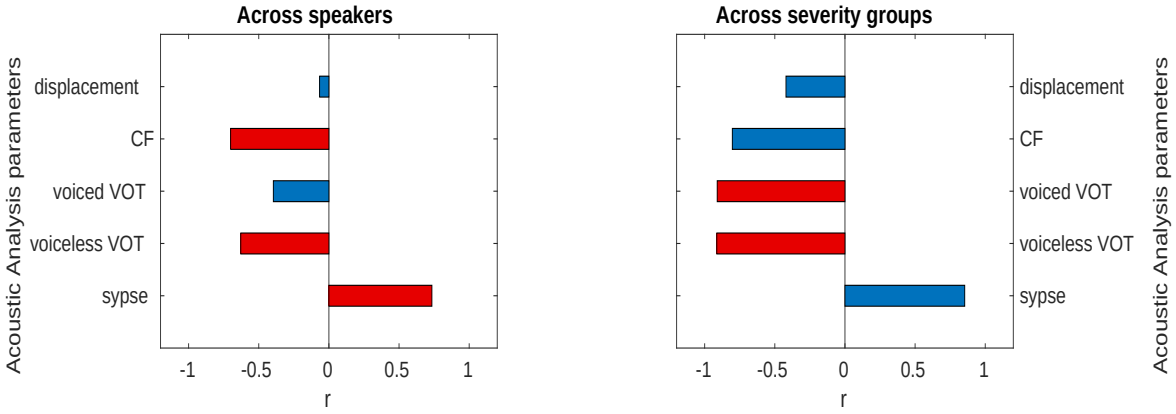

Figure 4.28: Correlation analysis of five acoustic parameters against the SAT-based ASR system. The **left** chart shows the correlation measured across all the speakers and the **right** chart shows the correlation measured across various severity groups.

It can be seen from figure 4.28 that for both the categories, "across speakers" and "across intelligibility", the correlation follows the expected directionality as defined by table 4.6. For example, both CF and voiceless-VOT tend to have a strong negative correlation as the ASR accuracy increases. Similarly sypse tends to have a strong positive correlation, which is expected as the rate of speaking for a dysarthric individual tends to increase with high intelligibility. Lastly, displacement variable was suggestive that the shift of formant space is more likely to be associated with lowest intelligibility and tends to balance out around a typical vowel quadrilateral for high intelligibility speakers.

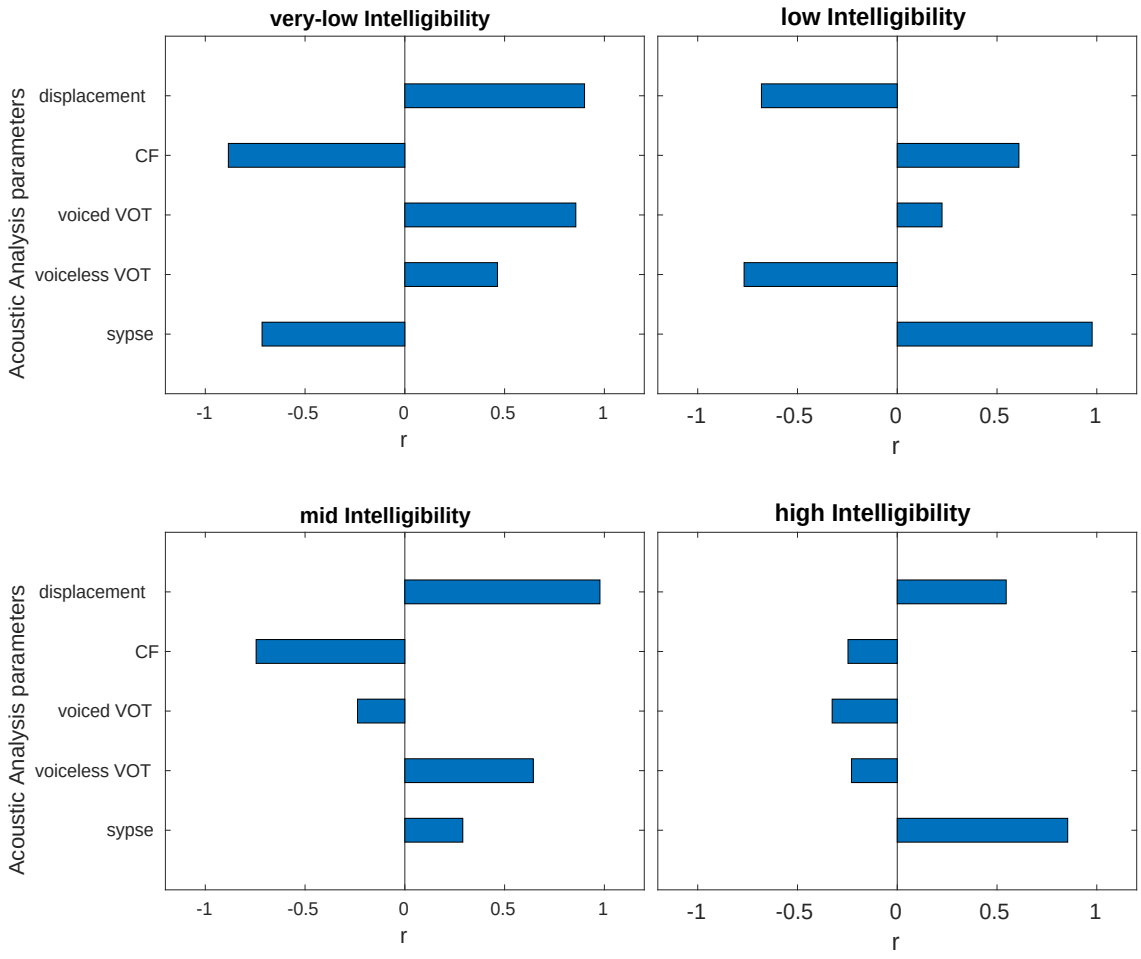

Figure 4.29: Correlation analysis of five acoustic parameters against the SAT-based ASR system for each of the intelligibility groups *very-low, low, mid, high.*

Although it will be inconclusive to report the correlation analysis for each intelligibility group due to speaker sparsity, for the sake of completeness the above analysis was also conducted within each intelligibility group and the results are presented in figure 4.29. The high intelligibility group of speakers follow the correlation trend dictated in table 4.6. There was however an exception for the directionality of displacement parameter, but it did not came up with a high degree of correlation.

On the other hand, there was a range of widespread inter-speaker variabilities observed in the other intelligibility groups. For example, both displacement and VOT showed strong

differences in relation to the ASR scores for the very-low and mid intelligibility group of speakers. The very-low group of speakers also showed a strong association of reduced speaking rate as the ASR accuracy increased within the group and there was a notable presence of malformed CF space within the low intelligibility group of speakers. This shows that for a speaker with dysarthria with reduced intelligibility, it is hard to predict that any one of the acoustic variables might be a major contributor to reduced ASR scores. The unexpected deviations in any of the acoustic variables need to be dealt on a speaker wise basis and there seems to be no methodology at present which can be globally applied to a group of speakers to reduce the negative effect of any acoustic variable for improved ASR performance.

4.2.4 Zeros of the z-Transform (ZZT) analysis for the vowel segments

Previous sections explored some of the standard methods for the analysis of dysarthric speech. This has given some useful insights into irregularities observed in temporal and frequency domains. Earlier research has addressed some of these issues to improve dysarthric ASR, with varying degrees of success. However, the underlying disorder is only implicitly handled by any modelling technique. Thus, one of the challenges in the analysis of dysarthric speech is to discover explicit patterns in the acoustic signal which are directly related to the dysarthric intelligibility, etiology or type. If such patterns can be discovered, this will help researchers design more structured measures to explicitly deal with such speech for improved ASR performance.

This section presents a new approach for looking at disordered speech signals, which will be based on finding the zeros of the z-transformed (ZZT) time-domain vowel segments. The idea for investigating this approach is not to give a comprehensive account of ZZT patterns observed in the UASPEECH database, but it is pursued with the aim of identifying alternate analytical approaches which could be more indicative of the underlying variabilities in dysarthric speech. It will be seen in the following sections and chapters that ZZT patterns of dysarthric speech can form the basis for one such investigative approach, based on the phase component of signal that will be (a) beneficial for robust classification of dysarthric severities and (b) suggest systematic methods for improving overall ASR performance for dysarthric speech.

4.2.4.1 ZZT Analysis of a basic signal: An example

In order to analyse the ZZT patterns of a real speech signal, it is important to understand the basic approach for finding the zeros of the z-transform for any arbitrary time-domain signal. This section will explain the fundamental approach with the aid of a simple example, which can be easily extended to real-time windowed signals of disordered speech. The z-transform of a sequence $x[n]$ is given as

$$\mathcal{Z}\{x[n]\} = \sum_{n=-\infty}^{\infty} x[n]z^{-n} = X(z) \tag{4.7}$$

where z is a complex number represented in polar form as $z = re^{j\theta}$, where r is the magnitude and θ is the phase. For practical analysis the above infinite length sequence is reduced to a finite length discrete time signal of length N. If the samples are represented

as $\{x(0), x(1), x(2), ..., x(N-1)\}$, then the z-transform and its corresponding roots (zero) equation is given by

$$\mathcal{Z}\{x[n]\} = \sum_{n=0}^{N-1} x[n] z^{-n} \tag{4.8}$$

$$Roots\{x(0), x(1), ..., x(N-1)\} = x[0] z^{-(N-1)} \prod_{k=1}^{N-1} (z - z_k), \qquad x[0] \neq 0 \tag{4.9}$$

In the above equation, usually the order of N can be around 400 for a 25ms window sampled at 16 kHz. According to the Abel-Ruffini theorem, there is no algebraic solution to find the roots of a polynomial of degree five or higher (Abel, 1824). Hence numerical methods are generally used to compute roots as these methods are independent of the degree of the polynomial. For the current study the *roots()* function defined in *MATLAB version R2016b* is used for the computation of the polynomial roots in equation 4.9.

Practically the z-transform given in equation 4.7 is useful if the infinite sum is expressable in a closed form by a simple mathematical formula. For analysing discrete time signals, the most important and useful z-transforms are those for which $X(z)$ is a rational expression of the form $P(z)/Q(z)$. The values of z which makes $X(z) = 0$ or $X(z) = \infty$ are defined as zeros and poles of $X(z)$, which are used to plot the Region of Convergence (ROC) in the z-plane for the sequence $x[n]$. Since a Fourier transform is the z-transform computed on a unit circle ($z = e^{j\omega}$), it only converges for the sequence $x[n]$ if the ROC of the z-transform includes the unit circle.

The above theoretical explanation can be summarised with an example. Consider an exponential sequence defined as

$$x[n] = \begin{cases} a^n, & 0 \leq n \leq N-1 \\ 0, & otherwise \end{cases} \tag{4.10}$$

For the above equation we have

$$X(z) = \sum_{n=0}^{N-1} a^n z^{-n} = \frac{1 - (az^{-1})^{-N}}{1 - az^{-1}} = \frac{z^N - a^N}{z^N (z - a)} \tag{4.11}$$

$$\underbrace{z = 0, a}_{\text{Poles}} \qquad \overbrace{z_k = a e^{j2\pi k/N}, k = 0, 1, 2, ...N-1}^{\text{Zeros}}$$

The zero at $k = 0$ cancels with the pole at $z = a$, thus creating a void in the ZP-plane. The void is termed "zero-gap" in the literature. The only other pole is at $z = 0$, which implies that ROC for the above transform is the complex plane $|z| > 0$. For the current analysis we are only interested in plotting the roots of the z-transform and studying the ZZT patterns. For the above exponential the ZZT plot looks like

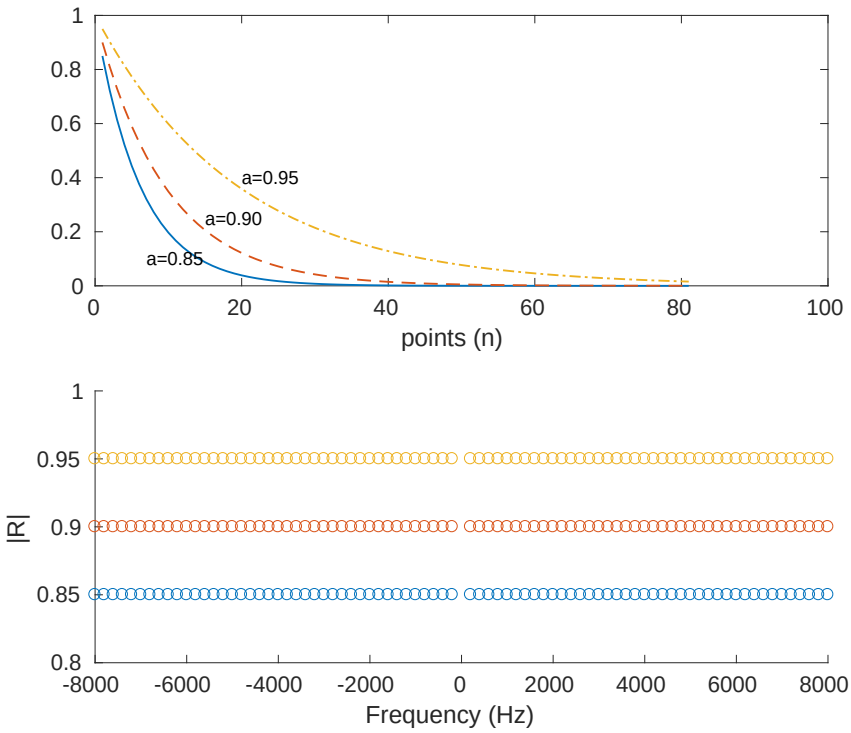

Figure 4.30: ZZT patterns for the exponential function a^n for varying values of a. The (i) top chart is the time domain signal and the (ii) bottom chart is the respective ZZT plot in polar format.

It should be noted that for all the values of a, there is a zero-gap gap created at the 0th frequency bin due to the pole-zero overlap. The idea of plotting the roots of the z-transform can now be easily extended to examine real speech segments, which will be covered in the next sections.

4.2.4.2 Relationship between ZZT, phase and articulation

The ZZT representation is completely characterised by the magnitude and phase of the complex roots. In this section we will explore the effect of phase on the articulation rate and ZZT patterns of a synthetic speech signal.

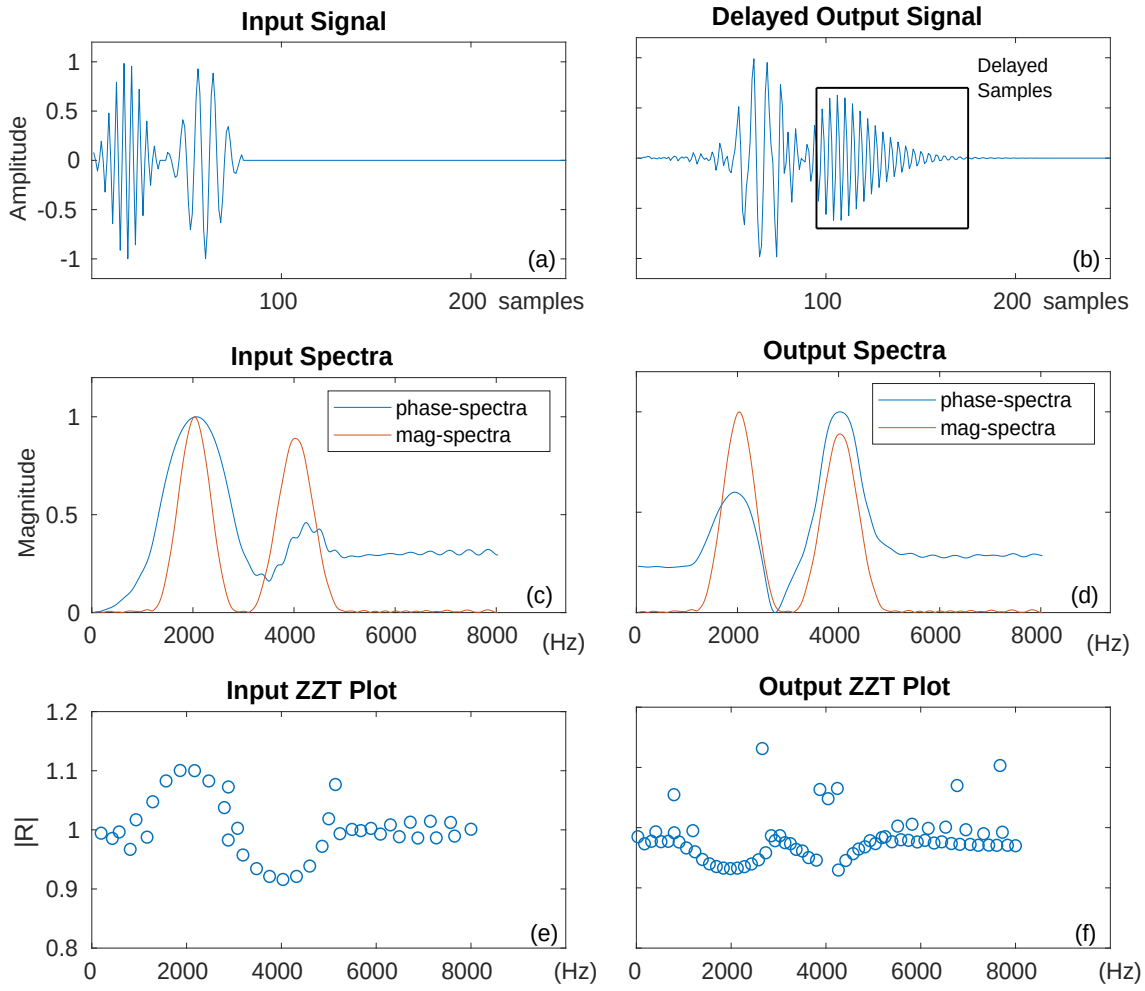

Figure 4.31: The effect of applying phase group delay filter on a time-domain signal. (a) Synthetic time-domain signal comprising two distinct frequencies at 4-kHz and 2-kHz, (b) Effect of applying phase group delay filter on one of the frequency components of the synthetic signal, (c) & (d) represent their magnitude and phase spectra and (e) & (f) represent the ZZT plot for the input and output time-domain signals.

The importance of emphasising phase early in this section is that dysarthric speech is often characterised by slow and imprecise articulatory movements. Therefore, there is a greater chance that its acoustics might show a larger degree of delay in the phase components when compared to typical speech. It will be worth exploring if such delayed phase response has any relationship to the ZZT patterns of dysarthric speech which provides observable evidence for the relationship between phase and articulatory delays.

We start by constructing a synthetic signal comprising of two sinusoidal components with frequencies of say 4-kHz and 2-kHz respectively. The different sinusoids are then concatenated one after another in time domain as depicted in part (a) of figure 4.31. The first envelope reflects the 4-kHz signal and the later envelope pertains to the 2-kHz sinusoid followed by silence.

The effect of phase delays are shown by simulating an arbitrary group delay filter, which is an all pass filter generally used for correcting phase distortions. For this we have used the *fdesign.arbgrpdelay()* function of *MATLAB version R2016b*. The order of the filter is set to 20 and the sampling frequency is assumed to be 16 kHz. We intentionally set the filters phase group delay at 6.25 milliseconds on the higher frequency component of 4-kHz and observe its effect. The application of the filter on the input signal produces the time-domain signal represented by part (b) of figure 4.31. It can be easily seen that due to the phase delay, the higher frequency component has moved back in time. It can also be seen that if a particular frequency component has a delay when passed through a filter system, then one of the causes might be related to phase group delays. This is exactly what we had speculated earlier about slow articulation and phase of dysarthric speech, where these delays might be greater than that of typical speech.

There were some other noteworthy observations from this simulation, which are highlighted below:

- The magnitude spectrum has always been the preferred part of the Fourier output and is widely used in most of the feature representation (MFCC, PLP etc.) of disordered speech. It can be seen that parts (c) and (d) of figure 4.31 shows the magnitude and phase spectrums for the input and output signal with high degree of correspondence. Both the representations show peaks at the expected frequencies of 4-kHz and 2-kHz. It is also observed in the later parts of this section that ZZT analysis conducted for dysarthric vowel segments gives suggestive evidence about the importance of phase in analysing dysarthric speech. This motivates us to explore if phase-based feature

representations of disordered speech are any better for representing dysarthric variabilities than the magnitude based spectrum. Later chapters in the thesis will explore this aspect from dysarthric ASR point of view and study the effect of different spectral representations.

- Lastly we will try to interpret the outcome of ZZT plots in parts (e) and (f) of figure 4.31 due to the application of phase group delay filter on the input signal. For this, it is important to understand the influence of poles and zeros on the frequency response of any system. We can recall that the frequency response magnitude of a system is represented as:

$$|X(z)| = \frac{\sum_{k=0}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}} \tag{4.12}$$

Since the stability of $X(z)$ relies on ROC to contain the unit circle at $|z| = 1$, the above equation is substituted with the complex exponential $z = e^{j\omega}$, which gives the frequency response magnitude in pole-zero format as:

$$\begin{aligned} |X(e^{j}\omega)| &= \frac{|b_0|}{|a_0|} \quad \frac{\prod_{k=1}^{M} |1 - c_k e^{-jw}|}{\prod_{k=1}^{M} |1 - d_k e^{-jw}|} \\ &= \frac{|b_0|}{|a_0|} \quad \frac{\prod_{k=1}^{M} |e^{jw} - c_k|}{\prod_{k=1}^{M} |e^{jw} - d_k|} \end{aligned} \tag{4.13}$$

In the above equation c_k and d_k represent the zeros and poles of the system and $|e^{jw} - d_k|, |e^{jw} - c_k|$ represent the distance of a particular frequency bin on the unit circle from the respective zero and pole. Equation 4.13 can be put in a simplified form as

$$|X(e^{j}\omega)| = \frac{|b_0|}{|a_0|} \quad \frac{\text{"}distance\ of\ e^{j\omega}\ from\ zeros\text{"}}{\text{"}distance\ of\ e^{j\omega}\ from\ poles\text{"}} \tag{4.14}$$

It is now easier to interpret from the above representation that poles near to the unit circle push the frequency response high and zeros near to the unit circle push the frequency response low. In lieu of the above explanation, we can observe from parts (e) and (f) of figure 4.31 that more zeros are present near to the unit circle of the output

signal due to application of the phase group delay on the input signal. This inadvertently introduces a limited passband, which might inhibit the full frequency response of the system. Hence, for dysarthric speech where phase delays might be influenced by slow articulation and other physiological insufficiencies, the phase distortions need a closer examination. This thesis will address one such instance of phase deviation in the next chapter and show its relationship to the intelligibility of dysarthria and how the knowledge can be utilised to improve ASR performance.

The remainder of this section will expand the above discussion in examining the ZZT patterns for disordered speech and corroborate the importance of phase related events in the analysis of dysarthric signals.

4.2.4.3 ZZT analysis of a typical vowel segment

In order to examine the ZZT patterns for dysarthric speech, the process is first studied for a typical speech token. It will enable us to draw a reference of comparison for distinguishing between ZZT plots of varying dysarthric intelligibility.

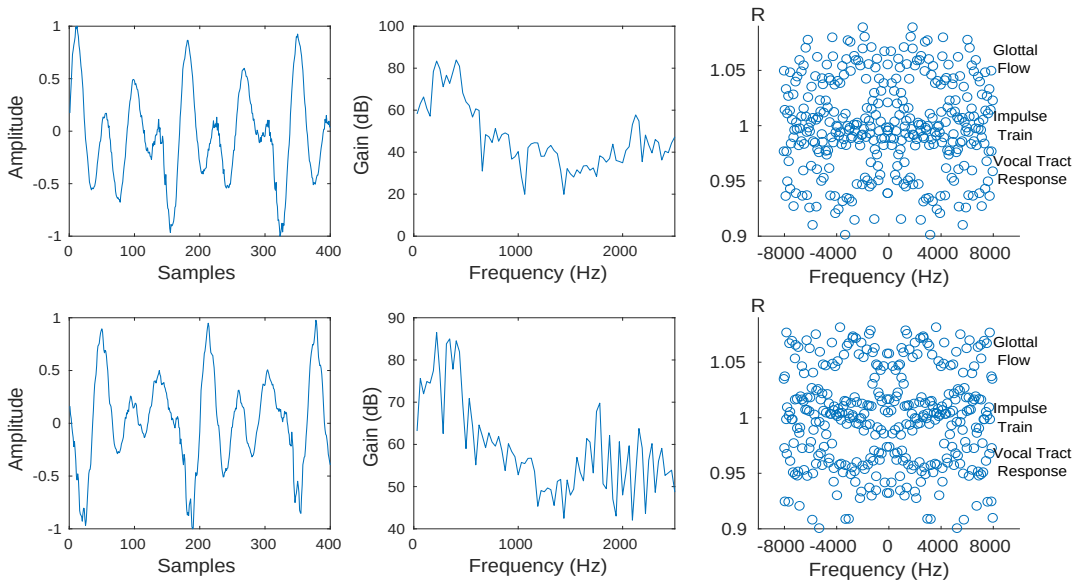

Figure 4.32: Analysis for the vowel /iy/ (top) and /uw/ (bottom) for a control speaker. It shows the **(left)** waveform representation for a 25 ms vowel segment, **(middle)** magnitude spectrum of a hanning-poisson windowed signal and **(right)** the ZZT plot.

Figure 4.32 shows the waveform, magnitude spectrum and the ZZT pattern for the manually selected vowel segments /iy/ and /uw/ for a control speaker. Since windowing is an essential step for acoustic analysis of any real-life speech data, a variety of window functions were tried out (not reported) and we have selected the **Hanning-Poisson** window for the computation of zero patterns as it gave the best resolution, in accordance with the expected theoretical output of the ZZT plot. When a ZZT distribution is plotted for the time domain convolved signal, it exhibits a butterfly like pattern, which is clearly demarcated into three distinctly visible areas on the z-plane.

A detailed explanation and interpretation of such a kind of ZZT analysis on synthetic and real speech signals can be found in the work carried out by Bozkurt, Couvreur, and Dutoit (2007). The authors presented that the the zeros above the unit circle pertain to the glottal pulse of the signal, the zeros below the unit circle pertain to the vocal tract filter response and the zeros along the approximate line of the unit circle correspond to the impulse train zeros. As explained earlier, a ZZT plot also produces void gaps along the horizontal axis where the poles and zeros coincide. It was presented by Bozkurt, Couvreur, and Dutoit (2007) that the presence of such void gaps is indicative of spectral dips on the spectrum that gives rise to the harmonics for the impulse train area, formant presence for the vocal-tract area, and for the glottal flow void gaps there is no clear understanding in the literature. Lastly, the authors also put emphasis on choosing particular window functions, like Gaussian or Hanning-Poisson and the centring of such analysis window at the glottal closure instant for getting a ZZT plot that closely matches the theoretical expectation.

4.2.4.4 ZZT analysis of a dysarthric vowel segment

In order to examine the ZZT patterns for dysarthric speech, one speaker from each of the intelligibility groups is selected. The ZZT analysis for each speaker is then conducted for the front-high vowel /iy/ in the production of the word **be**. The vowel segments are manually selected for each speaker's production, ensuring that the selected speech chunk is aligned at the glottal closure instants. As mentioned earlier, the aim of this section is not to conduct a detailed ZZT analysis of disordered speech for every vowel token, but rather explore any distinguishable cues in the z-plane, which might be representative of an underlying dysarthric variability or motivate novel ways to look at speech disorders. Figure 4.33 exhibits some noticeable observations and gives further points for consideration.

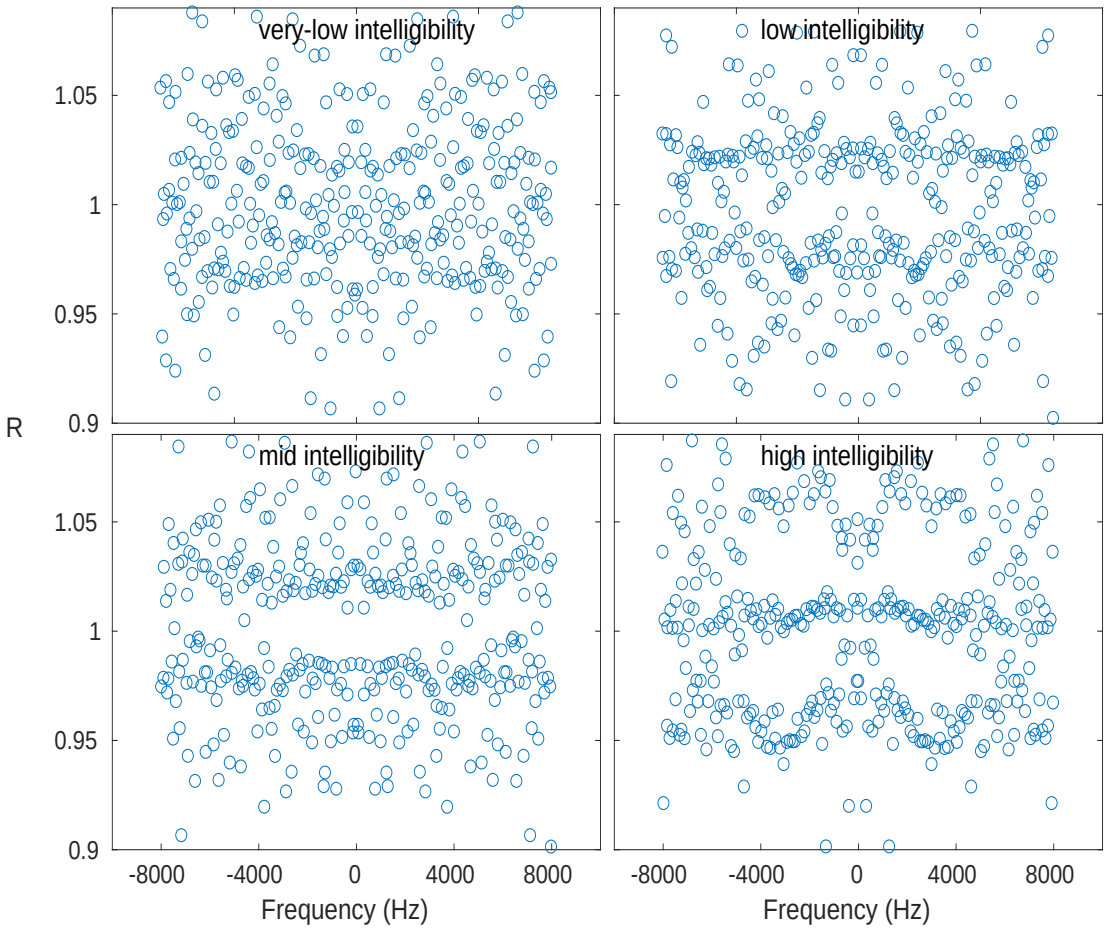

Figure 4.33: ZZT patterns for the front-high vowel /iy/ under various intelligibility groups.

Distribution of Zeros

The ZZT plot of figure 4.33 exhibits a distinguishable pattern for different levels of dysarthric intelligibility. The expected distribution of the roots for the mildly dysarthric and high intelligibility speakers closely resembles the control speaker representation shown in figure 4.32, i.e., it has a much clearer delineation between the zero regions related to the glottal pulse, impulse train and the vocal tract filter response. However, this distinction is skewed for severe group of speakers as shown in the upper half of figure 4.33. One speculation that can be drawn from such observations is the basis for establishing a link between articulatory, velo-pharyngeal and glottal insufficiencies manifest in dysarthric speech with the

ZZT distribution of the signal. For example, the ZZT plot of /iy/ in context with the stop consonant /b/ for the speaker with very-low intelligibility (top-left of figure 4.33) shows a skewed mapping of zeros, where the delineation between source and filter is indistinguishable. Hence it might give an unexpected region of zero-gaps near the circle below unity, which is not the expected formant structure for /iy/. This might be happening due to any physiological factor such as poor VOT due to bilabial insufficiency or poor formant structure due to a compromised filter system. However, all this is guesswork and there is no study in the literature which can give a systematic relationship between the ZZT patterns and articulatory or glottal events, albeit it is an area worth investigating for future research. This thesis will not look into these relational aspects and will focus on decoding the ZZT patterns from an ASR performance viewpoint.

ZZT Patterns of another vowel

The front-high vowel /iy/ was examined in figure 4.33 that gives us some expectation about how the ZZT patterns emerge for dysarthric speakers of varying intelligibility. It is practically infeasible to analyse every vowel token and speaker of the UASPEECH database for establishing a more robust understanding of the ZZT patterns of disordered speech. However, in order to establish a more convincing argument that the ZZT patterns of figure 4.33 is not an artefact for any particular vowel token, we extend the same experiment for the back-high vowel /uw/ with same speakers, spoken in context of the word **do**.

It can be seen that the ZZT patterns for the vowel /uw/ in figure 4.34 share a high degree of resemblance to the ZZT patterns of /iy/ in figure 4.33. It shows a discernible pattern of zeros for the high intelligible speakers and a more skewed distribution for speakers with lower intelligibility. It can thus be induced with a reasonable degree of confidence that such scatter of zero patterns is not by chance and it might be linked to the intelligibility of the disorder.

Theoretical Expectation for the ZZT distribution

The ZZT plot is completely characterised by the magnitude and phase of the complex number. From a speech processing perspective, we have less control over the region of convergence (magnitude), as speech is a mixed phase signal. This suggests that there might be some phase related acoustic events happening within severely disordered speech signals

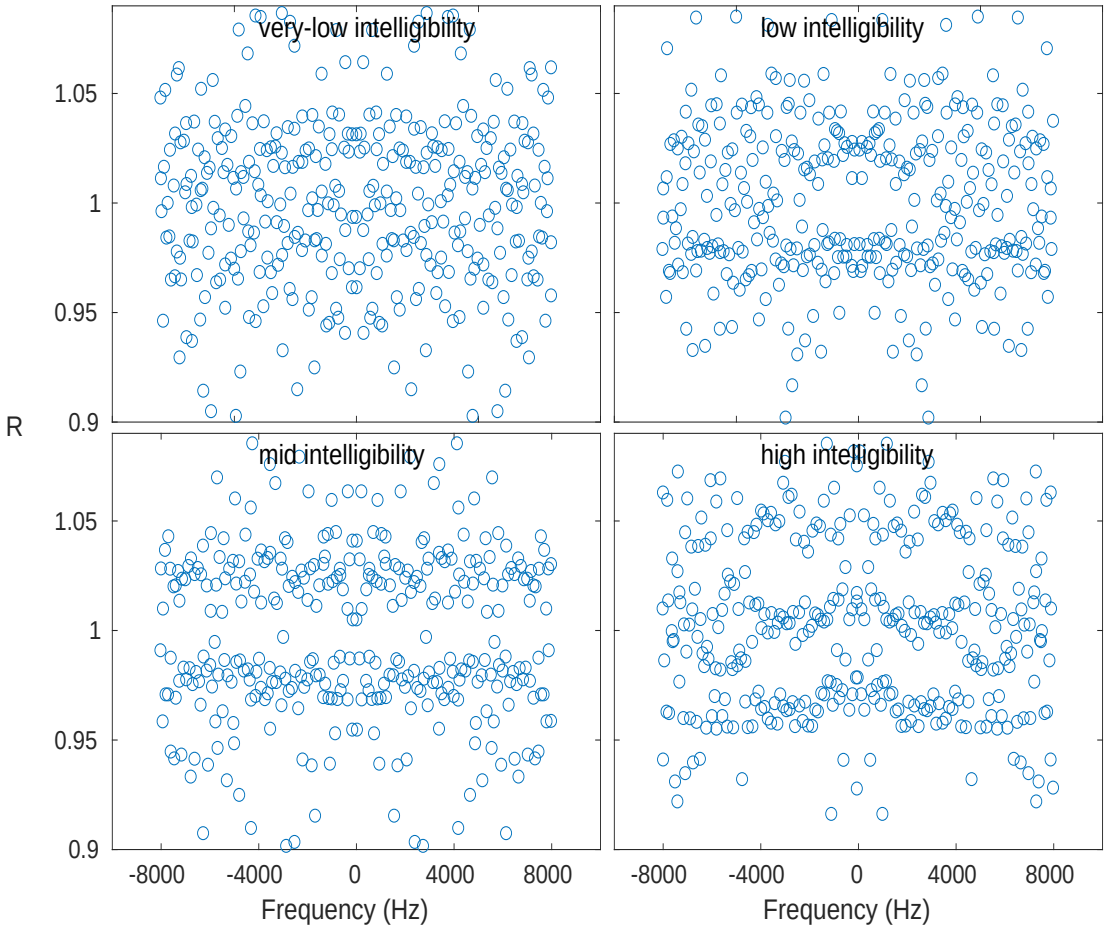

Figure 4.34: ZZT patterns for the back-high vowel /uw/ under various intelligibility groups.

that are producing a skewed map of the zeros in the z-plane. This might be one of the reasons for the poor recognition and perceptual understanding of speech signals with high degree of severity. This motivates us to think that if such phase distortions can be systematically manipulated in a controlled fashion to produce the expected ZZT distribution, it can be a useful step to reduce variabilities in dysarthric speech and produce a clear demarcation between source and filter components. As an example, consider the ZZT plot of speakers with very-low and high intelligibility for /iy/ and /uw/. It is the top-left & bottom-right charts in the figures 4.33 and 4.34. If we plot the unwrapped phase component of its complex roots and compare it against the control speaker, the difference is easy to visualise in figure 4.35. The plot is shown for the four corner vowels and is averaged across all the

utterances of the UAPSPEECH database for the examined speakers. The vowel segment is automatically extracted by taking a 25 ms frame from the centre of the vowel, whose location is derived by the process of forced alignment. It must be noted that the default ordering defined for the *roots()* function is used for each plot in figure 4.35. The phase slopes for the very-low and high intelligibility speakers seem to suggest a huge operational range within which the ZZT patterns for a variety of dysarthric speakers might fall.

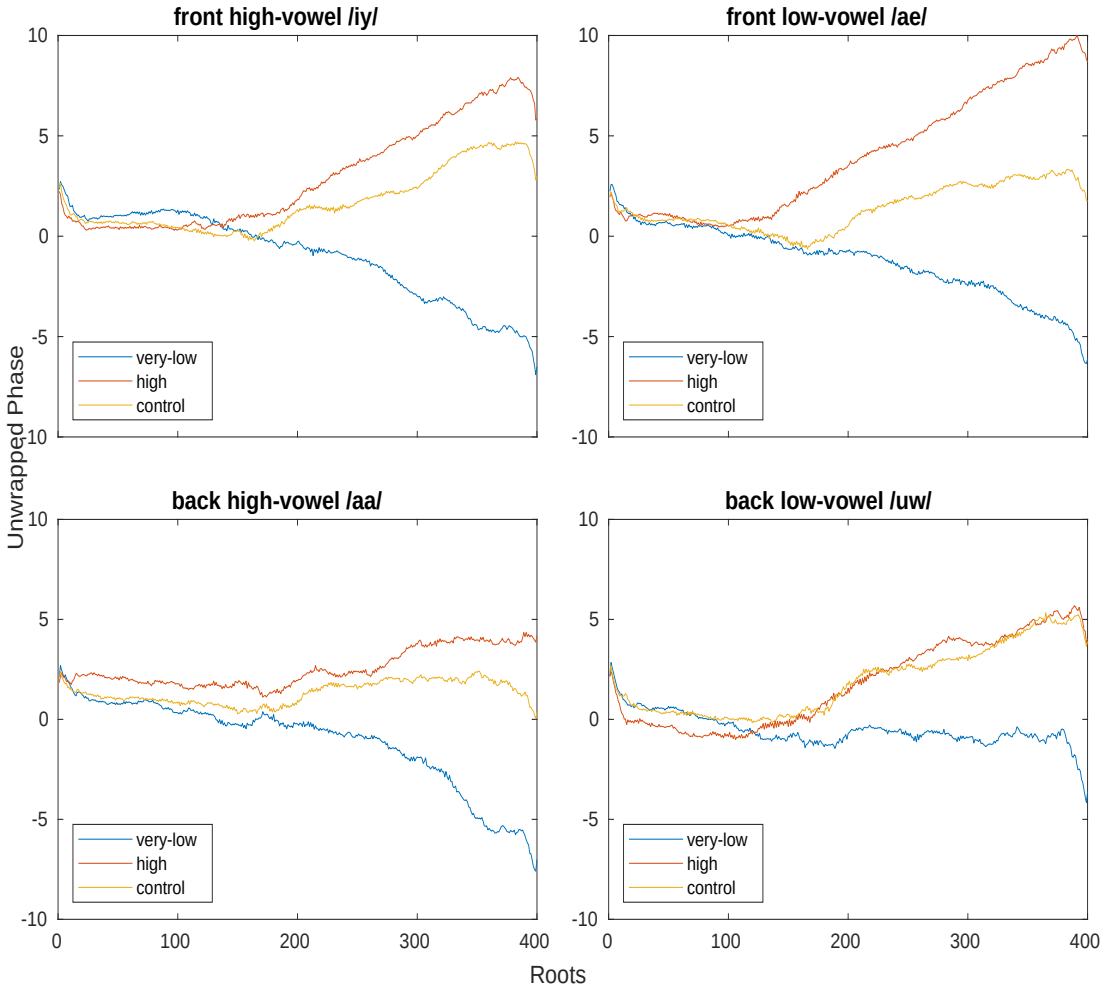

Figure 4.35: Plots for the unwrapped phase of the complex roots of the four corner vowels for a control speaker and dysarthric speakers with very-low and high intelligibility. The plot is averaged across all the utterances in the UASPEECH database for the examined speakers.

This range is by no means a generalisation for such phase deviations manifest in the dysarthric signal, but it rather seems to suggest some relationship between dysarthric speech signals and phase based acoustic events. This re-emphasises the notion that phase-based analysis might act as an important tool in understanding some of the speech aberrations, which are manifest in disordered speech with high degree of severity. We believe that reducing the aberrations observed in the ZZT distribution for speech with high degree of severity might not be useful from a perceptual point of view, as both magnitude and phase play a collectively important role in synthesizing intelligible speech. However, such phase-based corrections might act as a beneficial step for reducing some hidden artefacts in disordered speech. Such signals can act as better representatives of data that is easier to model by speech decoding algorithms.

Although the ZZT observations discussed earlier have given us useful insights into the importance of phase-based analysis in understanding some aspects of the disordered speech signals, the analysis will not be further explored in the Z domain in the remainder of the thesis. This is due to three main reasons. Firstly, since it is still one of the scantily explored areas in the field of dysarthric ASR, it is believed that more research is needed to further consolidate our observations and it is beyond the scope of this thesis to look into those aspects. There is hardly any work in the literature which looks at phase aspects of the dysarthric signal and if they bear any correlation with severity, etiology or type of dysarthria. Secondly, as discussed earlier, the plots shown in figure 4.35 are dependent on the internal ordering of the complex roots as dictated by MATLAB. It is not very clear at the moment what that internal ordering is and further research would be needed to understand the effect of the observed deviations in regard to the root order.

Lastly, although the averaging effect of the phase deviations observed in figure 4.35 gives a compelling picture of some underlying phase-based artefact that can be present in dysarthric speech of varying intelligibilities, it can be influenced by the alignment of the examined speech segment at the glottal closure instant. As an example, figure 4.36 shows a portion of the vowel segment that was manually extracted for one of the speakers with very-low intelligibility. The blue dotted area shows the approximate segment that is manually aligned at the glottal closure instant and the red dotted area shows a misaligned shift. It can be seen that the phase slope deviation is affected by the improperly aligned vowel segment at the glottal closure instant.

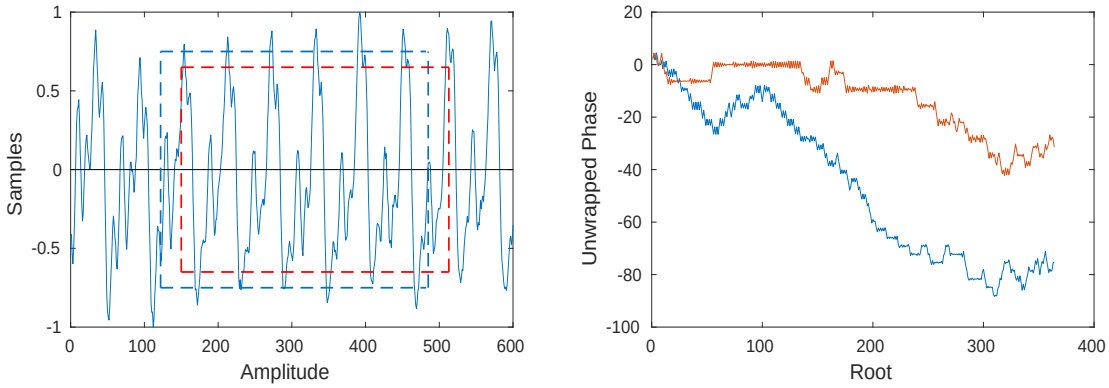

Figure 4.36: Effect of the incorrect glottal closure alignment on the unwrapped phase.

In future research, robust automatic methods for the alignment at the glottal closure instant could be explored to understand the effect of such deviations with a greater degree of accuracy, which could direct towards a more informed understanding of such a phenomenon.

4.2.5 Summary of the acoustic analysis

In section 4.2 of this chapter, an acoustic analysis was conducted on the UASPEECH database with the intent of finding an association between the acoustic variable(s) and the underlying dysarthric intelligibility, type or etiology. The initial investigation was conducted using several acoustic variables. The temporal disruptions were studied using measures like syllables per second (**sypse**) and VOT of voiceless/voiced stop consonants and the frequency aspect of the dysarthric speech was examined using qualitative and quantitative estimates in the F1-F2 plane. The investigation has given a better insight into the understanding of dysarthric speech. Many of the findings were coincidental with other similar studies on dysarthria and some new quantitative measures were also introduced.

The sypse measure showed a direct relationship between slow speaking rate and the underlying dysarthric intelligibility. The dysarthric speech in general was measured at more than 1.5 times slower than typical speech. It was observed that for the very-low intelligibility, syllable production was about twice as slow as the control group and the high intelligibility group was closer to typical speaking speed. The relationship between the syllable production rate and intelligibility were analogous to some other similar studies reported on dysarthria in general (Kent et al., 2000) and on other types of dysarthria (Blaney

and Wilson, 2000; Chenausky, MacAuslan, and Goldhor, 2011). However, the analysis also showed that the sypse trend does not hold between speakers for the lower intelligibility group, which is indicative of inter-speaker variabilities manifest in the acoustics of severely disordered speech. This might be one of the reasons that SAT based systems were more effective at modelling lower intelligibility group of speakers and high intelligibility speakers gave the best performance using SI-03 system prepared from typical speech with same vocabulary as the target test speakers.

The VOT measure also tends to exhibit escalated mean values for both voiced and voiceless plosives as the intelligibility decreases. Dysarthric speech had around 2.5 times more variability across the VOT estimates of all the stop consonants than the control group of speakers. Our findings are also in agreement with earlier research on dysarthria and VOT. For example Kent and Kim (2003a) reported a summary of key VOT studies that highlighted the link between longer VOT durations and dysarthric intelligibility. In addition, VOT can also be used as a useful measure for detecting phonemic and phonetic errors (Morris, 1989). The VOT analysis on UASPEECH exhibited the presence of such phonemic errors in the lower intelligibility group of speakers. An understanding of such phonemic distortions can be very important in delineating confusing phoneme pairs to build a customised user dictionary that will be more appropriate for particular dysarthric speakers. It can also aid in building robust acoustic models, for example, in the context of HMM-GMM systems, better phonetic decision trees can be constructed for optimal state clustering.

If the VOT and sypse measures were taken across various intelligibility groups, then the broad temporal rule that was evident is described as:

$$sypse \propto Intelligibility$$
$$VOT \propto \frac{1}{Intelligibility}$$

The above rule depicted by the sypse measure was also evident in the evaluation of F1-F2 space. The highly skewed plotting of the vowel quadrilateral in the F1-F2 plane was indicative of flaccid and limited range of tongue movement. The thesis improvised on the definition of F1-F2-(**area**) and introduced new measures of F1-F2-(**shape, displacement**). These measures were helpful to explicate a quantitative understanding of the F1-F2 plane and its relation to the underlying dysarthric intelligibility.

The first measure of F1-F2 area was computed as a compression factor (CF), which is the logarithmic ratio of the control and dysarthric formant spaces. It showed a better understanding of inter and intra speaker variabilities manifest in the dysarthric speech and showed no specific association with the perceptual correlate of intelligibility. A higher CF value was usually indicative of decreased intelligibility and skewed mapping of the vowel tokens. The second measure of shape gave a geometrical interpretation of the F1-F2 plot, which was either *convex, concave or flipped* in its presentation. In particular the concave/flipped arrangement was indicative of overlapping or confusing vowel tokens with reduced phonatory discrimination. Lastly, the F1-F2 displacement was also evaluated as a quantitative estimate that measured the displacement of dysarthric vowel space relative to typical. Any displacement observed above the mean central distance of an intelligibility group might be an indication of a notable formant shift for a particular speaker. Such displacement might suggest towards designing explicit filters or specific FFT parametrisation to capture the dynamics of the speech. Since this field of study is not the prime focus of the thesis, this topic will not be researched any further.

Although the area, shape and displacement are useful measures that highlight some underlying dysarthric phenomena, a more robust understanding can be attained by combining these variables together. For example, a concave F1-F2 space with high CF and displacement value is most likely indicative of a severe dysarthric speaker who exhibits compromised phonatory discrimination and has unbounded F1-F2 region. As another specific example, it was observed in figure 4.21 that the high intelligibility group shows a slight but consistent shift towards higher F1 and lower F2 that might affect the displacement factor. However, for the same group of speakers it was also observed that the CF factor remains very low indicating towards a well-defined F1-F2 region with non-overlapping phonetic tokens. It is shown in an earlier study that reduced intelligibility is more akin to overlaps of the corner vowels (Kim, Hasegawa-Johnson, and Perlman, 2011). In context of our analysis, the quantitative measures of shape and area can be jointly exploited to understand its relationship to intelligibility. As a hypothetical example, one can make a reasonable estimate that a flipped displacement of F1-F2 space with an increased CF value might indicate reduced intelligibility.

Another encouraging result was the outcome of the correlation analysis between the key acoustic variables investigated in this study and the corresponding ASR performance. It was noted that in most cases the results conformed to the expected directionality and

strength of the correlation measure across various speakers and intelligibility groups. This is promising as it not only indicates that such acoustic variables are useful from an analytical perspective, but it can be explored further to devise measures for improving ASR performance. For example, when observed across speakers, displacement did not come up as a strong correlating factor for ASR performance and *CF* showed up as a strong indicator. This is evident in the ASR results where high intelligibility speakers show better recognition despite subtle formant shifts as noted in figure 4.21. On the other hand, very-low intelligibility speakers who have escalated *CF* factor exhibit poor ASR scores.

Despite the encouraging insight into understanding the patterns of dysarthric speech, none of the acoustic variables examined exhibited a strong functional association with the intelligibility of dysarthria. It can be hypothesised that at present the irregularities of the dysarthric signal in the temporal and frequency domain are implicitly modelled by the speech systems. In a quest to devise an explicit methodology to rectify a dysarthric specific acoustic phenomenon, a novel way of comprehending dysarthric signals was explored. Instead of inspecting the speech signal in the frequency domain, the signal was directly examined in its original time-domain representation by studying the ZZT (Zeros of the Z-Transform) patterns of the vowel segments. The patterns in the ZZT plot revealed intriguing facts and exhibited a strong relationship to the underlying dysarthric intelligibility. The analysis shows that a phase based phenomenon was responsible for a skewed distribution of zeros patterns in speech with degrading intelligibility. It further pointed towards a functional relationship that might exist between the unwrapped phase component of the complex roots of ZZT and the underlying dysarthric intelligibility.

Concluding Notes

The ZZT analysis conducted here opens some interesting avenues, especially in the phase domain of dysarthric speech signals. It is compelling to think that if phase tends to reveal some important underlying dysartrhic artefacts, then it might also have the potential to encapsulate important acoustic cues that are unique to a particular dysarthric severity or variability. These aspects are explored in the remainder of this thesis that will:

- Attempt to develop a theory around the phase component of dysarthric speech signals and study its association with the underlying speech severity. It will also investigate whether such phase based artefacts can be systematically amended for improving the overall dysarthric ASR performance.
- Explore the usefulness of information that might be encoded in the phase component of speech signals and see if any phase based feature representation is better suited to model dysarthric variabilities for improving the ASR performance.

Chapter 5

Phase-based Analysis of Dysarthric Speech

In the previous chapter a new method for the analysis of dysarthric speech was explored by studying the ZZT patterns of its vowel segments. It was evident from the investigation that there is a relationship between the plots of the ZZT pattern and some underlying phase-based acoustic event in the acoustics of dysarthric speech. Figure 4.35 revealed that when the phase of the complex roots of the z-transform of the time domain signal was plotted, it exhibited a discernible operational range of phase deviations for the dysarthric signals across varying levels of intelligibility.

In this chapter we will extend the idea of phase deviations to investigate dysarthric vowel segments using Fourier transform analysis. We will study if phase deviation might have any relationship to the underlying speech impairment and if such association can be utilised for improving the overall ASR performance on dysarthric speech.

5.1 Phase-slope deviation

We begin by introducing a new metric that will encapsulate a quantitative notion of the phase slope deviations for analysing the dysarthric vowel segments. For any discrete signal $x[n]$, the N-point DFT is given by

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi kn}{N}} \tag{5.1}$$

where k is the k^{th} frequency bin of N uniformly spaced frequencies. The DFT of real numbers produces complex conjugate pairs at $ceil\{(N+1)/2\}$, so we discard the top half of the DFT output and use the lower half for processing information up to the Nyquist frequency. Since $X(k)$ is a complex quantity it can be decomposed into its polar form as

$$X(k) = |X(k)|e^{j\phi(k)} \tag{5.2}$$

where $|X(k)|$ is the magnitude spectrum and $\phi(k) = \sphericalangle X(k)$ is the wrapped phase spectrum, which is chaotic in nature. To make any meaningful interpretation from the phase spectrum it is generally unwrapped by adding multiples of $\pm 2\pi$ whenever the alignment between consecutive frequency bins exceeds π. This was performed using the *unwrap()* function defined in *MATLAB version R2016b*. Based on the definition of continuous phase spectrum we define a metric called Phase Slope Deviation (PSD) as

$$PSD(A,B) = \mathfrak{F}\big[\mathfrak{U}[\phi_A(\cdot)]\big] - \mathfrak{F}\big[\mathfrak{U}[\phi_B(\cdot)]\big] \tag{5.3}$$

where $\mathfrak{U}[.]$ is the unwrapped phase spectra, $\mathfrak{F}[.]$ is the slope of the first degree polynomial that fits the phase spectrum in the least square sense and A, B represents the phase data points at N discrete frequency bins that are compared for deviation. The range of $PSD(A,B)$ will lie between $(-\pi/2, \pi/2)$.

5.2 PSD analysis of dysarthric vowels

This section will investigate the effect of the phase-slope deviation defined by equation 5.3 between dysarthric and typical vowel tokens. In order to make an independent judgement in measuring the PSD, the analysis will be averaged across six different vowel groupings as shown in table 5.1.

Category	Tokens
front-vowels	/iy/ /ih/ /eh/ /ae/
back-vowels	/aa/ /ao/ /uh/ /uw/
high-vowels	/iy/ /ih/ /uh/ /uw/
low-vowels	/eh/ /ae/ /aa/ /ao/
diphthongs	/ey/ /ay/ /aw/ /ow/ /oy/
all-vowels	front + back + diphthongs

Table 5.1: Vowel categories with the list of phonetic tokens examined under it.

The analysis is conducted using a 512-point FFT with Hamming window applied over the entire vowel segment, which are pre-segmented using forced alignment. The final representation for each vowel grouping is averaged under the intelligibility categories defined in the UASPEECH corpus, viz., *very-low, low, mid and high.*

5.2.1 PSD and dysarthric intelligibility

The main steps involved in computing the PSD metric according to equation 5.3 between dysarthric and typical vowel tokens is given in the pseudo-code listing 5.1 and the slope differences evaluated for various vowel groupings and intelligibility categories is shown in figure 5.1. It seems to exhibit a strong association with the underlying dysarthric intelligibility. The plot shows a nearly linear relationship, where higher PSD value is indicative of lower intelligibility and vice versa. The deviations reported for various intelligibility groups are derived relative to similar vowel tokens examined for the typical speech data from the same corpus. The homogeneity of the dysarthric and typical data will ensure that similar vocabulary and recording conditions are present in both groups of speakers for each vowel token.

```
1. for each dysarthric speaker
2.  for each vowel token
3.       for each utterance
4.         compute the unwrapped phase
5.         compute the slope of the line that fits the phase
6.       end
7.       average slope across all the utterances
```

```
8.  end
9. end
10. repeat 1–9 for an average control vowel representation
11. for each dysarthric speaker
12.  for each vowel token
13.    compute PSD metric
14.  end
15. end
16. average across examined vowel groups
17. average across examined intelligibility groups
```

Listing 5.1: Main steps in computing the PSD metric between dysarthric and typical vowels.

It is also noted that the deviations in figure 5.1 seems to manifest an unbiased trend, as it appears to be independent of any particular vowel category. All this suggests that the results as predicted by the PSD metric are representative of some underlying acoustic artefact, which might be detrimental to dysarthric ASR performance.

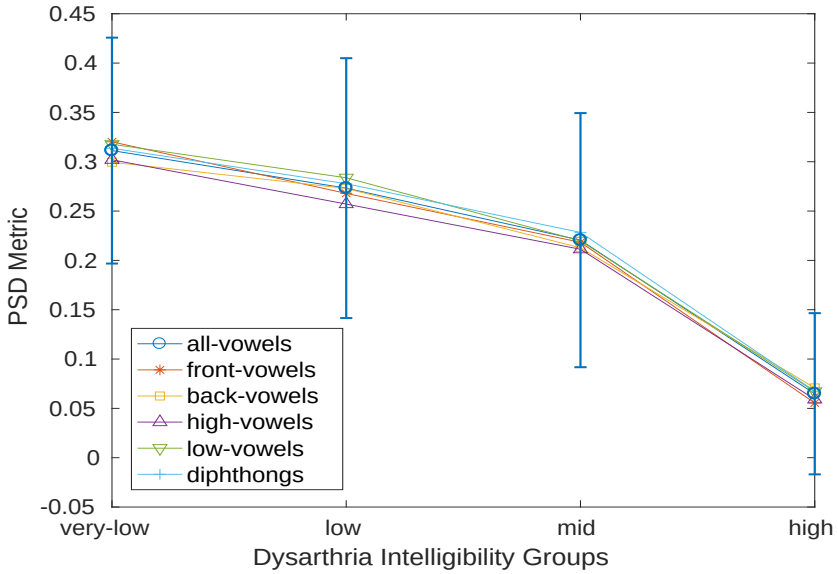

Figure 5.1: PSD analysis for speakers with dysarthria in UASPEECH database. The data spread is shown for the all-vowels category.

5.2.2 The behaviour of PSD on a secondary data source (VIVOCA)

It was seen in the last section that the PSD metric shows a linear correspondence with underlying intelligibility. In order to validate that the effect of the PSD metric is independent of any particular database, the same analysis was conducted on a different source of dysarthric data. The secondary data source was collected as a part of the VIVOCA project (Hawley et al., 2012; VIVOCA, 2012) at the University of Sheffield. The VIVOCA database comprises of 13 users with different dysarthric etiologies and varying levels of intelligibility. Table 5.2 shows a summary of the data for the VIVOCA users.

User	**Aetiology**	**Dysarthria Type**	**Vocabulary Size**	**Total Files**	**Intelligibility**
V2-1	CP	spastic	35	1225	A⋆
V2-2	TBI	spastic + ataxic	14	742	B
V2-3	CP	spastic	19	514	A
V2-4	CP	spastic	57	2956	A
V2-5	CP	spastic	35	1674	A
V2-6	CP	spastic	64	2821	A
V2-7	CP	spastic	100	4543	A
V2-8	MND	flaccid + spastic	28	933	A
V2-9	MND	flaccid	11	220	C
V2-10	PD	hypokinetic	6	145	A⋆
V2-7⋆	CP	spastic	20	934	A
V2-11	PD	hypokinetic	16	712	A
V2-12	MND	flaccid + spastic	13	432	A

Table 5.2: Summary of the VIVOCA users. The codes in the last column broadly indicate the intelligibility as: A(<20%), B(20%-50%) and C(>50%). Starred symbols for intelligibility are the result of informal listening tests and unstarred symbols are measured using the word-level intelligibility assessment procedure described in Hawley et al. (2012)

The VIVOCA data was collected over a duration of more than 5 years through different recording mediums (Hawley et al., 2012), viz., PDAs, PC and specialised hardware based

on Intel XScale family of processors (PXA270), especially built for the VIVOCA project by one of the industrial collaborators. All the recorded data was sampled at 8 kHz with a mono channel input. Since VIVOCA was aimed to provide bespoke speech solutions for dysarthric speakers with low intelligibility, there was no standard vocabulary used and it varied across all the speakers according to their individual needs (see table 5.2). Hence, there is no overlap between the vocabulary of UASPEECH and VIVOCA.

Although total words in the VIVOCA vocabulary is less than half and its speech material is around four times less than the UASPEECH database[1], it still provides a good variety of speakers with dysarthria. Also, in the way VIVOCA was implemented as a project, a large part of data was collected in more realistic conditions under which a dysarthric user is more likely to use any speech technology for communication (for e.g. home, place of work etc.). VIVOCA can thus be very useful dataset to test the efficacy of the PSD metric. It will not only investigate if the relationship between PSD and intelligibility is not an artefact of any particular database but it will also test the robustness of PSD using data from more diverse and realistic speaking environment.

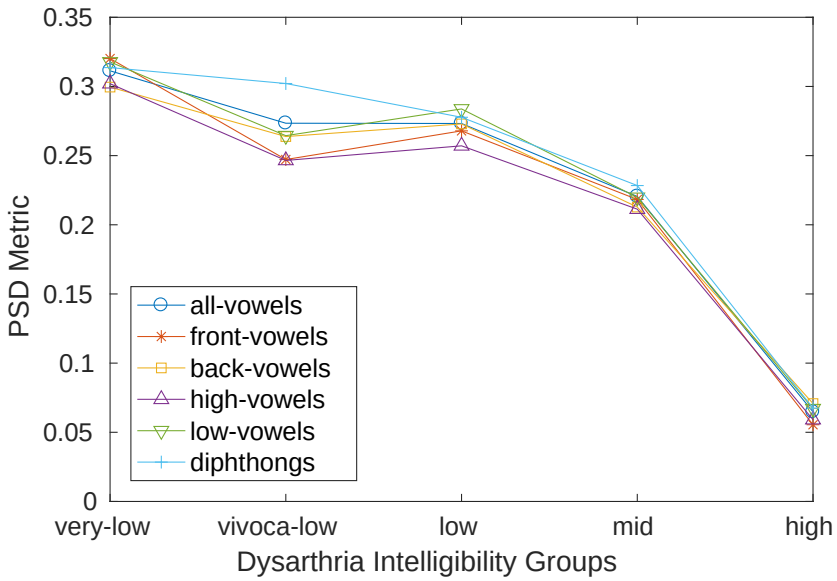

Figure 5.2: PSD analysis for the speakers with dysarthria in UASPEECH and VIVOCA databases.

[1]Due to the smaller size of the VIVOCA dataset, some of the vowel tokens were unavailable for particular speakers. The details of all the missing vowel tokens for individual speakers is given in the appendix table B.1.

The result of the PSD analysis for the VIVOCA dataset along with the earlier UASPEECH findings are presented in figure 5.2. It is an encouraging output where the average PSD effect for the VIVOCA speakers tend to converge within the range of severe speakers as predicted for UASPEECH database with a similar intelligibility profile of less than 50%. This is empirically suggestive that the PSD metric might be able to quantitatively explain the relationship between the extent of phase based deviations and the underlying speech intelligibility of the dysarthric speaker.

The above finding is also suggestive that the PSD metric is less sensitive towards missing information, whilst predicting the underlying intelligibility and can give reasonable estimates under sparse data conditions. This can be validated by the fact that, despite the unavailability of data for certain vowel tokens across various VIVOCA speakers (See appendix table B.1), the average PSD score was still predicted around the expected range for the lowest intelligibility group of speakers.

The PSD predictions for the VIVOCA database also seem to be independent of any particular vowel category examined, which confirms with the earlier findings on UASPEECH (Figure 5.1). Although there is a small deviation noted for the diphthong category of the vivoca-low group of speakers, it is most likely attributed to the unavailability of diphthong data for a large group of speakers, i.e., 12 out of 13 speakers had no data for /ow/ and 7 out of 13 speakers had no data for /aw/ (See appendix table B.1). The missing information can make the overall averaging effect biased towards the intelligibility of the examined speakers.

5.2.3 An operational understanding of PSD

It was shown in sections 5.2.1 and 5.2.2 that PSD metric seems to hold a strong association with the underlying dysarthric intelligibility, which seems to manifest a nearly linear model. The relationship of PSD was studied with broad categorical descriptions of intelligibility (very-low, low, mid, high). In order to further understand this association and draw any meaningful conclusion about the operational behaviour of PSD, the metric was computed on a speaker-wise basis. Figure 5.3 shows the PSD plot for each UASPEECH speaker which are colour coded according to their categorical description of intelligibility.

The scatter plot shows a negative correlation of $r = -0.87$ ($p < 0.01$) between the PSD metric and underlying quantitative estimate of intelligibility. The high degree of correlation emphasises the strength of PSD metric in predicting the underlying speaker intelligibility and the negative correlation is indicative of the fact that the deviation in PSD increases

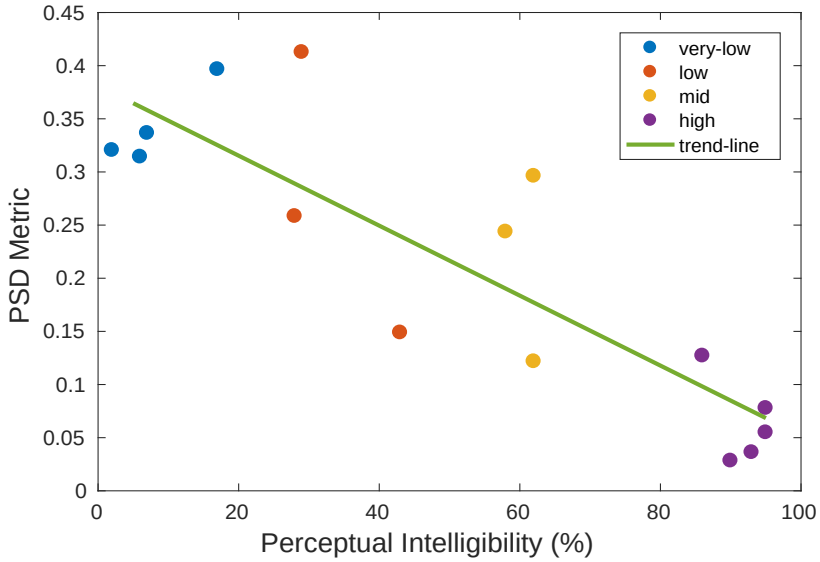

Figure 5.3: Relationship between PSD metric and intelligibility for speakers with dysarthria in UASPEECH. The coloured dots represent speakers from various intelligibility groups and the green line is the linear regression fit for the data points.

with decreasing intelligibility.

Figure 5.3 also shows a regression trend line plotted for the observed variables. It was plotted using the *polyfit()* function defined in *MATLAB version R2016b*. The parameters of the function were set to predict a linear model of best fit in the least square sense. If the trend line is used to predict the underlying average categorical intelligibility, an approximation of the operational range for the PSD metric can be hypothesised as shown in table 5.3.

Such regression analysis can be exploited to draw theoretical estimates for unknown measures like PSD metric or intelligibility about a given speaker. For example, if an estimate for intelligibility is available for a speaker then its approximate PSD score can be computed without analysing any user-specific acoustics and if sufficient user data is available for most of the vowels, it can estimate a quantitative measure for intelligibility. The regression trend shown in this case is represented by a simple linear model, which seems apt for the data. More sophisticated and accurate models can also be utilised if the data shows complex non-linear patterns. In either ways, it can enable us to plot a particular speaker as a point of singularity on similar charts as represented by figure 5.3.

Perceptual Correlate of Intelligibility	**Expected PSD Score**
High	< 0.15
Mid	$0.15 - 0.25$
Low	$0.25 - 0.30$
Very Low	> 0.30

Table 5.3: A hypothesised relationship between the PSD metric and the broad level categorical intelligibility classifications defined in UASPEECH.

This could be beneficial from an ASR perspective, as it can guide our understanding for systematically building speech models by clustering data of speakers from multiple sources who might be sharing similar acoustic properties. Alternatively, it can predict an acoustically driven real-time estimate of a speaker's intelligibility that can give a more streamlined approach for any speech based therapy and rehabilitation process.

In case of the VIVOCA database shown in table 5.2, it is difficult to draw any precise conclusion about the PSD metric or intelligibility for any speaker. It is partially attributed to (i) Missing vowel data for various speakers (Table B.1) that might give a slightly skewed prediction for the PSD metric or (ii) Inconclusive estimate for a speaker's intelligibility that is not based on a systematic assessment procedure of dysarthric speech (Enderby, 1983; Hartelius and Svensson, 1990; Yorkston and Beukelman, 1984). However, such missing information can always be gathered by either collecting more data or conducting formal listening tests.

In order to demonstrate the importance of the regression process, let us examine one speaker each from the three categorical intelligibility groups in VIVOCA, viz. A(<20%), B(20%-50%) and C(>50%). Let the speakers be V2-3 (A), V2-2 (B) and V2-9(C) with approximate average intelligibilities of 10%, 35% and 75%. In addition we also select the "only" speaker V2-4 that has data available for all the vowel tokens under examination. The task is to predict the PSD scores for the first three speakers and the perceptual measure of intelligibility for the last speaker and examine how it fits in relation to the UASPEECH plot of figure 5.3. The regression line shown in figure 5.3 has the following linear equation

$$y = -0.0033x + 0.3811 \tag{5.4}$$

where y is the PSD score and x is the quantitative estimate of intelligibility. Table 5.4 shows the predicted values of PSD and intelligibility based on equation 5.4.

Speaker	**Approximate Intelligibility**	**Known PSD Score**	**Predicted PSD Score**	**Predicted Intelligibility**
V2-3	A (10%)	-	0.3481	-
V2-2	B (35%)	-	0.2656	-
V2-9	C (70%)	-	0.1501	-
V2-4	-	0.3562	-	$\approx$ 8%

Table 5.4: PSD and intelligibility scores for the VIVOCA users as predicted by the regression equation 5.4.

A graphical representation of the estimated values is shown in figure 5.4.

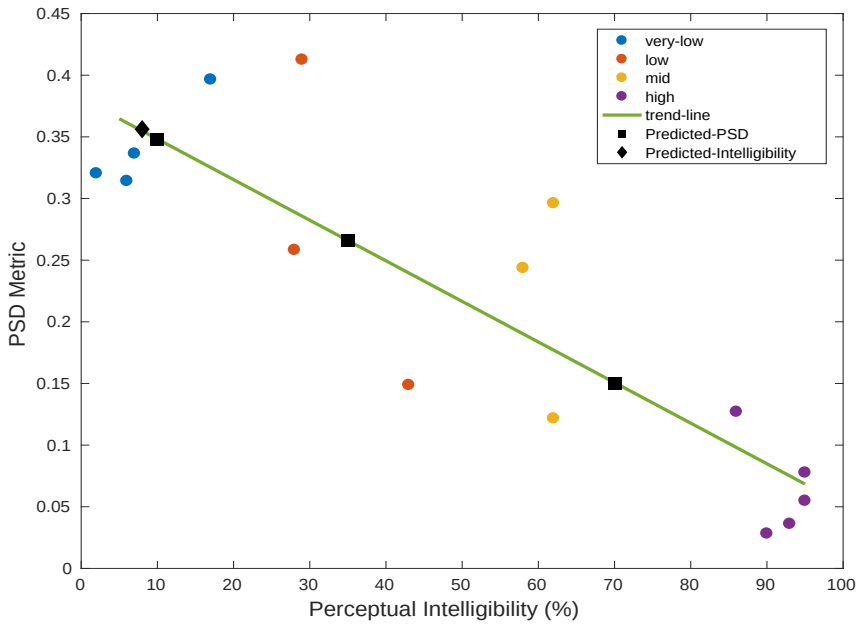

Figure 5.4: Predicted PSD and intelligibility estimates for VIVOCA speakers described in table 5.4. The estimated values are represented by solid squares and diamond and plotted against the UASPEECH speakers.

As per the PSD operational range defined in table 5.3, the first three speakers V2-3, V2-2 and V2-9 can be broadly classified as having very-low, low and mid level of intelligibility. On the other hand, for the speaker V2-4 with a PSD value of 0.3562, the intelligibility was predicted at approximately 8%. These theoretical estimates for the PSD and intelligibility values seem to nicely fit the practical expectation for the speakers under consideration.

5.2.4 Correcting PSD in dysarthric speech

In the preceding sections we have been able to draw a systematic inference about the relationship between the PSD metric and the perceptual notion of intelligibility. The examination conducted on two databases produced similar results, which exhibits a linear like relationship between the PSD scores and intelligibility. One of the propositions that can be drawn is to introduce a corrective approach which can reduce the effect of such phase deviations as measured by the PSD across various intelligibility groups. The efficacy of such corrections can then be directly tested by evaluating the ASR performance of various speech systems.

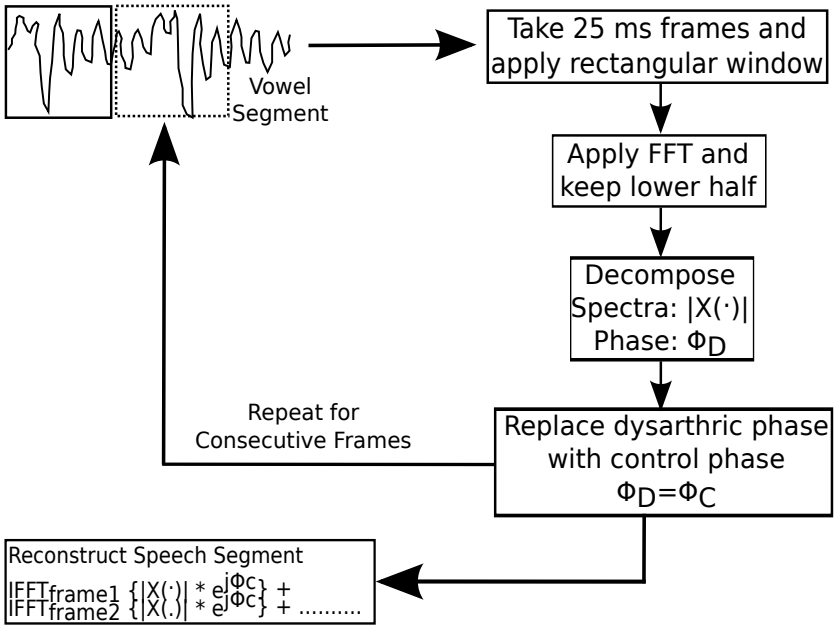

Figure 5.5: Correction measure to reduce the deviation effect of the PSD metric.

Historical evidence suggests that the human ear is not very good at resolving phase information (Helmholtz, 1912; Ohm, 1843) and for this reason most perceptually motivated

features like MFCC and PLP do not exploit the information in phase spectra. One might therefore think that any correction to PSD alignments would be futile. However, if PSD predicts undesirable acoustic artifacts relative to intelligibility, then its correction might alter some underlying variabilities in a way which might be useful from a representational perspective if not from a perceptual point of view.

One possible correction and test procedure is outlined in figure 5.5. It involves examining the short-time frames (25 ms) of a dysarthric vowel segment in a non-overlapping fashion. A rectangular window is applied to each frame and a fast Fourier transform was used to decompose into spectral and phase components of the linearly spaced frequency bins. The phase alignment for each vowel token of dysarthric speech was then replaced by the phase alignments from similar vowel tokens of typical speech that are averaged across all the UASPEECH control speakers. The phase transformation is applied to the vowel tokens across the entire UASPEECH database. Lastly, the spectral component and the new phase alignment is reconstructed using an inverse Fourier transform to get a new representation of the original utterance.

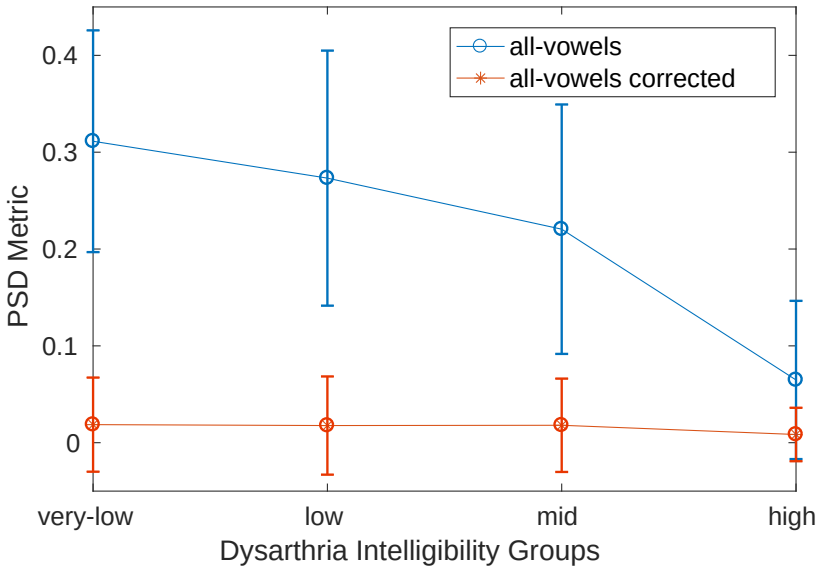

Figure 5.6: Corrected PSD alignment for various intelligibility groups. For clarity, it shows the correction effect for the "all-vowel" category only and similar results hold for other vowel and diphthong categories.

The effect of applying this correction on the "all-vowel" category is shown in figure 5.6. It shows a flattened response on the new set of dysarthric utterances that has the corrected phase alignments. The PSD metric for this new set of files falls within the most intelligible hypothesised range as shown in table 5.3. It should be noted that the flattened response due to the correction holds for all the individual vowel and diphthong categories and the purpose of showing only the "all-vowel" category is for clarity. As an example, appendix C exhibits the unwrapped phase alignment between the control and dysarthric speakers for a specific and average vowel representation that is used in the corrective procedure.

5.3 PSD effect on dysarthric ASR performance

The correction of the PSD, as shown in figure 5.6, could be evaluated perceptually and with a recognition experiment. Since the prime focus of the study is to develop methods to enhance the dysarthric ASR performance, we do not report any listening tests in the thesis.

5.3.1 Experimental setup

In order to have a direct comparison with the baseline results of section 4.1, the experimental design is kept exactly same. To the best of our knowledge, the results presented earlier (Section 4.1.2, Sehgal and Cunningham (2015)) are the best reported on this relatively large database (Kim et al., 2008) with a reasonably open vocabulary of 255 distinct words. For testing the effect of PSD corrected files we have picked up the speech systems which are practically more plausible to implement under real life scenarios, and are summarised in table 5.5. Both SI & SAT models are adapted using the hybrid MLLR-MAP approach.

System	**Training Dataset Used**
SD	UASPEECH-Dysarthria
SI-00	WSJ SI-84 + WSJCAM0
SI-02	UASPEECH-Dysarthria
SAT	UASPEECH-Dysarthria

Table 5.5: ASR systems that are re-tested to see the effect of PSD corrections.

5.3.2 Dysarthric ASR results

In order to estimate the effect of the PSD correction on the ASR output, the experiments for the speech systems mentioned in the last section are conducted under both supervised and unsupervised correction modes. It effectively means the amount of prior information that is exposed to guide the training process, which can take one of following forms:

- Supervised: The label alignments for the vowel tokens are pre-generated by the process of forced alignment. In addition vowel-specific PSD corrections are applied from control to dysarthric speakers. Under realistic setups such information is impractical to produce and is restricted to be generated under laboratory conditions. Hence, the ASR performance due to such corrections would rather aim at an oracle performance that can be attained.

- Semi-supervised: The label alignments for the vowel tokens are pre-generated by the process of forced alignment. There is no vowel-specific PSD correction applied, instead, a global PSD correction is applied for all the vowels. This correction step takes us one step closer to realistic usage as it aims to test the efficacy of ASR systems by ignoring any knowledge about any specific vowels on which corrections are applied.

- Unsupervised: No label alignments for the vowel tokens are available. The vowel segments are predicted using a vowel-detection classifier on which the global PSD correction is applied. This correction step uses the most minimalistic set of information. Usually such lack of information depicts a more realistic and practical setup. The ASR measure under such constraints will set the lower bound that the PSD correction can have on the ASR performance.

5.3.2.1 Supervised correction

The first set of experiments involved re-testing the four systems shown in table 5.5. The speech systems were trained and tested on the modified utterances of dysarthric data, where vowel specific PSD corrections were applied on the entire database. The detailed steps for implementing the corrective procedure is explained in section 5.2.4.

Figure 5.7 shows the ASR improvements for the PSD corrected speech. In nearly all the tested systems there were relative improvements across all the levels of dysarthric intelligibility. The largest relative gains were noticed for the SD (19.3%) and SI adapted systems

prepared using dysarthric speech data (14.22%) for the most severe group of users. The system adapted from typical speech data from other sources (SI-00) also showed an overall 3.44% improvement. It shows that the benefits of PSD correction are not only limited to homogeneous and dysarthric-only sources of information, but its efficacy can be extended to other sources also.

An interesting observation is that the ASR accuracies after the corrections seem to follow a similar pattern that the PSD metric exhibited with intelligibility as shown in figure 5.1. The corrective measure was found to be most effective for dysarthric speakers with the least intelligibility, since all the systems showed the maximum relative improvement for the least intelligible group as shown in figure 5.7. Thus, the ASR systems tested show an encouraging pattern, where the best performance is achieved where improvement is most needed.

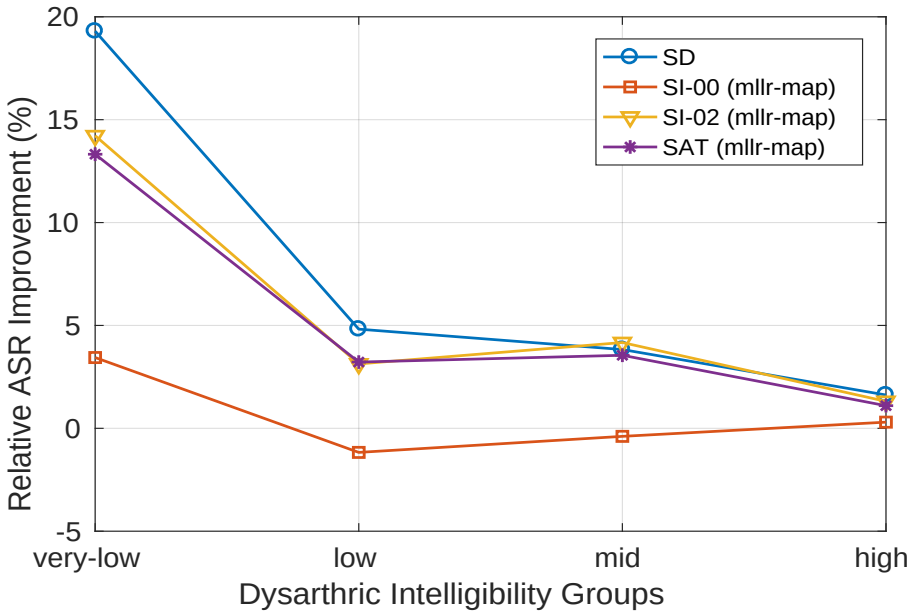

Figure 5.7: ASR improvement after the PSD corrections relative to the results presented in section 4.1.2. The four speech systems were re-tested on the speech files on which vowel-specific PSD transforms were applied.

The trend seems to converge for all the systems as it move towards the higher intelligibility dysarthric data (see figure 5.7). This was expected since it has been shown (Sehgal and Cunningham, 2015) that reduced severity of dysarthria is more closer to the control

group of speakers and models trained with typical speech may be the most appropriate approach for modelling.

Intelligibility	SD	SI-02	SAT	SI-00
very-low	23.52	27.36	28.71	20.61
	28.06 ††	31.26 ††	32.54 ††	21.32 †
low	62.48	62.92	62.98	57.89
	65.49 ††	64.89 ††	65.01 ††	57.22
mid	64.08	68.51	69.54	66.12
	66.54 ††	71.74 ††	72.02 ††	65.86
high	83.07	86.17	86.87	87.08
	84.42 ††	87.29 ††	87.83 ††	87.34

Table 5.6: Absolute ASR word accuracy averaged by various intelligibility groups. The top number in each cell represents the best results presented in table 4.3 and the shaded number is the result after the vowel-specific PSD correction was applied. Significant statistical gains are shown using a † ($p < 0.05$) or †† ($p < 0.01$).

Lastly, table 5.6 shows the absolute comparison between the results reported in this section and our earlier results in section 4.1.2 across all the tested ASR systems and intelligibility groups. Although SD modelling benefits most from PSD based corrections, SI-adapted and SAT systems trained with dysarthric data still had the best overall performance when averaged across all the intelligibility groups. A pairwise Cochran's Q test was conducted for each cell in table 5.6 to substantiate the findings from a statistical perspective. All the cells marked with a †† ($p < 0.01$) or † ($p < 0.05$) shows significant improvements.

5.3.2.2 Semi-supervised correction

In the semi-supervised corrective mode, the aim is to relax the PSD correction procedure and test its efficacy on the ASR output. This is achieved by applying a global PSD corrective transform averaged across all the 13 vowels instead of applying a vowel-specific transform. Similar to the corrections in the previous section, the transform is still applied to the pre-segmented vowel tokens, where their approximate temporal location is known.

Since a global transform requires applying a single transform across all the vowels instead

of multiple transforms, it has more practical potential to be used under realistic scenarios. Hence, it becomes imperative to evaluate the performance differences between the global and vowel-specific corrections applied earlier (figure 5.7).

A one-to-one comparison of the SD, SI-00(mllrmap), SI-02(mllrmap) and SAT(mllrmap) systems for both the transform procedures is shown in figure 5.8. All the four systems tend to show high degree of similarity between the global and vowel-specific transforms. A visual inspection of the charts reveal that the global PSD transform does not exhibit a substantial drop in ASR performance.

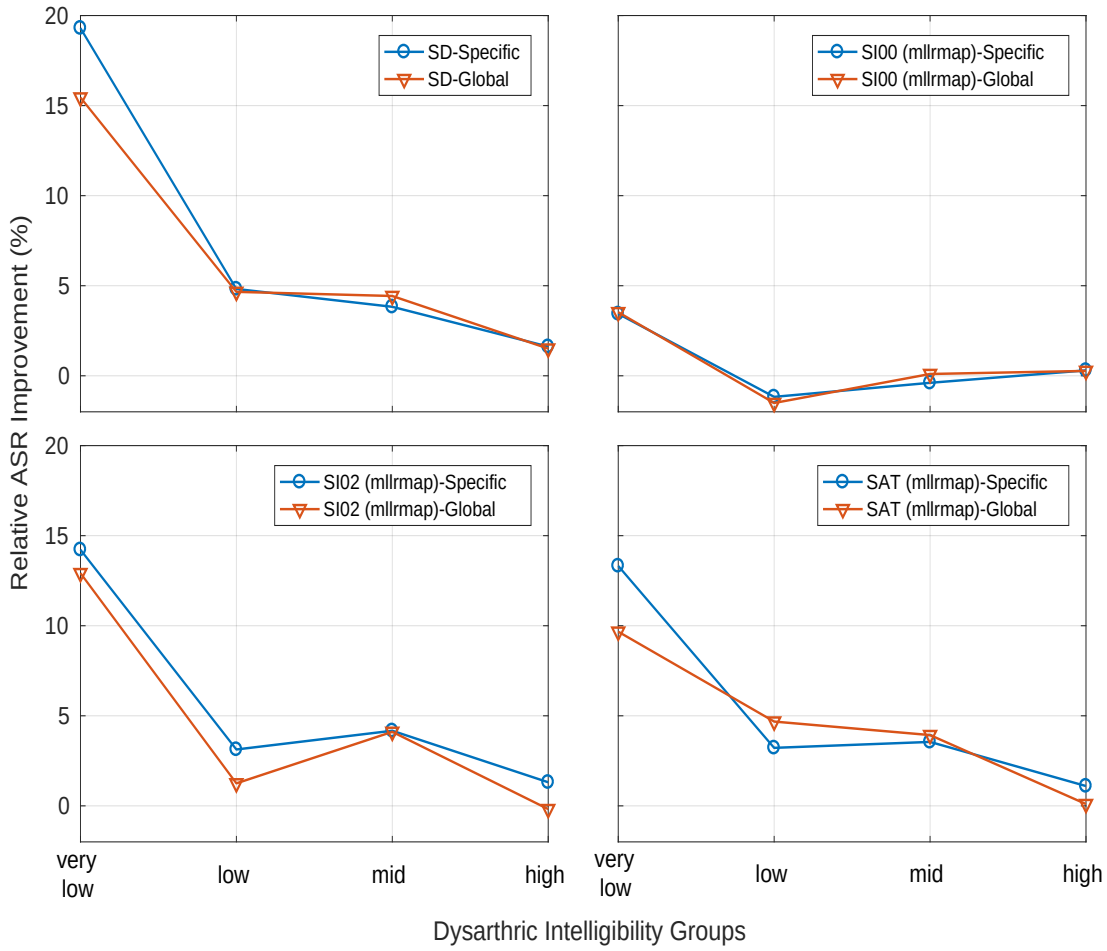

Figure 5.8: PSD correction comparison between specific vowel PSD transform (figure 5.7) and the global transform. Each chart exhibits a one-to-one comparison of a specific speech system where the x-axis represents the baseline result presented in section 4.1.2

Intelligibility	SD	SI-02	SAT	SI-00
very-low	28.06	31.26	32.54	21.32
	27.15 †	30.90	31.49 †	21.34
low	65.49	64.89	65.01	57.22
	65.38	63.70 †	65.93	57.02
mid	66.54	71.74	72.02	65.86
	66.92	71.33	72.27	66.19
high	84.42	87.29	87.83	87.34
	84.33	86.01 †	86.96 †	87.32

Table 5.7: Absolute ASR word accuracy averaged by various intelligibility groups. The top number in each cell represents the results presented in section 5.3.2.1 and the shaded number is the result after the global PSD correction was applied. The † indicates significant statistical gains ($p < 0.05$).

This was validated by running a Cochran's Q test between the results of the vowel-specific and global PSD corrections. Table 5.7 shows the absolute ASR scores for the two transform schemes along with the cells that show a significant difference (marked with a †). Out of the 16 possible combinations between the four systems and intelligibility groups, table 5.6 shows 12 out of 16 systems where PSD correction was better than no correction, whereas, table 5.7 highlights only 5 out of 16 such systems. It should be noted that at $p < 0.01$, it was found that the global transform (table 5.7) showed no significant difference across any of the speech systems and intelligibility groups. Hence a global transform can be thought of as equivalent to the computationally more expensive vowel-specific transform.

5.3.2.3 Unsupervised correction

The unsupervised correction is the most realistic and strictest mode of PSD transform that can be tested. In the unsupervised correction mode, the approximate time stamps for the respective vowel tokens are unknown, unlike earlier presentations, where they were pre-generated through the process of force alignment. The outcome of the unsupervised correction will also show real-life application scope of applying the PSD correction outside laboratory conditions.

In order to achieve this, it requires two steps, (i) preparing a classifier that can predict the vowels in a given utterance and (ii) applying a global PSD correction to the predicted vowel areas. This section will deal with both these aspects and report the results.

(i) Preparing a Vowel Prediction Classifier

The vowel/non-vowel classifier was constructed as an HMM-GMM system using the dysarthric training data from the UASPEECH corpus. The true phonetic alignment for each word utterance was replaced by the string of corresponding consonant (c) and vowel (v) labels. Table 5.8 shows an example of such a conversion

Example Word	**Phonetic Alignment**	**Vowel-Consonant String**
one	w ah n	c v c
juliet	jh uw l iy eh t	c v c v v c
whiskey	w ih s k iy	c v c c v
backspace	b ae k s p ey s	c v c c c v c
november	n ow v eh m b er	c v c v c c v

Table 5.8: Phonetic and corresponding vowel-consonant alignment for some example words.

The data was then processed as a 39 dimensional MFCC vector (12 static+c_0+Δ+$\Delta\Delta$) and analysed as 25 ms window with a 10 ms shift. The continuous density HMM is a word-internal tied-state tri-context model with clustering performed using decision trees, which follows a strict left-to-right topology with 32 Gaussian components used per state. Since it is a vowel-consonant (v-c) classifier setup, there are only three classes to start with, viz., *vowel*(v), *consonant*(c) and *silence*(sil), which gets expanded to triphone like context in the same way as it is done for the usual monophones. The exact monophone and triphone contexts that were generated during the training process included the 17 HMM classes c, v, sil, $c+c$, $c+v$, $c-c$, $c-v$, $v+c$, $v-c$, $v-v$, $c-c+c$, $c-c+v$, $c-v+c$, $c-v+v$, $v-c+c$, $v-c+v$, $v-v+c$, where for example $c-v+c$ implies the presence of any vowel token preceded and followed by any consonant.

Once the above HMM's were trained, it was used to recognise every UASPEECH word utterance as a blind sequence of $v's$ and $c's$ with no prior knowledge about the word itself. Since the task of the classifier is to solely capture the presence of vowels only and ignore the consonant sections, the recognition output was filtered to only retain the output timestamps

that has vowel as the central token and ignoring the others. A total of 8 triphone context sufficed this criteria, which are $c-v+c$, $c-v+v$, $v-v+c$, $v-v+v$, $v-v$, $v+v$, $v+c$, $c-v$.

```
0 3800000 sil -2932.056396 start_sil -50.000000
3800000 8300000 c+c -3751.847412 ccvcv -50.000000
8300000 9700000 c-c+v -1427.203369
9700000 10800000 c-v+c -1076.188599
10800000 11500000 v-c+v -722.165894
11500000 13800000 c-v -1917.229126
13800000 19000000 sil -3814.687744 end_sil -50.000000
```

Figure 5.9: A sample recognition output for a random utterance from the vowel-consonant classifier. The green sections indicate the presence of vowels in the utterance.

Figure 5.9 shows the recognition output from the HTK system of a random utterance using the vowel-consonant classifier. It illustrates the filtering process of picking up the timestamps that pertain to vowels only. In the above image the time segments marked as green indicate the presence of a vowel and all other areas were ignored when selecting the vowel-only sections.

Building a vowel-consonant classifier can be an arduous task due to the confusions that might arise between true vowels and voiced segments. In scope of this thesis, the aim of this section is not to build a highly efficient vowel detector, but it rather targets to build a classifier that is good enough to prove the efficacy of the PSD corrections in realistic, uncontrolled setups.

It is important to note that applying many PSD corrections to non-vowel segments can be detrimental for the ASR performance. This was concluded after a series of informal tests (not reported in the thesis) led to a fall in the ASR accuracy. Also, intuitively it can be thought that such inaccurate corrections can lead to poor model training and increase the effect of training/test mismatch. Hence, it becomes imperative to at least apply some basic measures that can check the efficacy of the classifier output with some degree of confidence. The force aligned data generated during the supervised correction mode can be the benchmark for comparing any classifier output, since it is the closest we can get in automatically predicting the location of the vowel tokens.

A frequency histogram plot is used to compare the distribution of the vowel segments

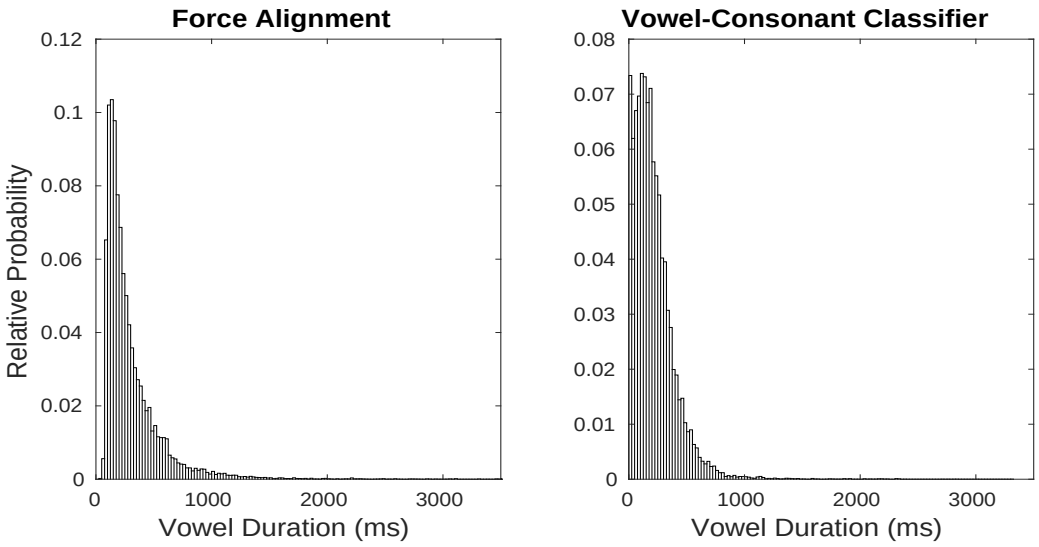

Figure 5.10: The histogram plot for the frequency distribution of vowel durations as predicted by the forced alignment process (left) and vowel-consonant classifier (right). The bin-width of the histograms was set to a fixed width of 25 ms.

predicted by the unsupervised classifier against the vowel segments generated by the forced alignment process (as used in supervised and semi-supervised PSD corrections). The comparative histograms are shown in figure 5.10. The y-axis represents the relative probability given by $\frac{c_i}{N}$, where c_i is the vowel frequency of the i^{th} bin and N is the total number of vowel tokens examined during the recognition process.

The vowel-consonant classifier predicted around 20% more vowel tokens than the force aligned timestamps. The average vowel duration of the classifier output was 214 ms (std.dev:182 ms) and the forced aligned timestamps is 272 ms (std.dev:177 ms). The visual inspection of the histograms in figure 5.10 shows a high degree of similarity between the vowel frequency durations of the vowel-consonant classifier and the force aligned timestamps. They both exhibit a positively skewed map of the histogram plots which seem close enough. In order to quantitatively measure the similarity of histogram shapes and the relative probability measure for the occurrence of various vowel durations associated with each bin, a χ^2 test was conducted on the two histograms. The first 68 bins were examined for the test, which gives a good coverage of vowel tokens ranging up to 1700 ms. This upper limit should be sufficient to conclude the test since we will cap our PSD correction experiments at 1500 ms to avoid capturing very long and possibly incorrect vowel segments. Hence, the

χ^2 test will use 67 degrees of freedom and the following test statistics was retrieved for the two histograms.

$$\chi^2 = 27.43$$
$$DF = 67$$
$$\alpha = 0.001$$
$$CV_{(0.001,67)} = 108.52$$

where CV is the critical value of the χ^2 distribution and $\chi^2_{67} < CV$. It implies that the null hypothesis is accepted in this case that suggests no statistically significant difference between the two histograms at $p < 0.001$.

In order to optimise the vowel prediction classifier, it is important to ignore irrelevant vowel segments that might be having a strong chance of being misclassified. Unfortunately there is no easy way to predict such differences between a true-positive and false-positive outputs. The best approach is to combine our understanding of the dysarthric speaking rates (see section 4.2.1) along with some basic statistics (mean, standard deviation etc.) collected by the classifier for each intelligibility group. Table 5.9 shows the approximate time zones that are used to extract the vowel segments for the global PSD corrections. The time regions for each intelligibility group are hypothesised based on the average measurements of vowel durations that were examined during the supervised and semi-supervised correction modes.

Dysarthric Intelligibility	**Vowel Duration (ms) for PSD conversion**
very-low	300 - 1500
low	300 - 1500
mid	250 - 750
high	250 - 650

Table 5.9: Operational duration range to extract the vowel segments for each of the intelligibility groups.

(ii) Results for the Unsupervised PSD corrections

The effect of unsupervised PSD correction on the dysarthric ASR performance was compared against the vowel-specific transforms of figure 5.8. It is important to reiterate that the benefits of such unsupervised corrections on the ASR output is highly dependent on the performance of the classifier that can optimally predict the timestamps of the possible vowel segments.

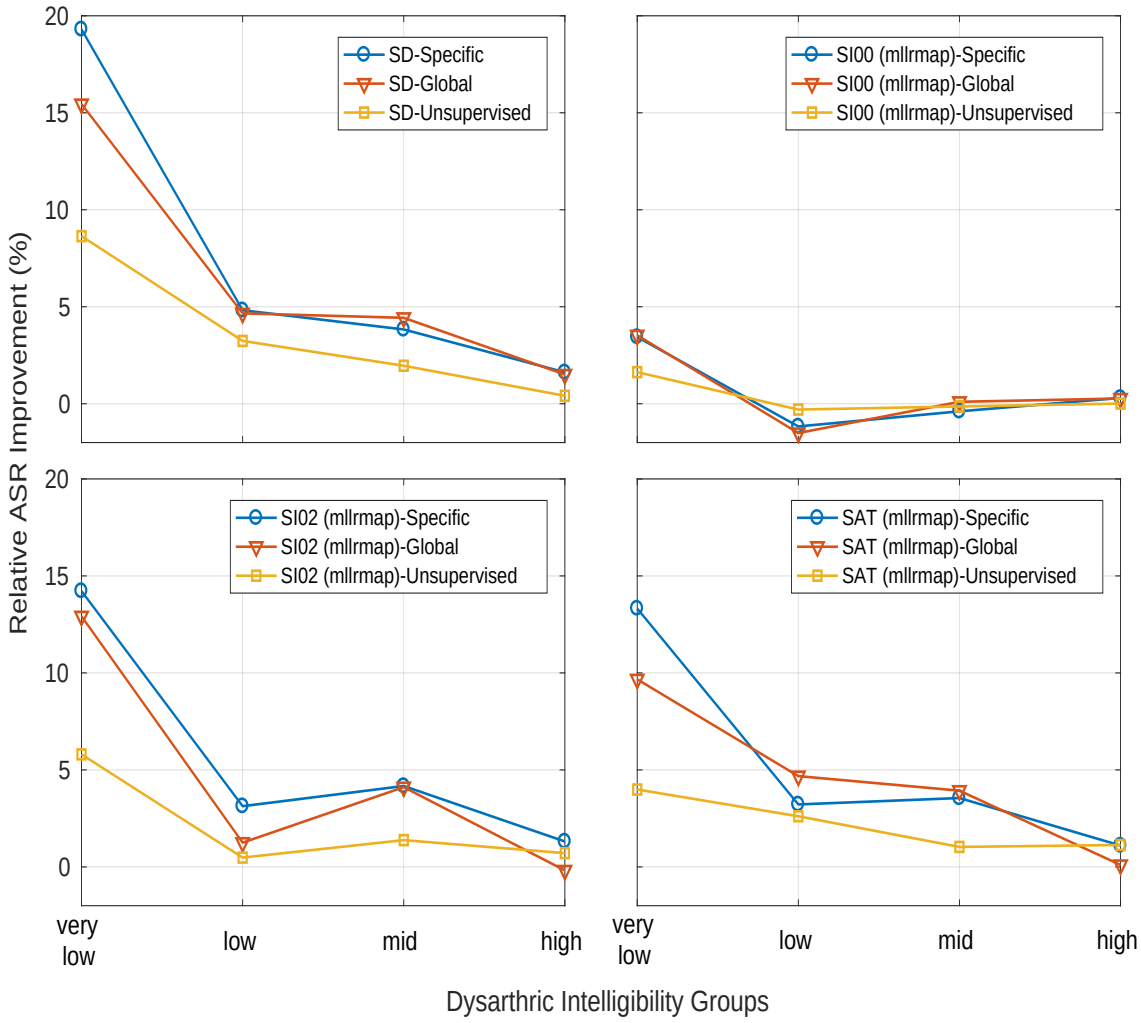

Figure 5.11: Comparison between specific, global and unsupervised vowel PSD transform. Each chart exhibits a one-to-one comparison of a specific speech system.

A large number of PSD corrections on the incorrectly segmented areas of the utterance might result in a cumulative negative impact on the ASR gains. The experiments in the current section are not targeted to construct an efficient and highly optimised vowel classifier, but it rather focuses on evaluating the effects of the unsupervised PSD correction using an average vowel classifier.

The outcome of this experiment will give a practical insight about the viability of applying such global phase-based corrective transforms in a realistic setup. A one-to-one comparison of the SD, SI-00(mllrmap), SI-02(mllrmap) and SAT(mllrmap) systems for all the three transform procedures is shown in figure 5.11. In all the four speech systems tested, the unsupervised correction mode shows no significant drop relative to the baseline results of section 4.1.2. Although the relative gains for the unsupervised corrections are less than that of vowel-specific and vowel-global transforms, it however shows a similar pattern that is more favourable for reduced dysarthric intelligibility.

Intelligibility	**SD**	**SI-02**	**SAT**	**SI-00**
very-low	23.52	27.36	28.71	20.61
	25.41 ††	28.84 ††	29.79 †	20.95
low	62.48	62.92	62.98	57.89
	64.50 ††	63.22	64.80 ††	57.72
mid	64.08	68.51	69.54	66.12
	65.34	69.45	70.26	66.03
high	83.07	86.17	86.87	87.08
	83.41	86.78 †	87.85 ††	87.09

Table 5.10: Absolute ASR word accuracy averaged by various intelligibility groups. The top number in each cell represents the baseline results presented in table 4.3 and the shaded number is the result after the global unsupervised PSD correction was applied. Significant statistical gains are shown using a † ($p < 0.05$) or †† ($p < 0.01$).

The results were validated by conducting a Cochran's Q test between the results of the unsupervised PSD correction and the baseline result represented by $y = 0$ in figure 5.11. Table 5.10 shows the absolute ASR scores along with the cells that show a significant difference (marked with a † or ††). It is noteworthy that the application of PSD transform

give statistically significant gains at $p < 0.01$ for most of the systems in the very-low and low intelligibility groups. These results are both encouraging and coincidental with our earlier findings, where PSD transforms are shown to be more effective in speech with high degree of pathological disorder. Hence an unsupervised transform shows promising results that might prove beneficial in uncontrolled and realistic setups for the application of dysarthric speech interfaces.

5.4 Conclusion

In this chapter a new metric called PSD was introduced. It is based on the slope deviations that are observed in the unwrapped phase spectra of the vowel segments. The PSD metric exhibited a strong and nearly linear correspondence with the underlying intelligibility when it was analysed across all the dysarthric vowel and diphthong segments. The PSD metric does not require any pre-training, which makes it independent of any database for its operation. Hence, PSD can make corpus independent predictions of a phase-based acoustic anomaly that might be manifest in a signal. In the current study the PSD analysis was conducted on two independent data sources, viz. the UASPEECH and VIVOCA corpora, and it displayed a strong association between the PSD scores and the expected intelligibility from a perceptual standpoint. It was also found that PSD was less sensitive towards missing phonetic data and can give reasonable approximations under sparse data conditions. PSD exhibits a reasonably good linear correlation with intelligibility that is useful for prediction. It can be used either as an acoustically driven predictor of intelligibility that can aid in a speaker's speech therapy assessment, or it can give an estimate of a speaker's PSD score that can help to deploy better modelling techniques by clustering similar speakers together.

Since PSD was found to be a predictor of some underlying acoustic artefact relative to intelligibility, a corrective procedure was applied to minimise the PSD aberration in the dysarthric vowel tokens. The correction involved replacing the phase alignment for vowel tokens of dysarthric speech with the phase alignment from similar vowel tokens of typical speech. The PSD correction alters the reconstructed signal. During the inverse Fourier transformation stage the new phase alignment is combined with the original magnitude. So even though we discard the phase information during the feature generation process of MFCCs, the time-domain signal has been modified. This results in changes to the magnitude spectrum that appear to be more amenable for improving the ASR performance.

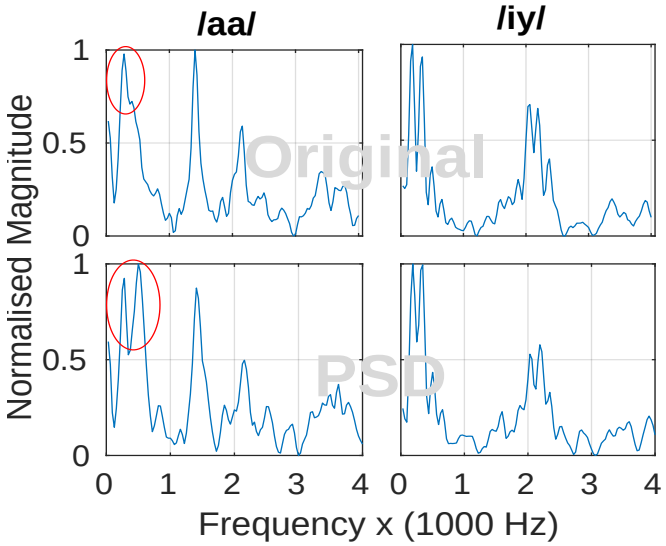

Figure 5.12: Comparison between vowels of the Original and PSD corrected file for the test word **copy** for a speaker with very-low intelligibility. The red ellipse shows an area of interest where the PSD correction seems to exhibits a finer resolution of the spectrum.

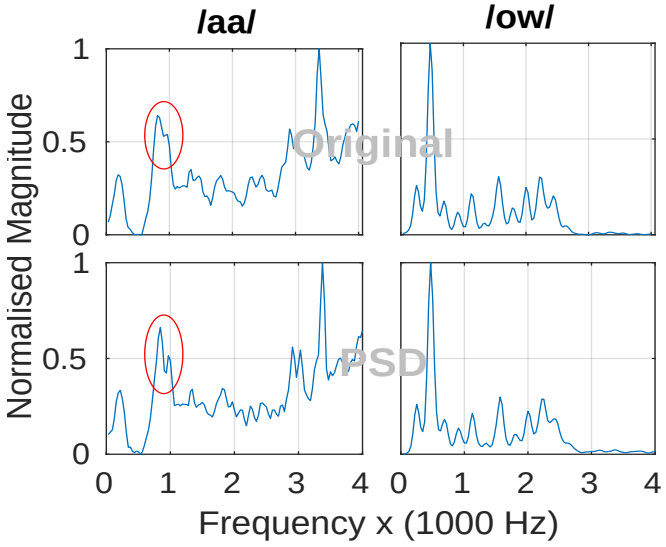

Figure 5.13: Comparison between vowels of the Original and PSD corrected file for the test word **bravo** for a speaker with low intelligibility. The red ellipse shows an area of interest where the PSD correction seems to exhibits a finer resolution of the spectrum.

As an example, figure 5.12 and figure 5.13 shows the spectra comparison of the original and PSD corrected files for the vowel segments of the words **copy** and **bravo** for two speakers with very-low intelligibility. The original file was misrecognised by the ASR system, whilst the same file is correctly recognised after the PSD correction was applied. One of the possible reasons that is evident from the two figures are highlighted in the small area represented by the red ellipse, which shows a better resolution of the possible formants in the initial vowel production /aa/ of the two words. Lastly, It is worth mentioning that although the corrections were applied in non-overlapping frames, it does not seem to have left any detrimental effect on the spectrum and overall the the PSD corrections seem to favour the ASR systems.

The PSD corrections gave significant ASR improvements across all the tested speech systems and dysarthric intelligibility groups. A broad level summary of the speech systems and intelligibility groups that benefitted from the PSD corrections relative to the baseline results of chapter 4 is given below in decreasing order:

$$\begin{aligned} \text{Speech Systems :} \quad & \text{SD} > \text{SI-02} > \text{SAT} > \text{SI-00} \\ \text{Intelligibility :} \quad & \text{very-low} > \text{low} > \text{mid} > \text{high} \end{aligned}$$

It should be noted that although speaker adapted systems that utilise all the available dysarthric speech (SAT, SI-02) had the best absolute scores for the majority of intelligibility groups, the speaker dependent (SD) system had the maximum relative gains (19.3%) using the PSD correction scheme. The relative benefit was also independent of the type of correction applied (*supervised, semi-supervised, unsupervised*), thus making the outcome more acceptable for different applications. This is an encouraging outcome since in real-life applications it is not always feasible to have data from multiple speakers due to physical and technical constraints in order to prepare SAT or SI-02 like systems. In most scenarios an SD model is the best choice of speech system that can be prepared for a given user and PSD correction tends to provide maximum performance benefits for the speakers with low intelligibility.

It is already mentioned that applying the PSD correction to other voiced regions instead of only vowels was detrimental for ASR performance, irrespective of the correction mode applied. Vowels are usually produced with an open vocal tract and its acoustics are fairly well defined. Hence, it is more easier to observe an underlying phenomenon and apply

any form of modification (PSD in our case), without adversely affecting the surrounding syllable or word. In contrast, consonants are more complex as they involve some sort of constriction along the vocal tract for it to manifest in the acoustics. They hardly ever form the nucleus of the syllable under consideration. For dysarthric speech consonant production has a greater chance to exhibit articulation errors due to the compromised filter system. The errors can be at the place or manner of articulation and it can be further accentuated by voicing errors. Due to this, it is extremely hard to understand any acoustic anomaly and suggest a corrective approach that will adequately attempt to address the underlying physiological weakness. At this stage there is a lack of understanding in associating any phase based phenomenon with voiced segments of dysarthric speech, and it will need further research to have a better understanding of the underlying mechanics.

The inclination of PSD to favour SD models can be attributed to both the data used for training and the automated way in which the PSD correction operates. Irrespective of which PSD scheme is used, it should be noted that all the three corrective methods rely on a process to predict valid vowel segments in an utterance where a possible correction could be applied. It has already been discussed that applying a lot of invalid corrections can degrade the overall performance benefits. In an SD system, the effect of good and bad PSD corrections is only limited to a single speaker, whereas, in SAT & SI-02 systems, the effect of bad corrections will have a cumulative impact of each speaker during the adaptation process. This could be one of the prime contributing factors for SD to receive maximum benefit from the PSD corrections in comparison to the other systems.

Lastly, the speakers with the lowest intelligibility benefit the most from the PSD correction procedure and the relative gains seem to decrease as we move towards the less severe end. The ASR outcome is in exact agreement to the PSD metric prediction for the underlying intelligibility. Since PSD effect shows maximum deviation for the least intelligible group, it is intuitive and highly likely that its associated phase correction will be the most beneficial for ASR performance. This outcome is promising from a practical perspective, since the maximum advantage of PSD correction is for speakers who will benefit most from it in realistic setups. In general, any specialised speech system is particularly designed for speakers with high degree of severity, since less severe speakers are more likely to benefit from any state-of-the-art commercial system. Hence, real time PSD corrections can show beneficial performance gains for speech systems that are especially designed for speakers with lowest intelligibility.

Chapter 6

Feature Representations based on Phase Spectrum

In the previous chapter, a new metric (PSD) was introduced that was based on the deviations observed in the unwrapped phase spectrum of vowels and diphthongs. In this chapter we extend the idea of useful information encoded in the phase component of the Fourier transform of dysarthric speech. It will be explored from a theoretical and practical standpoint that if phase based feature encoding on dysarthric speech show any properties that are beneficial for improving dysarthric ASR performance. It will also be explored if such phase based feature representations have any augmented benefits for improving dysarthric ASR when it is combined with the corrective properties of the PSD metric.

6.1 Phase-based feature representations for speech recognition

Fourier analysis is important in the front-end processing of speech signals for ASR. It breaks the complex speech signal into its fundamental constituents and encodes information for the observed frequencies in its respective magnitude and phase components. Despite the fact that both magnitude and phase parts are needed for the true representation of any speech signal, most of the conventional feature representations for ASR only exploit the magnitude spectrum of the Fourier analysis and the phase spectrum is mostly ignored. The reason for disregarding the phase spectrum could be either historically motivated that suggested the inadequacy of human ear to resolve phase information (Helmholtz, 1912; Ohm, 1843) or the difficulty involved in processing the phase spectrum due to its chaotic nature that results from random polarity and wrapping constraints of the phase between the range of $\pm\pi$.

One of the earliest references that show the importance of phase was shown in a systematic study conducted by Oppenheim and Lim (1981). The paper gives a summary of some key studies dating back to 1960's that showed the prominence of phase-only synthesis in analysing atomic crystal structures for measuring contours of the electron density. These results opened pathways for Fourier phase analysis into other applications. The paper illustrates examples where phase-only reconstruction of images and speech signals were more close to the original than magnitude-only reconstruction. It also showed that phase-only reconstruction of signals was better at preserving key "event locations" in the signal and had better correlation to the original signal.

In context of discrete speech signals it is known that one can apply Hilbert transform to recover the magnitude spectrum of any signal from its phase spectrum within a scale factor, if the underlying signal is either minimum or maximum phase (Oppenheim and Schafer, 1989). Since most of the speech signals of interest are mixed phase, such transformation might not be straightforward. Hence, the remainder of this section will outline some key representations of speech signals that are based only on the processing of phase spectrum. These phase-only feature representations will be used in later sections of this chapter to investigate if they are beneficial in evaluating the performance of dysarthric ASR in comparison to standard magnitude-based features like MFCC.

6.1.1 Group Delay Function

The processing of the phase spectrum is a difficult task due to wrapping constraints of the spectrum between the values of $\pm\pi$. The wrapping exhibits the phase spectrum as a chaotic curve with random fluctuations. In order to overcome this problem, the phase spectrum can be computed as the negative first order derivative from the unwrapped phase spectrum. It is known as the **group delay function** that is mathematically denoted as:

$$\tau(\omega) = -\frac{d(\phi(\omega))}{d\omega} \tag{6.1}$$

The above equation represents the *"rate of change in the phase spectrum"*, where $\phi(\omega)$ is the continuous unwrapped phase spectrum. The unwrapping of phase involves adding multiples of $\pm 2\pi$ whenever the alignment between consecutive frequency bins exceeds π.

```
for each frequency bin;
        if( (φ(n) − φ(n−1)) < −π )
                φ(n) = φ(n) + 2π

        if( (φ(n) − φ(n−1)) > π )
                φ(n) = φ(n) − 2π
```

Listing 6.1: Pseudo-code illustration for unwrapping the phase

The above listing shows the pseudo-code for unwrapping the phase where $\phi(n)$ is the phase at the n^{th} frequency bin. The unwrapping process generally employs extra steps to take into account the direction of phase shift also. In the current thesis, the unwrapping process was however performed using the *unwrap()* function defined in *MATLAB version R2016b*.

Once the unwrapped phase spectrum is determined, it is possible to compute the group delay function of equation 6.2 by applying the definition of derivatives. The group delay function can thus be defined as:

$$\tau(\omega) = -\frac{\phi(n) - \phi(n-1)}{f(n) - f(n-1)} \tag{6.2}$$

where $\phi(n)$ and $f(n)$ are the unwrapped phase and frequency at the n^{th} bin respectively. The unwrapping process can completely be avoided and the phase spectrum can also be

computed directly from the time-domain signal (Oppenheim and Schafer, 1989) as:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \tag{6.3}$$

where $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x(n)$ and $nx(n)$ respectively and R and I denote the real and imaginary parts of the complex output.

6.1.1.1 Properties of Group Delay Function

The motivation behind representing the speech signal using the phase spectrum instead of the more commonly used magnitude spectrum lies in the properties of the group delay function. The theory and properties with its practical applications are discussed in greater detail in the earlier studies (Murthy and Yegnanarayana, 1991, 2011). The current section will however briefly highlight the key properties along with some simple illustrations.

(I) Property of Additivity: The convolution of any time-domain signal is additive in the group delay phase spectra. This is in contrast to the magnitude spectra, which is multiplicative in its presentation. For example, if $X(t)$ is the convolution of two signals $X_1(t)$ and $X_2(t)$ given as:

$$X(t) = X_1(t) * X_2(t) \tag{6.4}$$

The Fourier transform for the above will be

$$X(e^{j\omega}) = X_1(e^{j\omega})X_2(e^{j\omega}) \tag{6.5}$$

From the basic properties of complex number we can deduce the following for equation 6.5 :

$$|X(e^{j\omega})| = |X_1(e^{j\omega})||X_2(e^{j\omega})| \tag{6.6}$$

$$\begin{aligned} arg[X(e^{j\omega})] &= arg[X_1(e^{j\omega})] + arg[X_2(e^{j\omega})] \\ \lhd[X(e^{j\omega})] &= \lhd[X_1(e^{j\omega})] + \lhd[X_2(e^{j\omega})] \\ \phi_X(e^{j\omega}) &= \phi_{X1}(e^{j\omega}) + \phi_{X2}(e^{j\omega}) \end{aligned} \tag{6.7}$$

Now applying the definition of group delay function (equation 6.2) in the above phase relationship, we get

$$\tau_X(e^{j\omega}) = \tau_{X1}(e^{j\omega}) + \tau_{X2}(e^{j\omega}) \tag{6.8}$$

where $\tau_{X1}(e^{j\omega})$ and $\tau_{X2}(e^{j\omega})$ represent the group delay functions of $X_1(e^{j\omega})$ and $X_2(e^{j\omega})$ respectively. Equations 6.6 and 6.8 clearly shows the multiplicative and additive behaviour of the magnitude and group delay spectra.

At this stage we can recall that the transfer function of any speech signal in pole-zero format is written as:

$$X(e^j\omega) = \frac{b_0}{a_0} \frac{\prod_{k=1}^{M}(e^{jw} - c_k)}{\prod_{k=1}^{M}(e^{jw} - d_k)} \tag{6.9}$$

The additive property of the group delay function can be applied to the above equation to get:

$$\tau_X(e^{j\omega}) = \tau_{zeros}(e^{j\omega}) - \tau_{poles}(e^{j\omega}) \tag{6.10}$$

Hence, the inherent additive nature of group delay function has a more direct application in contrast to the magnitude spectrum, which requires a logarithmic domain to work on.

(II) Property of Higher Resolution: This is one of the most important properties of the group delay function that makes it an attractive form for representing the spectral structure of a speech signal. It tends to exhibit higher resolving power in differentiating closely spaced peaks in the spectrum.

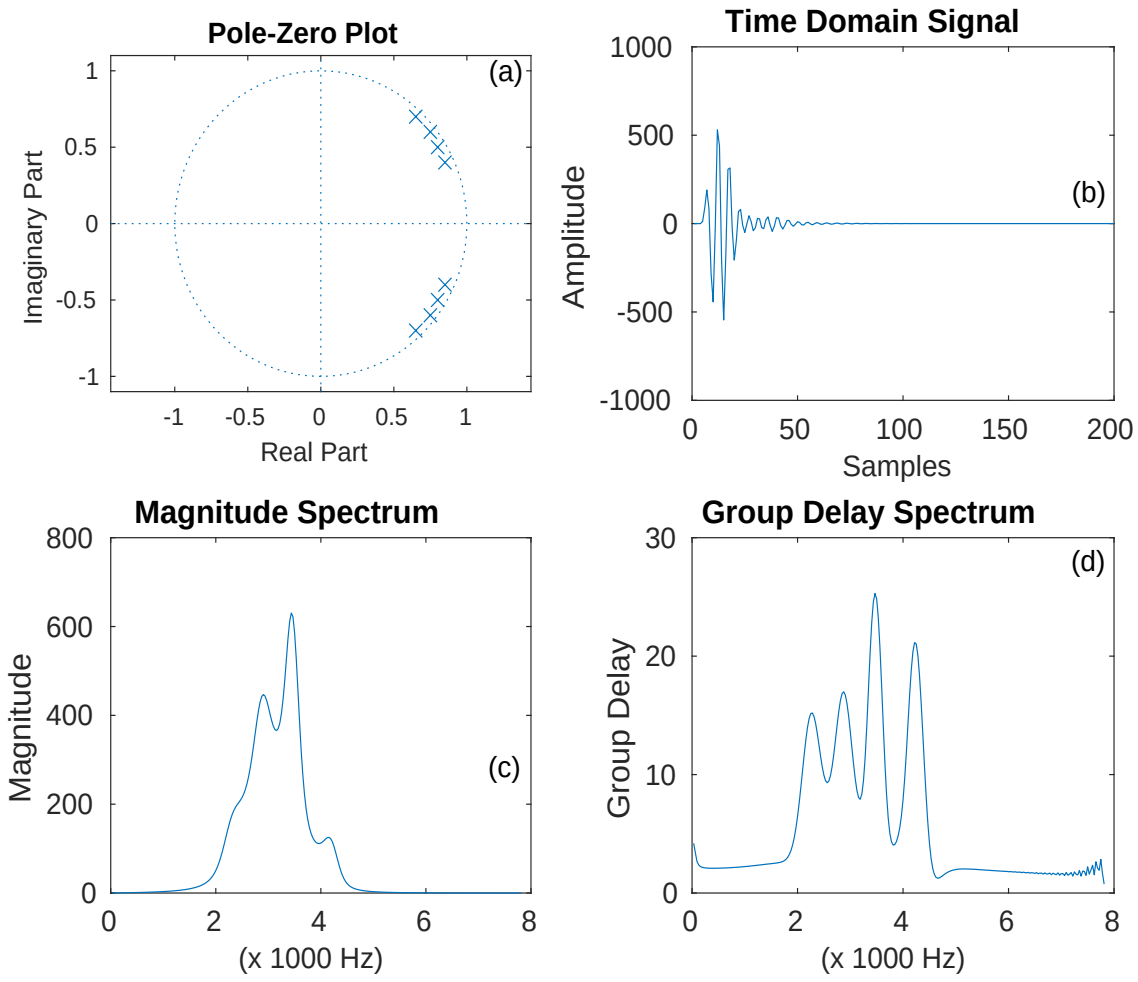

Figure 6.1: The figures demonstrate the high resolution capability of the group delay spectrum. Part (a) shows the pole-zero plot for four poles that are located very close to the unit circle and occur in complex conjugate pairs, (b) gives the first 200 samples of an approximate time-domain signal which comprises of expected frequencies predicted by the presence of poles and (c), (d) show the magnitude and group delay spectrum respectively.

It is easier to understand the phenomenon of higher resolution by an example as illustrated in figure 6.1. The part-(a) represents a pole-zero plot of four poles along with their complex conjugate pairs. The poles have been intentionally placed in close proximity to the unit circle as this will push the frequency response high around that band of frequencies,

which should be depicted as prominent peaks in the spectrum. In addition, the poles are also placed very close to each other, which can better aid in understanding the resolving power of the magnitude and group delay spectra. It can be easily seen that the magnitude spectrum represented in part-(c) can only coarsely define the four expected peaks in the spectrum. There is a very prominent peak at the around 3.5 kHz, two weak peaks at around 3 kHz and 4.5 kHz followed by a very faint crest just noticeable at around 2.5 kHz. In contrast, the group delay spectra represented in part-(d) resolves the peaks in the spectrum with much greater resolution and accuracy at the expected frequencies. Since group delay spectrum exhibits higher resolving properties, the speech features generated from them might hold a better chance in distinguishing important acoustic events.

This can especially be beneficial for dysarthric speech, where the underlying acoustics is generally convolved with severity or etiology based disfluencies that can be difficult to process. It is seen in previous chapters that such unexpected artefacts are directly proportional to the underlying intelligibility. Hence, one might expect greater benefits of speech feature representations based on group delay spectrum for speakers with lower intelligibility. The outcome of this is reported in the following sections of this chapter.

The powerful properties of the group delay function can prove beneficial for an effective representation of signals. It has been advantageous for various signal processing related tasks like digital filtering and pole-zero decomposition (Murthy and Yegnanarayana, 1989). The high resolution property of the group delay function has been effective in estimating an accurate spectrum under noisy conditions (Yegnanarayana and Murthy, 1992) and robust formant extraction (Murthy, Murthy, and Yegnanarayana, 1989; Murthy and Yegnanarayana, 1991; Yegnanarayana, 1978). The benefits of the group delay function in speech technology are covered in detail in an article by Murthy and Yegnanarayana (2011).

6.1.1.2 Problem with Group Delay Function

Despite the advantages of the group delay spectrum, it however comes with a caveat, which can be detrimental for front-end processing in ASR. We know that a speech signal is characterised by its spectral envelope that results from the filter response of the vocal tract and

the fine harmonic structure due to the excitation source. The aim of any front-end processing is to disregard the effect of fine structure and encode the spectral shape. Typically speech is a mixed phase signal where poles are well within the unit circle and the zeros can be within or outside the unit circle.

The problem with the processing of group delay function is that if there are zeros which occur too close to the unit circle, these can result in huge spikes in the group delay spectrum. The spikes tend to dominate the spectral shape and shadow the true locations of the formants and this makes the spectrum not very useful for feature generation purpose. The occurrence of spikes in the spectrum results when the denominator term $|X(\omega)|^2$ in equation 6.3 gets smaller, i.e., when the distance between the zero location and the corresponding frequency bin on the unit circle reduces.

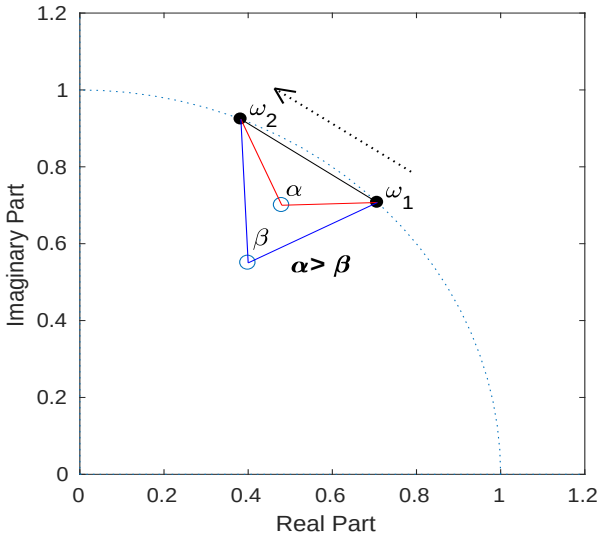

Figure 6.2: A simple illustration showing that the angle subtended by the chord joining two adjacent frequency bins is greater for the zero location near to the unit circle ($\alpha > \beta$).

A simple graphical illustration showing the main reason for spikes is shown in figure 6.2. We know from the properties of triangles that the largest interior angle is always opposite to the longest side. Hence, it is easy to assert that the angle subtended by the common chord joining the adjacent frequency bins (ω_1, ω_2) will be greater for zero locations nearer to the unit circle ($\alpha > \beta$). It thus leads to two possible conclusion; (i) rate of change of phase (equation 6.2) is greater for zeros near to the unit circle and (ii) greater angle (α)

will always have shorter distance to the frequency bins on the unit circle resulting in higher value of $|X(\omega)|^2$ (equation 6.3).

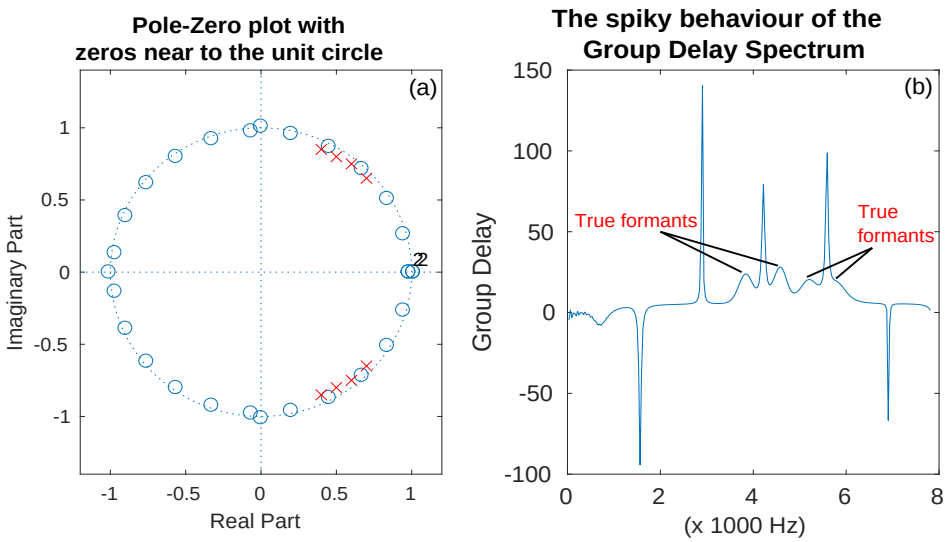

Figure 6.3: Spikes in the group delay spectrum due to introduction of random zeros around the unit circle. The spikes obscure the true location of the formants in the spectrum.

Figure 6.3 shows the effect of zeros near to the unit circle on the group delay spectrum. The pole-zero plot shown in part-(a) shows four closely spaced conjugate poles and random zeros are introduced around and near to the unit circle. It can be easily seen that the high resolution property of the group delay spectrum demonstrated earlier is now masked by unexpected spikes in the spectrum. The spikes tend to obscure the true location of the formants, thus making any meaningful interpretation of the phase spectrum difficult and misleading.

6.1.2 Modified Group Delay Function (MODGDF)

It was shown in the previous section that how the advent of spikes can make the processing of group delay spectrum difficult and inhibit its practical applicability. The spikes are introduced by the smaller values of $|X(\omega)|^2$ in equation 6.3. The modified group delay function (Murthy and Gadde, 2003) was formulated to reduce the effect of spikes in order to maintain the dynamic range of the spectrum. It was shown that by introducing $|S(\omega)|$, which is a cepstrally smoothed version of $|X(\omega)|$, very low values can be avoided in the denominator of equation 6.3. The modified group delay function is defined as:

$$\tau_{MODGDF}(\omega) = \left(\frac{\tau_X(\omega)}{|\tau_X(\omega)|}\right)\left(|\tau_X(\omega)|\right)^{\alpha} \tag{6.11}$$

where

$$\tau_X(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}} \tag{6.12}$$

where $S(\omega)$ is the cepstrally smoothed version (Yegnanarayana and Murthy, 1992) of $X(\omega)$. In addition, the parameters α,γ can be empirically controlled to reduce the effect of spikes in the modified group delay function.

6.1.3 Product Spectrum (PS)

The product spectrum is an alternate form of group delay representation that includes information from both the magnitude and phase spectrum. It is defined as the product of the group delay function and the power spectrum (Zhu and Paliwal, 2004) denoted as:

$$\begin{aligned} \tau_{PS}(\omega) &= |X(\omega)|^2\, \tau(\omega) \\ &= X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega) \end{aligned} \tag{6.13}$$

As a consequence of the definition of product spectrum, the denominator term of $|X(\omega)|^2$ that was responsible for the spikes in the group delay spectrum is cancelled out. This can be a useful representation, since it exploits the benefits of both the power spectrum and the phase spectrum without any need for applying smoothing techniques.

6.1.4 Cepstral coefficients based on MODGDF and PS representations

The MODGDF and PS representations of the group delay function will be used in the remainder of this chapter to extract the new speech features. It will thus be explored if the feature representations based on the phase spectrum are better at characterising dysarthric speech instead of the magnitude spectrum.

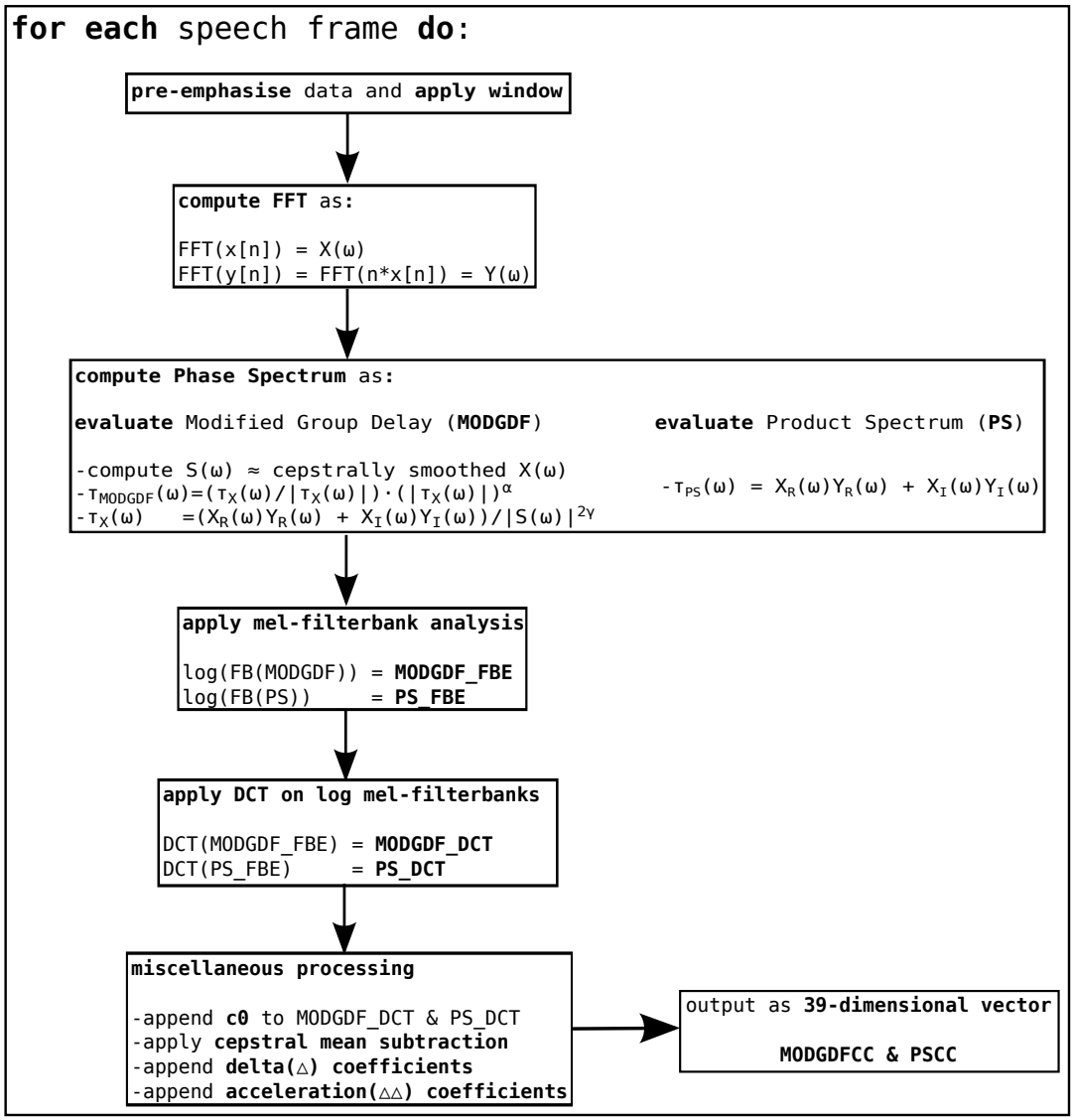

Figure 6.4: Main steps for the generation of phase based cepstral coefficients.

The main steps to generate the phase-based cepstral features is outlined in figure 6.4.

The empirical evaluation of the new representation is covered in the following sections. For MODGDF generation, there is a lack of any theoretical justification for the optimal values for α, γ. The smoothing parameters were thus determined by using the brute-force search with values tested between $0.05 - 0.95$ with a step increment of 0.05. The one that gave the best result was retained. For the experiments using MODGDF, the smoothing parameters were thus set at $\alpha = 0.95$ and $\gamma = 0.20$. A 26-band mel spaced triangular filters was used for the filterbank analysis. Lastly, the choice of window function was also carefully selected. It was observed that certain window operations affected the group delay spectrum to a greater degree by the introduction of spurious spikes, whilst others produced a much smoother spectrum, keeping the resolution of the formants intact. It was found that both Gaussian and Hanning-Poisson window produced the smoothest group delay spectrum and the later was chosen as the application window. Our choice of window function for phase-based spectrum generation has also been corroborated in other studies (Bozkurt, Couvreur, and Dutoit, 2007). The Hanning-Poisson window is defined as:

$$w(n) = \frac{1}{2}\left(1 - cos\left(\frac{2\pi n}{N-1}\right)\right) e^{\frac{-\alpha|N-1-2n|}{N-1}} \tag{6.14}$$

where α controls the exponential slope. For all the experiments reported later on phase features, the value was set to $\alpha = 2.5$.

MODGDF & PS based cepstral coefficients will be referred to as **MODGDFCC** and **PSCC** in the remainder of the thesis.

6.2 Phase based features for dysarthric speech

The conceptual understanding of the group delay spectrum along with its properties form a convincing and sufficient basis to extend the idea of phase based features for representing dysarthric speech. To the best of our knowledge, there is no work in the literature that explores the possibility of phase feature representation for evaluating the performance of ASR on dysarthric speech. This section will briefly examine some of the useful properties of a feature representation and compare some noticeable and important differences manifest in the magnitude and phase based speech features of dysarthric speech signals.

6.2.1 Frequency representation using phase spectrum

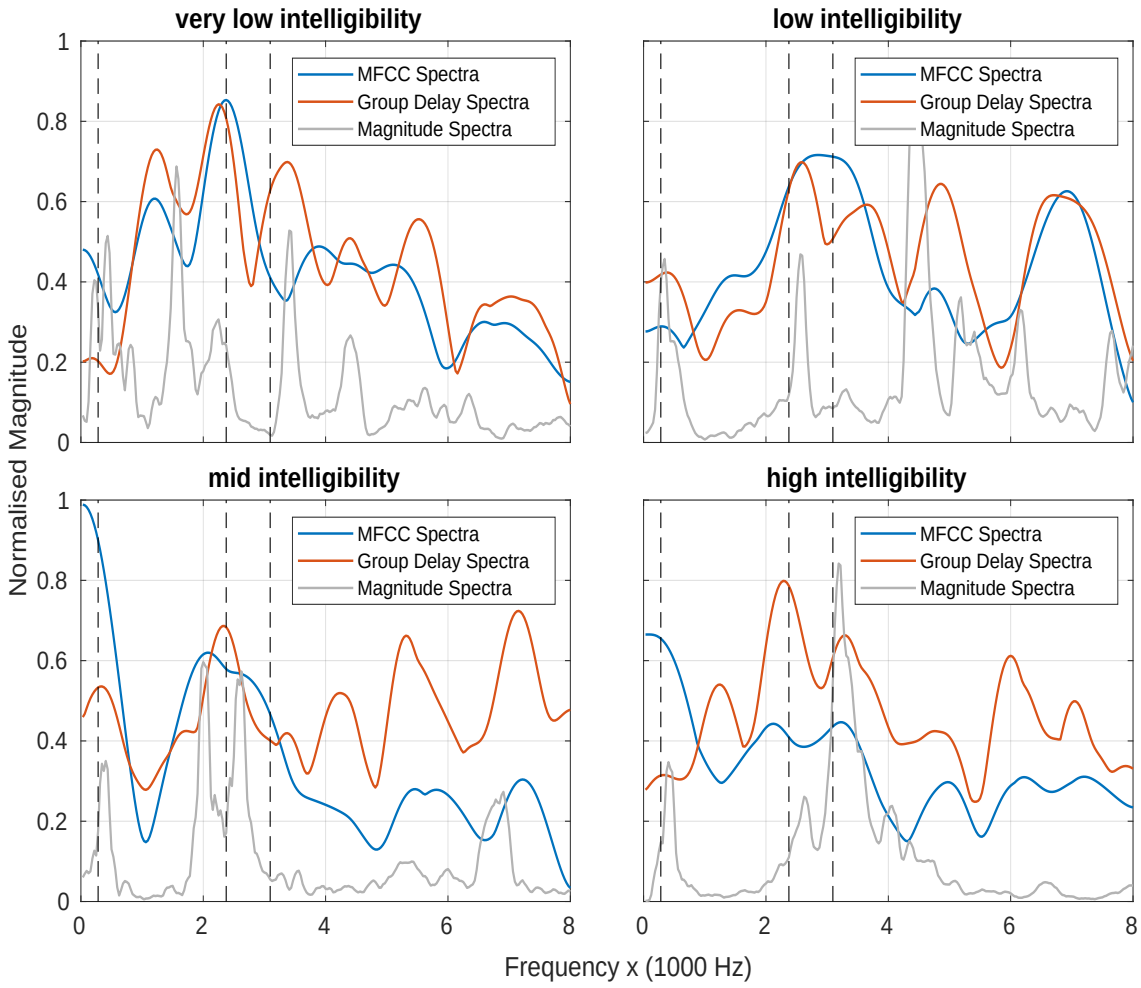

Figure 6.5: The figure shows the long term average spectra for the centre of the vowel /iy/ in the word **be**. It compares the MFCC and group delay based spectral representations along with the actual magnitude spectrum for a speaker chosen from each of the intelligibility groups in UASPEECH database. The comparison was done using around 1000 speech samples extracted from the centre of the vowel /iy/. The specific speakers selected for each intelligibility group are very-low→M04, low→F02, mid→M05, high→F05. The dotted line represents the approximate location of the first three formants for a typical vowel /iy/ given in Wells (1962).

For consistency, the analytical work conducted here uses the same set of files that were used in section 4.2.4 for the ZZT analysis of vowel tokens. The spectrum comparison was conducted for the front high vowel /iy/ in the production of the word **be**. A comparison between the MFCC and Group Delay based spectra is shown in figure 6.5 along with the actual magnitude spectrum for a speaker chosen from each of the intelligibility groups. Around 1000 samples of voiced segment was selected from the centre of the vowel /iy/ in generating the two spectra for each speaker. It can be easily seen in figure 6.5 that across all the intelligibility groups, the group delay spectrum shows a better correspondence of the expected peaks in comparison to the standard spectrum. For example, it can be seen for both mid and high intelligibility speakers, the expected locations for the first three formants (represented by dashed lines) are more closely aligned with the group delay spectra instead of the MFCC spectra. Since the distribution of formants is affected in speakers with lower intelligibility, a slight shift from the typically expected format locations is expected. Despite this, the peaks noticed in the magnitude spectrum show a higher degree of agreement to the group delay spectrum instead of MFCC for speakers with reduced intelligibility. For example, in the very-low intelligibility speaker the peak in the magnitude spectrum around 3 kHz matches with a peak in the corresponding group delay spectra, whilst the standard spectrum does not show the expected peak around this frequency region. A similar observation is also noticed for the low intelligibility speaker around 2.5 kHz, where the group delay spectrum shows a much finer resolution of the peak. It can be emphasised here that both the spectra are generated from the MFCC and PSCC based cepstral coefficients. The overall outcome is promising as it can be seen that the high resolution properties of the group delay spectrum discussed in section 6.1.1.1 can prove to be beneficial in the processing of disordered signals.

6.2.2 Better class separability

The aim of any feature representation technique for speech recognition is to capture sufficient discriminatory information about the individual phonetic tokens. However, generating such an optimal feature set can be more difficult in dysarthric speech due to the high degree of inter- and intra-speaker variability, data sparsity issues and malformed phonetic space (Blaney and Wilson, 2000; Kent et al., 2000; Morris, 1989). In the current section, the analysis of class (phonetic) separability is studied by examining certain aspects of speech production that are more likely to generate consonant articulation errors in dysarthric

speech. The most common distortions reported in the literature include detrimental effects to place of articulation, manner of articulation and voicing (Kent et al., 1990; Kim et al., 2010a; Riddel et al., 1995).

For example, in a study on 50 speakers with Cerebral palsy it was found that speech production errors were primarily noticed in voicing and place of articulation in comparison to manner of articulation (Platt et al., 1980; Platt, Andrews, and Howie, 1980). Another recent study on 7 native American speakers with Cerebral palsy and varying degree of intelligibility examined the presence of only voicing errors in speakers with high intelligibility and place, manner and voicing errors collectively present in lower intelligibility speakers. It has been examined that the place errors are mostly noted for labiodental, dental and alveolar sounds and manner category errors are more manifest in fricatives and affricates (Platt et al., 1980; Platt, Andrews, and Howie, 1980). In another study by Antolik and Fougeron (2013), place articulation errors due to incomplete closure of alveolar and velar stops was noticed for speakers with amyotrophic lateral sclerosis and devoicing of voiced consonants were found problematic in dysarthric speakers with Parkinson's disease and cerebellar ataxia.

It is beyond the scope of the current study to examine all the articulation and voicing errors and we limit our analysis to the aspects of speech production errors that are found to be more prominent in speakers with dysarthria, viz., place of articulation and voicing. In the current section, the analysis of class (phonetic) separability is studied by examining the following four sounds as shown in table 6.1.

Examined Tokens	**Token Type**	**Targeted Error**
/s/ , /sh/	Alveolar vs Post-Alveolar fricative	Place of articulation
/t/ , /d/	Voiceless vs Voiced Stop	Voicing

Table 6.1: Phonetic tokens examined for certain speech production errors.

The fricative sounds /s/ and /sh/ are examined in context of the following vowel /iy/ and the stop sounds /t/ and /d/ are examined in context of the following vowel /uw/. The example utterances of the word **see**:**she** are selected for the alveolar:post-alveolar fricatives and **two**:**do** are selected for the voiceless:voiced stops from the UASPEECH database for a speaker each from the four intelligibility groups. Both MFCC and phase group delay cepstral coefficients are examined for representing the above fricative and stop syllables. It is emphasised that only PSCC representation will be shown for brevity and MODGDFCC

gives similar results.

Since the cepstral representation is generally encoded in higher dimensional space (39 commonly), the problem is first approached as a dimensionality reduction task that will assist in data visualisation in a lower dimensional subspace. The efficacy of magnitude and phase based cepstral coefficients will then be examined as a clustering separation problem to discriminate between the underlying acoustic phonetic tokens ([**/s/, /sh/, & /iy/**] and [**/t/, /d/, & /uw/**] in our case). Principal Component Analysis (PCA) is used for the dimensionality reduction task. A complete mathematical description of the PCA is detailed in the book by Jolliffe (2002), however a simple summary is given here.

Let us consider a matrix $\mathbf{X}$ of MFCC or PSCC representation of speech with n feature observations and m cepstral coefficients given as:

$$\mathbf{X} = \left.\overbrace{\begin{pmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1m} \\ s_{21} & s_{22} & s_{23} & \cdots & s_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & s_{n3} & \cdots & s_{nm} \end{pmatrix}}^{m \text{ cepstral coefficients}}\right\} n \text{ features} \tag{6.15}$$

where s_{ij} is the j^{th} cepstral coefficient for the i^{th} feature. PCA uses an orthogonal transformation to reduce an n-dimensional space into an r-dimensional space where $r < n$. It reduces the dimension of the above matrix $\mathbf{X}$ by treating it as an eigendecomposition problem defined as:

$$\underbrace{(\mathbf{X}^T\mathbf{X})}_{\text{m x m matrix}} \mathbf{W} = \lambda \mathbf{W} \tag{6.16}$$

where $\mathbf{W}$ is the eigenvector, also known as "loadings" and λ is the diagonal matrix of eigenvalues that are used to describe the data. Each column of $\mathbf{W}$ represents a principal component that accounts for the variability in the data in decreasing order of importance. Hence, the first column of $\mathbf{W}$ will have the largest eigenvalue representing the greatest variance. In order to reduce the dimensionality, a transformation is computed on the original data matrix $\mathbf{X}$ as:

$$\mathbf{T} = \mathbf{X}\mathbf{W} \tag{6.17}$$

where **T** represents the PCA scores. As each column of **W** represents a principal component, in the current analysis we will only pick up the first two columns of the matrix that will result in a **two-dimensional** representation of the original cepstral features **X**.

The two dimensions are chosen for presentation purpose only as it will assist in visualising the representation of the two different cepstral representations (MFCC, PSCC) and compare its discriminatory capabilities for dysarthric signal.

Figures 6.6 - 6.9 show the two dimensional projection of the MFCC and PSCC features for the syllable fricatives **she** and **see** and the syllable stops **two** and **do** for dysarthric speech of varying intelligibilities. If we first look at the very-low intelligibility speaker (part-(a) of figure 6.6 and figure 6.8), it can be easily seen that the MFCC projection exhibits overlapping clusters across the entire length of the syllable fricative and stop. In addition to indistinguishable clusters between [/s/, /sh/, /iy/] and [/t/, /d/, /uw/], there is a marked zone towards the middle of the utterances that shows the effect of coarticulation that is not easily discernible for the respective two sounds. Since the features are not discriminatory in the MFCC representation, it can make the acoustic modelling task a challenge resulting in incorrect clustering of Gaussian distributions, especially in an HMM-GMM system. In contrast, the PSCC representation for the same speech shows much better discriminatory capabilities. It shows the presence of more tightly bound clusters which are easily separable in the acoustic space. Also, the coarticulatory effect of the two syllable fricatives and stops tend to be non-overlapping and are well defined within its own domain.

The above explanation for overlapping clusters and coarticulation for the MFCC representation seems to extend for the low (part-(b) of figure 6.6 and figure 6.8) and mid (part-(a) of figure 6.7 and figure 6.9) intelligibility speakers too. It can be easily seen that the PSCC representation of the same utterance shows much better discriminatory capabilities by representing tightly bound clusters for the phones [/s/, /sh/ & /iy/] and [/t/, /d/ & /uw/], which are easily separable in the acoustic space. Although the difference between MFCC and PSCC representation seem to reduce for the high intelligibility group (part-(b) of figure 6.7 and figure 6.9), it can still be observed that the PSCC clusters for /s/ and /sh/ exhibit non-overlapping clusters in contrast to MFCC that still shows a noticeable overlap between the cloud of distinct points.

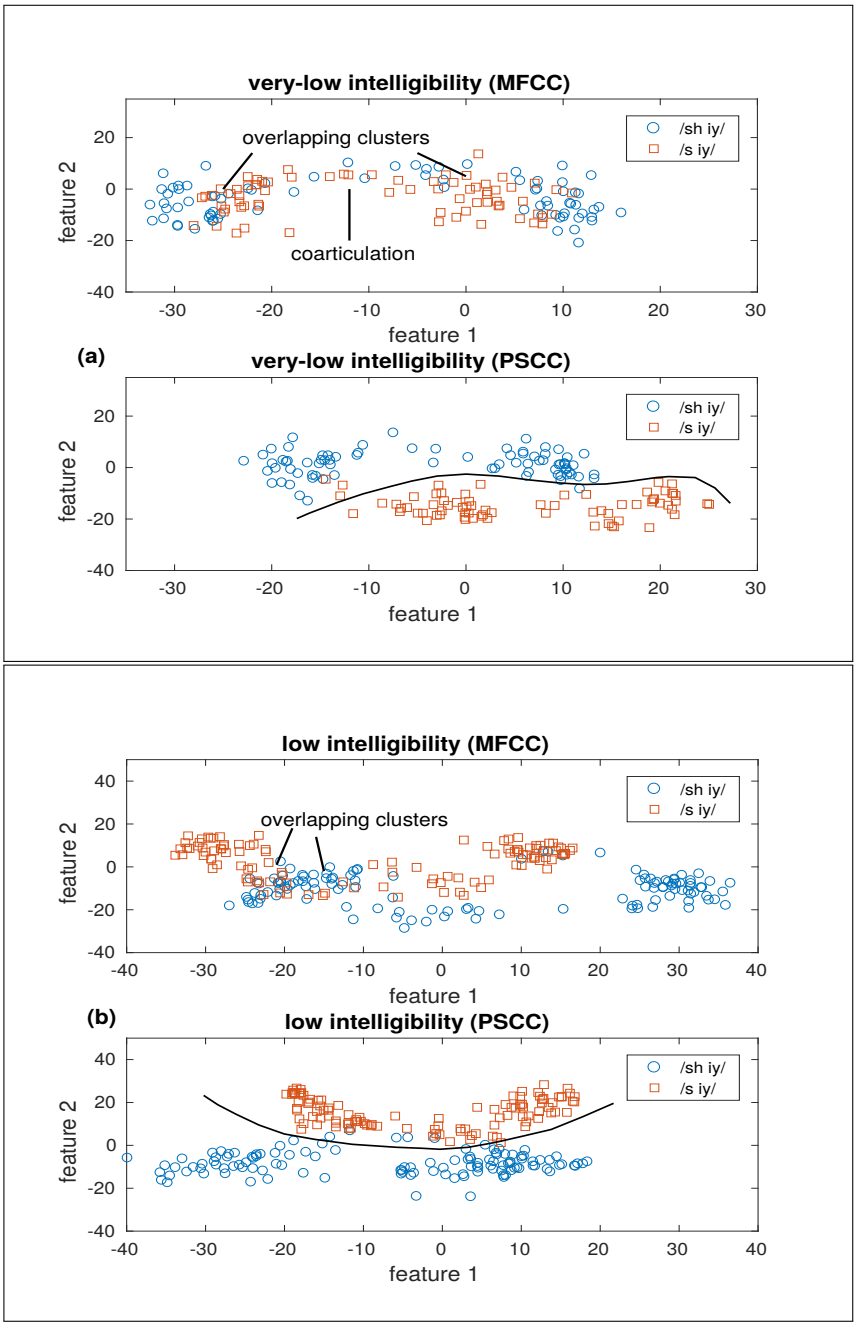

Figure 6.6: The above figures shows a two dimensional PCA representation for the MFCC and PSCC features derived from the syllable fricatives **she** & **see**. The plots are shown for a speaker each from the (a) very-low and (b) low intelligibility groups where each point represents a frame.

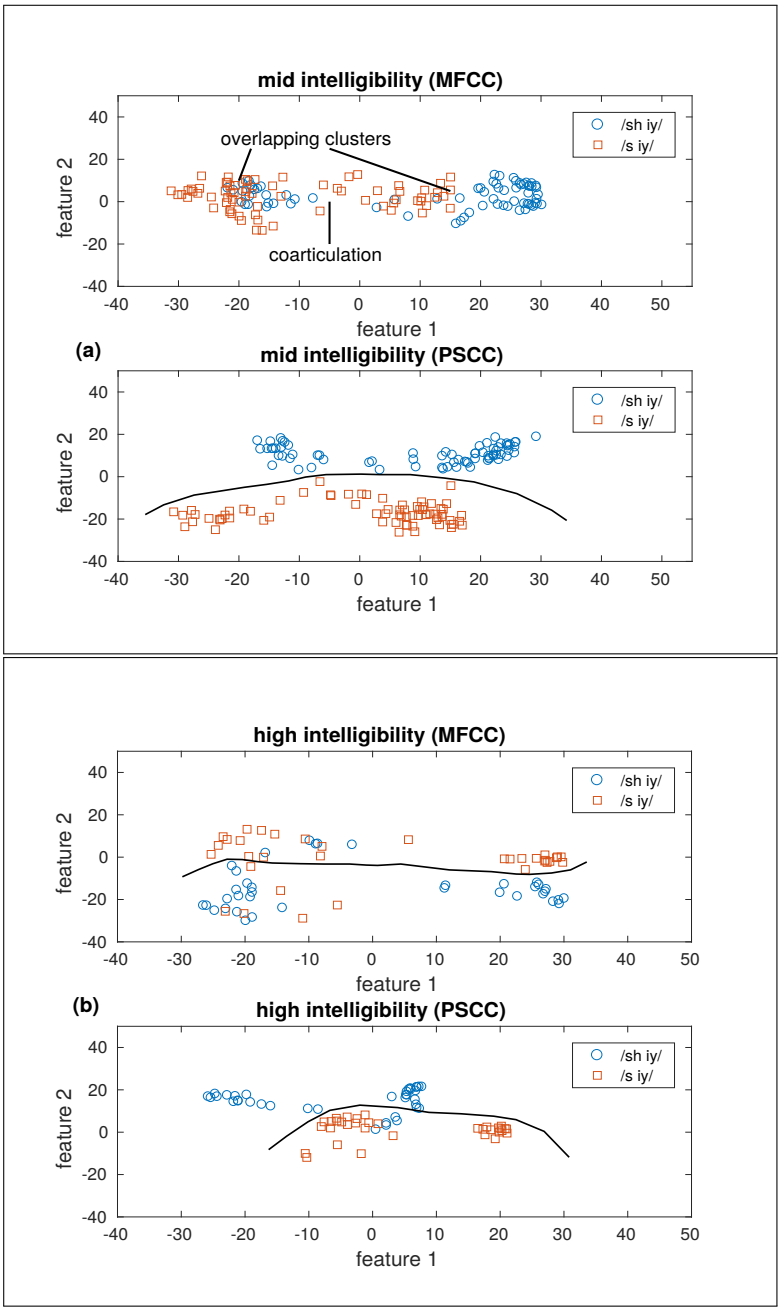

Figure 6.7: The above figures shows a two dimensional PCA representation for the MFCC and PSCC features derived from the syllable fricatives **she** & **see**. The plots are shown for a speaker each from the (a) mid and (b) high intelligibility groups where each point represents a frame.

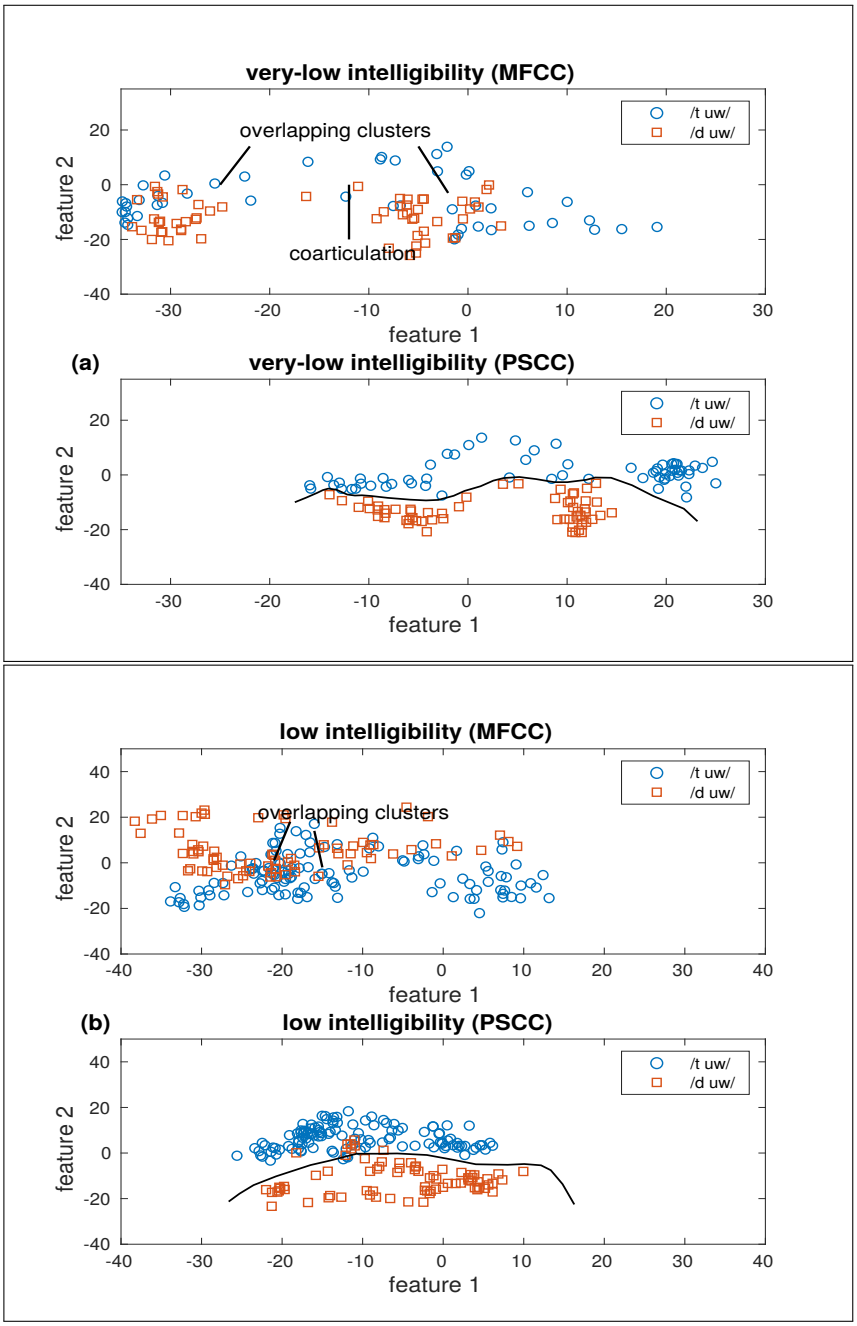

Figure 6.8: The above figures shows a two dimensional PCA representation for the MFCC and PSCC features derived from the syllable stops **two** & **do**. The plots are shown for a speaker each from the (a) very-low and (b) low intelligibility groups where each point represents a frame.

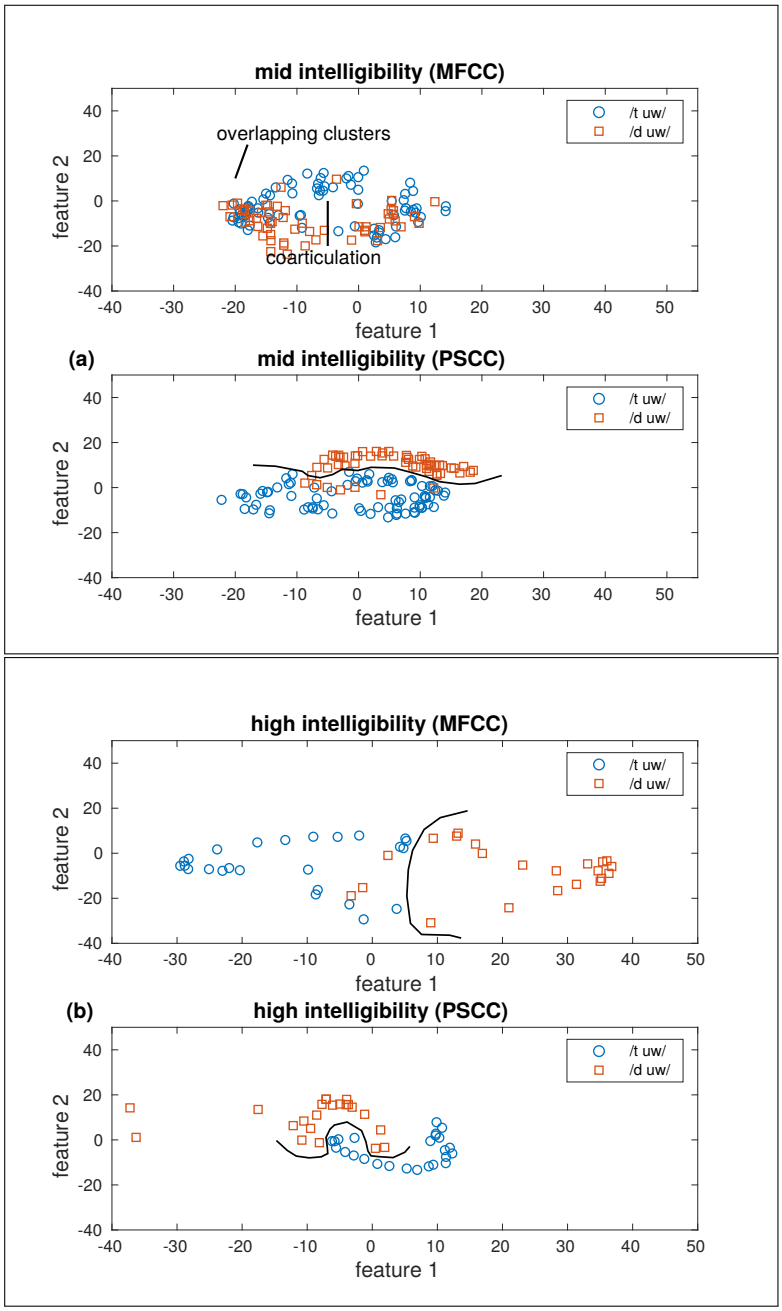

Figure 6.9: The above figures shows a two dimensional PCA representation for the MFCC and PSCC features derived from the syllable stops **two** & **do**. The plots are shown for a speaker each from the (a) mid and (b) high intelligibility groups where each point represents a frame.

It is observed in the previous presentation and discussion that the PSCC tends to exhibit better discriminatory capabilities for representing unique phoneme clusters in comparison to the standard MFCC based representation. Due to better discrimination, it was also able to show promising alternatives that might be capable of handling certain speech production errors, like place of articulation and voicing that are more exhibit as a characteristic of dysarthric speech. One of the problem with MFCC is its sensitivity to noise that can inadvertently introduce ripples in the spectral valley and degrade performance (Tyagi and Wellekens, 2004; Zhao and Wang, 2013). Group Delay based cepstral representation show better robustness to convolution & white noise and is more suited for inter-speaker class separations (Hegde, Murthy, and Gadde, 2007; Murthy, Hegde, and Rao, 2004). In the current study, we have exploited the idea of phase based spectrum to the analysis for dysarthric signals and find it to be a promising alternative for better phoneme discriminatory capabilities for disordered speech. It implies that if dysarthric speech can be regarded as a noisy channel that is often convolved with a wide variation of non-speech sounds, PSCC and MODGDFCC based cepstral representation might prove more effective than the standard MFCC.

In the current section two important properties of feature representation have been explored in context of disordered speech signals. The **high resolution** aspect of the group delay spectrum was explored in section 6.2.1 that shows promising signs to better model the resonances of the vocal tract of dysarthric speech. It was also shown in section 6.2.2 that PSCC/MODGDFCC representations of disordered signal were more coherent and optimal at defining phonetic clusters in the acoustic space, thus exhibiting refined **class separability** property for better characterising the underlying speech. There is sufficient theoretical and practical evidence to suggest the beneficial aspects of the phase based spectrum. In the following sections the effect of the group delay representation will be empirically evaluated on dysarthric speech by measuring its relative ASR performance on various speech systems.

6.3 Experiments on phase vs magnitude based features of dysarthric speech

In this section experiments will be conducted to compare the performance between the standard MFCC and phase based PSCC & MODGDFCC representations.

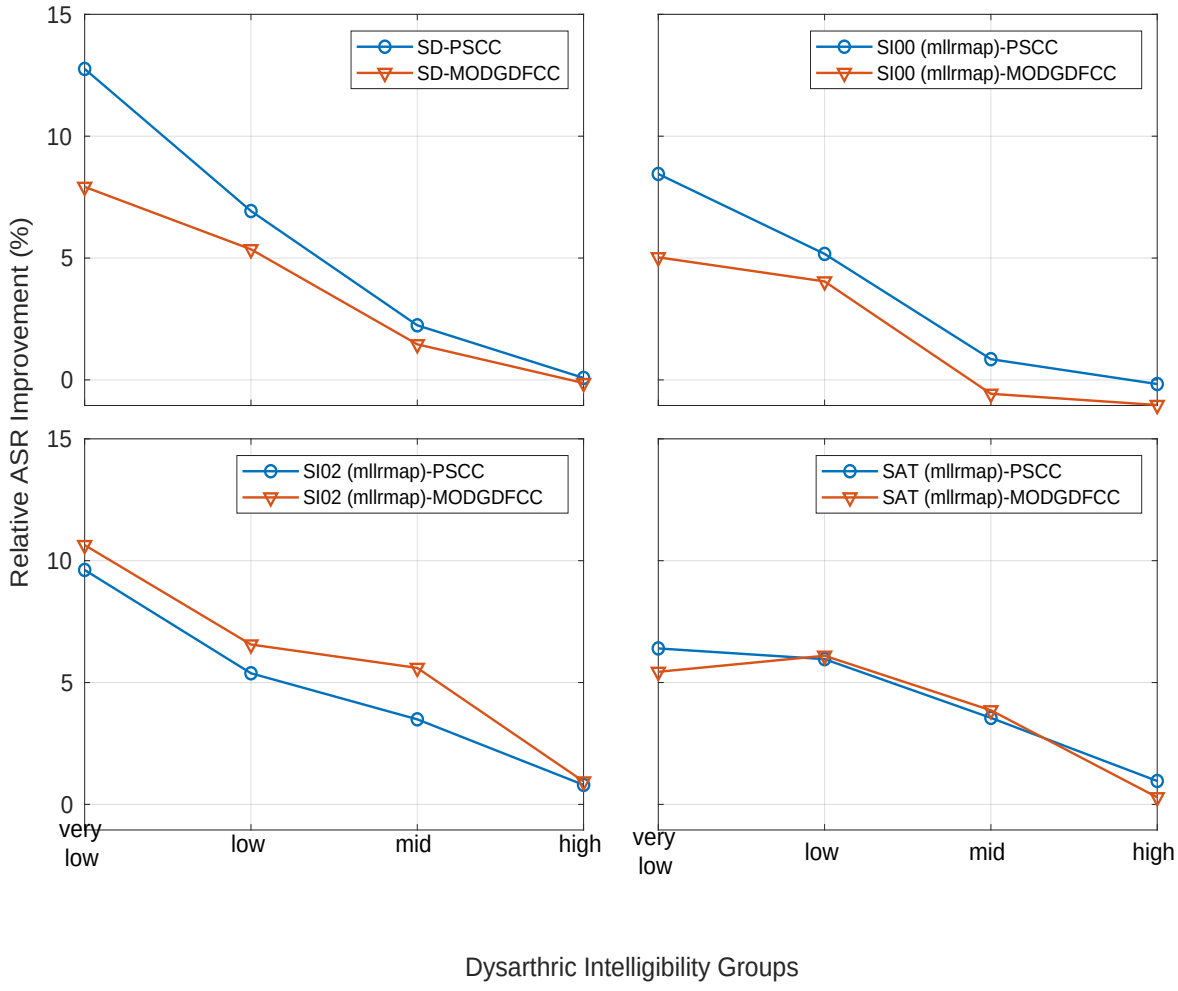

Figure 6.10: Relative ASR gains in comparison to the baseline results presented in section 4.1.2 for the phase based feature representation of dysarthric speech. The figures show the results for PSCC and MODGDFCC based cepstral coefficients.

The results are shown in figure 6.10, where the x-axis is representative of the standard

MFCC results presented in table 4.3. The plots show the relative ASR gains across the four tested speech systems (SD, SI-00 (mllrmap), SI-02 (mllrmap), SAT (mllrmap)). It is evident that the phase based feature representations of dysarthric speech show gains across all the tested systems. Both PSCC and MODGDFCC feature representations are highly effective for modelling dysarthric speech with greater degree of pathological disorder in comparison to standard magnitude based MFCC representation. The relative benefit to ASR performance is reduced in less severe cases. The outcome of the experiment tends to coincide with the earlier PSD findings for predicting intelligibility or improving ASR performance, where PSD corrections were found to be more effective for dysarthric speech with lowest intelligibility.

Intelli-gibility	**PSCC Features**				**MODGDFCC Features**			
	SD	**SI-02**	**SAT**	**SI-00**	**SD**	**SI-02**	**SAT**	**SI-00**
very-low	23.52	27.36	28.71	20.61	23.52	27.36	28.71	20.61
	26.52 ††	30.00 ††	30.55 ††	22.36 ††	25.33 ††	30.27 ††	30.28 ††	21.65 †
low	62.48	62.92	62.98	57.89	62.48	62.92	62.98	57.89
	66.81 ††	66.30 ††	66.72 ††	60.89 ††	65.82 ††	67.05 ††	66.83 ††	60.23 ††
mid	64.08	68.51	69.54	66.12	64.08	68.51	69.54	66.12
	65.52 †	70.90 ††	72.02 ††	66.69	65.02	72.34 ††	72.23 ††	66.23
high	83.07	86.17	86.87	87.08	83.07	86.17	86.87	87.08
	83.14	86.86 †	87.71 ††	86.93	82.96	86.98 †	87.12	86.19

Table 6.2: Absolute ASR word accuracy averaged by various intelligibility groups. The top number in each cell represents the best baseline results presented earlier in table 4.3 using standard MFCC features. The shaded number is the result of using phase based feature representation for the MFCC's. Significant statistical gains are shown using a † ($p < 0.05$) or †† ($p < 0.01$).

In order to investigate the benefit to ASR performance of phase based features, a pairwise Cochran's Q test was conducted for MODGDFCC/Standard-MFCC and PSCC/Standard-MFCC feature representations. Table 6.2 shows the absolute ASR scores for the two feature representations. The cells that exhibit significant gains are marked with a †† ($p < 0.01$) or † ($p < 0.05$).

Out of the 16 possible combinations between the four systems and intelligibility groups,

PSCC based feature representation shows significant gains in 13 systems and MODGDFCC based feature representation shows significant gains in 12 systems. It is noteworthy that for both the feature representations, all the systems showed highly significant gains for the *very-low* and *low* intelligibility groups. This is an encouraging outcome, since the majority of dysarthric speech systems are primarily targeted to benefit users with a high degree of speech disorder. Hence, feature representation based on group delay spectra of pathological speech can prove to be significantly beneficial for robust acoustic modelling.

It was noted that PSCC was significantly better (†) at speaker dependent modelling over MODGDFCC for speakers with lowest intelligibility. The selection between PSCC and MODGDFCC seems to be a matter of choice and can be dependent on particular applications. MODGDFCC also comes with an additional constraint of finding optimal values of (α, γ), which can be dependent on the underlying dataset, whereas, PSCC is free from such constraints and can benefit from the information in both the magnitude and phase spectrum.

6.4 PSD enhanced phase based feature representation for dysarthric ASR

It was seen in chapter 5 that PSD was not only effective at quantitatively predicting the underlying intelligibility, but a systematic correction of PSD was beneficial to the ASR performance on dysarthric speech. In the previous section it was shown that the group delay based feature representations were significantly better at characterising the acoustics of the dysarthric speech in comparison to the standard MFCC. The two set of experiments might look unrelated, however, both utilise a common source of information that emanates from the phase component of the Fourier transformations. The former uses deviations of the continuous phase spectrum and the latter exploits the group delay phase spectrum with a common goal of improving dysarthric ASR performance.

Hence the experiments outlined in this section are a logical extension of our previous work that will combine the beneficial properties of PSD correction and PSCC/MODGDFCC based feature representations. It is hypothesised that if phase based features are generated on speech utterances with corrected PSD slopes, then the average effect might have a favourable impact on the dysarthric ASR performance.

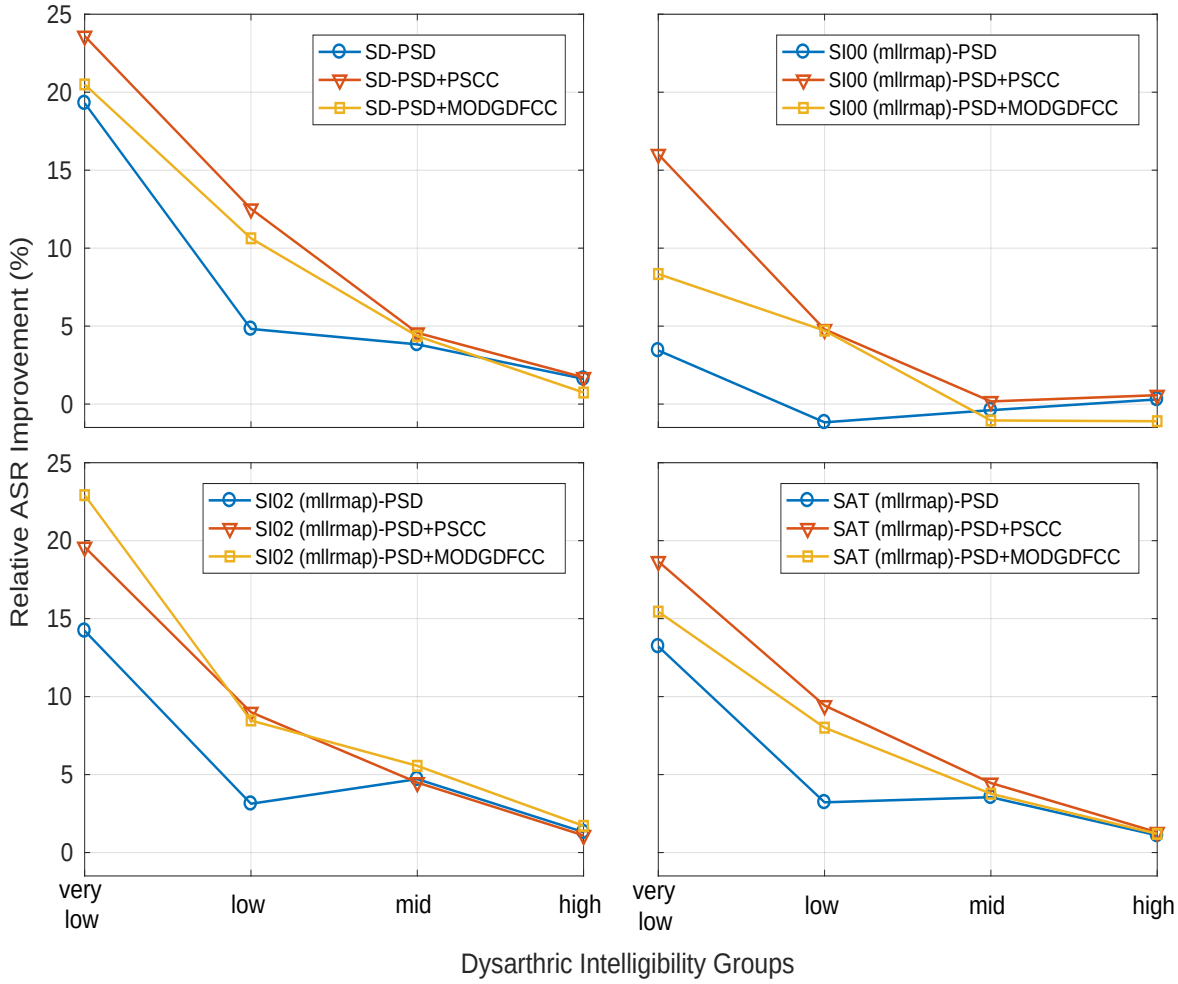

Figure 6.11: Relative ASR gains for the PSD corrections on standard MFCC along with PSD enhanced phase based feature representation (PSCC, MODGDFCC) of dysarthric speech.

The results are shown in figure 6.11, where the x-axis is representative of the baseline results presented in table 4.3. The PSD results are as presented in section 5.3.2.1 for the supervised correction mode. The plots show the relative ASR gains across the four tested speech systems (SD, SI-00 (mllrmap), SI-02 (mllrmap), SAT (mllrmap)). It is evident that the combined effect of **PSD + PSCC/MODGDFCC** for dysarthric speech shows noticeable gains in comparison to the individual benefit of PSD or PSCC/MODGDFCC application seen earlier. As it was anticipated, the outcome of the combined approach

Intelligibility	PSD + PSCC Features				PSD + MODGDFCC Features			
	SD	**SI-02**	**SAT**	**SI-00**	**SD**	**SI-02**	**SAT**	**SI-00**
very-low	28.06	31.26	32.54	21.32	28.06	31.26	32.54	21.32
	29.07 †	32.72 ††	34.08 †	23.92 ††	28.34	33.64 ††	33.16	22.33 †
low	65.49	64.89	65.01	57.22	65.49	64.89	65.01	57.22
	70.30 ††	68.38 ††	68.92 ††	60.67 ††	69.12 ††	68.25 ††	68.03 ††	60.73 ††
mid	66.54	71.74	72.02	65.86	66.54	71.74	72.02	65.86
	67.01	71.58	72.65	66.23	66.88	72.32	72.17	65.43
high	84.42	87.29	87.83	87.34	84.42	87.29	87.83	87.34
	84.49	87.12	88.00	87.58	83.09	87.04	87.92	86.12

Table 6.3: Absolute ASR word accuracy averaged by various intelligibility groups. The top number in each cell represents the best results of the PSD correction presented in section 5.3.2.1 using standard MFCC features. The shaded number is the result of using PSD correction with phase based feature representation for the MFCC's. Significant statistical gains are shown using a † ($p < 0.05$) or †† ($p < 0.01$).

tends to follow similar pattern of the earlier experiments and looks highly effective for modelling dysarthric speech with greater degree of pathological disorder.

In order to investigate the ASR performance of the combined approach, a pairwise Cochran's Q test was conducted for PSD/PSD+PSCC and PSD/PSD+MODGDFCC approaches. Table 6.3 shows the absolute ASR scores for the two feature representations. The cells that exhibit significant gains are marked with a †† ($p < 0.01$) or † ($p < 0.05$). There were highly significant gains noted in nearly all the speech systems tested under the *very-low* and *low* intelligibility groups. Once again the results present an encouraging outcome with notable impact on the ASR performance of highly severe dysarthric speech with marginal benefits for the *mid* and *high* intelligibility groups.

6.5 Conclusion

The work in this chapter is inspired from the ZZT and PSD applications presented in earlier chapters. It has shown that phase of a signal might preserve useful acoustic information necessary for improved ASR performance. The phase information was extracted in the form of the group delay spectrum from the phase component of the Fourier transformations

and it was encoded using cepstral coefficients in the form of **MODGDFCC** and **PSCC** representations. The justifications presented in this chapter argued that such phase based representation of dysarthric speech is motivated by a strong theoretical framework that gives compelling evidence to represent such signals in these alternative forms. For example, the extremely important **high resolution** and **class separability** properties of the group delay spectrum extended these beneficial effects to the recognition of dysarthric speech. The properties of the group delay spectrum showed strong evidence that phase based feature representations are more suited to characterise the resonances of the vocal tract, and exhibited better phone discrimination capabilities in dysarthric signals.

Despite significant ASR results and compelling evidence that phase-based feature representation is more amenable for dysarthric signals, it still leaves us with important questions, such as how these alternate features represent the signal that leads to higher recognition accuracy for dysarthric speech.

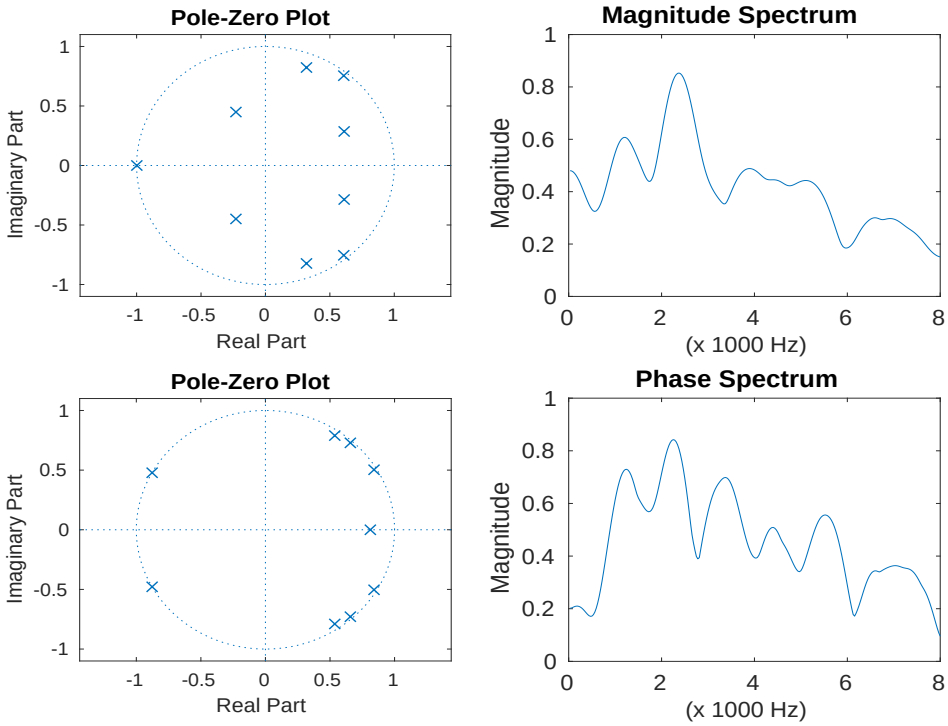

Figure 6.12: Pole locations of the magnitude and phase spectra for the vowel /iy/.

One of the ways in which this question might be answered is by looking at poles of

the magnitude and phase spectra to see if they give any informative cues. By way of an example, figure 6.12 shows the approximate pole locations inferred from the two spectra for one particular instance of the vowel /iy/ for a dysarthric speaker. This suggests that the poles of a phase spectrum might be closer to the unit circle, hence reflecting the prominent peaks in the spectrum. This aspect could help to preserve the acoustically relevant features for distinguishing vowels. In order to substantiate this observation, further research is needed to explore appropriate methods for inferring accurate pole locations, and then to apply these over multiple utterances generated by multiple speakers.

In the current chapter, the first set of experiments empirically corroborated the efficacy of phase based cepstral coefficients over the standard MFCC features. Both PSCC and MODGDFCC representations showed statistically significant gains in nearly all the speech systems and intelligibility groups. It is important to highlight that the results were highly significant ($p < 0.01$) for the *very-low* and *low* intelligibility groups and SD based speech systems showed maximum performance benefit. The outcome is in agreement with the PSD findings of chapter 5 that also favoured speakers with lowest intelligibility and performed best for SD systems. Another interesting observation was the outcome of the SI-00 (prepared from large amount of typical speech) system. It exhibited significant gains using either phase representations, which is an encouraging result as it showcases the importance of PSCC/MODGDFCC over MFCC based representation, whilst using disparate and heterogeneous datasets for dysarthric speech recognition.

The individual work conducted on both PSD and phase representations of dysarthric signals gave significant improvements across various speech systems and intelligibility groups. The last set of experiments used in this study explored the compound impact of PSD and PSCC/MODGDFCC work. In this study the PSD corrections were first applied to the dysarthric speech that was later featured using PSCC/MODGDFCC representations. The results of this composite approach were compared against the best PSD results of chapter 5 that used vowel-specific transforms for the PSD corrections. There were performance benefits across all the speech systems and intelligibility groups with highly significant gains for the *very-low* and *low* intelligibility groups. Hence, the combined approach can accentuate the beneficiary properties of PSD and PSCC/MODGDFCC to maximise the ASR performance for speakers where it is most needed.

The main phase-related work conducted in this thesis are based on PSD corrections, PSCC/MODGDFCC representation and the amalgamation of PSD + PSCC/MODGDFCC.

All three approaches exhibited a prominent pattern in their implementation, i.e., the maximum relative benefit was projected for the SD speech systems and favoured for speakers with severe dysarthria.

Chapter 7

Discussions and Future Work

The thesis begins by providing a background of dysarthria from an anatomical perspective and discussed the broad categories of such neurological speech impairments. It was apparent from research reports that there is an ever growing need to improve human-to-machine interaction for people with dysarthria and speech was classified as one of the principal medium to provide a natural and faster mode of interaction.

The study systematically explored standard state-of-the-art ASR techniques for model training and adaptation. It was determined that hybrid adaptation approaches like MLLR-MAP was better than MLLR and MAP only adaptations for modelling of dysarthric speech. SAT based training, which has the inherent capability to reduce the inter-speaker variabilities were statistically more effective to model dysarthric speech with low to mid level of intelligibility. Since dysarthric speech with high intelligibility was found to be more similar to typical, it was determined that ASR systems prepared from typical data with homogeneous vocabulary and recording conditions (SI-03) were statistically more effective than ASR systems trained using a typical database with disparate acoustic profile (SI-00). The results of the ASR systems described in chapter 4 on the UASPEECH database was found to be statistically better than any of the earlier published results in the literature. The results, thus formed the baseline for comparison with all the proposed approaches in the thesis for improving ASR performance.

The thesis also explored conventional acoustic analysis approaches to understand the characteristics of dysarthric speech and its relationship to intelligibility. Temporal measures like sypse and VOT indicated more than two times slower speaking rate and exhibited greater variability for dysarthric speakers with reduced intelligibility. The analysis further

pointed towards the presence of inter-speaker variabilities and speech production errors, such as phonemic errors, manifest in dysarthric speakers. In addition, frequency based analysis was also conducted to study the F1-F2 plane. The standard metric of F1-F2 area was examined along with two new measures of shape and displacement. The area was computed using the log ratios between the formant profiles of dysarthric and control speakers and was termed as the compression factor (CF). F1-F2 shape was classified as being convex, concave or flipped and displacement was introduced to evaluate the directional shift of the F1-F2 quadrilateral space. It was observed that a high CF and displacement value along with a concave/flipped arrangement of vowels in the F1-F2 space indicated reduced phonatory discrimination with overlapping vowel tokens. This behaviour was generally more evident in speakers with reduced intelligibility.

The thesis further explored a new analytical approach to find acoustic evidence in dysarthric speech that conveys a functional association with the intelligibility of the underlying speech. The acoustic analysis informed the development of a new metric that was beneficial for computing quantitative estimates about the intelligibility of dysarthria. The outcome measures of the metric was systematically exploited to improve the ASR performance of various dysarthric speech systems.

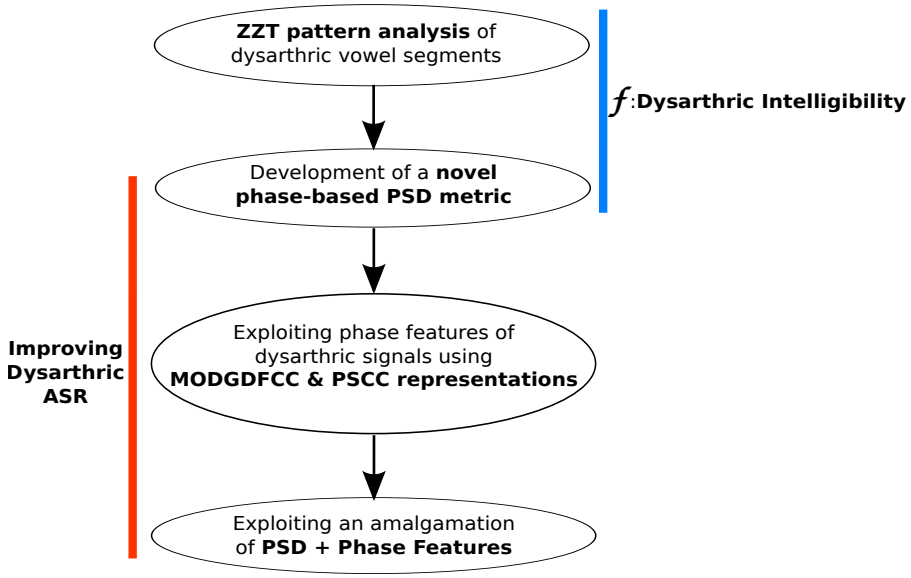

Figure 7.1: Proposed framework for predicting intelligibility and improving ASR performance.

The study proposed a series of coherent approaches that are outlined in figure 7.1 and discussed below along with an overall summary of results.

ZZT and PSD analysis of dysarthric speech

A new approach for the acoustic analysis of dysarthric speech was investigated by studying the Z-transform of a time-domain dysarthric vowel segment that was windowed using the Hanning-Poisson function. It was displayed as a two dimensional plot of the **ZZT (Zeros of the Z-Transform)** patterns that emerge from the underlying complex roots. The expected typical speech pattern usually manifests itself in three distinct zones in the z-plane, where the zeros above the unit circle represents the glottal pulse, zeros below the unit circle pertain to the vocal tract filter response and zeros around the unit circle correspond to the impulse train. The distribution of zeros was observed to be in nearly the typical range for the *mild* and *high* intelligibility group of dysarthric speakers, however it showed a highly skewed mapping of zeros for the *low* and *very-low* intelligibility speakers. One of the factors contributing to the skewed distribution of zeros was found to be related to some underlying phase based acoustic event that was more prominent in speakers with lower intelligibility. This understanding was later confirmed when the unwrapped phase component of the complex roots was plotted. The slope of the unwrapped phase for a speaker with lower intelligibility showed significant deviation relative to the slope of a typical or high intelligibility dysarthric speakers.

The above observations suggested the possibility of a phase related phenomenon that might encapsulate some underlying dysarthric artefact. The idea of phase deviations was further extended in chapter 5. The study proposed a new metric, called the **Phase Slope Deviation (PSD)** that was used to examine the effect of phase deviations in the dysarthric vowel segments. PSD uses the unwrapped phase component of the Fourier transform of any discrete signal under consideration to compute the deviation estimate. The metric was computed on two disparate data sources, viz. UASPEECH and VIVOCA, and it exhibited a strong and nearly linear relationship between the acoustically derived PSD scores and underlying dysarthric intelligibility. It was also found to be very effective under sparse and missing data conditions. For example, it was found that nearly 23% of the vowel tokens had no data for the VIVOCA corpus when examined for 13 speakers across 13 vowel tokens under consideration. Despite the missing information, PSD predicted the expected intelligibility range of VIVOCA speakers in the expected zones for the perceptual correlates of

intelligibility. Also, since PSD has no pre-involvement with any particular data source, it should be independent of any particular corpus. The proposed metric also displays a functional association with the underlying dysarthric intelligibility. If a fair degree of acoustic data is available for a dysarthric speaker, PSD can predict an approximate intelligibility zone that can aid in better dissemination of any speech based therapy. Alternatively, if the intelligibility of a speaker is known, an approximate PSD score can be computed, which can help to devise better acoustic or pronunciation modelling techniques by examining possibly homogeneous speakers with a similar acoustic profile.

Improving Dysarthric ASR

It has been evident from the ZZT and PSD based analysis that a phase related phenomenon seems to be intrinsically embedded in the acoustics of the dysarthric speech and that it has a strong correspondence to the underlying speech intelligibility. In the current study a series of phase motivated methodologies were suggested with each having its own merits for improving the ASR performance. A summary of the proposed methods are:

- PSD Corrections: The PSD metric was used to give a quantitative estimate of the phase deviations that are manifest in a dysarthric vowel segment. As the deviation was observed to a greater degree in speech with lowest intelligibility, corrective methods were applied to the dysarthric vowel segments to reduce the overall effect of PSD. The corrections were applied in a *supervised, semi-supervised and unsupervised* mode that differed in the amount of prior information that was available for applying a suitable vowel phase modification.

- Phase Based Features: This work was inspired by the ZZT and PSD analysis of the dysarthric vowel segments. It was based on the notion that the phase component of a Fourier transform of disordered speech has some inherent property that might convey or encode important acoustic cues relevant to the underlying intelligibility. Cepstral features were thus prepared from the group delay spectrum and extracted using the modified group delay (MODGDFCC) and product spectrum (PSCC) functions. The phase cepstral representation of dysarthric speech showed better **formant resolving** and **class separability** properties relative to the cepstral features extracted from the magnitude spectrum. The effect of phase based features on dysarthric ASR is an independent approach to improve the performance of dysarthric speech recognition

systems. It has no direct relationship to the PSD corrective measures applied for improving the ASR performance.

- PSD + Phase Based Features: Both PSD and phase based features (MODGDFCC, PSCC) of dysarthric speech are inspired from the potential information that might be encapsulated in the phase component of a speech signal. The former exploits the relationship between the phase deviation and severity of dysarthric speech and the later exploits the theoretical and practical properties of the phase spectrum that show better potential to represent dysarthric speech. This is an amalgamative approach that utilises the beneficial aspects of the otherwise two independent techniques for improving the ASR performance of dysarthric speech.

It is reiterated that the SD system refers to a speaker dependent system, SI-02 uses all the dysarthric data for preparing the base model for speaker specific adaptation, SI-00 uses typical speech data from the WSJ SI-84 and WSJCAM0 corpus for preparing the base model for speaker specific adaptation and the SAT model uses the speaker adaptive training regime with SI-02 models as the starting canonical model set. The baseline results produced in section 4.1.2 gives an absolute gain of 11.05% (20.42% relative) over the last published best results in the literature that was evaluated on a large dysarthric vocabulary set of 255 competing words (Christensen et al., 2012). The baseline results have already shown the significance of using hybrid adaptation techniques where SAT systems were especially important for modelling speech with decreasing intelligibility.

A collective summary of the results applying the proposed methods in this study is compiled in figure 7.2. All the suggested methods have shown significant benefits relative to our baseline results, albeit, the combined approach of using the PSD corrections along with the phase features produce the best overall gains in nearly all the systems and intelligibility groups. It should be noted that all the proposed methods are the most effective for the speakers with lowest intelligibility. This can be easily seen from the y-axis scale of individual graphs, which is approximately halved for each intelligibility group from very-low to high. This is a very promising outcome from a practical application perspective. Since most of the dysarthric speech systems are targeted for speakers with lowest intelligibility, the suggested methods can prove highly beneficial in situations where real-time improvement is most needed. Out of the four speech systems presented in this study, the speaker dependent and the dysarthric data based SI-02 and SAT systems are the most productive from a practical

and performance perspective.

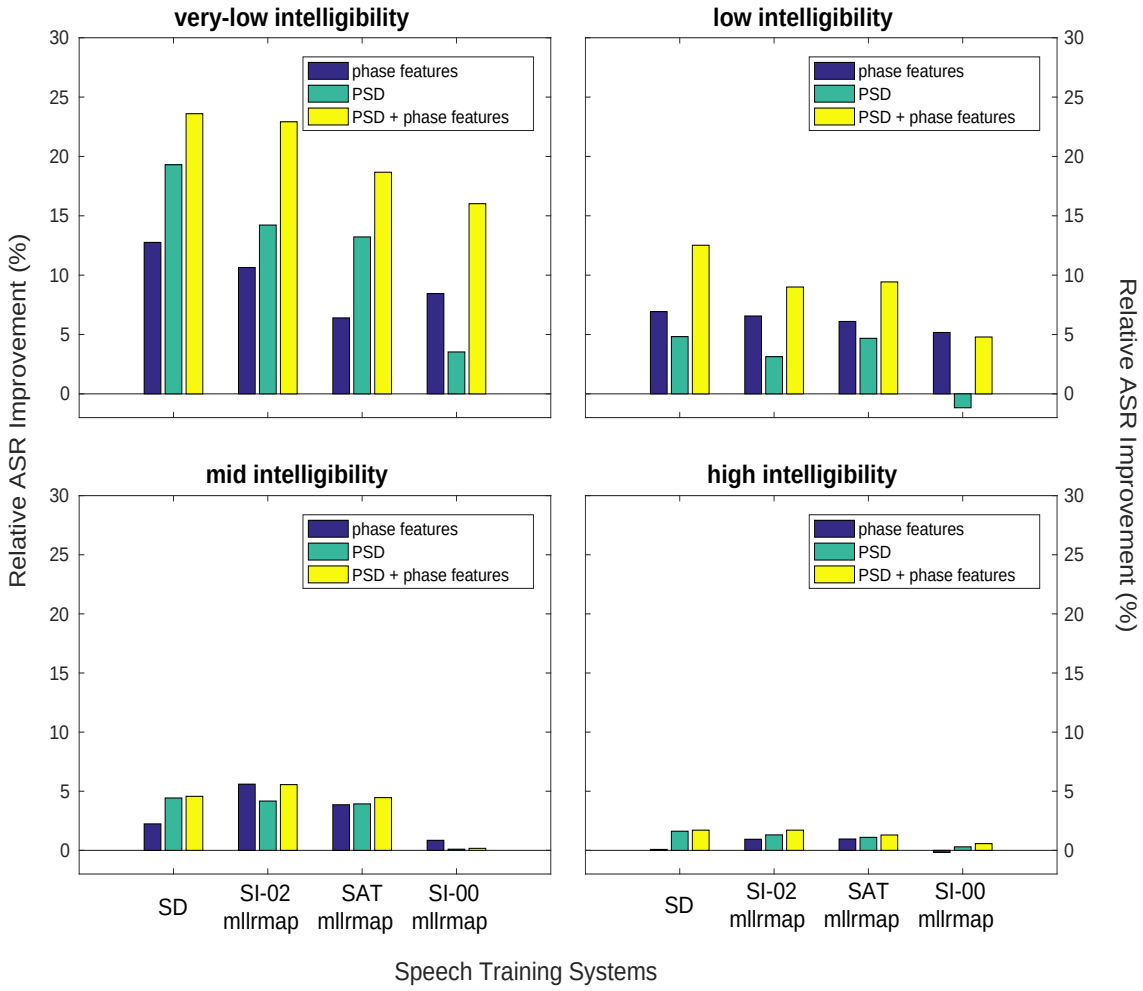

Figure 7.2: ASR gains relative to the best baseline results presented in section 4.1.2 for the three proposed methods in the study. The average results are presented for speakers in each intelligibility group of the UASPEECH database.

The PSD correction method has shown significant gains across all these systems and was the most effective for the SD system in comparison to the adaptation methods. This is predominantly due to the negative impact of the incorrect PSD corrections, which has a single speaker effect in the SD system and manifests as a cumulative effect of all the speakers involved in the adaptation process. In the current study the effect of PSD corrections on

the ASR performance was evaluated using the *supervised, semi-supervised and unsupervised* modes that varies in the amount of prior information that is exposed for the corrections to apply to a particular vowel segment. For example, in a supervised mode the approximate location of vowel segments in an utterance was predicted by the process of forced alignment and in the unsupervised mode, the possible location of vowel(s) was predicted by using a vowel-consonant classifier. Although PSD corrections are the most effective under supervised and semi-supervised modes, the unsupervised correction has shown significant gains for speakers with reduced intelligibility using the SD, SI-02 and SAT systems. This is a very beneficial outcome, as it makes PSD more versatile and independent for its applications in realistic practical settings. The success of the unsupervised PSD corrections is largely dependent on the strength of the vowel-consonant classifier to predict precise vowel locations. This is not an easy task, since true vowel locations in an utterance can often get confused with other voiced segments of speech that can lead to a fall in the ASR performance due to the application of erroneous PSD corrections.

In this study the representation of dysarthric speech as PSCC and MODGDFCC phase features also came out as a significantly effective alternative to standard MFCC features. The success of phase based feature representation was primarily attributed to the unique properties of the group delay spectrum. It was emphasised that better frequency representation and class separation attribute of the phase spectrum was advantageous to better characterise dysarthric speech in comparison to the standard MFCC's. Another important merit of phase based feature representation was that it was found to be robust at effectively characterising varying degrees of dysarthric severities. It was especially seen that PSCC/MODGDFCC features showed refined phone discriminatory capabilities, whilst the standard MFCC features were adversely affected by the high degree of variability manifest in the dysarthric speech with reduced intelligibility. The processing of dysarthric speech as phase representations shows a similar beneficial trend as noted for the method of PSD corrections. The phase representations of PSCC and MODGDFCC tends to marginally favour the SD system over the adaptive counterparts for the very-low and low intelligibility group of speakers, but is more effective for the adaptive systems as the dysarthric intelligibility increases. This can be attributed to the properties of the phase spectrum discussed in the study that might hold a greater chance of producing better base models for the SI-02 and SAT systems as the underlying acoustic variability reduces.

Another important outcome of the study is the recognition results for the SI-00 system. Ideally, for the preparation of any dysarthric speech system, the SI-00 models is the least preferred choice for adaptation. This is primarily due to the huge amount of acoustic dissimilarity that exists between the training and adaptation datasets, which might produce sub-optimal speaker specific models. The findings in the current study report that SI-00 adapted systems show significant gains for the very-low and low intelligibility group of speakers using nearly all the three suggested methods. This is highly encouraging as it widens the scope of potential data sources that can be used to build dysarthric speech systems for speakers with the lowest intelligibility. Since there is already a scarcity of openly available dysarthric databases for research and development purposes, the suggested methods of PSD correction, PSCC/MODGDFCC feature representations and its combination can extend its beneficial properties on varied datasets.

An alternative perspective for a good dysarthric speech system

For the preparation of any dysarthric speech system, choice of vocabulary and its size plays an important role. It involves an iterative communication process between the therapist and the dysarthric user to establish the best possible set of words, which can be effectively used for communication. In the current study, the grammar network that was built for all the tested speech systems allowed the recognition of one out of the possible 255 competing words where each word had an equally likely chance to get recognised. The efficacy of ASR systems was evaluated using the standard metric of Word Accuracy that can be defined for word level recognition as:

$$WAcc = \left(\frac{N_R}{N_T}\right) \cdot \ 100 \tag{7.1}$$

where N_R are the utterances that were successfully recognised and N_T is the total number of utterances. However, the above metric only operates on the output of a single ASR system and does not take account of its improvement relative to another ASR system. In designing dysarthric speech systems it is important to test the effectiveness of multiple systems in relation to each other and select the one that might be more amenable for a particular speaker with a specific set ot vocabulary. Hence, it is sometimes desirable to measure the effectiveness of any improved speech system in more tangible terms that reflect the possible increase in the likeliness of a word to get recognised. In reflection of this, one can define:

$$PWI = \left(\frac{N_{proposed} - N_{baseline}}{v} \right) \cdot 100 \tag{7.2}$$

where PWI is the improvement factor of recognition for each word in the vocabulary using the proposed system relative to any baseline system. $N_{proposed}$ and $N_{baseline}$ are the total number of utterances that were successfully recognised using the two systems and v is the total vocabulary size. As an illustration, if we pick the proposed system as the combined approach of PSD Corrections + PSCC/MODGDFCC feature representation and the baseline system as the initial hybrid adaptation systems defined in section 4.1.2, then the per word recognition improvement using the above equation is shown in figure 7.3.

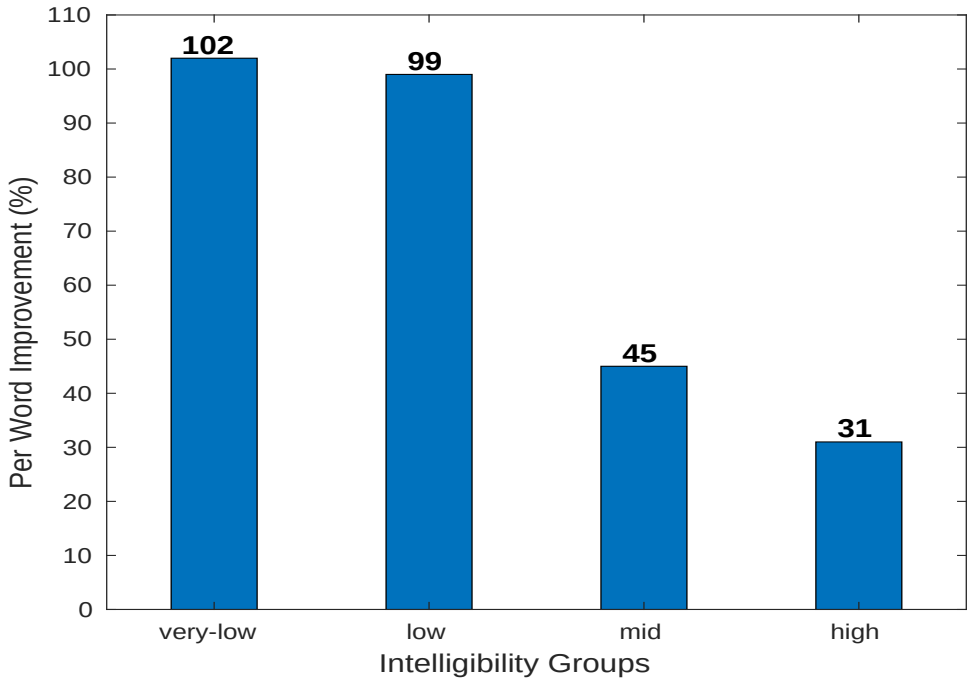

Figure 7.3: Per word recognition improvement where the proposed system uses the combined approach of PSD+PSCC/MODGDFCC and the baseline system is the one suggested in section 4.1.2 using hybrid adaptation method.

It can be seen that the proposed system increases the recognition improvement for each word in the vocabulary (255 in our case) by around 100% for the speakers with lowest intelligibility. The benefit tends to reduce with increasing dysarthric intelligibility, albeit, it

still provides a positive improvement factor across all the intelligibility groups. This aspect of looking at ASR performance can give an alternate perspective to select speech systems relative to each other. In practical setups, for any dysarthric speech system, the size of v is generally in the range of 5-50, which can give a more realistic estimate using equation 7.2.

Adaptive methods seem to be amongst the best techniques that can be used to prepare efficient dysarthric speech systems like SAT, SI-02, especially for speech with low degree of intelligibility. It however comes with an overhead of requiring more dysarthric data for preparing good base systems that are suited for speaker specific adaptation. Speaker dependent systems on the other hand can be prepared with a more simpler setup and require data on a per speaker basis. In the current study, it was shown that all the proposed methods were more beneficial for SD systems than any alternative. It was also shown that the benefits of PSD, phase features and its combination can be achieved by using the same amount of available information and the proposed methods did not prove to be resource or data hungry, which gives promising prospects to addresses the problem of data sparsity to some extent.

Future Work

The study conducted in the thesis has attempted to explore novel approaches based on phase of a signal for better comprehension of dysarthric speech. The suggested methods were beneficial from the perspective of acoustic analysis and ASR improvement. This has also opened new pathways that are worth exploring from research and application viewpoint.

It was examined in chapter 4 that a relationship might exist between the unwrapped phase component of the complex roots of ZZT and the underlying intelligibility. Since speaking rate measure like **sypse** is also found to have a reasonably linear association with intelligibility, it will be interesting to explore if such phase slope deviations have any relationship with rate of speech in general instead of dysarthric speech. One of the way in which the study can be extended is to examine control group of speakers (especially the ones with a slow speaking rate) and investigate the phase deviation. If any phase-based effect is present, it might indicate towards a more generic association to the rate of speech instead of a dysarthric specific anomaly causing the deviation.

In order to conclude anything meaningful from such a kind of analysis, further research is also needed to make the method of plotting unwrapped phase component of the complex roots of ZZT more robust as it currently lacks refinement and is sensitive to the ordering of

roots. It is also influenced by the alignment of the examined speech segment at the glottal closure instant. Future research can direct towards exploring automatic alignment methods to give a more informed understanding of such a phenomenon.

PSD exhibited a strong functional relationship with the underlying intelligibility that could be associated using regression trends. It can be valuable if such trends can be utilised to predict intelligibility based on PSD scores and the other way round. This can assist in setting up a more structured speech therapy routine and aid in designing better speech systems by clustering acoustically similar speakers together. The metric can be further exploited to investigate if PSD holds any relationship at a phonetic level, which might give useful insights into understanding the effect of speech impairment on specific phonemes in relation to others. Also, since PSD corrections proved helpful to improve the efficacy of ASR performance, it will be worth researching if such corrections can be systematically utilised in the speech synthesis domain to improve the perceptual quality of dysarthric speech.

For improving ASR performance, PSD corrections were found to be beneficial for only vowels and was detrimental when applied to other voiced tokens. It was possibly due to articulation and voicing errors that are more manifest in consonants than vowels for dysarthric speech. It requires further research to understand how phase based phenomenon (such as PSD) can be studied and utilised for other voiced segments.

It was also found that phase based features were better at characterising dysarthric speech in comparison to magnitude based features for improving the ASR performance. In light of these findings it will be worth extending the study of phase representations to investigate its efficacy for noise robustness and other environmental factors that can affect dysarthric ASR performance. In addition, since both phase and magnitude spectrum give a complete description of a speech signal rather than any single one, it will be worth researching whether any joint feature representation of phase and magnitude can add any significant benefit for the ASR performance of dysarthric speech.

The phase based approaches investigated in this thesis were found to be significantly beneficial for the classical approaches like HMM-GMM to improve the ASR performance of dysarthric speech. It will be interesting to see if class separation properties of the phase spectrum and/or the corrections of the PSD can further assist the discriminative training paradigms to optimise the classification margins for the input-output mappings. It will be of particular interest to extend the suggested approaches into deep architectures like DNN-HMM that intrinsically aim to optimise a generative-discriminative modelling framework.

Appendix A

Acoustic Analysis of UASPEECH

A.1 F1-F2 vowel quadrilaterals for UASPEECH dysarthric speakers

The vowel quadrilaterals for various dysarthric speakers in the UASPEECH database. For clarity the speakers are grouped according to their intelligibility groups, i.e., **very-low, low, mid and high**. Each quadrilateral is plotted against the average F1-F2 representation of the control speakers in the UASPEECH database.

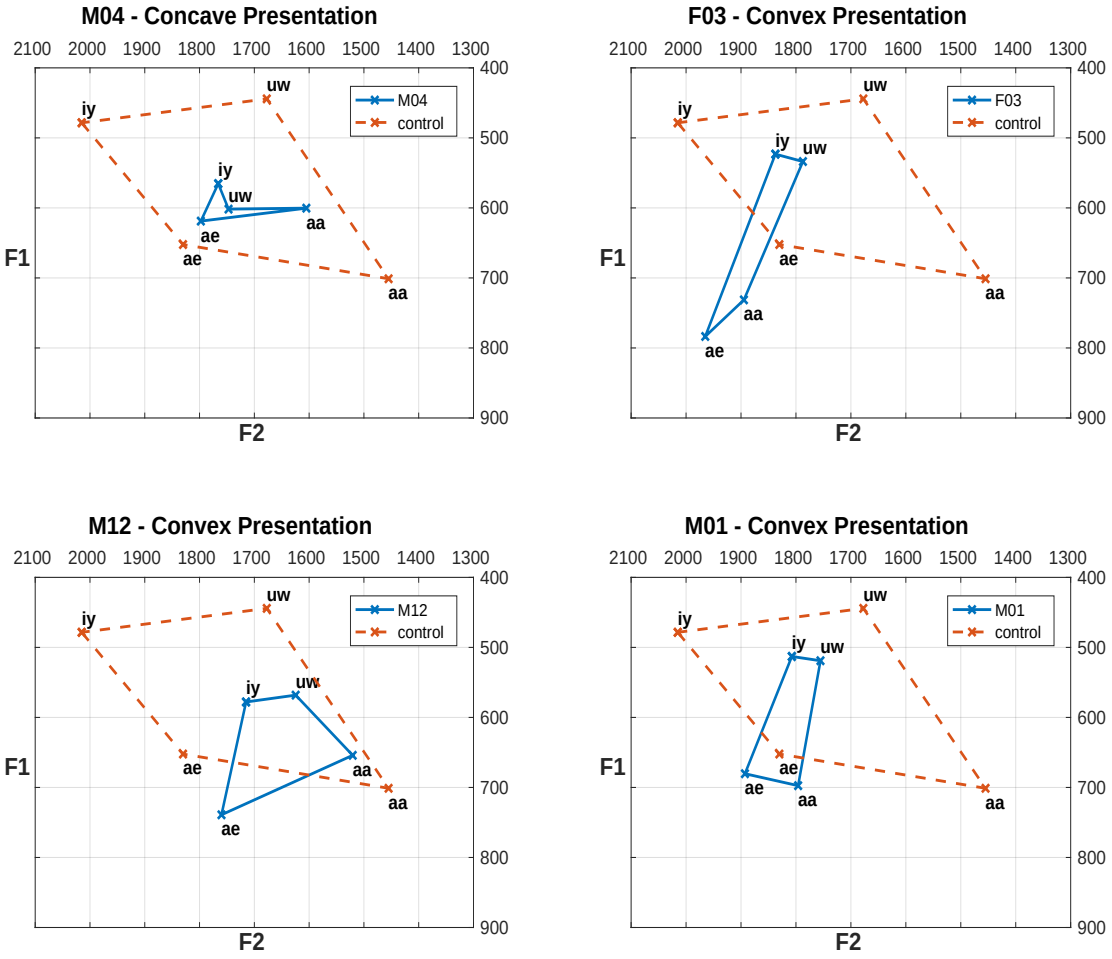

Figure A.1: F1-F2 vowel quadrilateral for speakers with **very-low** intelligibility. The red polygon represents the average vowel quadrilateral for the control speakers in UASPEECH database. The speakers are arranged in their increasing order of severity from top-left to bottom-right.

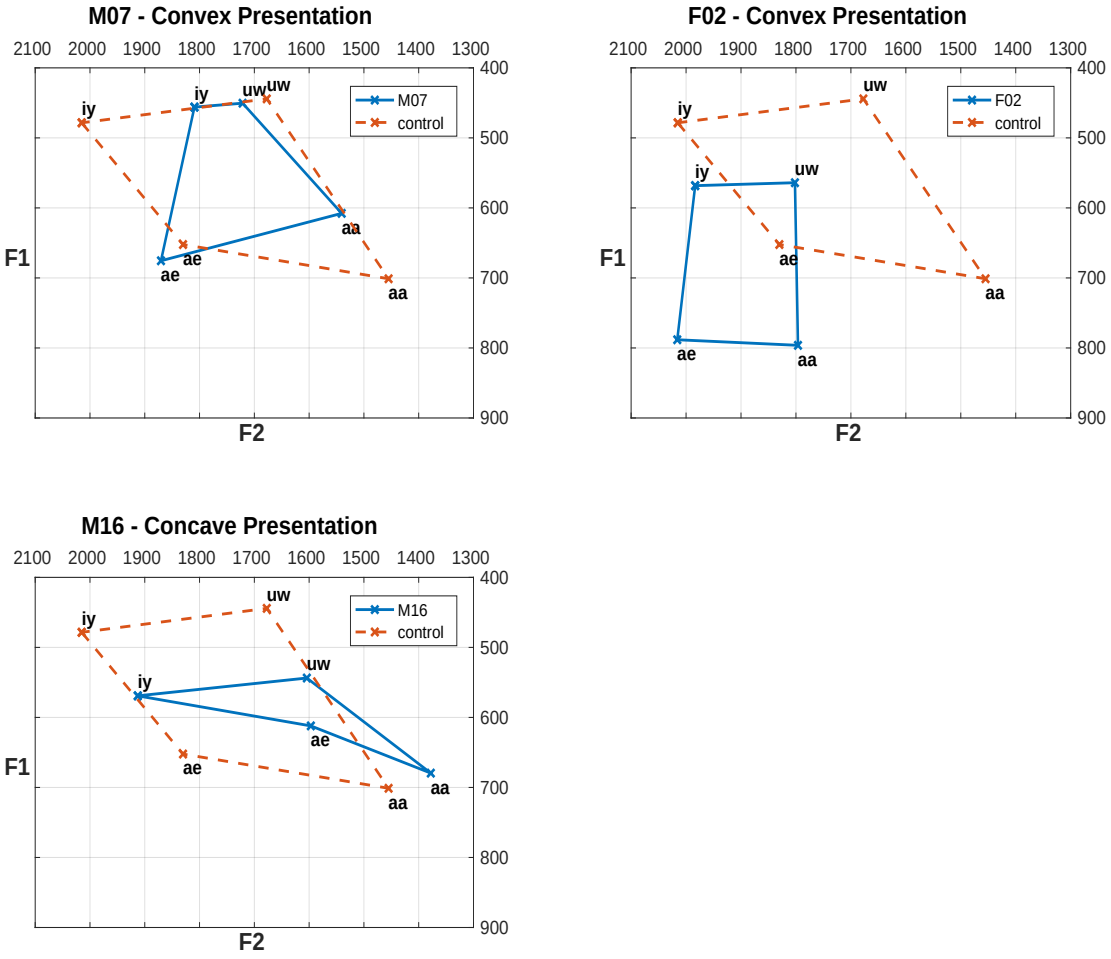

Figure A.2: F1-F2 vowel quadrilateral for speakers with **low** intelligibility. The red polygon represents the average vowel quadrilateral for the control speakers in UASPEECH database. The speakers are arranged in their increasing order of severity from top-left to bottom-right.

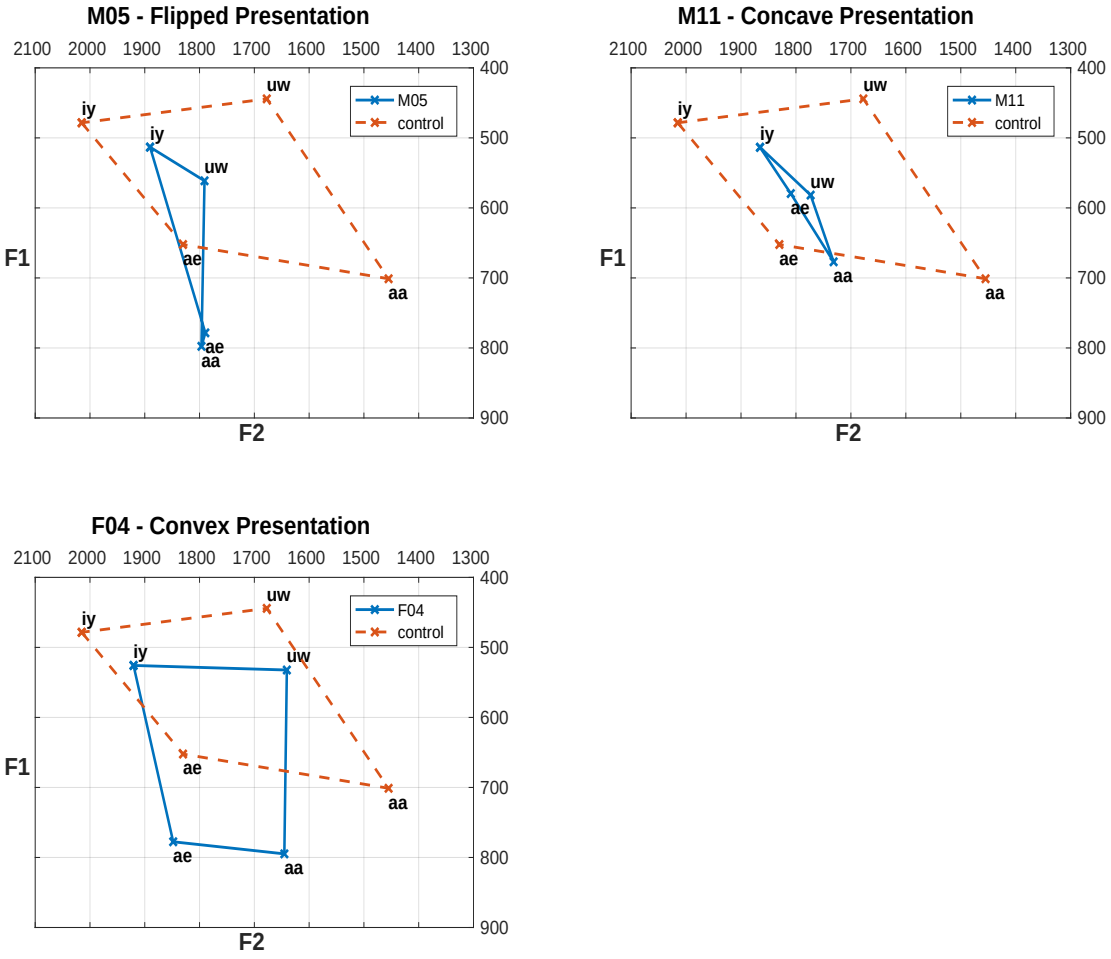

Figure A.3: F1-F2 vowel quadrilateral for speakers with **mid** intelligibility. The red polygon represents the average vowel quadrilateral for the control speakers in UASPEECH database. The speakers are arranged in their increasing order of severity from top-left to bottom-right.

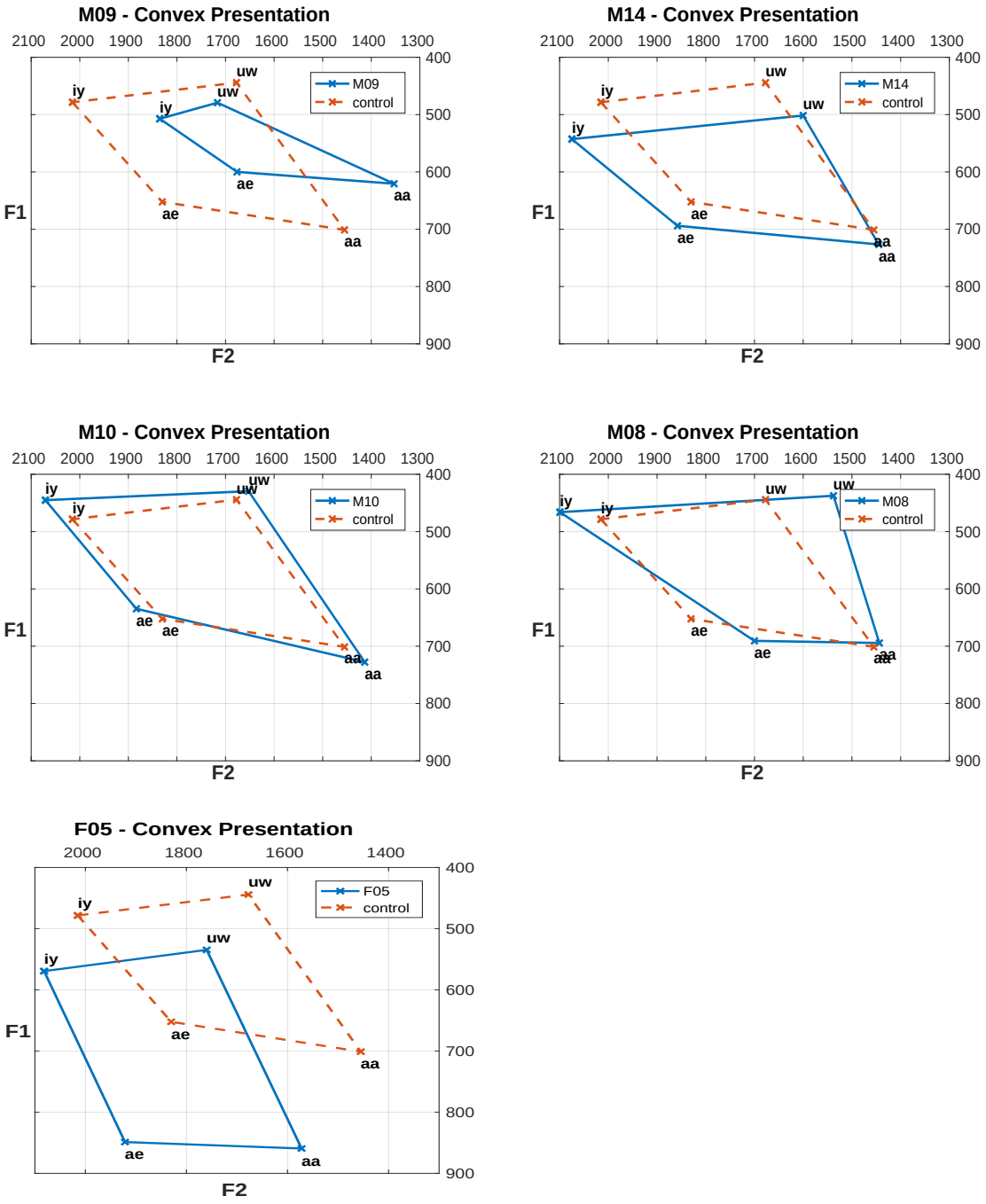

Figure A.4: F1-F2 vowel quadrilateral for speakers with **high** intelligibility. The red polygon represents the average vowel quadrilateral for the control speakers in UASPEECH database. The speakers are arranged in their increasing order of severity from top-left to bottom-right.

Appendix B

The VIVOCA Data Source

B.1 Missing vowel tokens

Due to the smaller size of the VIVOCA database and bespoke vocabulary requirements for individual users, there are some vowel tokens for which no example was recorded from a particular speaker. Table B.1 shows all the missing vowel tokens for each speaker. The table cells that are coloured as red were unavailable for a particular speaker-vowel pair. As an example, majority of VIVOCA speakers have no data for the diphthong /oy/.

	iy	ih	eh	ae	aa	ao	uh	uw	ey	ay	aw	ow	oy
V2-1													
V2-2													
V2-3													
V2-4													
V2-5													
V2-6													
V2-7													
V2-8													
V2-9													
V2-10													
V2-7⋆													
V2-11													
V2-12													

Table B.1: The availability of data for the vowel tokens of VIVOCA speakers. The red blocks indicate that there was no speech utterance for the specific speaker-vowel pair.

Appendix C

Phase Alignment for Control and Dysarthric Speakers

C.1 Unwrapped phase alignment for control and dysarthric speakers

The figure below shows the unwrapped phase alignment of the front-low vowel /ae/ averaged across control speakers and for two dysarthric speakers from the very-low and high intelligibility groups. A similar alignment holds for all the individual vowel and diphthong categories and the purpose of showing only for a single vowel is for clarity. The figure also shows the global unwrapped phase alignment, which is averaged across all the 13 vowel tokens for the control speakers.

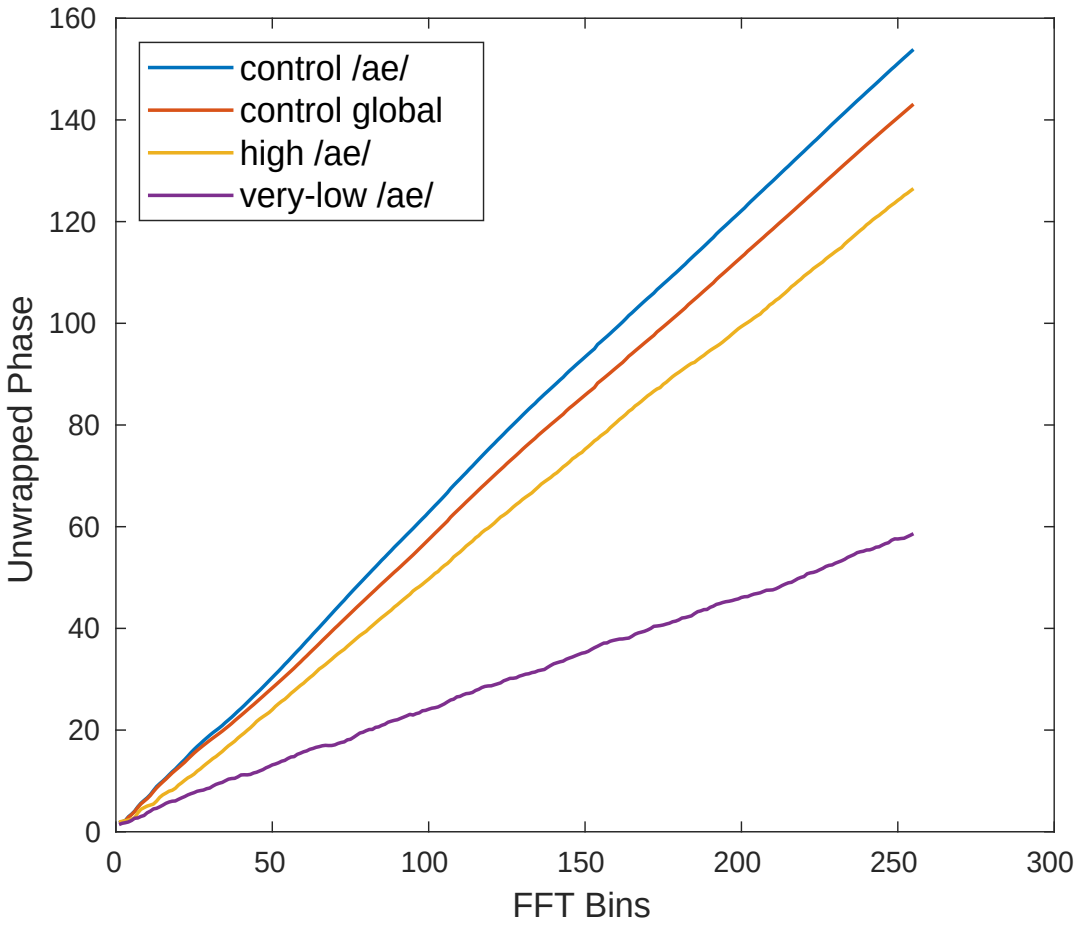

Figure C.1: Unwrapped phase alignment for control and dysarthric speakers with very-low and high intelligibility. The alignment is presented for the front vowel token /ae/ along with the global alignment that is averaged across all the vowel tokens for control speakers.

Appendix D

Standard Deviational Ellipses

D.1 Standard Deviational Ellipses for control and dysarthric intelligibility groups

The standard deviational ellipses for the control and dysarthric intelligibility groups (**very-low, low, mid and high**) in the UASPEECH database. Each plot shows the data spread in both F1 and F2 directions for the vowels and diphthongs. The plots also exhibits the angle of rotation for each ellipse relative to the data under consideration for the mean points represented for each phonetic token.

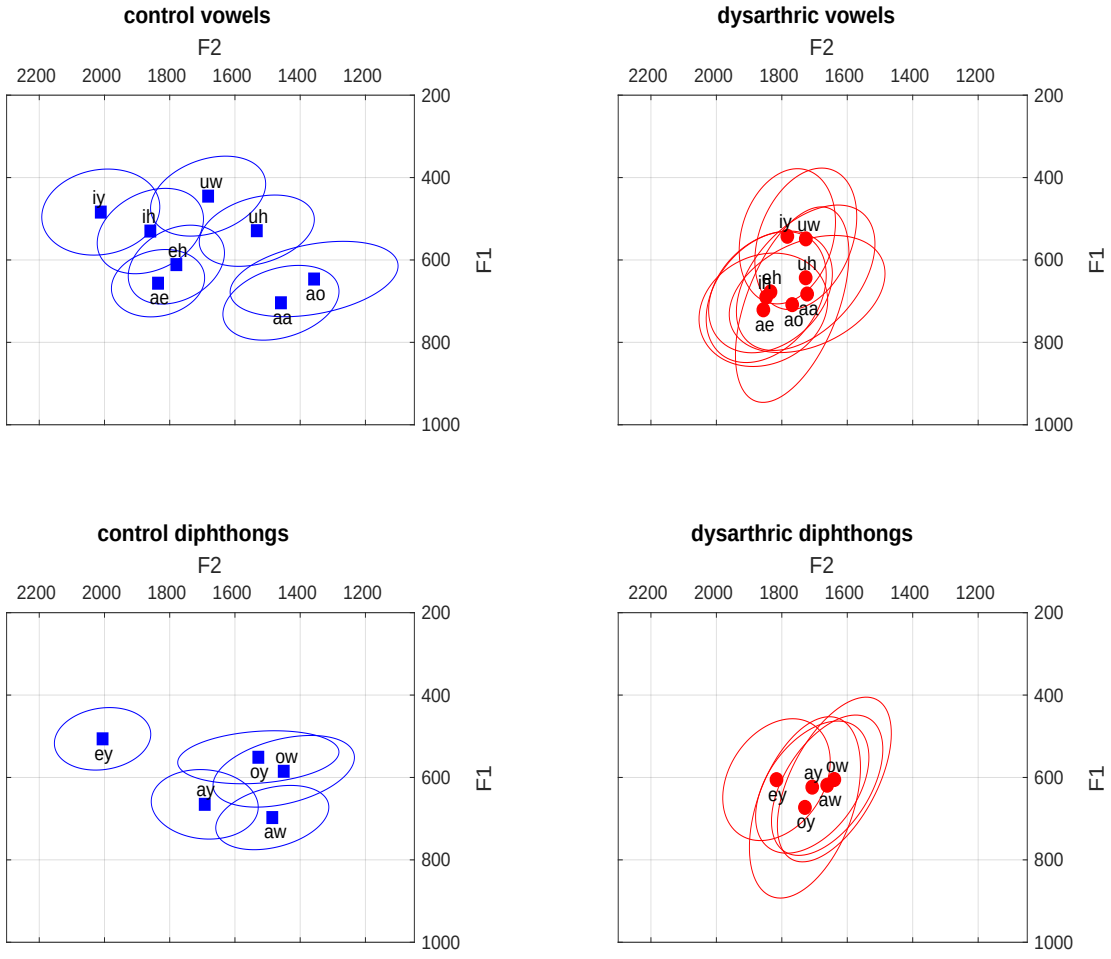

Figure D.1: Standard deviational ellipse for the control and very-low intelligibility group. The top and bottom graphs show the variation for vowels and diphthongs respectively.

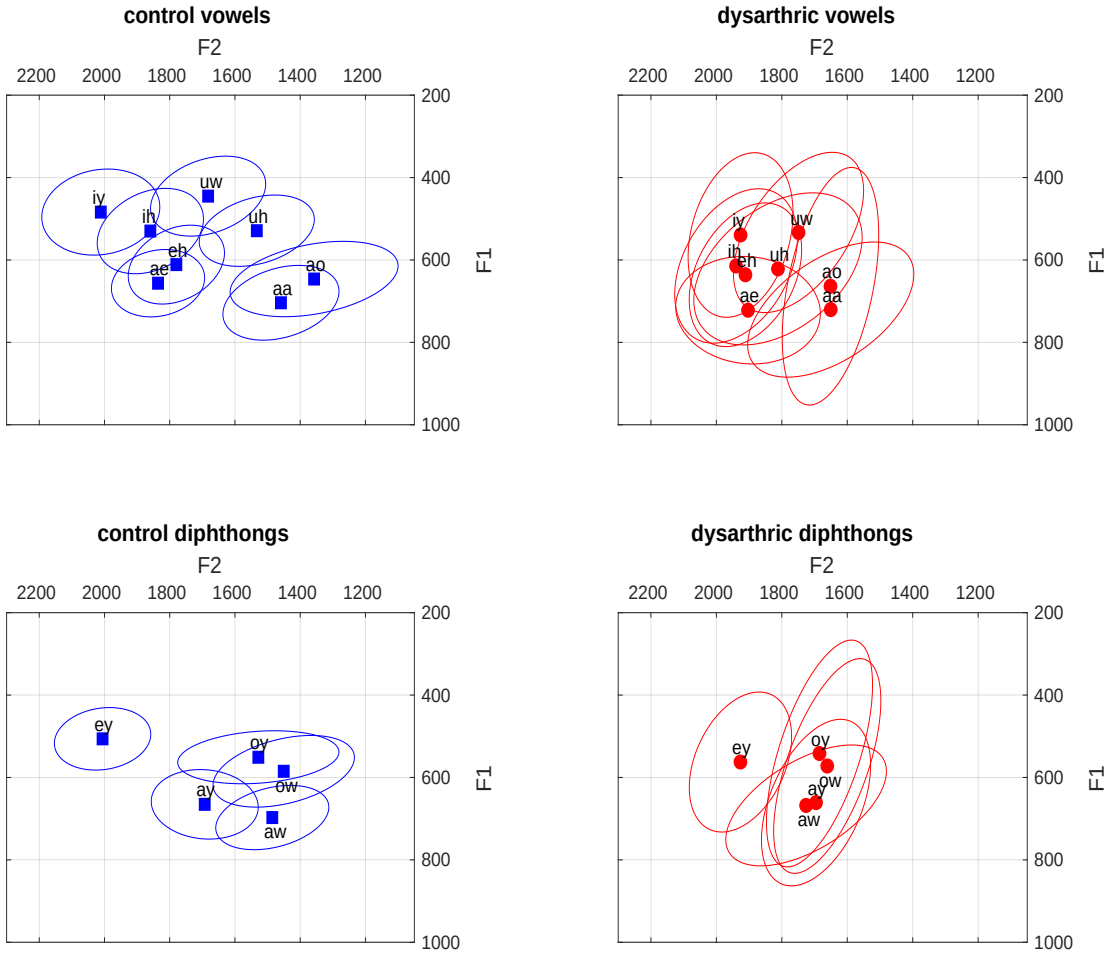

Figure D.2: Standard deviational ellipse for the control and low intelligibility group. The top and bottom graphs show the variation for vowels and diphthongs respectively.

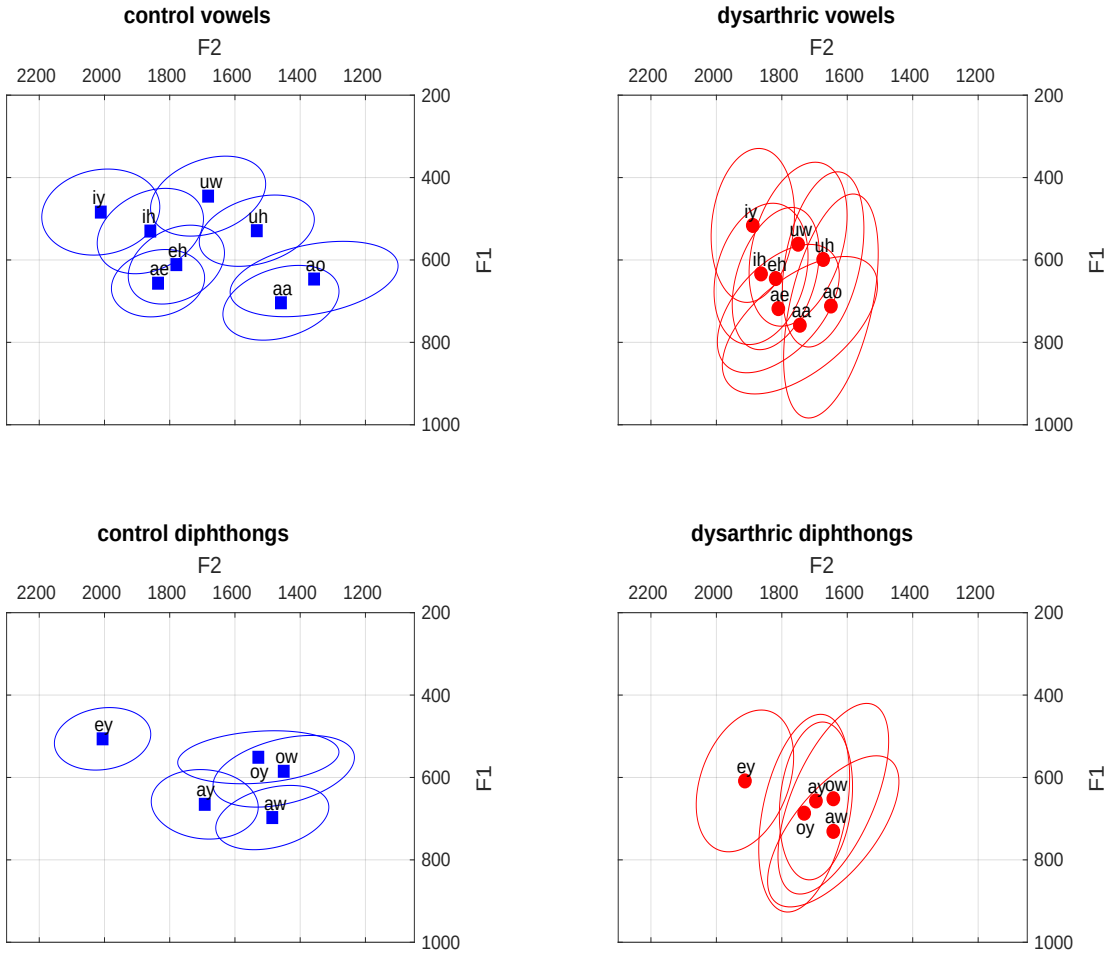

Figure D.3: Standard deviational ellipse for the control and mid intelligibility group. The top and bottom graphs show the variation for vowels and diphthongs respectively.

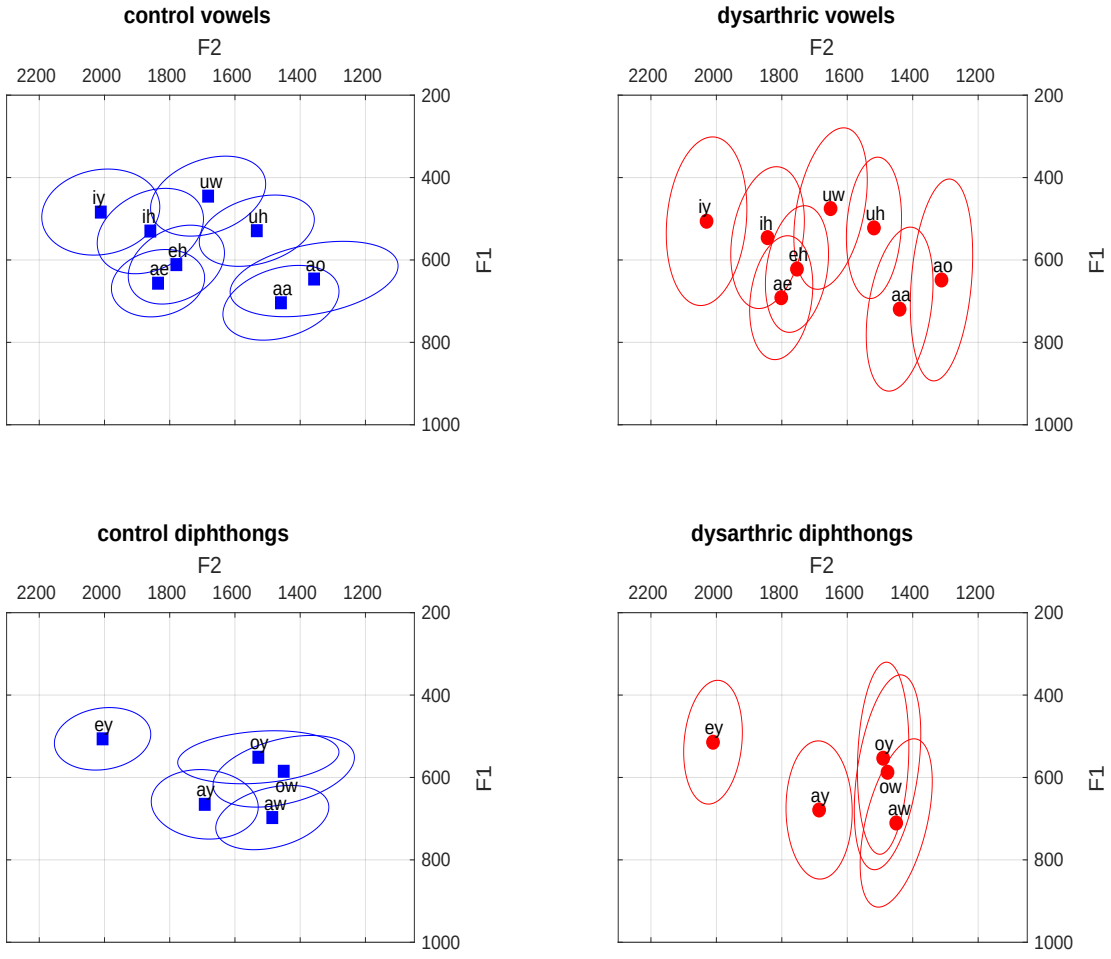

Figure D.4: Standard deviational ellipse for the control and high intelligibility group. The top and bottom graphs show the variation for vowels and diphthongs respectively.

Bibliography

Abel, N.H. (1824). "Mmoire sur les quations algbriques, ou l'on dmontre l'impossibilit de la rsolution de l'quation gnrale du cinquime degr". In: *Oeuvres compltes de Niels Henrik Abel*. 2nd ed. Vol. 1. Grndahl and Sn Forlag, pp. 28–33.

Ackermann, H. and Ziegler, W. (1991). "Cerebellar voice tremor: an acoustic analysis". In: *Journal of Neurology, Neurosurgery and Psychiatry* 54.1, pp. 74–76.

Adde, L. and Svendsen, T. (2011). "Pronunciation variation modeling of non-native proper names by discriminative tree search". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 4928–4931.

Adde, L. et al. (2010). "A minimum classification error approach to pronunciation variation modeling of non-native proper names". In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, pp. 2282–2285.

Ahadi, SM and Woodland, PC (1997). "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models". In: *Computer Speech & Language* 11.3, pp. 187–206.

Anastasakos, A. et al. (1994). "Adaptation to new microphones using tied-mixture normalization". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-94.* Vol. 1, pp. 433–436.

Anastasakos, T., McDonough, J., and Makhoul, J. (1997). "Speaker adaptive training: a maximum likelihood approach to speaker normalization". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-97.,* vol. 2, pp. 1043–1046.

Anastasakos, T. et al. (1996). "A compact model for speaker-adaptive training". In: *Fourth International Conference on Spoken Language, ICSLP 96., Proceedings.* Vol. 2, pp. 1137–1140.

Andrei, V., Paleologu, C., and Burileanu, C. (2011). "Implementation of a real-time text dependent speaker identification system". In: *Proceedings of the 6th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2011.*

Antolik, T.K. and Fougeron, C. (2013). "Consonant distortions in dysarthria due to Parkinson's disease, Amyotrophic Lateral Sclerosis and Cerebellar Ataxia". In: *INTERSPEECH*, pp. 2152–2156.

Armstrong, L., Jans, D., and MacDonald, A. (2000). "Parkinson's disease and aided AAC: Some evidence from practice". In: *International Journal of Language and Communication Disorders* 35.3, pp. 377–389.

Aronson, A.E. et al. (1992). "Rapid voice tremor, or flutter, in amyotrophic lateral sclerosis." In: *The Annals of otology rhinology and laryngology* 101.6, pp. 511–518.

Atal, B. S. (1974). "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification". In: *The Journal of the Acoustical Society of America* 55.6, pp. 1304–1312.

Atal, B. and Schroeder, M." (1970). ""Adaptive predictive coding of speech signals"". In: *j-bell-sys-tech* 49, pp. 1973–1986.

Ataxia Org. (2013). `http://www.ataxia.org.uk/`. Online; accessed on: 09-September-2013.

Bahl, L. R. et al. (1986). "Maximum mutual information estimation of hidden Markov model parameters for speech recognition". In: *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 49–52.

Ball, L. J., Beukelman, D. R., and Pattee, G. L. (2004). "Acceptance of Augmentative and Alternative Communication Technology by Persons with Amyotrophic Lateral Sclerosis". In: *Augmentative and Alternative Communication* 20.2, pp. 113–122.

Bamberg, P. G. (1990). "Adaptable phoneme-based models for large-vocabulary speech recognition". In: *Tutorial and Research Workshop on Speech Characterization in Speech Technology, SCST, Edinburgh: European Speech Communication Association*, pp. 1–9.

Bartkova, K. and Jouvet, D. (2006). "Using Multilingual Units for Improved Modeling of Pronunciation Variants". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Proceedings*. Vol. 5, pp. V1037–V1040.

Baum, L. E. et al. (1970). "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains". In: *Annals of Mathematical Statistics* 41, pp. 164–171.

Bellman, R.E. (1953). *Al introduction to the theory of dynamic programming*. R-245. RAND Corporation.

Beukelman, D. R. et al. (2007). "AAC for adults with acquired neurological conditions: A review". In: *AAC: Augmentative and Alternative Communication* 23.3, pp. 230–242.

Bhat, C., Vachhani, B., and Kopparapu, S. (2016a). "Improving Recognition of Dysarthric Speech Using Severity Based Tempo Adaptation". In: *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings*. Springer International Publishing, pp. 370–377.

— (2016b). "Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper Spectral Estimation". In: *Interspeech*, pp. 228–232.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford University Press, Inc.

Blaney, B. and Wilson, J. (2000). "Acoustic variability in dysarthria and computer speech recognition". In: *Clinical Linguistics & Phonetics* 14.4, pp. 307–327.

Bloch, S. and Wilkinson, R. (2004). "The understandability of AAC: A conversation analysis study of acquired dysarthria". In: *AAC Augmentative Alternative Communication* 20.4, pp. 272–282.

Bogert, B., Healy, M., and Tukey, J. (1963). "The quefrency alanysis of time series for echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking". In: *Proc. Symp. on Time Series Analysis*. Chap. 15, pp. 209–243.

Bonafonte, A., Ros, X., and Marifio, J.B. (1993). "An efficient algorithm to find the best state sequence in HSMM." In: *EUROSPEECH*. ISCA.

Bowman, S. and Livescu, K. (2010). "Modeling pronunciation variation with context-dependent articulatory feature decision trees". In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, pp. 326–329.

Bozkurt, B., Couvreur, L., and Dutoit, T. (2007). "Chirp group delay analysis of speech signals." In: *Speech Communication* 49.3, pp. 159–176.

Brown, M. and Rabiner, L. (1982). "Dynamic time warping for isolated word recognition based on ordered graph searching techniques". In: *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing.* Vol. 7, pp. 1255–1258.

Brown, P.F. et al. (1992). "Class-based n-gram models of natural language". In: *Comput. Linguist.* 18.4, pp. 467–479.

Bunton, K. et al. (2007). "Listener agreement for auditory-perceptual ratings of dysarthria". In: *Journal of Speech, Language, and Hearing Research* 50.6, pp. 1481–1495.

Chandrakala, S. and Rajeswari, N. (2017). "Representation Learning Based Speech Assistive System for Persons with Dysarthria". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.9, pp. 1510–1517.

Chen, F. and Kostov, A. (1997). "Optimization of dysarthric speech recognition". In: *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* Vol. 4, pp. 1436–1439.

Chen, K-T. et al. (2000). "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression". In: *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000*, pp. 742–745.

Chenausky, K., MacAuslan, J., and Goldhor, R. (2011). "Acoustic Analysis of PD Speech". In: *Parkinson's Disease* 2011, p. 13. DOI: 10.4061/2011/435232.

Chesta, C., Siohan, O., and Lee, C.H. (1999). "Maximum A Posteriori Linear Regression for Hidden Markov Model Adaptation". In: *Sixth European Conference on Speech Communication and Technology, EUROSPEECH*, pp. 211–214.

Choi, J.-H. and Chang, J.-H. (2012). "On using spectral gradient in conditional MAP criterion for robust voice activity detection". In: *Proceedings - 2012 3rd IEEE International Conference on Network Infrastructure and Digital Content, IC-NIDC 2012*, pp. 370–374.

Chow, Yen-Lu (1990). "Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-best algorithm". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing.* Vol. 2, pp. 701–704.

Christensen, H. et al. (2012). "A comparative study of adaptive, automatic recognition of disordered speech". In: *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*. Vol. 2, pp. 1774–1777.

Claes, T. et al. (1998). "A Novel Features Transformation for Vocal Tract Length Normalization in Automatic Speech Recognition". In: *IEEE Transaction on Speech and Audio Processing* 6.6, pp. 549–557.

Clarke, M. and Price, K. (2012). "Augmentative and alternative communication for children with cerebral palsy". In: *Paediatrics and Child Health (United Kingdom)* 22.9, pp. 367–371.

Cochran, W. G. (1950). "The Comparison of Percentages in Matched Samples". In: *Biometrika* 37.3/4, pp. 256–266. URL: `http://www.jstor.org/stable/2332378`.

Cohen, A. and Graupe, D. (1980). "Speech recognition and control system for the severely disabled". In: *Journal of Biomedical Engineering* 2.2, pp. 97–107.

Coleman, C. and Meyers, L. (1991). "Computer recognition of the speech of adults with cerebral palsy and dysarthria". In: *Augmentative and Alternative Communication* 7.1, pp. 34–42.

Dahl, George et al. (2011). "Large Vocabulary Continuous Speech Recognition With Context-Dependent DBN-HMMS". In: *Proc. ICASSP, Prague*. IEEE.

Darley, F.L., Aronson, A.E., and Brown, J.R. (1969a). "Clusters of deviant speech dimensions in the dysarthrias". In: *Journal of Speech and Hearing Research* 12, pp. 462–496.

— (1969b). "Differential diagnostic patterns of dysarthria". In: *Journal of Speech and Hearing Research* 12, pp. 246–269.

Davis, K.H., Biddulph, R., and Balashek, S. (1952). "Automatic Recognition of Spoken Digits". In: *The Journal of Acoustical Society of America* 24.6, pp. 637–642.

Davis, S. and Mermelstein, P. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 28.4, pp. 357–366.

De Bodt, Marc S., Hernndez-Daz Huici, Maria E., and Van De Heyning, Paul H. (2002). "Intelligibility as a linear combination of dimensions in dysarthric speech." In: *Journal of Communication Disorders* 35.3, pp. 283–292.

Delatycki, M. B., Williamson, R., and Forrest, S. M. (2000). "Friedreich ataxia: an overview." In: *Journal of Medical Genetics* 37.1, pp. 1–8.

Deller, J.R., Hsu, D., and Ferrier, L.J. (1991). "On the use of hidden Markov modelling for recognition of Dysarthric speech". In: *Computer Methods and Programs in Biomedicine* 35.2, pp. 125–139.

Dempster, A.P., Laird, N.M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal Of The Royal Statistical Society, Series B* 39.1, pp. 1–38.

Deng, L. (2014). "A Tutorial Survey of Architectures, Algorithms, and Applications for Deep Learning". In: *APSIPA Transactions on Signal and Information Processing.*

Deng, L and Li, X. (2013). "Machine Learning Paradigms for Speech Recognition: An Overview". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.5, pp. 1060–1089.

Deng, Y. et al. (2009). "Disordered speech recognition using acoustic and sEMG signals". In: *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, pp. 644–647.

Digalakis, V.V., Rtischev, D., and Neumeyer, L.G. (1995). "Speaker adaptation using constrained estimation of Gaussian mixtures". In: *IEEE Transactions on Speech and Audio Processing* 3.5, pp. 357–366.

Dowden, P. (1997). "Augmentative and alternative communication decision making for children with severely unintelligible speech". In: *Augmentative and Alternative Communication* 13.1, pp. 48–59.

Duffy, J.R. (2005). *Motor Speech Disorders : Substrates, Differential Diagnosis, and Management.* Second. Elsevier Mosby.

Dworkin, J.P. and Hartman, D.E. (1979). "Progressive speech deterioration and dysphagia in amyotrophic lateral sclerosis: Case report". In: *Archives of Physical Medicine and Rehabilitation* 60.9, pp. 423–425.

Eigentler, A. et al. (2011). "The scale for the assessment and rating of ataxia correlates with dysarthria assessment in Friedreichs ataxia." In: *Journal of Neurology* 259.3, pp. 1717–1720.

Elenius, K. and Blomberg, M. (1986). "Electronic speech recognition: techniques, technology, and applications". In: *Voice input for personal computers*. Ed. by Geoff Bristow. McGraw-Hill, Inc., pp. 361–372.

Enderby, P. (1988). "The assessment of dysarthria: A challenge to more than the ears." In: *Journal of Clinical Rehabilitation* 2.4, pp. 267–273.

Enderby, P. et al. (2013). *Communication Matters Research Matters: an AAC Evidence Base*. `https://communicationmatters.org.uk/sites/default/files/downloads/projects/aac_evidence_base/2013_AAC_Evidence_Base_Beyond_the_Anecdote.pdf`. Online; accessed on: 08-Mar-2017.

Enderby, P.M. (1983). *Frenchay Dysarthria Assessment*. Pro-Ed.

Espaa-Bonet, C. and Fonollosa, J.A.R. (2016). "Automatic speech recognition with deep neural networks for impaired speech". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10077 LNAI, pp. 97–107.

Esposito, S.J., Mitsumoto, H., and Shanks, M. (2000). "Use of palatal lift and palatal augmentation prostheses to improve dysarthria in patients with amyotrophic lateral sclerosis: A case series". In: *Journal of Prosthetic Dentistry* 83.1, pp. 90–98.

Evans, J.D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove: Brooks/Cole Pub. Co.

Evidente, V. G. H. and Adler, C. H. (2010). "An update on the neurologic applications of botulinum toxins". In: *Current Neurology and Neuroscience Reports* 10.5, pp. 338–344.

Fedorova, N. V. and Chigir, I. P. (2007). "Use of the dopamine receptor agonist Mirapex in the treatment of Parkinsons disease." In: *Neuroscience and Behavioral Physiology* 37.6, pp. 539–546.

Feenaughty, L., Tjaden, K., and Sussman, J. (2014). "Relationship between acoustic measures and judgments of intelligibility in Parkinsons disease: A within-speaker approach". In: *Clinical Linguistics & Phonetics* 28.11, pp. 857–878.

Ferrier, L.J. et al. (1992). "A case study of a dysarthric speaker using the DragonDictate speech recognition system". In: *Journal of Computer Users in Speech and Hearing* 8.1-2, pp. 33–52.

Fletcher, H. (1940). "Auditory Patterns". In: *Reviews of Modern Physics* 12.1, pp. 47–65.

Forgie, J.W. and Forgie, C.D. (1959). "Results Obtained From a Vowel Recognition Computer Program". In: *The Journal of Acoustical Society of America* 31.11, pp. 1480–1489.

Fosler-Lussier, E. (2003). "A tutorial on pronunciation modeling for large vocabulary speech recognition". In: *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*. Vol. 2705, pp. 38–77.

Fosler-Lussier, E., Amdal, I., and Kuo, H-K.J. (2005). "A framework for predicting speech recognition errors". In: *Speech Communication* 46.2, pp. 153–170.

Fox, C.M. and Boliek, C.A. (2012). "Intensive voice treatment (LSVT LOUD) for children with spastic cerebral palsy and dysarthria". In: *Journal of Speech, Language, and Hearing Research* 55.3, pp. 930–945.

Fried-Oken, M. (1985). "Voice recognition device as a computer interface for motor and speech impaired people". In: *Archives of Physical Medicine and Rehabilitation* 66.10, pp. 678–681.

Gales, M.J.F. (1998a). "Cluster adaptive training for speech recognition". In: *Fifth International Conference on Spoken Language Processing*.

— (1998b). "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition". In: *Computer Speech and Language* 12, pp. 75–98.

— (2000). "Cluster adaptive training of hidden Markov models". In: *IEEE Transactions on Speech and Audio Processing* 8.4, pp. 417–428.

Gales, M.J.F., Pye, D., and Woodland, P.C. (1996). "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation". In: *Fourth International Conference on Spoken Language, ICSLP 96, Proceedings*. Vol. 3, pp. 1832–1835.

Gales, M.J.F. and Woodland, P.C. (1996). "Mean and variance adaptation within the {MLLR} framework". In: *Computer Speech & Language* 10.4, pp. 249–264.

Garofolo, J. S. et al. (1993). *TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*.

Gauvain, J.L. and Lee, C.H. (1992). "MAP Estimation of Continuous Density HMM: Theory and Applications". In: *Proceedings of DARPA Speech and Natural Language Workshop*, pp. 185–190.

Gauvain, J.L. and Lee, C.H. (1994). "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains". In: *IEEE Transactions on Speech and Audio Processing* 2, pp. 291–298.

Gemmeke, J.F. et al. (2014). "Dysarthric vocal interfaces with minimal training data". In: *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 248–253.

Geppener, V.V., Simonchik, K.K., and Haidar, A.S. (2007). "Design of speaker verification systems with the use of an algorithm of Dynamic Time Warping (DTW)". In: *Pattern Recognition and Image Analysis* 17.4, pp. 470–479.

Gescheider, G.A. (1976). *Psychophysics: Method and Theory*. L. Erlbaum Associates.

Gillick, L. and Cox, S. (1989). "Some statistical issues in the comparison of speech recognition algorithms". In: *In Proc. of ICASSP*, pp. 532–535.

Godin, C and Lockwood, P (1989). "DTW schemes for continuous speech recognition: a unified view". In: *Computer Speech and Language* 3.2, pp. 169–198.

Good, I.J. (1953). "The population frequencies of species and the estimation of population parameters". In: *Biometrika* 40.3/4, pp. 237–264.

Goronzy, S. and Kompe, K. (1999). "A MAP-like weighting scheme for MLLR speaker adaptation". In: *Sixth European Conference on Speech Communication and Technology, EUROSPEECH*, pp. 5–8.

Gunawardana, A. and Byrne, W. (2001). "Discriminative speaker adaptation with conditional maximum likelihood linear regression". In: *In Proc. Eurospeech*.

HDA Org. (2013). `http://hda.org.uk/professionals/carepathway.html`. Online; accessed on: 09-September-2013.

Halpern, C. et al. (2007). "Deep brain stimulation in neurologic disorders". In: *Parkinsonism & Related Disorders* 13.1, pp. 1–16.

Hammerlin, G. and Hoffmann, K.H. (1991). *Numerical Mathematics*. Readings in mathematics. U.S. Government Printing Office.

Harries, J.R. and Lawes, W.E. (1957). "The advantages of glossopharyngeal breathing." In: *British medical journal* 2.5055, pp. 1204–1205.

Hartelius, L., Buder, E.H., and Strand, E.A. (1997). "Long-term phonatory instability in individuals with multiple sclerosis". In: *Journal of Speech, Language and Hearing Research* 40.5, pp. 1056–1072.

Hartelius, L. and Svensson, P. (1990). *Dysartritest: manual*. Psykologiförl.

Hasegawa-Johnson, M. et al. (2006). "HMM-BASED and SVM-BASED Recognition of the Speech of Talkers With Spastic Dysarthria". In: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. Vol. 3, pp. III–III.

Hawley, C. et al. (2003). "Prevalence of traumatic brain injury amongst children admitted to hospital in one health district : a population-based study". In: *Injury* 34.4, pp. 256–260.

Hawley, M. S. (2002). "Speech Recognition as an Input to Electronic Assistive Technology". In: *The British Journal Of Occupational Therapy* 65.1, pp. 15–20.

Hawley, M. S. et al. (2007). "A speech-controlled environmental control system for people with severe dysarthria." In: *Med Eng Phys* 29.5, pp. 586–593.

Hawley, M. et al. (2012). "A Voice-Input Voice-Output Communication Aid for People With Severe Speech Impairment". In: *Neural Systems and Rehabilitation Engineering, IEEE Transactions on* PP.99, p. 1.

Hazen, T.J. et al. (2005). "Pronunciation modeling using a finite-state transducer representation". In: *Speech Communication* 46.2, pp. 189–203.

He, X. and Chou, W. (2003). "Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs". In: *in Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 556–559.

He, X., Deng, L., and Chou, W. (2008). "Discriminative learning in sequential pattern recognition". In: *IEEE Signal Processing Magazine* 25.5, pp. 14–36.

Hegde, R. M., Murthy, H. A., and Gadde, V.R.R. (2007). "Significance of the Modified Group Delay Feature in Speech Recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.1, pp. 190–202.

Helmholtz, H.L.F. (1912). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Fourth. Longmans Green and Co. Original work published 1877.

Hermansky, H. (1990). "Perceptual Linear Predictive (PLP) Analysis of Speech". In: *J. Acoust. Soc. Am.* 57.4, pp. 1738–52.

Hinchcliffe, A. (2007). *Children With Cerebral Palsy: A Manual for Therapists, Parents and Community Workers*. SAGE Publications. URL: http://books.google.co.uk/books?id=3foRWmuYUfgC.

Hinton, G.E., Osindero, S., and Teh, Y. (2006). "A Fast Learning Algorithm for Deep Belief Nets". In: *Neural Comput.* 18.7, pp. 1527–1554.

Hinton, G.E. and Salakhutdinov, R.R. (2006). "Reducing the dimensionality of data with neural networks". In: *Science* 313.5786, pp. 504–507.

Hinton, Geoffrey, Vinyals, Oriol, and Dean, Jeffrey (2015). "Distilling the Knowledge in a Neural Network". In: *NIPS Deep Learning and Representation Learning Workshop*.

Holmes, J.N. and Holmes, W.J. (2001). *Speech Synthesis and Recognition*. Second. Taylor & Francis.

Holter, T. and Svendsen, T. (1999). "Maximum likelihood modelling of pronunciation variation". In: *Speech Communication* 29.24, pp. 177–191.

Humphries, J. J., Woodland, P.C., and Pearce, D. (1996). "Using accent-specific pronunciation modelling for robust speech recognition". In: *Fourth International Conference on Spoken Language Processing Processing, ICSLP, Proceedings*. Vol. 4, pp. 2324–2327.

Humphries, J.J. and Woodland, P.C. (1998). "Use of accent-specific pronunciation dictionaries in acoustic model training". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 1, pp. 317–320.

Hunter, L., Pring, T., and Martin, S. (1991). "The use of strategies to increase speech intelligibility in cerebral palsy: an experimental evaluation." In: *The British journal of disorders of communication* 26.2, pp. 163–174.

Hussain, G. and Manyam, B. V. (1997). "Mucuna pruriens proves more effective than L-DOPA in Parkinson's disease animal model". In: *Phytotherapy Research* 11.6, pp. 419–423.

Hustad, K. C. (2005). "Effects of speech supplementation strategies on intelligibility and listener attitudes for a speaker with mild dysarthria". In: *AAC: Augmentative and Alternative Communication* 21.4, pp. 256–263.

Hustad, K. C (2008). "The relationship between listener comprehension and intelligibility scores for speakers with dysarthria." In: *Journal of Speech, Language and Hearing Research* 51.3, pp. 562–573.

Hustad, K. C. and Cahill, M. A. (2003). "Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech." In: *American journal of speechlanguage pathology American SpeechLanguageHearing Association* 12.2, pp. 198–208.

Hustad, K. C. and Garcia, J. M. (2005). "Aided and unaided speech supplementation strategies: Effect of alphabet cues and iconic hand gestures on dysarthric speech". In: *Journal of Speech, Language, and Hearing Research* 48.5, pp. 996–1012.

Hustad, K. C. and Gearhart, K. J. (2004). "Listener attitudes toward individuals with cerebral palsy who use speech supplementation strategies". In: *American Journal of Speech-Language Pathology* 13.2, pp. 168–181.

Hux, K. et al. (2000). "Accuracy of three speech recognition systems: Case study of dysarthric speech". In: *Augmentative and Alternative Communication* 16, pp. 186–196.

Itakura, F and Saito, S. (1970). "A statistical method for estimation of speech spectral density and formant frequencies". In: *Electron. Commun. Jpn.* 53A, pp. 36–43.

Itakura, F. and Umezaki, T. (1987). "Distance measure for speech recognition based on the smoothed group delay spectrum". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 1257–1260.

Jayaram, G. and Abdelhamied, K. (1995). "Experiments in dysarthric speech recognition using artificial neural networks". In: *Journal of Rehabilitation Research and Development* 32.2, pp. 162–169.

Jelinek, F. (1976). "Continuous speech recognition by statistical methods". In: *Proceedings of the IEEE* 64.4, pp. 532–556.

Jelinek, F. and Mercer, R.L. (1980). "Interpolated estimation of Markov source parameters from sparse data". In: *In Proceedings of the Workshop on Pattern Recognition in Practice*, pp. 381–397.

Johansson, K.M., Nygren-Bonnier, M., and Schalling, E. (2012). "Effects of glossopharyngeal breathing on speech and respiration in multiple sclerosis: A case report". In: *Multiple Sclerosis* 18.6, pp. 905–908.

Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer.

Joy, N.M. and Umesh, S. (2018). "Improving Acoustic Models in TORGO Dysarthric Speech Database". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Juang, B-H., Hou, W., and Lee, C-H. (1997). "Minimum classification error rate methods for speech recognition". In: *Speech and Audio Processing, IEEE Transactions on* 5.3, pp. 257–265.

Juang, B-H. and Katagiri, S. (1992). "Discriminative learning for minimum error classification". In: *IEEE Transactions on Signal Processing* 40.12, pp. 3043–3054.

Juneja, A., Deshmukh, O., and Espy-Wilson, C. (2002). "An event-based acoustic-phonetic approach for speech segmentation and E-set recognition". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 4.

Jurafsky, D. and Martin, J.H. (2000). *Speech and Language Processing*. First. Prentice Hall.

Kain, A. and Santen, J. van (2009). "Using speech transformation to increase speech intelligibility for the hearing- and speaking-impaired". In: *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 3605–3608.

Kain, A.B. et al. (2007). "Improving the intelligibility of dysarthric speech". In: *Speech Communication* 49.9, pp. 743–759.

Kaiser, J., Horvat, B., and Kacic, Z. (2002). "Overall risk criterion estimation of hidden Markov model parameters". In: *Speech Communication* 38.34, pp. 383–398.

Kalinina, L. V. et al. (2000). "Botox in combined treatment of cerebral palsy". In: *Zhurnal Nevropatolgii i Psikhiatrii im.S S Korsakova* 100.12, pp. 60–63.

Katz, S. (1987). "Estimation of probabilities from sparse data for the language model component of a speech recognizer". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 35.3, pp. 400–401.

Katzenschlager, R. and Lees, A.J. (2002). "Treatment of Parkinson s disease : levodopa as the first choice". In: *Journal of Neurology* 215.0, pp. 19–24. URL: `http://www.ncbi.nlm.nih.gov/pubmed/12375059`.

Kaufman, C.B., Mink, J.W., and Schwalb, J.M. (2010). "Bilateral deep brain stimulation for treatment of medically refractory paroxysmal nonkinesigenic dyskinesia." In: *Journal Of Neurosurgery* 112.4, pp. 847–850.

Kent, R. D. et al. (1989). "Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects". In: *Clinical Linguistics Phonetics* 3.4, pp. 347–358.

Kent, R.D. and Kim, Y.J. (2003a). "Toward an acoustic typology of motor speech disorders." In: *Clinical Linguistics and Phonetics* 17.6, pp. 427–445.

— (2003b). "Toward an acoustic typology of motor speech disorders". In: *Clinical Linguistics & Phonetics* 17.6, pp. 427–445.

Kent, R.D. et al. (1990). "Impairment of Speech Intelligibility in Men with Amyotrophic Lateral Sclerosis". In: *Journal of Speech and Hearing Disorders* 55.4, pp. 721–728.

Kent, R.D. et al. (1999a). "Acosutic studies of dysarthric speech: Methods, Progress and Potential". In: *Journal of Communication Disorders* 32.3, pp. 141–186.

— (1999b). "Acoustic studies of dysarthric speech: Methods, progress, and potential". In: *Journal of Communication Disorders* 32.3, pp. 141–186.

Kent, R.D. et al. (2000). "What dysarthria can tell us about the neural control of speech". In: *Journal of Phonetics* 28.3, pp. 273–302.

Kessens, J.M., Cucchiarini, C., and Strik, H. (2003). "A data-driven method for modeling pronunciation variation". In: *Speech Communication* 40.4, pp. 517–534.

Kim, H., Hasegawa-Johnson, M., and Perlman, A. (2011). "Vowel contrast and speech intelligibility in dysarthria". In: *Folia Phoniatrica et Logopaedica* 63.4, pp. 187–194.

Kim, H. et al. (2008). "Dysarthric speech database for universal access research". In: *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, pp. 1741–1744.

Kim, H. et al. (2010a). "Frequency of consonant articulation errors in dysarthric speech". In: *Clinical linguistics and phonetics* 24.10, pp. 759–770.

Kim, M., Oh, Y.R., and Kim, H.K. (2007). "Non-native pronunciation variation modeling using an indirect data driven method". In: *IEEE Workshop on Automatic Speech Recognition Understanding, ASRU*, pp. 231–236.

Kim, M. et al. (2017). "Regularized Speaker Adaptation of KL-HMM for Dysarthric Speech Recognition". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.9, pp. 1581–1591.

Kim, M.J. and Kim, H. (2012). "Combination of multiple speech dimensions for automatic assessment of dysarthric speech intelligibility". In: *13th Annual Conference of the International Speech Communication Association*, pp. 8–12.

Kim, S.K. et al. (2010b). "Toward detecting voice activity employing soft decision in second-order conditional map". In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, pp. 3082–3085.

Kim, Y., Kent, R.D., and Weismer, G. (2011a). "An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria." In: *Journal of Speech, Language and Hearing Research* 54.2, pp. 417–429.

— (2011b). "An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria". In: *Journal of Speech Language and Hearing research* 54.2, pp. 4177–429.

Kotler, A-L. and Stonell, N.T. (1997). "Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment". In: *Augmentative and Alternative Communication* 13.2, pp. 71–80.

Krogh, A. and Riis, S.K. (1999). "Hidden neural networks". In: *Neural Computation* 11.2, pp. 541–563.

Kuhn, R. et al. (1998). "Eigenvoices for Speaker Adaptation". In: *International Conference on Spoken Language Processing*.

Kuhn, R. et al. (2000). "Rapid speaker adaptation in eigenvoice space". In: *IEEE Transactions on Speech and Audio Processing* 8.6, pp. 695–707.

Ladefoged, P. (1993). *A course in phonetics*. Harcourt Brace Jovanovich College Publishers.

— (1996). *Elements of Acoustic Phonetics*. University of Chicago Press.

Lalitha, V., Prema, P., and Mathew, L. (2010). "A Kepstrum based approach for enhancement of dysarthric speech". In: *Image and Signal Processing (CISP), 2010 3rd International Congress on*. Vol. 7, pp. 3474–3478.

Lang, A.E. and Fishbein, V. (1983). "The "pacing board" in selected speech disorders of Parkinson's disease." In: *Journal of Neurology Neurosurgery and Psychiatry* 46.8, pp. 789–791.

Lang, B.R. (1967). "Modification of the palatal lift speech aid". In: *The Journal of Prosthetic Dentistry* 17.6, pp. 620–626.

Langhorne, P., Bernhardt, J., and Kwakkel, G. (2011). "Stroke rehabilitation". In: *The Lancet* 377.9778, pp. 1693–1702.

Lee, L. and Rose, R.C. (1996). "Speaker normalization using efficient frequency warping procedures". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96, Conference Proceedings,* vol. 1, pp. 353–356.

Leggetter, C.J. and Woodland, P.C. (1995a). "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression". In: *Proc. ARPA Spoken Language Technology Workshop*, pp. 110–115.

— (1995b). "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models". In: *Computer Speech & Language* 9.2, pp. 171–185.

Leung, K-Y. et al. (2005). "Speaker Verification Using Adapted Articulatory Feature-based Conditional Pronunciation Modeling". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Proceedings.* Vol. 1, pp. 181–184.

Li, X., Jiang, H., and Liu, C. (2005). "Large margin HMMs for speech recognition". In: *(ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.* Vol. 5, pp. v–513–v–516.

Lieu, C. A. et al. (2010). "A water extract of Mucuna pruriens provides long-term amelioration of parkinsonism with reduced risk for dyskinesias". In: *Parkinsonism and Related Disorders* 16.7, pp. 458–465.

Light, J. et al. (2007). "Children's ideas for the design of AAC assistive technologies for young children with complex communication needs". In: *Augmentative and Alternative Communication* 23.4, pp. 274–287.

Lin, Q. and Che, C. (1995). "Normalizing the vocal tract length for speaker independent speech recognition". In: *Signal Processing Letters, IEEE* 2.11, pp. 201–203.

Lisker, L. and Abramson, A.S. (1964). "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements". In: *WORD* 20.3, pp. 384–422.

Liu, Chaojun, Jiang, Hui, and Rigazio, L. (2005). "Maximum relative margin estimation of HMMS based on N-best string models for continuous speech recognition". In: *2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 420–425.

MND Org. (2013). `http://www.mndassociation.org/`. Online; accessed on: 09-September-2013.

MS Org. (2013). `http://www.mssociety.org.uk/`. Online; accessed on: 09-September-2013.

MSA Org. (2013). `http://www.msatrust.org.uk/`. Online; accessed on: 09-September-2013.

Magimai-Doss, M. and Bourlard, H. (2005). "On the adequacy of baseform pronunciations and pronunciation variants". In: *Proceedings of the First international conference on Machine Learning for Multimodal Interaction*, pp. 209–222.

Magnuson, T. and Blomberg, M. (2000). "Acoustic analysis of dysarthric speech and some implications for automatic speech recognition". In: *TMH-QPSR* 41.1, pp. 019–030.

Mahler, L.A. and Ramig, L.O. (2012). "Intensive treatment of dysarthria secondary to stroke". In: *Clinical Linguistics and Phonetics* 26.8, pp. 681–694.

Maier, A. et al. (2009). "PEAKS - A system for the automatic evaluation of voice and speech disorders". In: *Speech Communication* 51.5, pp. 425–437.

Marck, M. A. van der et al. (2009). "Multidisciplinary care for patients with Parkinson's disease". In: *Parkinsonism and Related Disorders* 15.SUPPL. 3, S219–S223.

Marshall, R.C. and Jones, R.N. (1971). "Effects of a palatal lift prosthesis upon the speech intelligibility of a dysarthric patient". In: *The Journal of Prosthetic Dentistry* 25.3, pp. 327–333.

McDonough, J. et al. (1998). "Speaker Normalization With All-Pass Transforms". In: *In Proc. ICSLP*.

McKeever, S.L. and Miller, R.M. (2002). "Glossopharyngeal breathing to improve functional vital capacity and speech production in a patient with flaccid dysarthria". In: *Journal of Medical Speech-Language Pathology* 10.4, pp. 307–311.

Mcdonough, J., Schaaf, T., and Waibel, A. (2002). "On maximum mutual information speaker-adapted training". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. Vol. 1.

Menendez-Pidal, X. et al. (1996). "The Nemours database of dysarthric speech". In: *Fourth International Conference on Spoken Language, ICSLP 96. Proceedings.* Vol. 3, pp. 1962–1965.

Mengistu, K.T. and Rudzicz, F. (2011). "Adapting acoustic and lexical models to dysarthric speech". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 4924–4927.

Mitchell, T.M. (1997). *Machine Learning*. 1st ed. McGraw-Hill, Inc.

Mohamed, A.R., Dahl, G., and Hinton, G. (2009). "Deep belief networks for phone recognition". In: *in Proceedings of the NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*.

Mohamed, A.R., Yu, D., and Deng, L. (2010). "Investigation of Full-Sequence Training of Deep Belief Networks for Speech Recognition". In: *Interspeech 2010*. International Speech Communication Association.

Morales, O.C. and Cox, S. (2007). "Modelling confusion matrices to improve speech recognition accuracy, with an application to dysarthric speech". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 1, pp. 277–280.

Morales, S.O.C and Cox, S.J. (2009). "Modelling errors in automatic speech recognition for dysarthric speakers". In: *EURASIP J. Adv. Signal Process*, 2:1–2:14.

Morris, R.J. (1989). "VOT and Dysarthria: A Descriptive Study". In: *Journal of Communication Disorders* 22.1, pp. 23–33.

Murray, J. and Goldbart, J. (2009). "Augmentative and alternative communication: a review of current issues". In: *Paediatrics and Child Health* 19.10, pp. 464–468.

Murthy, H.A. and Gadde, V. (2003). "The modified group delay function and its application to phoneme recognition". In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*. Vol. 1, pp. I–68–71.

Murthy, H.A., Hegde, R.M., and Rao, G.V.R. (2004). "The modified group delay feature: A new spectral representation of speech". In: *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*.

Murthy, H.A., Murthy, K.V.M., and Yegnanarayana, B. (1989). "Formant extraction from phase using weighted group delay function". In: *Electronics Letters* 25.23, pp. 1609–1611.

Murthy, H.A. and Yegnanarayana, B. (1991). "Formant extraction from group delay function". In: *Speech Communication* 10.3, pp. 209–221.

— (2011). "Group delay functions and its applications in speech technology". In: *Sadhana* 36.5, pp. 745–782.

Murthy, K.V.M. and Yegnanarayana, B. (1989). "Effectiveness of representation of signals through group delay functions". In: *Signal Processing* 17.2, pp. 141–150.

Myers, C.S. and Rabiner, L.R. (1981a). "A Comparitive Study Of Several Dynamic Time-Warping Algorithms For Connected-Word Recognition". In: *The Bell System technical journal* 60.7, pp. 1389–1409.

Myers, Cory S. and Rabiner, Lawrence R. (1981b). "Connected Digit Recognition Using a Level-Building DTW Algorithm". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-29.3, pp. 351–363.

Mythri, R. B., Harish, G., and Bharath, M. M. (2012). "Therapeutic potential of natural products in parkinson's disease". In: *Recent Patents on Endocrine, Metabolic and Immune Drug Discovery* 6.3, pp. 181–200.

Nadas, A., Nahamoo, D., and Picheny, M.A. (1988). "On a model-robust training method for speech recognition". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 36.9, pp. 1432–1436.

Nakagawa, S. (1984). "Comparison Of O(n) DP And Augmented Continuous DP Matching For Connected Spoken Word Recognition". In: *Proceedings - International Conference on Pattern Recognition*. Vol. 2, pp. 1236–1239.

Neiman, R. F., Mountjoy, J. R., and Allen, E. L. (1975). "Myasthenia gravis focal to the larynx. Report of a case". In: *Archives of Otolaryngology* 101.9, pp. 569–570.

Ney, H. (1984). "Use Of a One-Stage Dynamic Programming Algorithm For Connected Word Recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-32.2, pp. 263–271.

Ng, R.W.M. and Hirose, K. (2012). "Syllable: A self-contained unit to model pronunciation variation". In: *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4457–4460.

Noyes, J.M., Haigh, R., and Starr, A.F. (1989). "Automatic speech recognition for disabled people". In: *Applied Ergonomics* 20.4, pp. 293–298.

Noyes, Jan M. and Frankish, Clive R. (1992). "Speech recognition technology for individuals with disabilities". In: *Augmentative and Alternative Communication* 8.4, pp. 297–303.

Ohm, G. S. (1843). "Ueber die Definition des Tones, nebst daran geknpfter Theorie der Sirene und hnlicher tonbildender Vorrichtungen". In: *Annalen der Physik*.

Oliveira, A.S.B. and Pereira, R.D.B. (2009). "Amyotrophic lateral sclerosis (ALS): Three letters that change the people's life. For ever". In: *Arquivos de Neuro-Psiquiatria* 67.3 A, pp. 750–782.

Olivier, O. et al. (1996). "A fuzzy acoustic-phonetic decoder for speech recognition". In: *The 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA, October 3-6, 1996*.

Onishi, Y. and Iso, Ken-ichi (2003). "Speaker adaptation by hierarchical EigenVoice". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings. (ICASSP '03)*. Vol. 1, pp. 576–579.

Ono, T. et al. (2005). "Collaboration of a dentist and speech-language pathologist in the rehabilitation of a stroke patient with dysarthria: a case study." In: *Gerodontology* 22.2, pp. 116–119.

Ons, B., Gemmeke, J.F., and Van hamme, H (2014). "The self-taught vocal interface". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2014, pp. 1–16.

Oppenheim, A. V. and Lim, J. S. (1981). "The Importance of Phase in Signals". In: *IEEE Proceedings* 69, pp. 529–541.

Oppenheim, A.V. and Schafer, R.W. (1989). *Discrete-time signal processing*. International. Prentice Hall.

PD Org. (2013). `http://www.parkinsons.org.uk/`. Online; accessed on: 09-September-2013.

PSP Org. (2013). `http://www.pspassociation.org.uk/`. Online; accessed on: 09-September-2013.

Paja, M.S. and Falk, T.H. (2012). "Automated dysarthria severity classifictaion for improved objective intelligibility assessment of spastic dysarthria". In: *13th Annual Conference of the International Speech Communication Association*, pp. 1–4.

Patel, R. (2000). *Identifying Information-bearing Prosodic Parameters in Severely Dysarthric Vocalizations*. Canadian theses. Thesis (Ph.D.)–University of Toronto.

Paul, D.B. and Baker, J.M. (1992). "The Design for the Wall Street Journal-based CSR Corpus". In: *Proceedings of the Workshop on Speech and Natural Language*. HLT '91, pp. 357–362.

Pitz, M. et al. (2001). "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space". In: *In Proc. of the EUROSPEECH*, pp. 2653–2656.

Platt, L. J. et al. (1980). "Dysarthria of Adult Cerebral Palsy: I. Intelligibility and Articulatory Impairment". In: *Journal of Speech, Language, and Hearing Research* 23.1, pp. 28–40.

Platt, L.J., Andrews, G., and Howie, P.M. (1980). "Dysarthria of adult cerebral palsy: II. Phonemic analysis of articulation errors". In: *Journal of Speech, Language, and Hearing Research* 23, pp. 41–55.

Polur, P.D. and Miller, G.E. (2005a). "Effect of high-frequency spectral components in computer recognition of dysarthric speech based on a Mel-cepstral stochastic model". In: *Journal of Rehabilitation Research and Development* 42.3, pp. 363–371.

— (2005b). "Experiments with fast Fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model". In: *Neural Systems and Rehabilitation Engineering, IEEE Transactions on* 13.4, pp. 558–561.

Polur, Prasad D. and Miller, Gerald E. (2006). "Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals". In: *Medical Engineering & Physics* 28.8, pp. 741–748.

Povey, D. and Woodland, P.C. (2002). "Minimum Phone Error and I-smoothing for improved discriminative training". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 1, pp. I–105–I–108.

Povey, D. et al. (2011). "The Subspace Gaussian Mixture model-A Structured Model for Speech Recognition". In: *Computer Speech and Language* 25.2, pp. 404–439.

Pylkknen, J. and Kurimo, M. (2004). "Duration Modeling Techniques for Continuous Speech Recognition". In: *8th International Conference on Spoken Language Processing (Interspeech 2004), Jeju Island, Korea, October 4-8, 2004*, pp. 385–388.

RCP (2003). *Rehabilitation following acquired brain injury: National Clinical Guidelines*. URL: `http://www.rcplondon.ac.uk/sites/default/files/documents/rehabilitation-followingacquired-brain-injury.pdf`.

— (2004). *National clinical guidelines for stroke*. `http://www.sorcan.ca/Resources/General/NCG.pdf`.

RCSLT (2006). *Communicating Quality 3: RCSLT's Guidance on Best Practice in Service Organisation and Provision*. Royal College of Speech & Language Therapists. URL: http://books.google.co.uk/books?id=udcuAAAACAAJ.

— (2009). *Resource Manual for Commissioning and Planning Services for SLCN*. http://www.rcslt.org/speech_and_language_therapy/commissioning/aac_plus_intro. Online; accessed on: 13-May-2015.

Rabiner, L. and Juang, B.H. (1993). *Fundamental of Speech Recognition*. First. Prentice Hall.

Rabiner, L.R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2, pp. 257–286.

Rabiner, L.R. and Schafer, R.W. (2007). *An Introduction to Digital Speech Processing*. Now the essence of knowledge.

Rabiner, L.R. and Schmidt, C.E. (1980). "Application Of Dynamic Time Warping To Connected Digit Recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-28.4, pp. 377–388.

Raghavendra, P., Rosengren, E., and Hunnicutt, S. (2001). "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems". In: *AAC: Augmentative and Alternative Communication* 17.4, pp. 265–275.

Ramig, L.O. et al. (1995). "Comparison of two forms of intensive speech treatment for Parkinson disease". In: *Journal of speech and hearing research* 38.6, pp. 1232–1251.

Ramig, L.O. et al. (1996). "Intensive speech treatment for patients with Parkinson's disease: Short- and long-term comparison of two techniques". In: *Neurology* 47.6, pp. 1496–1504.

Ramig, L.O. et al. (2001). "Intensive voice treatment (LSVT) for patients with Parkinson's disease: A 2 year follow up". In: *Journal of Neurology Neurosurgery and Psychiatry* 71.4, pp. 493–498.

Rao, K. S. and Yegnanarayana, B. (2004). "Modeling syllable duration in Indian languages using neural networks". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 5, pp. V–313–16.

Rasipuram, R. and Magimai.-Doss, M. (2013). *KL-HMM and Probabilistic Lexical Modeling*. Idiap-RR Idiap-RR-04-2013. Idiap.

Ray, J. (2002). "Orofacial myofunctional therapy in dysarthria: a study on speech intelligibility." In: *The International journal of orofacial myology : official publication of the International Association of Orofacial Myology* 28, pp. 39–48.

Reichl, W. and Ruske, G. (1995). "Discriminative Training For Continuous Speech Recognition". In: *Proc. 1995 Europ. Conf. on Speech Communication and Technology*, pp. 537–540.

Riddel, J. et al. (1995). "Intelligibility and Phonetic Contrast Errors in Highly Intelligible Speakers With Amyotrophic Lateral Sclerosis". In: *Journal of Speech, Language, and Hearing Research* 38.2, pp. 304–314.

Roberts, P.E. (1985). *Speech Recognition Technology for Dysarthric Speech.* URL: `http://www.wseas.us/e-library/conferences/skiathos2002/papers/447-342.pdf`.

Robinson, T. et al. (1995). "WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition". In: *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95.,* vol. 1, pp. 81–84.

Rong, P. et al. (2012a). "Relationship between kinematics, F2 slope and speech intelligibility in dysarthria due to cerebral palsy." In: *Clinical Linguistics Phonetics* 26.9, pp. 806–822.

— (2012b). "Relationship between kinematics, F2 slope and speech intelligibility in dysarthria due to cerebral palsy". In: *Clinical Linguistics & Phonetics* 26.9, pp. 806–822.

Rudzicz, F (2007). "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech". In: *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility.* Assets '07, pp. 255–256.

Rudzicz, F. (2009). "Phonological features in discriminative classification of dysarthric speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009.* Pp. 4605–4608.

— (2010). "Towards a noisy-channel model of dysarthria in speech recognition". In: *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies.* SLPAT '10, pp. 80–88.

— (2011). "Articulatory Knowledge in the Recognition of Dysarthric Speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4, pp. 947–960.

— (2012). "Using articulatory likelihoods in the recognition of dysarthric speech". In: *Speech Communication* 54.3, pp. 430–444.

Rudzicz, F. (2013). "Adjusting dysarthric speech signals to be more intelligible". In: *Computer Speech & Language* 27.6, pp. 1163–1177.

Rudzicz, F., Namasivayam, A.K., and Wolff, T. (2012). "The TORGO database of acoustic and articulatory speech from speakers with dysarthria." In: *Language Resources and Evaluation* 46.4, pp. 523–541.

Russell, M. and Cook, A. (1987). "Experimental evaluation of duration modelling techniques for automatic speech recognition". In: *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 12, pp. 2376–2379.

Sajjan, S. C. and Vijaya, C. (2012). "Comparison of DTW and HMM for isolated word recognition". In: *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*, pp. 466–470.

Sakoe, Hiroaki (1979). "Two-Level DP-Matching - A Dynamic Programming-Based Pattern Matching Algorithm For Connected Word Recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-27.6, pp. 588–595.

Sakti, S., Markov, K., and Nakamura, S. (2008). "Probabilistic Pronunciation Variation Model Based on Bayesian Network for Conversational Speech Recognition". In: *Universal Communication, International Symposium on* 0, pp. 405–410.

Sakti, S. et al. (2010). "Korean pronunciation variation modeling with probabilistic Bayesian networks". In: *Universal Communication Symposium (IUCS), 2010 4th International*, pp. 52–57.

Samudravijaya, K., Singh, S.K., and Rao, P.V.S. (1998). "Pre-recognition measures of speaking rate". In: *Speech Communication* 24.1, pp. 73–84.

Sanders, E. et al. (2005). "Automatic recognition of dutch dysarthric speech: a pilot study." In: *INTERSPEECH*.

Sapir, S. et al. (2001). "Effects of intensive phonatory-respiratory treatment (LSVT) on voice in two individuals with multiple sclerosis". In: *Journal of Medical Speech-Language Pathology* 9.2, pp. 141–151.

Sapir, S. et al. (2002). "Speech loudness and quality 12 months after intensive voice treatment (LSVT) for Parkinson's disease: A comparison with an alternative speech treatment". In: *Folia Phoniatrica et Logopaedica* 54.6, pp. 296–303.

Sapir, S. et al. (2003). "Effects of Intensive Voice Treatment (the Lee Silverman Voice Treatment [LSVT]) on Ataxic Dysarthria: A Case Study". In: *American Journal of Speech-Language Pathology* 12.4, pp. 387–399.

Sapir, S. et al. (2007). "Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: Acoustic and perceptual findings". In: *Journal of Speech, Language, and Hearing Research* 50.4, pp. 899–912.

Saralar, M., Nock, H., and Khudanpur, S. (2000). "Pronunciation modeling by sharing Gaussian densities across phonetic models". In: *Computer Speech and Language* 14.2, pp. 137–160.

Sehgal, S. and Cunningham, S. (2015). "Model adaptation and adaptive training for the recognition of dysarthric speech". In: *6th Workshop on Speech and Language Processing for Assistive Technologies*.

Selouani, S.-A. et al. (2012). "Using speech rhythm knowledge to improve dysarthric speech recognition". In: *International Journal of Speech Technology* 15.1, pp. 57–64.

Seong-Jin, Y., Yung-Hwan, O., and Gyung, C.H. (1997). "Improved lexicon modeling for continuous speech recognition". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. Vol. 3, pp. 1827–1830.

Seong, W.K., Park, J.H., and Kim, H.K. (2012b). "Dysarthric speech recognition error correction using weighted finite state transducers based on context dependent pronunciation variation". In: *Proceedings of the 13th international conference on Computers Helping People with Special Needs - Volume Part II*, pp. 475–482.

— (2012a). "Dysarthric speech recognition error correction using weighted finite state transducers based on context-dependent pronunciation variation". In: *Proceedings of the 13th international conference on Computers Helping People with Special Needs - Volume Part II*. Vol. 7383, pp. 475–482.

Sha, F. and Saul, L.K (2007). "Large margin hidden Markov models for automatic speech recognition". In: *Advances in Neural Information Processing Systems*, pp. 1249–1256.

Sharma, H.V. and Hasegawa-Johnson, M. (2010). "State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition". In: *Proceedings of the NAACL*

HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies, pp. 72–79.

Sharma, H.V. and Hasegawa-Johnson, M. (2013). "Acoustic model adaptation using in-domain background models for dysarthric speech recognition". In: *Computer Speech and Language* 27.6, pp. 1147–1162.

Shimura, E and Kakehi, K (2011). "The effect of delayed auditory feedback on the speech quality of non-hypokinetic type dysarthrias". In: *Japan Journal of Logopedics and Phoniatrics* 52.3, pp. 233–241.

Shin, J.W. et al. (2008). "Voice activity detection based on conditional MAP criterion". In: *IEEE Signal Processing Letters* 15, pp. 257–260.

Shinoda, K. and Chin-Hui, L. (1997). "Structural MAP speaker adaptation using hierarchical priors". In: *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 381–388.

Shreekanth, T., Udayashankara, V., and Chandrika, M. (2015). "Duration Modelling Using Neural Networks for Hindi TTS System Considering Position of Syllable in a Word". In: *Procedia Computer Science* 46, pp. 60–67.

Shuanghu, B. et al. (1998). "Building class-based language models with contextual statistics". In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. 1, pp. 173–176.

Sjlander, K. (2004). *The Snack Sound Toolkit*. `http://www.speech.kth.se/snack/`.

Skelly, R., Lindop, F., and Johnson, C. (2012). "Multidisciplinary care of patients with Parkinson's disease". In: *Progress in Neurology and Psychiatry* 16.2, pp. 10–14.

Skodda, S., Visser, W., and Schlegel, U. (2010). "Short- and long-term dopaminergic effects on dysarthria in early Parkinsons disease." In: *Journal of neural transmission Vienna Austria 1996* 117.2, pp. 197–205.

— (2011). "Acoustical Analysis of Speech in Progressive Supranuclear Palsy". In: *Journal of Voice* 25.6, pp. 725–731.

Smith, M. and Kurian, M. A. (2012). "The medical management of cerebral palsy". In: *Paediatrics and Child Health (United Kingdom)* 22.9, pp. 372–376.

Spyros, M. et al. (1997). "Practical Implementations of Speaker-Adaptive Training". In: *DARPA Speech Recognition Workshop*.

Stevens, K.N. (2000). *Acoustic Phonetics*. Current Studies in Linguistics. Mit Press.

Stonell, N.T. et al. (1998). "Computerized speech recognition: influence of intelligibility and perceptual consistency on recognition accuracy". In: *Augmentative and Alternative Communication* 14.1, pp. 51–56.

Strik, H. and Cucchiarini, C. (1999). "Modeling pronunciation variation for ASR: A survey of the literature". In: *Speech Communication* 29.24, pp. 225–246.

Stroke Org. (2013). `http://hda.stroke.uk/`. Online; accessed on: 09-September-2013.

Tejaswi, S. and Umesh, S. (2017). "DNN acoustic models for dysarthric speech". In: *2017 23rd National Conference on Communications, NCC 2017*.

Theodoros, D.G. et al. (1999). "The effects of the Lee Silverman Voice Treatment program on motor speech function in Parkinson disease following thalamotomy and pallidotomy surgery: A case study". In: *Journal of Medical Speech-Language Pathology* 7.2, pp. 157–160.

Thompson-Ward, E.C., Murdoch, B.E., and Stokes, P.D. (1997). "Biofeedback rehabilitation of speech breathing for an individual with dysarthria". In: *Journal of Medical Speech-Language Pathology* 5.4, pp. 277–288.

Tjaden, K. and Liss, J. (1995). "The role of listener familiarity in the perception of dysarthric speech". In: *Clinical Linguistics Phonetics* 9.2, pp. 139–154.

Tjalve, M. and Huckvale, M. (2005). "Pronunciation variation modelling using accent features". In: *9th European Conference on Speech Communication and Technology*, pp. 1341–1344.

Tohkura, Y. (1986). "Weighted cepstral distance measure for speech recognition". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 761–764.

Tolba, H. and El Torgoman, A.S. (2009). "Towards the improvement of automatic recognition of dysarthric speech". In: *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, pp. 277–281.

Tourtellotte, W.W. et al. (1982). "Parkinson's disease: Cogentinr with sinemetr, a better response". In: *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 6.1, pp. 51 –55.

Toy, N. and Joubert, K. (2008). "Listeners' attitudes: speech supplementation strategies for improving effectiveness of speakers with mixed dysarthria as a result of motor neuron disease." In: *The South African journal of communication disorders.Die Suid-Afrikaanse tydskrif vir Kommunikasieafwykings* 55, pp. 63–76.

Tsakalidis, S., Doumpiotis, V., and Byrne, W. (2003). "Discriminative Linear Transforms for Feature Normalization and Speaker Adaptation in HMM Estimation". In: *in Proc. ICSLP*, pp. 2585–2588.

Tsao, Y., Lee, S-M., and Lee, L-S. (2005). "Segmental eigenvoice with delicate eigenspace for improved speaker adaptation". In: *IEEE Transactions on Speech and Audio Processing* 13.3, pp. 399–411.

Turner, G.S., Tjaden, K., and Weismer, G. (1995). "The Influence of Speaking Rate on Vowel Space and Speech Intelligibility for Individuals With Amyotrophic Lateral Sclerosis". In: *Journal of Speech, Language, and Hearing Research* 38.5, pp. 1001–1013.

Tyagi, V. and Wellekens, C. (2004). *On desensitizing the Mel-cepstrum to spurious spectral components for robust speech recognition.* Tech. rep.

Uebel, L. F. and Woodland, P.C. (2001). "Discriminative linear transforms for speaker adaptation". In: *ITR-Workshop on Adaptation Methods for Speech Recognition, ISCA*, pp. 61–64.

VIVOCA (2012). *Voice Input Voice Output Communication Aid.* URL: `https://www.sheffield.ac.uk/cast/projects/vivoca`.

Vachhani, B. et al. (2017). "Deep autoencoder based speech features for improved dysarthric speech recognition". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1854–1858.

Vaidya, A. B. et al. (1978). "Treatment of Parkinson's disease with the cowhage plant - Mucuna pruriens Bak". In: *Neurology India* 26.4, pp. 171–176.

Van Der Graaff, M. et al. (2009). "Clinical identification of dysarthria types among neurologists, residents in neurology and speech therapists." In: *Journal of European Neurology* 61.5, pp. 295–300.

Van Nuffelen, G. et al. (2009). "The effect of rate control on speech rate and intelligibility of dysarthric speech". In: *Folia Phoniatrica et Logopaedica* 61.2, pp. 69–75.

Van Nuffelen, G. et al. (2010). "Effect of rate control on speech production and intelligibility in dysarthria". In: *Folia Phoniatrica et Logopaedica* 62.3, pp. 110–119.

Vorperian, H.K. and Kent, R.D. (2007). "Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data". In: *Journal of Speech, Language, and Hearing Research* 50.6, pp. 1510–1545.

Walker, Francis O (2007). "Huntingtons disease". In: *The Lancet* 369.9557, pp. 218–228.

Wan, V. and Carmichael, J. (2005). "Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data". In: *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, pp. 3321–3324.

Wang, L. and Woodland, P.C. (2003). "Discriminative adaptive training using the MPE criterion". In: *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU '03*, pp. 279–284.

— (2004). "MPE-Based Discriminative Linear Transform for Speaker Adaptation". In: *International Conference on Acoustics, Speech, and Signal Processing*, pp. 321–324.

Wang, Xu, Bing-xi, Wang, and Qi, Ding (2004). "A bilinear transform approach for vocal tract length normalization". In: *Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th.* Vol. 1, pp. 547–551.

Weismer, G. et al. In:

Weismer, G. et al. (2001). "Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders". In: *Folia Phoniatrica et Logopaedica* 53.1, pp. 1–18.

Wells, J.C. (1962). "A study of the formants of the pure vowels of British English". MA thesis. University of London.

Wen, X., Liu, J., and Liu, R. (2003). "Improved algorithm for speaker verification". In: *Qinghua Daxue Xuebao/Journal of Tsinghua University* 43.1, pp. 51–54.

Wenke, R.J., Theodoros, D., and Cornwell, P. (2008). "The short- and long-term effectiveness of the LSVT for dysarthria following TBI and stroke". In: *Brain Injury* 22.4, pp. 339–352.

— (2011). "A comparison of the effects of the Lee Silverman voice treatment and traditional therapy on intelligibility, perceptual speech features, and everyday communication

in nonprogressive dysarthria". In: *Journal of Medical Speech-Language Pathology* 19.4, pp. 1–24.

Wester, M. (2003). "Pronunciation modeling for ASR - knowledge-based and data-derived methods". In: *Computer Speech & Language* 17.1, pp. 69–85.

Whitehill, T. L. et al. (2011). "Effect of LSVT on lexical tone in speakers with Parkinson's disease". In: *Parkinson's Disease*.

Witt, P.D. et al. (1995). "Do palatal lift prostheses stimulate velopharyngeal neuromuscular activity?" In: *Cleft Palate-Craniofacial Journal* 32.6, pp. 469–475.

Witten, I.H. and Bell, T. (1991). "The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression". In: *IEEE Transactions on Information Theory* 37.4, pp. 1085–1094.

Woodland, P.C. (2001). "Speaker Adaptation for Continuous Density HMMs: A Review". In: *ITRW on Adaptation Methods for Speech Recognition*, pp. 11–19.

Wrench, A. (1999). *The MOCHA-TIMIT articulatory database*. URL: `http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html`.

Xiong, W. et al. (2016). "Achieving Human Parity in Conversational Speech Recognition." In: *CoRR* abs/1610.05256.

Xu, L. and Ke, M. (2012). "Research on isolated word recognition with DTW-based". In: *2012 7th International Conference on Computer Science Education (ICCSE)*, pp. 139–141.

Yakcoub, M.S., Selouani, S., and O'Shaughnessy, D. (2008). "Speech assistive technology to improve the interaction of dysarthric speakers with machines". In: *Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on*, pp. 1150–1154.

Yamamoto, H., Isogai, S., and Sagisaka, Y. (2001). "Multi-Class Composite N-gram language model for spoken language processing using multiple word clusters". In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. ACL '01, pp. 531–538.

Yang, L. et al. (2011). "Improvement algorithm of DTW on isolated-word recognition". In: *2011 IEEE International Conference on Computer Science and Automation Engineering*. Vol. 3, pp. 319–322.

Yegnanarayana, B. (1978). "Formant extraction from linear-prediction phase spectra". In: *The Journal of Acoustical Society of America* 63.5, pp. 1638–1640.

Yegnanarayana, B. and Murthy, H.A. (1992). "Significance of group delay functions in spectrum estimation". In: *IEEE Transactions on Signal Processing* 40.9, pp. 2281–2289.

Yilmaz, E. et al. (2016). "Combining non-pathological data of different language varieties to improve DNN-HMM performance on pathological speech". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 218–222.

— (2017). "Multi-stage DNN training for automatic recognition of dysarthric speech". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2685–2689.

Yorkston, K. M., Spencer, K. A., and Duffy, J. R. (2003). "Behavioral management of respiratory/phonatory dysfunction from dysarthria: a flowchart for guidance in clinical decision making." In: *Journal of Medical SpeechLanguage Pathology* 11.2, pp. xxxix–xi.

Yorkston, K.M. and Beukelman, D.R. (1984). *Assessment of Intelligibility of dysarthric speech*. Austin, TX : Pro-Ed.

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

Yorkston, K.M. et al. (2012). *Management of Speech and Swallowing Disorders in Degenerative Diseases*. Pro-Ed.

Young, S.J., Russell, N.H., and Thornton, J.H.S (1989). *Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems*. Tech. rep.

Yu, K., Gales, M., and Woodland, P.C. (2009). "Unsupervised Adaptation With Discriminative Mapping Transforms". In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.4, pp. 714–723.

Yu, K. and Gales, M.J.F. (2006). "Discriminative cluster adaptive training". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.5, pp. 1694–1703.

Yu, K., Gales, M.J.F., and Woodland, P.C. (2008). "Unsupervised discriminative adaptation using discriminative mapping transforms". In: *IN PROC. ICASSP*, pp. 4273–4276.

Zhao, X. and Wang, D. (2013). "Analyzing noise robustness of MFCC and GFCC features in speaker identification". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7204–7208.

Zhu, D. and Paliwal, K.K. (2004). "Product of power spectrum and group delay function for speech recognition". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1, pp. I–125–8.

Zyski, B. J. and Weisiger, B.E. (1987). "Identification of dysarthria types based on perceptual analysis." In: *Journal of Communication Disorders* 20.5, pp. 367–378.