Automatic Recognition and Assessment of Dysarthric Speech

Chitralekha Ganapati Bhat

Centre for Language Studies

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Automatic Recognition and Assessment of Dysarthric Speech

Chitralekha Ganapati Bhat

Chitralekha Ganapati Bhat
Automatic Recognition and Assessment of Dysarthric Speech

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Chitralekha Ganapati Bhat
Cover: Chitralekha Ganapati Bhat (created using AI)
Printing: DPN Rikken/Pumbo

ISBN: 9789465150314
DOI: 10.54195/9789465150314
Free download at: https://doi.org/10.54195/9789465150314

RADBOUD
UNIVERSITY
PRESS

Automatic Recognition and Assessment of Dysarthric Speech

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 24 maart 2025
om 10.30 uur precise

door

Chitralekha Ganapati Bhat
geboren op 8 april 1978
te Sagar (India)

Promotor:
Dr. W.A.J. Strik

Copromotoren:
Dr. C. Cucchiarini
Dr. S. K. Kopparapu (Tata Consultancy Services, India)
Prof. dr. R.W.N.M. van Hout

Manuscriptcommissie:
Prof. dr. M.T.C. Ernestus
Prof. dr. ing. E. Nöth, (Friedrich-Alexander-Universität, Duitsland)
Prof. dr. B. Möbius (Universität des Saarlandes, Duitsland)
Prof. dr. P.A.M. Gerrits (Universiteit Utrecht)
Dr. L.J. Beijer (Hogeschool van Arnhem en Nijmegen)

Automatic Recognition and Assessment of Dysarthric Speech

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Monday, March 24, 2025
at 10.30 am

by

Chitralekha Ganapati Bhat
born on April 8, 1978
in Sagar (India)

Supervisor:
Dr. W.A.J. Strik

Co-supervisors:
Dr. C. Cucchiarini
Dr. S. K. Kopparapu (Tata Consultancy Services, India)
Prof. dr. R.W.N.M. van Hout

Manuscript Committee:
Prof. dr. M.T.C. Ernestus
Prof. dr. ing. E. Nöth, (Friedrich-Alexander-Universität, Germany)
Prof. dr. B. Möbius (Saarland University, Germany)
Prof. dr. P.A.M. Gerrits (Utrecht University)
Dr. L.J. Beijer (HAN University of Applied Sciences)

Contents

Acknowledgment

Completing this doctoral journey has been one of the most challenging yet rewarding experiences of my life. It would not have been possible without the guidance, encouragement, and support of many incredible individuals who have stood by me every step of the way.

First and foremost, I extend my deepest gratitude to my supervisor, Dr. W.A.J. Strik. Our journey began at the SLPAT, 2016 conference in San Francisco, where we first discussed the possibility of pursuing my PhD under his guidance. It was a difficult decision for me due to my inability to relocate from Mumbai. Dr. Strik's patience and understanding, particularly when I mentioned the option of enrolling as an external candidate at Radboud University, were pivotal. His willingness to explore this opportunity for me has changed my life. During the pandemic, his kind and supportive words helped me regain confidence when I felt low. Working with Dr. Strik has been truly inspiring; his gentle guidance, dedication to quality, and unwavering support have left an indelible mark on my academic journey. I am immensely grateful for his mentorship.

Dr. Sunil K. Kopparapu has been a guiding light since the beginning of my career as a researcher. His advice, encouragement, and belief in my potential have been steadfast through thick and thin, especially during times when our articles faced rejection. His ability to channel my energy in positive ways has helped me grow both as a researcher and as a person. I cannot thank him enough for his unwavering support and mentorship. I am also deeply thankful to Dr. Catia Cucchiarini and Prof. Dr. R.W.N.M. van Hout for their critical reviews and invaluable feedback on our articles. Their expertise and constructive criticism during research group meetings significantly improved the quality of my work. Their patience and support have been a cornerstone of my progress.

I would also like to take this opportunity to express my heartfelt gratitude to Prof. Preeti Rao, my supervisor at the Digital Audio Processing Lab at IIT Bombay, where my journey into speech technology research began, and to Dr. Say-Wei Foo, who guided me during my MSc at NTU, Singapore. Their mentorship has profoundly shaped my career path and continues to inspire me.

I would also like to extend my heartfelt gratitude to the manuscript committee members Prof. Ernestus, Prof. Nöth, Prof. Möbius, Prof. Gerrits and Dr. Beijer for taking the time to carefully read my thesis and find improvements. I look

forward to seeing (some of) you in person in Nijmegen in March.

To my parents, especially my father, who never stopped asking me when I would finish—thank you for keeping me accountable and motivated. To my husband and daughter, who have witnessed every high and low of this journey, your unwavering faith and strength have been my pillars. To my sister, whose belief in me has been unshakable, I am forever grateful. My deepest gratitude goes to my parents-in-law, whose support has been invaluable in enabling me to achieve this milestone.

A special note of appreciation goes to my co-author and partner in crime, Bhavik. From the moment we celebrated the acceptance of our first peer-reviewed paper on the automatic recognition of dysarthric speech at Interspeech 2016, you have been an integral part of this journey. Despite the challenges of navigating new territory, including the atypical nature of dysarthric speech and the intricacies of deep neural networks, we persevered together. Thank you for your camaraderie and dedication.

I would also like to thank Dr. Biswajit Das and Dr Ashish Panda, with whom I co-authored papers. Your collaboration enriched my work and broadened my perspective.

To my friends who have rooted for me from day one, and to my colleagues at TCS Research, thank you for your constant encouragement and support. Your belief in me has been a source of strength, and I deeply value the camaraderie we share.

Finally, I dedicate this achievement to everyone who has been instrumental in this journey. Each of you has played a unique role in shaping my path, and I am profoundly grateful for your presence in my life. Thank you for making this milestone possible.

Chapter 1

Introduction

1.1 General Introduction

Speech is a fundamental aspect of human communication, serving as a primary means for conveying thoughts, emotions, and information within society. Its significance lies in its ability to facilitate interpersonal relationships, share knowledge, and contribute to social cohesion. Effective speech communication plays a crucial role in education, business, healthcare, and other domains, underscoring its importance in human interaction. What's truly remarkable is that the act of speaking becomes so automatic that it often escapes conscious thought. Nonetheless, the sheer intricacy and complexity of speech sets it apart from all other human movements. The generation of an utterance entails the careful selection, sequencing, and articulation of pertinent information in a highly time-sensitive manner. Furthermore, as these sequential processes unfold, the system integrates sensory feedback, which is essential for the execution of a skilled task, such as speech production (Asan et al., 2022) [8].

Speech production constitutes an intricately complex motor activity wherein respiratory, laryngeal, and supraglottal vocal tract articulators collaborate in a highly synchronized manner. Typically, every speech outcome involves the coordinated movement of multiple articulators. The production of an isolated vowel necessitates a seamless interaction of the jaw, tongue, lips, larynx, and respiratory system. The foundation of this intricate motor process lies in the speech motor control system. This system adeptly integrates auditory, somatosensory, and motor information, which is represented in the temporal, parietal, and frontal cortex, respectively, along with associated sub-cortical structures. This integration allows for the production of fluid and comprehensible speech, whether the task involves generating a simple nonsense syllable or pronouncing a single meaningful word (Kotz & Schwartze, 2016; Kearney & Guenther, 2019) [111, 89]. Kearney & Guenther (2019) [89] present a historical view describing the neural mechanisms of speech-motor control. The initiation of a speech sound, whether it be a commonly produced phoneme, syllable, or word, commences with the activation of its neural representation within

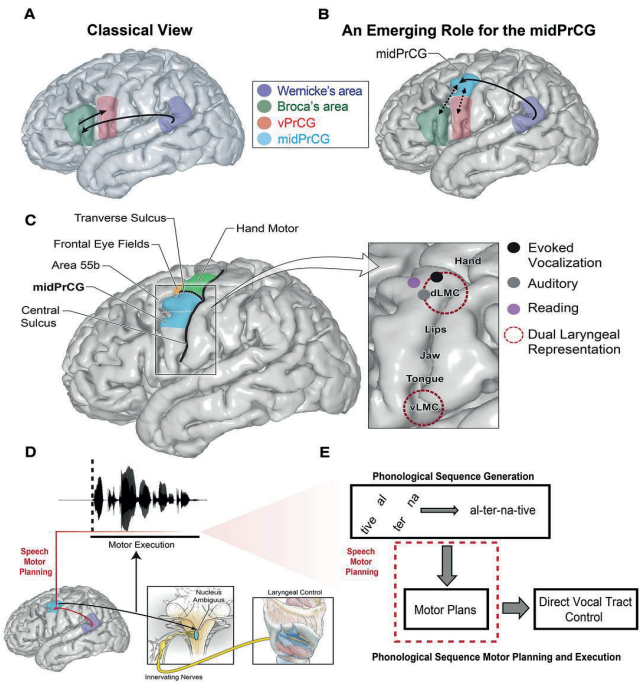

Figure 1.1: Neurological and anatomical foundations of speech [184]

a hypothesized *speech sound map* located in the left ventral premotor cortex. Figure 1.1, from a more recent study by Silva et al. (2022) [184] emphasizes the neurological and anatomical foundations of speech. It shows the various brain regions involved in the planning, formulation, and execution of speech, along with the vocal tract organs responsible for shaping sounds like the larynx, tongue, and lips. The study challenges the classic language model that posits that a single frontal region, Broca's area, is responsible for speech motor planning and proposes that the middle precentral gyrus (midPrCG), is crucial for speech production, bridging hand and orofacial cortical regions. These studies highlight the complexity of the speech production mechanism, damage to which can significantly impact human interactions and communication and, in turn, the quality of life of the person affected.

Motor speech disorders refer to disruptions in the systems and mechanisms governing the movements essential for speech production. These disorders arise from disturbances in muscular control, manifesting as weaknesses, slowness, or incoordination in the speech mechanism due to central nervous system damage. This term encompasses concurrent neurogenic disorders affecting various or all the fundamental processes of speech production, including respiration, phonation, resonance, articulation, and prosody.

Dysarthria, stemming primarily from brain damage, often due to conditions like stroke, is a group of disorders that hinder speech intelligibility. It is characterized by weakness, slowness, lack of coordination, and imprecise movements in the speech muscles. Dysarthria is categorized into progressive and non-progressive types. Progressive dysarthrias are observed in conditions such as Parkinson's disease, Huntington's disease, multiple sclerosis, motor neuron disease, etc. While some cases may exhibit delayed decline, individuals with progressive dysarthria typically experience a gradual decrease in muscle function over time. Conversely, non-progressive dysarthrias resulting from conditions like stroke and traumatic brain injury (TBI) may show improvement in muscle function with appropriate treatment (Qualls, 2012) [160]. More than 80% of people with motor neuron disorders (MND) experience dysarthria. The exact incidence of dysarthria remains uncertain and varies depending on the underlying cause. Approximately 90% of individuals with Parkinson's disease (PD) experience dysarthria at some point during the illness. In patients with amyotrophic lateral sclerosis (ALS), dysarthria may precede limb weakness by 3 to 5 years and affects around 70% of those with limb weakness. In a study involving stroke patients, 28% were found to have both aphasia and dysarthria, while 24% had dysarthria alone. Among children with neuromuscular diseases, the prevalence of dysarthria was reported to be 31.5%. Additionally, it is estimated that 10% to 60% of individuals with traumatic brain injury (TBI) develop dysarthria. The statistics reported here are based on a study by Jayaraman & Das (2023) [78]. Typically, 25-30% of people with MND have dysarthria as a first or predominant sign in the early stage of the disease [1]. A detailed study on dysarthric speech and research on speech technology for dysarthric speech is presented in Chapter 2. The sheer number of people affected by dysarthria indicate that Artificial Intelligence (AI)-based objective and automatic systems for assessing and recognizing dysarthric speech would go a long way in integrating persons with dysarthria into society by addressing the therapy and assisted living needs.

Our research aims to explore the new horizons and applications of speech technology, enabling people with dysarthria to avail themselves of the same opportunities accorded to persons with normal speech, thereby improving their quality of life. The following sections in this chapter briefly outline the research work incorporated into this thesis, with a focus on research questions that form the basis of this work.

1.2 The Present Research: Objective Assessment and Recognition – System

In this thesis, we explore two different aspects of the application of speech technology to dysarthric speech, namely (1) Automatic intelligibility assessment and (2) Automatic recognition of dysarthric speech. These aspects find applications in assisted speech therapy, serving as a second opinion to speech pathologists and enabling them in therapy planning. Secondly, efficient automatic speech recognition

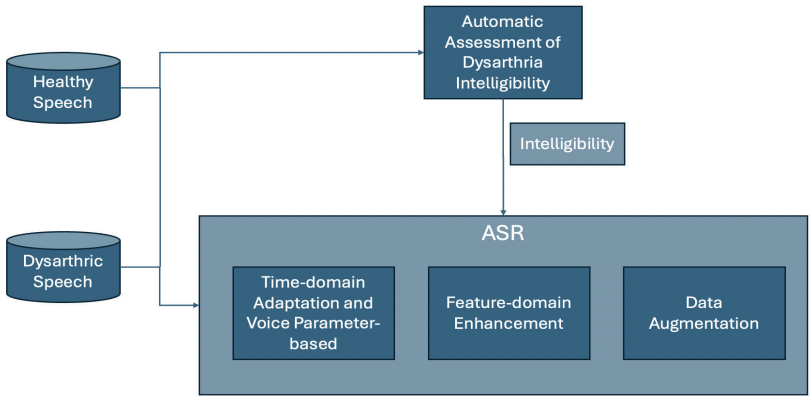

Figure 1.2: An Intelligent System for Automatic Intelligibility and Speech Recognition of Dysarthric Speech

is aimed at building systems for assisted living. We believe that intelligibility assessment has the potential to aid in the automatic recognition of dysarthric speech. Figure 1.2 depicts a system that can be built using this research work. Figure 1.2 is representative of the components of the research work conducted and described in this thesis. An extensive survey of the research carried out in the interdisciplinary area of speech technology and dysarthric speech is presented in Chapter 2. Starting with a nominal amount of dysarthric speech and healthy speech data, which is essential to benchmark the performance of the ASR systems in various scenarios, we design modules for Automatic Intelligibility Assessment (AIA) and Recognition of dysarthric speech. The intelligibility information from the AIA module acts as an important input in designing the ASR modules. Four different techniques for improving ASR performance for dysarthric speech have been explored as a part of this thesis, namely 1. Time domain adaptation, 2. Harnessing acoustic parameters, 3. Speech enhancement and, 4. Data augmentation. Each of these techniques was explored in terms of their applicability to dysarthric speech.

As is common with AI research involving pathological speech, the availability of valid speech data restricts the application of Deep Neural Network-based techniques to evaluate or recognize dysarthric speech. Through this work, we explore mechanisms to overcome this limitation and significantly improve the state-of-the-art (SOTA) for automatic assessment and recognition of dysarthric speech.

1.3 Research Questions and Outline

This dissertation attempts to gain insights into the development of techniques for objective measurement of the intelligibility of dysarthric speech and automatic

recognition of dysarthric speech. To this end, this dissertation evaluates both automatic assessment and recognition of dysarthric speech by addressing three research questions:

1. RQ1: What is the status of research into the interdisciplinary area of dysarthric speech and speech technology and which knowledge gaps should be addressed? **Chapter 2**

2. RQ2: How can we efficiently and automatically assess the intelligibility of dysarthric speech? **Chapter 3**

3. RQ3: How can we improve the automatic recognition of dysarthric data in terms of word error rate? **Chapter 4, 4 and, 6**.

   Additionally, building a dysarthric speech corpus is challenging. Therefore, speech researchers are limited by data availability to implement the latest machine learning techniques for automatic recognition of dysarthric speech. Considering this challenge, how can we adopt the latest techniques in speech technology for automatic recognition of dysarthric speech?

To address the first research question, we carry out an extensive literature survey of speech technology that has been developed for dysarthric speech in **Chapter 2**, starting with the dysarthric speech corpora chronicled, acoustic studies that are critical in understanding the characteristics of dysarthric speech, followed by automatic intelligibility assessments and finally research work that outlines Automatic Speech Recognition (ASR) systems. This survey highlights the development of the interdisciplinary research area of speech technology and dysarthric speech over the last two decades. It aims to serve as a basis for future research work in this area.

Note: Chapters 3 to 6 each contain two published research papers. These papers have been categorized as part of the same chapter because they address the same research question by applying similar techniques; i.e. research question 2 in Chapter 3, and research question 3 in Chapters 4, 5, and 6, respectively.

In **Chapter 3**, we present two studies that delve into automatic intelligibility assessment of dysarthric speech. Speech intelligibility is considered to be an indicator of the severity of dysarthria. The objective is to enable a speech pathologist to understand the patient's status, especially in case of a progressive neuromotor disease. The outcome of objective assessment methods also serves as an electronic medical record (ERM). The first study by Bhat et al. (2017) [15] uses an Artificial Neural network (ANN)-based classification of dysarthric speech with a high correlation with subjective assessment. The second study by Bhat & Strik (2020) [12] uses a sophisticated Bidirectional Long-short Term Memory (BLSTM)-based system and Transfer Learning to classify sentence-level dysarthric speech into intelligible and non-intelligible categories.

In **Chapters 4, 5, and 6**, we explore various mechanisms of improving ASR performance in terms of word error rate (WER). The literature review in Chapter 2 suggests that our research is pioneering in the application of these techniques.

We present two studies in **Chapter 4**, (1) Multi-tapered spectral representations of dysarthric speech along with specific voice parameters were used in a DNN-HMM framework (Bhat et al., 2016a) [13] (2) A mechanism to adapt the tempo of the sonorant part of dysarthric speech to match that of normal speech. We leverage the knowledge of the severity of dysarthria to design a system for this purpose (Bhat et al., 2016b) [14]. In both the above studies we harness an understanding of the acoustic characteristics of dysarthric speech to improve the ASR performance.

In **Chapter 5**, we discuss feature-domain enhancement of dysarthric speech using auto-encoders. The objective of both the studies is to enhance dysarthric speech in such a way, as to be able to use an ASR, trained mainly on healthy data for automatic recognition of dysarthric speech. Two different designs of autoencoders are presented; the first one by Vachhani et al. (2017) [195] uses only healthy speech to generate the bottleneck layer, whereas the second one (Bhat et al., 2018) [16] is a Time-delay Neural Network autoencoder that is modeled on a denoising autoencoder with healthy speech representing clean speech and dysarthric speech representing noisy speech. We analyze severity-based tempo adaptation followed by autoencoder-based speech feature enhancement. ASR performance in speaker-dependent and speaker-independent training scenarios was examined.

In **Chapter 6**, we address the dearth of speech data for dysarthric speech. We propose the use of traditional data augmentation techniques along with data augmentation specifically designed to emulate dysarthric speech. These techniques are applied to healthy speech. While the ASR performance showed improvements for the speaker-independent scenarios, maximum improvements were observed when speaker adaptation was used. Two studies (Vachhani et al., 2018; Bhat et al., 2022) [196, 17] have been presented in this chapter. In Vachhani et al. (2018)[196], we discuss augmentation techniques using only healthy speech data transformed in the time domain. In Bhat et al. (2022) [17], we design an end-to-end DNN-based system trained on augmented dysarthric data, with augmentation techniques that specifically use the characteristics of dysarthric speech for augmentation.

It is to be noted that Chapters 2 to 6 are each a compilation of peer-reviewed publications in related areas co-authored by the author of the current thesis. All research papers were published as self-contained publications. This indicates that there is an overlap in information about the research area (particularly in the introduction) and the data and techniques used. We decided not to edit these aspects to maintain the integrity of the original publications. In Chapter 7, we discuss our contributions and the impact of our research on the research questions raised in Section 1.3.

Chapter 2

Speech Technology for Automatic Recognition and Assessment of Dysarthric Speech: An Overview

This chapter is based on the following publications:

Bhat, C., and Strik, H. (2025) Speech Technology for Automatic Recognition and Assessment of Dysarthric Speech: An Overview
Journal of Speech, Language, and Hearing Research, Pages 1-31.

Purpose: In this chapter, we present an extensive overview of recent developments in the area of dysarthric speech research. One of the key objectives of speech technology research is to improve the quality of life of its users, as evidenced by the focus of current research trends on creating inclusive conversational interfaces that cater to pathological speech, out of which dysarthric speech is an important example. Applications of speech technology research for dysarthric speech demand a clear understanding of the acoustics of dysarthric speech as well as of speech technologies, including machine learning and deep neural networks for speech processing.
Method: We review studies pertaining to speech technology and dysarthric speech. Specifically, we discuss dysarthric speech corpora, acoustic analysis, intelligibility assessment, and automatic speech recognition. We also delve into deep learning approaches for automatic assessment and recognition of dysarthric speech.
Ethics Committee or Institutional Review Board did not apply to this study.
Conclusion: Overcoming the challenge of limited data and exploring new avenues in data collection, AI-powered analysis, and teletherapy hold immense potential for significant advancements in dysarthria research. To make longer and faster strides, researchers typically rely on existing research and data on a global scale. Therefore, it is imperative to consolidate the existing research and present it in a form that can serve as a basis for future work. In this chapter, we have reviewed the contributions of speech technologists to the area of dysarthric speech with a focus on acoustic analysis, speech features, and techniques used. By focusing on the existing research and future directions, researchers can develop more effective tools and interventions to improve communication, quality of life, and overall well-being for people with dysarthria.

Keywords: Dysarthric speech, speech corpora, acoustics characteristics, Intelligibility, ASR

2.1 Introduction

Speech production is one of the most complex human motor skills and involves both linguistic units and acoustic events. Motor speech problems caused by neurological difficulties can be congenital or acquired, impacting one or several speech subsystems, namely, respiratory, phonatory, and articulatory. For intelligible speech production, the muscles and muscle groups in these subsystems must be well coordinated in time and space. Manifestations of dysarthria may include restricted lip, tongue, and jaw movement, abnormal speech rate or volume, breathy or hoarse voice, drooling, and swallowing difficulty. Congenital dysarthria can be caused by an inherited condition, such as cerebral palsy (CP), which affects the muscles used for speech production. Dysarthria acquired later in life may result from stroke, brain injury, tumors, infection, or progressive neurological diseases such as amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS), or Parkinson's disease (PD). An overview study by Poole & Vogel (2020) [154] lists the following six types

of dysarthria.

- Flaccid dysarthria: damage to the cranial nerves or regions of the brainstem and midbrain. Distinguishing features include breathy voices, short phrases, increased nasal resonance, and imprecise articulation.
- Spastic dysarthria: damage to the motor regions in the cortex on both sides of the brain; distinguishing features are strained voice, monotonicity, and slow rate.
- Ataxic dysarthria: damage to pathways connecting the cerebellum and other brain regions. Distinguishing features include irregular articulatory errors, equal and excessive stress on syllables, and inappropriate pitch and loudness variations.
- Hypokinetic dysarthria: rigidity and bradykinesia resulting from impairment of the basal ganglia control circuit. The distinguishing features include reduced loudness, rapid speech rate, sound repetition, and reduced stress.
- Hyperkinetic dysarthria: involuntary movements associated with impairment of the basal ganglia control circuit. It is characterized by unpredictable speech mechanism movements.
- Unilateral upper motor neuron (UUMN) dysarthria is caused by unilateral damage to UMNs. Distinguishing features include a hoarse voice, imprecise articulation, and slow rate.

The factors causing dysarthria as well as the characteristics of dysarthria vary and extend across a wide range of possibilities, posing challenges to machine-based objective assessments and analyses of dysarthric speech. Furthermore, mixed dysarthria, which involves the features of two or more types of dysarthria, and co-morbidities make it more complex. Recently, we have witnessed an increase in the availability, adaptation, and popularity of speech-enabled interfaces, especially in the assisted and smart living domains. Speech is a more convenient alternative to other machine interfaces, such as remote controls, keyboards, or PC mice, given that persons with dysarthria are often faced with physical inabilities too (Rudzicz, 2010) [165]. While traditional off-the-shelf automatic speech recognition (ASR) systems perform well for unimpaired speech, this is not the case with atypical dysarthric speech owing to the inter-speaker and intra-speaker deviations in the acoustic space as well as the sparseness of speech data that can be used for training the ASR algorithms. However, we want persons suffering from dysarthria to benefit as much as possible from current technological advances in automatic speech processing. A plethora of studies and research work on dysarthric speech from a speech signal processing technology perspective have paved the way for techniques that have improved automatic and objective assessments, such as the automatic intelligibility assessment (AIA) and ASR for dysarthric speech.

This review article examines how speech technology can aid individuals with dysarthria. We focus on speech corpora, acoustic analysis, AIA, and ASR for dysarthric speech. The objective is to provide an extensive overview of what has been done so far to pave the way for future work on the automatic assessment of dysarthric speech. A search was carried out on Google Scholar, ScienceDirect, IEEE Xplore, SCOPUS, and ACM using keywords specific to the topic discussed in each section of the current review article. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart for the selection of review literature is shown in 2.1. The studies reported were chosen based on their relevance from a speech technology perspective. In the case of similar studies by the same authors/group of researchers, preference was given to the most recently published article. The availability of well-documented speech corpora is crucial for speech technology-based research. In the Dysarthric Speech Corpora section, we present an overview of the existing dysarthric speech corpora that are available and have been used for the development of speech technologies for dysarthric speech. We have also described the clinical evaluation criteria typically used for dysarthria evaluation within these corpora. Keywords used were: 'dysarthric speech corpus', 'dysarthric speech database', and 'dysarthric speech data'. The acoustic analysis of speech signals is a valuable tool for understanding the complex processes underlying speech production and can provide important insights into speech disorders. Understanding the acoustic characteristics of dysarthric speech is crucial in the design of automatic speech processing techniques. Therefore, we delve into studies of the acoustic analysis of dysarthric speech in Section 2.3. Keywords used (not limited to) were: 'dysarthric speech acoustics', 'dysarthric speech acoustic studies', 'dysarthric speech acoustic characteristics', 'dysarthric speech acoustic analysis', and 'dysarthric speech characteristics'. Traditionally, the severity of dysarthria has been evaluated through perceptual evaluations by human experts such as speech-language pathologists. However, such evaluations can be subject to inter- and intra-rater variability, which can affect the reliability and validity of assessments. In Section 2.4, we discuss research on the features and techniques used for the automatic severity level assessment of dysarthric speech. Keywords used (not limited to) were: 'dysarthric speech intelligibility/severity', 'automatic assessment of dysarthria intelligibility/severity', and 'automatic assessment of dysarthric speech intelligibility/severity'. An overview of the research trends in ASR for dysarthric speech is discussed in Section 2.5. Keywords used (not limited to) were: 'dysarthric speech recognition', 'automatic recognition of dysarthric speech', and 'automatic speech recognition for dysarthric speech'.

Each section comprises subsections that are categorized based on the techniques used to achieve the objective described in that section. Within the sections, we mainly present the studies in chronological order. Finally, in Section 2.6, we look into the various possibilities of research avenues for dysarthric speech.

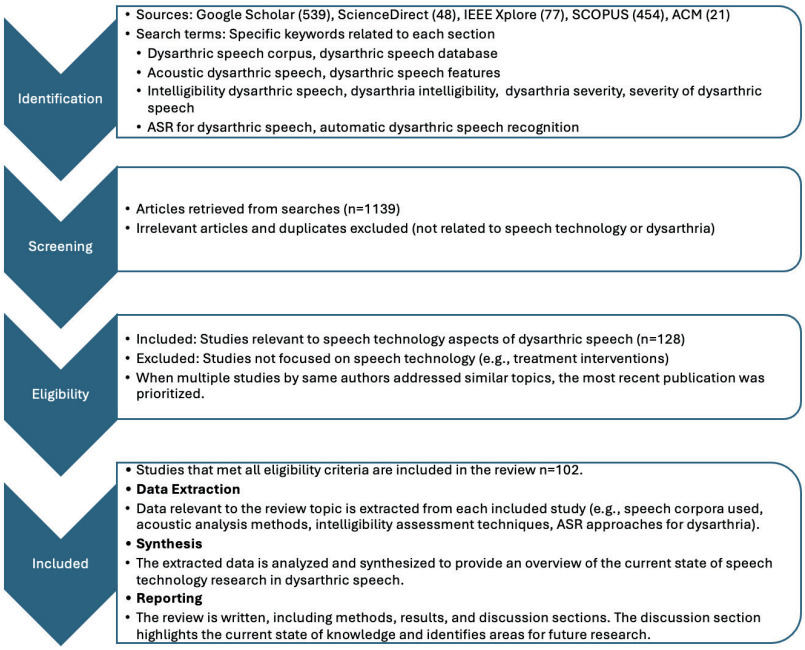

Figure 2.1: PRISMA flowchart for the selection of review literature

2.2 Dysarthric Speech Corpora

To build a high-performance speech technology-based system, suitable data are imperative. However, owing to muscle weakness and fatigue in dysarthric speakers, collecting speech from these subjects can be time-consuming and tedious, especially for speakers with severe dysarthria. Additionally, since dysarthria can stem from a variety of neurological disorders including mixed dysarthria caused by multiple comorbidities, the characterization of dysarthric speech is complex. This complexity translates into challenges during the design of a data collection process. Speech corpora typically include acoustic data as well as transcriptions. In addition, dysarthric speech corpora often provide speaker information, such as gender, age, and dysarthria severity level or speech intelligibility. This section explores dysarthric speech corpora, which are valuable resources used extensively for developing ASR and assessment techniques for dysarthric speech. Details regarding speakers, cause of dysarthria, clinical evaluation criteria used wherever applicable, severity levels, and speech material for the databases described are outlined in Table 2.1. A brief description of the clinical evaluation criteria typically used for dysarthria evaluation within these corpora has been provided in the Clinical Evaluation Scales for Dysarthric Speech section.

2.2.1 Dysarthric Speech Corpora

The Whitaker Database of Dysarthric (Cerebral Palsy) Speech, which comprises the speech of six speakers with CP and one healthy speaker, is an earlier and relatively small collection of dysarthric speech (Deller et al., 1993) [33]. The Nemours database (Menendez-Pidal et al.,1996) [126] comprises the Frenchay Dysarthria Assessment (FDA) [37] of each speaker given by a speech pathologist, perception data for each sentence from listening tests by five listeners, and audio data. The corpus also provides word- and phoneme-level transcription for each sentence. The universal access (UA) dysarthric speech corpus (Kim et al., 2008) [95] comprises video data in addition to audio data to allow the exploration of multimodal dysarthric speech signals. Speaker-wise speech intelligibility provided in this corpus was computed using word transcription tasks performed by untrained human listeners. The TORGO (Rudzicz et al., 2011) [168] database of dysarthric articulation consists of aligned acoustics and measured three-dimensional articulatory features from speakers' dysarthria, as well as matched controls. A three-dimensional electromagnetic articulograph (EMA) was used to analyze speech production. The EMA system features automatic calibration and allows for high-precision, three-dimensional recordings of articulatory movements, both within the vocal tract and externally, providing a comprehensive view of speech-related activity. Detailed physiological information is expected to enable the explicit learning of hidden articulatory parameters automatically via statistical pattern recognition. The above databases are most commonly used for dysarthric speech research in American English. Dysarthric speech corpora have been created to study dysarthria in languages other than English as well. Two French corpora, namely, the Dr. Claude Chevrie- Muller corpus and the Aix-Neurology-Hospital corpus, were described by Fougeron et al. (2010) [44]. A Korean dysarthric speech corpus was built as part of the Quality of Life Technology (QoLT) project that focuses on developing speech technologies for people with articulation disabilities (Choi et al., 2011) [25]. A Cantonese corpus focusing on the investigation of the articulatory and prosodic characteristics of Cantonese dysarthric speech is presented by Wong et al. (2015) [205]. Yılmaz et al. (2016) [213] describe a Dutch dysarthric speech database containing mildly to moderately dysarthric speech from patients with PD, traumatic brain injury (TBI), or cerebrovascular accidents. To study dysarthric speech in the Indian context, a Tamil dysarthric speech corpus of 22 speakers across age groups was created (Celin et al., 2016) [2]. Spanish (Orozco-Arroyave et al., 2014) [141], Czech (Rusz et al., 2011) [169] and German (Skodda et al., 2011) [185] corpora were collected to study dysarthric speech in patients with PD. EasyCall corpus (Turrisi et al., 2021) [194] is a dysarthric speech corpus in Italian compiled with the primary objective of serving as a valuable resource for the advancement of assistive technologies based on Automatic Speech Recognition (ASR) for individuals with dysarthria. The article by Marini et al. (2021) [122] describes IDEA, a database of Italian dysarthric speech produced by 45 speakers affected by eight different pathologies such as ALS, MS,

PD, TBI, and stroke as can be seen in Figure 2.2. While most corpora comprise data collected in clinical settings, Nicolao et al. (2016) [138] describe *the home-Service corpus*, a British English corpus of realistic dysarthric data collected over time, in the home environment. This corpus showcases speech data collected from five speakers with severe dysarthria as part of their daily interactions with their devices. Each of the above databases was designed for a specific purpose with a broad perspective on improving the lives of people with dysarthria. Speech tasks used to assess dysarthric speech play a significant role in the outcome of analysis and recognition of dysarthric speech. From Table 2.1 it can be observed that some datasets use sustained phonation, reading, monologue, or diadochokinetic (DDK) evaluation, but some others include only sustained phonation or reading of a short text. Different speech tasks can challenge people with dysarthria to varying degrees. Reading simple sentences might be easier than carrying out a conversation or delivering a monologue. The type of speech task selected for research on automatic dysarthria analysis plays a dual role. Firstly, the task complexity can influence how clearly the characteristics of dysarthria manifest themselves in the speech samples. Secondly, the chosen task determines the kind of speech data collected, which subsequently shapes the development and training of automatic analysis models.

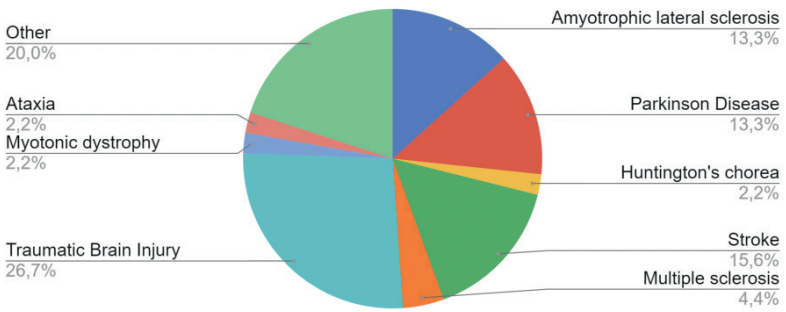

Figure 2.2: Percentage of pathologies present in IDEA database

2.2.2 Clinical Evaluation Scales for Dysarthric Speech

Clinicians rely on standardized evaluation scales to assess dysarthria. These scales go beyond simple perception of speech difficulties. Some provide a detailed breakdown of speech intelligibility, articulation, voice quality, and other aspects, while others focus on the functional impact of dysarthria. These scales are crucial for accurate diagnosis, planning treatment, and monitoring progress. The clinical evaluation scales used in the corpora mentioned in Table 2.1 are discussed below.

2.2.2.1 FDA:

The Frenchay Dysarthria Assessment (Enderby, 1983) [37] is a standardized test used by speech therapists to diagnose dysarthria. It assesses 28 key speech aspects categorized into 8 areas, including reflexes, breathing, and tongue movement, and assigns scores to provide a detailed picture of the type and severity of dysarthria. This detailed evaluation helps therapists pinpoint the specific type and severity of dysarthria for each patient and develop targeted treatment plans.

2.2.2.2 UPDRS:

The Unified Parkinson's Disease Rating Scale (Stebbins & Goetz, 1998) [186] is a multi-part tool that goes beyond simply assessing speech. While not a dedicated dysarthria evaluation, it offers valuable insights into speech problems commonly seen in PD. Divided into four sections, Part III, the clinician-scored motor examination, holds particular importance for speech. This section evaluates various aspects of movement, including a sub-section focused on speech characteristics like initiation, loudness, voice quality, articulation, and fluency. The UPDRS Part III score, also known as UPDRS III, ranges from 0 to 108, with 0 indicating a symptom-free state and higher scores reflecting increasing motor impairment. Within this score, speech is ranked from 0 to 4, with 0 indicating no speech problems and 4 representing complete unintelligibility. Although the UPDRS does not provide a detailed dysarthria analysis, it serves as a helpful tool for clinicians to identify and monitor speech difficulties in PD patients.

2.2.2.3 H&Y:

The Hoehn & Yahr scale (Hoehn & Yahr, 1967) [65] focuses specifically on the staging of PD, categorizing patients based on the severity of their motor symptoms. The scale assigns stages ranging from 1 (minimal symptoms) to 5 (advanced disease with dependence on a caregiver). While the H&Y scale does not directly assess speech, it can provide a general indication of the potential for speech difficulties based on the overall disease severity. Clinicians may choose to use the H&Y stage alongside other speech-specific assessments like the UPDRS or FDA for a more comprehensive picture.

2.2.2.4 APAC:

The Assessment of Phonology and Articulation for Children (M. J. Kim et al., 2007) [99] differs from the previously discussed scales (FDA, UPDRS, H&Y) in its target population and focus. Unlike those designed for adults with neurological conditions, the APAC is specifically tailored for assessing speech sound production (phonology and articulation) in young children. This clinician-administered tool evaluates a child's ability to produce individual sounds, syllables, words, and sentences. It helps

identify any developmental delays or disorders affecting a child's speech clarity. The APAC provides valuable information for speech-language pathologists to diagnose specific speech sound errors and guide appropriate therapy approaches for improving a child's communication skills.

2.2.2.5 NIEPMD scale:

National Institute of Empowerment of Persons with Multiple Disabilities (Celin et al., 2016) [2], an institute for students with multiple disorders run by the Government of India. The scale is as described below:

- 0- Normal
- 1- Can understand with difficulty, however the clinician/listener feels the speech is not normal
- 2- Can understand with difficulty, occasionally needs repetitions
- 3- Can understand with concentration but a sympathetic listener needs two or three repetitions
- 4 -Can understand with difficulty
- 5- Can understand with effort if the context is known
- 6- Cannot understand at all even if the context is known

2.2.2.6 TOM:

The Therapy Outcome Measure (Enderby & John, 1999) [38] is designed to track progress and measure the effectiveness of therapy interventions. This clinician-administered tool assesses a patient's abilities and limitations across four key domains: impairment, activity, participation, and well-being. The TOM's strength lies in its ability to be used throughout the course of therapy. By periodically reassessing these four domains, clinicians can monitor a patient's progress, identify areas needing further intervention, and demonstrate the effectiveness of the chosen treatment approach.

To summarize Section 2.2, it is visible from Table 2.1 that dysarthric speech data, irrespective of the corpus, have generally been collected from very few speakers. The distribution of speakers and speech data across severity levels is often skewed, with large amounts of data tending to be speech with a lower dysarthria severity. This calls for more inclusive speech corpora. However, collecting data from speakers with severe dysarthria is a challenging task. In addition, a longitudinal study conducted over several years is needed to observe the changes in speech quality and intelligibility, along with their correlation to disease progression, medication, and therapy in order to identify requirements and build suitable technology that can help people suffering from dysarthria.

2.3 Acoustic Studies of Dysarthric Speech

Acoustic assessment of speech has often been recommended to supplement perceptual methods, considering that these methods are more objective than the more subjective perceptual assessments performed by different experts (Kent, 1996) [91]. Acoustic analyses of speech signals can potentially describe the speech subsystem and the correlates of the perceptual evaluation of speech. Typically, an acoustic study of dysarthria aims to understand and quantify the acoustic characteristics of dysarthric speech and their correspondence to dysarthria type, as mentioned in Section 2.1, dysarthria severity, and the underlying cause. Careful examination of the qualitative features of dysarthric speech may reveal a phenomenon that would eventually lead to the formulation of a hypothesis that can be tested using more standard quantitative techniques. However, acoustic studies of dysarthric speech are challenging, considering the complexity of neuromotor disruption occurring across the speech subsystem. In this section, we present studies that explore the key acoustic aspects of speech, such as loudness, pitch, duration, and speech rate, and their nature with respect to dysarthric speech. Typically, a study on acoustic parameters of dysarthria explains aspects such as perceptual speech intelligibility and dysarthria severity, type, or compares and contrasts with unimpaired speech. The following subsections present acoustic studies that address these aspects.

2.3.1 Dysarthria Intelligibility

Speech intelligibility is used as an indication of the severity of a speech disorder (Maier et al., 2009) [120], making this an important aspect of the study. The studies presented in this subsection explore the acoustic characteristics of speech related to intelligibility, specifically for dysarthric speech.

The study (Liu & Tseng, 1996) [118] used minimal phonetic contrasts and found that aspiration-unaspiration, affricate-fricative, and vowel frontness contrasts accounted for 99% of the variance in intelligibility judgments. The analysis focused on several acoustic features, of which first formant ($F1$), difference in first and second formant ($F2-F1$), voice onset time (VOT), nasality, and burst spectrum, were identified as effective in differentiating dysarthric from normal speech. A statistical model built using these features achieved an impressive 74.8% accuracy in predicting speech intelligibility. In the study of Bunton et al. (2001), [19], the relationship between fundamental frequency ($F0$) variability and speech intelligibility in persons with dysarthria of two different types was explored. The results indicated a strong correlation between sentence-level $F0$ variations and speech intelligibility and suggested that the type of dysarthria was significant when estimating the contribution of $F0$ to intelligibility. Prosodic features in severe dysarthric speech were examined by Patel (2002) [148] to understand prosodic control and its usage by dysarthric speakers to communicate intentions, thereby providing insights into speech intelligibility. The authors found that $F0$ contour and, to a lesser extent,

syllable duration were responsible for encoding prosodic information in dysarthric speech. An exploration by De Bodt et al. (2002) [32] to understand the effects of four main dimensions of speech production, namely voice quality, articulation, nasality, and prosody, as well as overall intelligibility employing perceptual judgments revealed that intelligibility can be expressed as a linear combination of these weighted perceptual dimensions and that articulation was the strongest contributor to intelligibility. In the study by Tjaden and Wilding (2004), [190], acoustic measures such as articulatory rate, sound pressure level, vowel space area (VSA), first-moment difference measures, and second formant ($F2$) trajectory characteristics for diphthongs were examined as indicators of dysarthric speech intelligibility by analyzing dysarthric speech with reduced articulatory rate (slow) and increased loudness. Intelligibility estimates for dysarthric speech were found to be correlated with articulatory rate and in turn with VSA, first-moment difference measures, and loudness, but with no consistent correlation to $F2$ slope measures for diphthongs or the acoustic measures of supraglottal behavior. In contrast, the study by Y. Kim et al. (2009) [106] examined the relationship between dysarthric speech intelligibility and $F2$ slope and reported significantly reduced $F2$ slopes compared to healthy speakers for almost all words, recommending $F2$ slope as a quantitative metric for dysarthric speech intelligibility prediction. The research work by Ijitona et al. (2017) [70] examined the relationship between acoustic and perceptive cues indicative of intelligibility such as pitch, intensity, and duration by modifying these parameters. A perceptual evaluation of the modified dysarthric speech indicated that amplification of any of the three stress markers improved the perceptual outcome significantly.

2.3.2 Type of Dysarthria

Acoustic metrics have also been used to investigate the nature of specific types of dysarthria. Ataxic dysarthria has been characterized by a slow speaking rate, relatively great variability in VOT, a tendency toward equalized vowel/syllable duration within utterances, and an unusually large $F0$ range across utterances (Kent et al., 2000) [215]. Hypo-kinetic dysarthria displays normal or faster-than-normal speaking rates, relatively high mean $F0$, decreased $F2$ extents and slopes and decreased $F0$ variability (Goberman et al., 2005) [54]. In yet another work (Y. Kim et al., 2011) [107], the authors discussed the individual acoustic variables that contribute to the identification of dysarthria type, severity, and the underlying cause. Articulation rate, voiceless interval duration, and intensity range contributed to etiology classification. In contrast, articulation rate and $F0$ range contributed to type classification. $F2$ slope, $F0$ range, and vowel space made significant contributions to severity classification. The articulatory implementation of dysarthria types has an impact on the emergent flow of the syllable stream and consequently on the perceived rhythm of speech, allowing quantification of rhythmic patterns as a means to classify the various types of dysarthria. Liss et al. (2009) [116] examined traditional

rhythm metrics, along with novel metrics that combined successive vocalic and consonant segments, which emerged as an important predictor to classify dysarthrias. Liss et al. (2010) [117] demonstrated the potential of envelope modulation spectra (EMS) for differentiating dysarthric from healthy speech. EMS is used to quantify rhythmic problems in dysarthria and can be computed automatically, enabling the objective classification of speakers. The study showed good accuracy in distinguishing control groups, different dysarthria subtypes, and healthy speakers while being able to capture temporal irregularities independent of linguistics.

2.3.3 Comparative Studies (Dysarthric and Healthy Speech)

Comparative studies play a key role in understanding the typical characteristics of dysarthric speech and its deviation from unimpaired speech. An extensive study on the acoustics of dysarthric speech has been presented by Kent et al. (1999) [93]. The authors examined various aspects of acoustic analysis such as the acoustic measures, acoustic–articulatory relationship, and their applicability to dysarthric speech. Special attention was given to describing phonation, voice quality, vocal tract function for vowel and consonant articulation, velopharyngeal function, laryngeal and supralaryngeal activity, sound segment timing, prosody, and paralinguistic variables using various measures of $F0$, formant frequencies $F1$, $F2$ and $F3$, VOT, spectral variations; duration of syllables and other sound segments; intensity/energy; and speaking rate. Dorze et al. (1994) [35] is a comparative study on the intonation and speech rate of dysarthric and non-dysarthric speakers, using declarative and interrogative sentences. The average difference in $F0$ between the last syllables of interrogative and declarative sentences was used as a metric for intonation measurement. It was found that this intonation measure of dysarthric speech was significantly lower for dysarthric speech, while speech rate was dependent on the subject group, sentence type, and sentence set. Acoustic features that contribute to sentence accent production in dysarthric speech and healthy speech were explored (Mendoza Ramos et al., 2020) [125]. Accent production was attributed to the duration parameters and the contrast in frequency and intensity between the target syllable and the previous syllable, rather than to the contrast with the rest of the sentence.

More recently, the effects of compression on various acoustic aspects of dysarthric speech such as mean harmonic-to-noise ratio, voiced and unvoiced frames, statistics of pitch, VSA, jitter, shimmer, cepstral peak prominence, and goodness of pronunciation were studied in the context of speech degradation on telepractice platforms (Tran et al., 2022) [192]. This article highlights the importance of studying dysarthric speech from a speech technology perspective. Such advancements can lead to improved speech recognition and communication assistance, ultimately empowering people with dysarthria.

Summarizing Section 2.3, acoustic analyses can complement perceptual evaluations and are particularly valuable as sources of quantitative assessment for clinical

evaluations of dysarthria as well as for therapy planning and tracking. Additionally, an in-depth acoustic study on dysarthria types will enable researchers to apply their understanding of the latest technological advances in speech processing to dysarthric speech processing. However, there is a lacuna in the area of acoustic studies pertaining to the automatic identification of the underlying cause of dysarthria since it is an extremely complex challenge compounded by mixed dysarthria. The reviewed studies from this section are presented in Table 2.2

2.4 Dysarthria Severity-level Identification

Identification of dysarthria type and severity level is the first step to planning speech therapy or interventions for the rehabilitation of a person exhibiting dysarthria. Objective assessment of dysarthria severity helps identify the bias in perceptual evaluations, which tend to be subjective and are also time-consuming and expensive. Furthermore, an understanding of severity has contributed to improved ASR performance for dysarthric speech as seen in the works of M. J. Kim et al. (2013) [102] and Mustafa et al. (2015) [133]. Xue (2023) [211] makes a clear distinction between comprehensibility and intelligibility of speech. Comprehensibility is assessed by considering contextual and context-independent information derived from speech and language, whereas speech intelligibility is evaluated from the acoustics of speech alone without taking context into account. This study delved into the development of valid subjective and objective procedures for measuring the intelligibility of dysarthric speech. Speech intelligibility has been used as an indicator of the severity of speech disorders in general (Maier et al., 2009) [120]. Although there is no standard measure of speech severity in dysarthria, estimates of speech intelligibility are often used to index the extent to which the speech mechanism gets impacted by the neuromotor disease (Kent et al., 1989) [90]. Intelligibility and speaking rate are the key quantitative measures used to identify dysarthria severity levels (Dahmani, et al., 2013) [31]. Assessment of the severity level of dysarthric speech is crucial to analyze the outcome of a surgical procedure or to determine the patient's status in disease progression (e.g., PD) or in clinical decision-making.

In recent decades, research has shifted toward computer-based dysarthria assessment, aiming for repeatable, reliable evaluations at minimal cost. Two main approaches are at the forefront:

1. Non-reference-based assessment: This approach analyzes various speech features independent of the intended message as well as without using healthy speech data for machine learning (ML).

2. Reference-based intelligibility estimation: These methods compare dysarthric speech to a healthy speech reference signal to quantify deviations. These algorithms use healthy speech to train models and to measure the error rate on dysarthric speech as an indirect indicator of intelligibility decline, utilizing

both non–deep learning and deep learning techniques and various feature sets and classifiers.

2.4.1 Non-reference-based Approaches

Hummel (2011) [68] applied the P.563 standard for blind speech quality assessment to dysarthric speech. In addition, speech features related to perceptual qualities of dysarthric speech were explored, and it was shown that kurtosis of the spectral flatness of the linear prediction (LP) filter performed well in both speaker-dependent (SD) and speaker-independent (SI) evaluations. Falk et al. (2012) [41] investigated the kurtosis of the LP residual, long- and short-term temporal dynamics, nasality, and prosody features as predictors of intelligibility. A composite measure, which is a weighted linear combination of the above measures, was designed for intelligibility prediction. In this measure, different weights were found to be beneficial for different dysarthria severity classes. The algorithm proposed by Berisha et al. (2013) [11] used acoustic cues at different time scales at phonetic, segmental, and suprasegmental levels, resulting in features that represent the distorted rate and timing of speech, unnatural loudness variation, unnatural pitch or formant variation, articulatory imprecision, and omissions or distortions of specific consonants and vowels. Support vector machine (SVM) classifiers were trained using dysarthric speech features that were extracted at sentence and vowel or consonant levels, followed by an ensemble learning technique to decide the intelligibility and thereby the severity level.

In summary, these studies applied different speech quality assessment standards and explored various speech features for the prediction of dysarthria severity level using intelligibility prediction in dysarthric speech. The studies showed that kurtosis of spectral flatness, LP residual, temporal dynamics, nasality, prosody, and acoustic cues at different time scales were some of the useful features for intelligibility prediction. Some studies have also proposed composite measures and ML algorithms for improved prediction.

2.4.2 Reference-based Approaches

In the recent scientific literature, we observe a trend toward moving away from language-specific methods and prioritizing language-independent AIA. Initial works on automatic intelligibility prediction focused on appropriate acoustic representations of dysarthric speech followed by the application of classifiers. Most of the recent works use deep learning–based intelligibility prediction. The subsections below present studies using non-deep learning– and deep learning–based intelligibility prediction.

2.4.2.1 Non-Deep Learning Methods

In the study of Middag et al. (2011) [130], a combination of two approaches was used to predict the intelligibility of speakers from the Flemish pathological speech corpus containing 48 dysarthric speakers' data. A speaker verification-based approach in which a Gaussian mixture model (GMM-based) super-vector in combination with a phonological feature set that related directly to the articulatory dimensions of speech was used. Intelligibility prediction models built using support vector regression (SVR) and late fusion resulted in the best outcome. Dysarthria severity was quantified by computing the distance between the speaker-specific GMM and the SI universal background model (UBM). The speaker GMM is derived from the speaker's Mel-frequency cepstral coefficient (MFCC) feature vectors, while the UBM is trained on a large data set of speech from healthy individuals. Dysarthria severity is typically calculated as the distance metric between individual dysarthric speaker models and reference models constructed from healthy speech data. In the study of M. J. Kim and Kim (2012) [101], severity classification of dysarthric speech from the QoLT database was performed using a subset of features related to phonetic quality, prosody, and voice quality using a linear-kernel SVM classifier.

The research work by Dahmani et al. Dahmani et al. [31] explored the Gaussian–Bayes classification technique to classify dysarthric speech from the Nemours database into three classes using rhythm metrics based on acoustic measures of the duration of vocalic and consonant intervals in continuous speech. In the work of Martínez et al. (2013) [124], a feature set of i-vectors derived from perceptual linear predictors (PLPs) comprising a T matrix spanning a space trained on the main variabilities of dysarthric speech was assumed to contain information about the speech intelligibility. These features were used to predict speech intelligibility using SVR predictors and are evaluated on the UA dysarthric speech corpus. A set of prosodic features selected using a linear discrimination analysis (LDA) and an SVM classifier was used to classify dysarthric speech from the Nemours database into four classes (Kadi et al., 2013) [80]. Prosodic, voice quality, and pronunciation aspects at the sentence level were used as features, followed by a post-classification posterior smoothing scheme to evaluate the TORGO dysarthric speech into binary intelligibility labels (intelligible and not intelligible) in the study of J. Kim et al. (2015) [97]. Furthermore, feature-level fusions and subsystem decision fusion were used to arrive at a final intelligibility decision. The work by Nerendra and Alku (2018) [134] presents a binary classification of dysarthric speech, carried out using glottal parameters, and acoustic features as well as a combination of the two feature sets at word level, sentence level, and nonword level using dysarthric speech from the TORGO database and SVM classifiers. Although the latter combination framework performed the best overall, glottal features performed the best for nonword utterances indicating that glottal features contain relevant information related to speech disorder classification. In the study of Wang et al. (2018) [202], the effectiveness of acoustic and articulatory speech features from patients with ALS and

healthy controls was used to automatically predict intelligible speaking rate, which is a multiplication product of speech intelligibility and speaking rate, using ML. The model achieved good accuracy with high R2 and low root-mean-square error using a feature selection method and SVM regression. The selected features included both acoustic and articulatory measures, with a higher proportion from the articulatory system. The results supported the effectiveness of this approach, with tongue movement data yielding the best individual performance. The model was theoretically content-independent due to its reliance on low-level features. In the study of Hernandez, Kim, and Chung (2020) [62], the authors explored prosody, voice quality, and MFCC features separately as well as in combination. Several classifiers such as Random Forest (RF), SVM, and neural network–based were employed for the classification of dysarthric speech from the TORGO and the QoLT databases. It was observed that while a combination of MFCC and prosody features, along with neural network classifiers, performed best overall for high-severity dysarthric speech, MFCC features provided the best accuracy for both the English and Korean databases. A subsequent study by Hernandez, Yeo, et al. (2020) [63] investigated the efficacy of incorporating rhythm-based metrics into dysarthria detection and severity assessment in addition to prosodic features. RF, SVM, and feed-forward multilayer perceptron (MLP) classifiers were employed. The findings revealed a significant improvement in accuracy with the inclusion of rhythm metrics for both detection and severity assessment. This improvement was observed in both Korean and English data sets, with a larger relative increase in accuracy for the Korean QoLT data set compared to the English TORGO data set. The study by Janbakhshi et al. (2020) [73] proposed using spectrotemporal subspaces, extracted from speech recordings. Spectral subspaces were based on the frequency content of the speech, while temporal subspaces were based on how the speech signal changed over time. Grassmann discriminant analysis (GDA) was used to classify the speech recordings as healthy or dysarthric. GDA is an ML algorithm that is specifically designed to work with data that lie on a manifold, which is a curved space that cannot be flattened out into an Euclidean plane. The results showed that the method using temporal subspaces achieved significantly better accuracy than the method using spectral subspaces. This method also outperformed other state-of-the-art methods for automatic dysarthric speech detection. Neumann et al. (2021) [136] proposed a cloud-based platform for remotely assessing and monitoring ALS progression through speech and video analysis. Acoustic metrics included timing and frequency aspects, along with conversation-specific measures such as syllable rate and speaking rate variability. Visual metrics were centered on facial features and were calculated using a three-step process involving face detection, landmark extraction, and metric computation based on specific landmarks. A nonparametric Kruskal–Wallis evaluation with healthy controls and patients with ALS revealed statistically significant differences in both acoustic and visual features between the groups. Least absolute shrinkage and selection operator–based regression analysis to investigate the predictive power of the extracted acoustic and visual

metrics further emphasized the multimodal approach. The study by Al-Qatab et al. (2021) [3] investigated the potential of acoustic features and feature selection for a severity-based classification of dysarthric speech. Four acoustic features such as prosody, spectral, cepstral, and voice quality were investigated, along with seven feature selection methods, namely, interaction capping, conditional information feature extraction, conditional mutual information maximization, double input symmetrical relevance, joint mutual information, conditional redundancy, and relief. Additionally, six classification algorithms such as SVM, LDA, artificial neural network (ANN), classification and regression tree, naive Bayes, and RF were examined. Although the study concluded that there is no best method for improving the classification accuracy of an ASR system, they highlighted the effectiveness of prosody features for the classification of dysarthric speech.

The cited research works above aim to predict the intelligibility or classify the severity of dysarthric speech using various feature sets such as phonological features, prosodic features, voice quality features, glottal parameters, and acoustic features. The studies used different ML algorithms such as SVM, SVR, Gaussian–Bayes, and LDA for the prediction. The best results in each study were achieved by using a combination of different feature sets and decision fusion techniques as shown in Table 2.2.

2.4.2.2 Neural-network-based Intelligibility Prediction

Wang et al. (2016) [201] explored the feasibility of automatically detecting ALS from presymptomatic speech samples using ML. They investigated two ML approaches, namely, SVM and deep neural network (DNN) on a data set combining speech acoustics and articulatory data from patients with ALS and healthy controls. The results using leave-one-out cross-validation indicated the promise of this approach, with further improvement observed when incorporating articulatory motion information, with DNN outperforming SVM in all configurations. Tu et al. (2017) [193] proposed a DNN-based interpretable model for objective assessment of dysarthric speech that provided users with an estimate of severity as well as a set of explanatory features. Bhat et al. (2017) explored a nonlinguistic approach to the automatic assessment of severity levels of dysarthric speech, using audio descriptors that are traditionally used to define the timbre of musical instruments, along with multitapered spectral estimation for classification. An ANN was trained to classify speech into various severity levels within the UA dysarthric speech corpus and the TORGO database. The study by An et al. (2018) [6] investigated the feasibility of automatically detecting ALS from speech samples using convolutional neural networks (CNNs). In addition to predefined, hand-crafted features and ANN, CNN-based representation learning has been used. Time-domain CNNs achieved the best performance at the sample level, while frequency-domain CNNs achieved the best performance at the person level (when considering multiple speech samples from the same person). In the study of Chandrashekar et al. (2020a) [23], joint

spectrotemporal features from the mel-scale spectrogram were used for dysarthria severity estimation. Intelligibility estimation was carried out using ANN and CNN. The performance of the time-frequency CNN configuration proved to be the best, as they captured both spectral and temporal variations in the audio signal. The authors demonstrated that the time-frequency CNN that jointly captured spectral as well as temporal information was superior to the time or frequency CNN that captured either temporal or spectral information and not both. In yet another work, authors explored the use of perceptually enhanced Fourier transform spectrograms and constant Q transform spectrograms using CNN classifiers to assess word-level and sentence-level intelligibility of dysarthric speech from the UA Speech and the TORGO databases (Chandrashekar et al., 2020b) [22].

In the work of Bhat and Strik (2020) [12], bidirectional long short-term memory (BLSTM) networks are used for binary classification of intelligibility of dysarthric speech. The performance of the classifier using speech parameters such as MFCC, log filter banks, and i-vectors has been compared. Furthermore, a transfer learning (TL) approach was investigated, in which dysarthria intelligibility level was predicted using acoustic models that were pre-trained on unimpaired speech giving the best performance for dysarthria intelligibility prediction. In the study of Joshy and Rajan (2022) [77], authors compared the performances of various DNN architectures along with generic speech features as well as dysarthria-specific speech features to classify dysarthric speech based on the severity from the UA Speech and the TORGO databases. Architectural choices such as DNN, CNN, gated recurrent units, and long short-term memory (LSTM) networks were explored. Speech features such as (a) MFCCs and constant Q cepstral coefficients; (b) speech disorder-specific features computed from prosody, articulation, phonation, and glottal functioning; and (c) low-dimensional speech representation such as i-vectors are investigated. It was found that the DNN classifier using MFCC-based i-vectors outperformed other systems. In the study of Hall et al. (2022) [59], the performance of deep learning–based automatic dysarthric intelligibility assessment models was explored, with the objective to generalize to new speakers and arrive at the optimal setup by identifying the acoustic features useful for this purpose such as MFCC and spectral representations along with their various configurations. Xu et al. (2023) [209] investigated the application of DNNs for classifying dysarthric speakers. While DNNs are promising for dysarthria classification, they lack clinical interpretability. To address this, they proposed a DNN model with a bottleneck layer that was trained to jointly classify dysarthria and extract clinically relevant acoustic features. This method allowed researchers to balance classification accuracy with interpretability. The model was evaluated on two dysarthria subtypes and achieved good performance. Additionally, the Shapley additive explanation was employed to analyze the contribution of each interpretable feature to the classification decisions. The results demonstrated that the proposed model could be tuned to prioritize either classification accuracy or clinical interpretability. In a more recent study, the potential of leveraging dysarthric speech data sets in

Netherlandic Dutch to enhance the efficacy of ASR systems for dysarthric speech in Flemish Dutch was explored (Xue et al., 2023) [212], Flemish Dutch and Netherlandic Dutch were treated as the dominant and nondominant variants of the same pluricentric language, namely, Dutch. The evaluation of ASR models involved the utilization of two distinct intelligibility metrics: orthographic transcriptions and overall intelligibility assessments.

Potential bias in automatic dysarthria classification due to recording environment characteristics using the UA Speech and the TORGO databases was investigated by Schu et al. (2023) [173]. They hypothesized that the classification tasks conducted on these databases rely more on environmental noise rather than dysarthric speech features. To test this hypothesis, utterance-level signal-to-noise ratios were estimated, and state-of-the-art dysarthria classifiers such as SVM, CNN, speech representation learning, and MLP were trained and validated on both speech and nonspeech segments. Several classifiers achieved comparable or superior dysarthria classification performance using only nonspeech segments compared to speech segments. These findings highlight the importance of recording quality in developing and evaluating dysarthria classification methods. Additionally, these results encourage the development of novel classification approaches robust to adverse recording conditions.

Table 2.3 provides a summary of the studies reviewed in the section. In conclusion, a common approach in research on dysarthric speech intelligibility is to extract and analyze relevant acoustic features, through which researchers aim to gain insights into the speech patterns of individuals with dysarthria and to develop methods for improving automatic speech intelligibility estimation. It can be noted that while the earlier works involved extensive feature engineering, representing prosodic, voice quality, phonological, and articulatory aspects of speech followed by feature selection and classified using a classifier such as SVM/SVR, recent explorations involve the speech features such as MFCC, log filter banks, i-vectors, and spectrotemporal features being used in a neural network framework. While the performance of DNN-based speech intelligibility prediction of dysarthric speech is good, they are demanding in terms of data availability, tuning of the network, and computing requirements.

2.5 Automatic Speech Recognition

The neurological damage affecting speech-motor functions also impacts physical activities associated with the motor neurons. Typical human interaction with gadgets and devices involves typing into a keyboard. Keyboard input using hand movements is slowed down by a factor of 150–300 in severe cases of dysarthria in comparison with regular users (Hosom et al., 2003) [66]. However, dysarthric speech is slow by a factor of 10–17 as compared to regular speech, at about 15 words per minute in the most severe cases (Rudzicz, 2010) [165]. Also, it has been found that dysarthric

speakers exhibit good prosodic control, which in turn aids communication efficiency (Patel, 2002) [148]. However, due to the atypical nature of speech, traditional techniques and off-the-shelf ASR become unusable for persons with dysarthria. This indicates the need for research into techniques and speech features that provide improved performance for ASR for dysarthric speech, which we will discuss in this section.

Several techniques have been employed to improve ASR performance for dysarthric speech such as speaker adaptation, lexical model adaptation, feature engineering, acoustic space enhancement, and DNN—either individually or in combination. Attempts have been made to synthesize dysarthric speech data to be able to augment the existing data and avail the benefits of advanced neural network architectures for ASR. Table 2.4 outlines the studies reviewed in this section, along with the research outcomes in terms of performance metrics.

2.5.1 Speaker Adaptation-based ASR

Speaker adaptation involves using dysarthric speech from individual speakers during the training process. It has been seen that speaker adaptation yields good ASR performance, the caveat being that a specific speaker's speech needs to be a part of the training data. Initial investigations into ASR for dysarthric speech involved applying the techniques traditionally used for improving ASR performance such as maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptation to arrive at an SD ASR with a limited amount of dysarthric speech data (Christensen et al., 2012; Mengistu & Rudzicz, 2011; Sehgal & Cunningham, 2015; Sharma & Hasegawa-Johnson, 2010) [180, 128, 27, 174]. A hidden Markov model (HMM-based) ASR along with MLLR and MAP adaptation was used to build SD acoustic models using the TORGO corpus (Mengistu & Rudzicz, 2011) [128]. PLP features along with lexical model adaptation achieved the best performance in terms of word recognition accuracy. The SD pronunciation lexicons consisted of multiple pronunciations for some words that reflect the particular pronunciation deviation of each dysarthric subject from the canonical forms. In the study of Christensen et al. (2012) [27], authors explored strategies such as MLLR and MAP adaptation. They used speech data from the dysarthric domain for MAP adaptation of a typical speech acoustic model. In addition, they also examined SI and SD systems. PLP features with first and second derivatives were used. The evaluation of these systems was carried out on the UA Speech corpus and showed that significant improvement in dysarthric speech recognition was achieved using MAP adaptation for most speakers with low to moderate severity levels. In the study of Sharma and Hasegawa-Johnson (2013) [181], a background model of the general speech characteristics of a particular dysarthric speaker was used to select a suitable acoustic model from unimpaired speakers using an interpolation-based technique to maximize the matching between the dysarthric speaker's acoustic model and unimpaired speaker models. Prior knowledge of the dysarthric speaker's sever-

ity level was used to select an initial model for adaptation followed by traditional adaptation methods, thereby reducing the word error rate (WER) for the QoLT dysarthric speech data (M. J. Kim et al., 2013) [102]. Authors explored the factors that influence ASR performance for dysarthric speech, concluding that factors specific to dysarthria such as intelligibility and severity had a significant influence on recognition accuracy (Mustafa et al., 2014) [132]. An exploration of various combinations of MFCC features was conducted and narrowed down to the conventional 12 coefficients MFCC features without the use of delta and acceleration features, which were used, along with an ANN, to build an SI ASR (Shahamiri & Binti Salim, 2014) [177]. Authors concluded that building acoustic models using only a select set of speakers for the training process based on acoustic closeness to the target speaker, rather than the entire speaker pool data, provided better ASR performance (Christensen et al., 2014) Christensen et al. [26]. A convolutive bottleneck network (CBN) was used for dysarthric speech feature extraction. The pooling operations of the CBN resulted in features that were more robust toward the small local fluctuations in dysarthric speech and that outperformed the traditional MFCC feature–based recognition (Nakashika et al., 2014) [135]. Authors used voice parameters such as jitter and shimmer, along with multitaper MFCC followed by feature space MLLR (fMLLR)–based speaker adaptation, to reduce the WER for the UA dysarthric speech data (Bhat et al., 2016b) [13]. They also analyzed the ASR performance from the perspective of speaker severity. M. J. Kim et al. (2017) [103]proposed a Kullback–Leibler divergence-based HMM (KL-HMM) approach to capture these variations. Phoneme posteriors from a DNN acoustic model inform the emission probabilities in the KL-HMM states. A novel speaker adaptation method, combining L2 (L2 norm or Euclidean norm) and confusion-reducing regularization, further enhanced discriminability. Evaluations on a dysarthric/non-dysarthric speaker data set demonstrated significant performance improvements over conventional DNN speaker adaptation. In the study by Yılmaz et al. [220], a multistage DNN training scheme is used to model dysarthric speech. Only a small amount of in-domain training data showed considerable improvement in the recognition of dysarthric speech. In the study of Takashima et al. (2020) [187], authors proposed a two-step model adaptation approach, in which an ASR model was first adapted to the general speaking style of multiple dysarthric speakers, and then the adapted model was further adapted for the target speaker, demonstrating a better performance in ASR than a conventional one-step adaptation approach. A novel adaptation network specifically designed to fine-tune Wav2Vec2 models using fMLLR features was proposed by Baskar et al. (2022) [10]. This network exhibited flexibility, allowing it to handle other speaker-adaptive features such as x-vectors as well. A comprehensive experimental analysis demonstrated consistent improvements across all dysarthria severity levels in the UA Speech data set, with a WER of 57.72% for high-severity cases. Furthermore, experiments on a German data set substantiated the generalizability of this approach across diverse linguistic domains. Deep embedding features derived from a spectrotemporal subspace analy-

sis using singular value decomposition of the speech spectrum were used to improve ASR performance based on speaker adaptation (Geng et al., 2022) [52]. These facilitated auxiliary feature-based adaptation for hybrid DNN/time delay neural network (TDNN) and end-to-end Conformer architectures achieving a WER of 25.05% on the UA Speech test set. The intention was to address the spectrotemporal variations observed in dysarthric and elderly speech, characterized by articulatory imprecision, reduced volume/ clarity, slower speaking rates, and increased dysfluencies

To summarize, speaker adaptation for dysarthric speech has been shown to improve ASR performance. Factors such as intelligibility and severity of dysarthria have a significant impact on recognition accuracy. Initial work in this area applied traditional techniques such as MLLR and MAP adaptation to limited amounts of dysarthric speech data. The studies explored various feature extraction techniques, such as PLP features, MFCC features, and voice parameters such as jitter and shimmer. SD and SI ASRs were also built using HMM-based ASRs and ANNs. Interpolation-based techniques and fMLLR-based speaker adaptation were also used to improve performance based on speaker severity.

2.5.2 Enhancement of Dysarthric Speech

Enhancement of dysarthric speech either in the time domain or feature domain is yet another mechanism used to improve ASR performance. Acoustic space modification carried out through temporal and frequency morphing improved automatic dysarthric speech recognition and subjective evaluation (Rudzicz, 2013) [167]. An investigation of the relationship between durations of the voiced sections within dysarthric speech utterances and dysarthria severity level revealed that temporal manipulations of dysarthric data would lead to improved intelligibility. Leveraging this insight, Bhat et al. (2016a) [14]applied temporal adaptation of dysarthric speech based on severity level. This approach improved ASR performance for dysarthric speech at each severity level. Recent studies leverage the latest DNN architectures to improve ASR performance. Feature-level enhancement of dysarthric speech was carried out (Bhat et al., 2018; Vachhani et al., 2017) [195, 16]. Deep autoencoder (DAE) was trained with only unimpaired speech data and the bottleneck layer was used to enhance dysarthric speech data from the UA speech corpus. Deep autoencoder (DAE) was trained with only unimpaired speech data, and the bottleneck layer was used to enhance dysarthric speech data from the UA Speech corpus (Vachhani et al., 2017) [195]. A TDNN denoising autoencoder (TDNN-DAE) was used to enhance the dysarthric speech features. Unimpaired speech and the corresponding tempo-adapted dysarthric speech were used to train the TDNN-DAE. This methodology showed significant improvement in ASR performance when evaluated for both SI and speaker adaptation–based ASR systems (Bhat et al., 2018) [16]. The study by M. J. Kim, Cao, and Wang (2018) [104] investigated the effectiveness of multiview representation learning using canonical correlation analysis (CCA) for dysarthric speech recognition. A representation of

acoustic data was learned using CCA from the multiview data, namely, acoustic and articulatory. The findings demonstrated that incorporating articulatory information through CCA improves dysarthric speech recognition. The improvement applied to various speech recognizers including GMM-HMM, DNN-HMM, and LSTM-HMM, with LSTM-HMM achieving the overall best performance. Prananta et al. (2022) [157] explored a novel approach for enhancing dysarthric speech recognition by leveraging MaskCycleGAN methods, typically used for voice conversion. This technique involves applying MaskCycleGAN to modify dysarthric speech features, followed by dysarthric speech recognition. The performance was compared with time stretching–based enhancement alone and showed marginal improvements for mid- and high-severity dysarthric speech with high dependency on the temporal structure of dysarthric speech, indicating that research efforts should be focused on sequence-to-sequence–based architectures. In the study of Hernandez et al. (2022) [64], the authors used various self-supervised acoustic representation learning techniques such as speech features extracted from Wav2Vec2.0 (Baevski et al., 2020) [9], Hubert (Hsu et al., 2021) [67], and the multi-lingual speech representation (XLSR) models (Conneau et al., 2021) [29], to evaluate ASR performance for English, Spanish, and Italian dysarthric speech with the XLSR models performing the best, indicating that the multilingual data that contained more variations of similar phonemes could be a closer representation of dysarthric speech.

There have been various studies on improving the recognition of dysarthric speech using different methods such as formant resynthesis, acoustic space modification, temporal adaptation, and DNN architectures. Some of the studies showed significant improvement in the recognition of dysarthric speech using feature-level enhancement using DAE and TDNN-DAE, and MaskCycleGAN-based methods. The most recent studies also show the potential of self-supervised acoustic representation learning techniques such as speech features extracted from Wav2Vec2.0 and XLSR models for improving the recognition of dysarthric speech across different languages.

2.5.3 Data Augmentation

Data augmentation is yet another methodology that has gained popularity for improving dysarthric speech recognition. Time and tempo-stretching of healthy speech-based data augmentation for improving speech recognition were investigated (Vachhani et al., 2018) [196]. In order to address ASR performance on severe dysarthric speech, SD acoustic models based on phoneme-level speech tempo ratio between typical and speaker-specific dysarthric speech were created to augment existing dysarthric speech (Xiong et al., 2019) [207]. Two separate augmentation policies involving speed, tempo, and vocal tract length perturbation applied on healthy and dysarthric speech showed significant improvement in ASR performance (Geng et al., 2020) [51]. Both speed perturbation and tempo modification involve altering the speed of dysarthric speech. However, while speed perturbation affects the pitch

of the speech, tempo modification maintains the original pitch. A transformation of healthy speech to dysarthric speech using voice conversion–based techniques involving speaking rate modification, pitch modification, and spectral feature transformation using adversarial training has been employed to simulate training data using healthy speech (Celin et al., 2020) [21]. Speech vision is a dysarthria-specific ASR system that uses visual speech features combined with data augmentation with synthetic dysarthric acoustic visuals and leveraging TL for improved ASR performance (Shahamiri, 2021). In the study of Yue et al. (2022) [176], the authors explored raw-waveform acoustic modeling to overcome any loss of information that happened due to hand-crafted features for dysarthric speech recognition. Parametric CNNs along with data augmentation were used to address the data sparseness issue. Significant improvement in ASR performance for the TORGO data was achieved using parametric CNNs and multistream acoustic modeling. Two-stage data augmentation comprising traditional static data augmentation, a TDNN-DAE–based and a dynamic data augmentation scheme using modifications specifically designed for dysarthric speech were used to improve ASR performance in the study of Bhat et al. (2022) [17]. A deep convolution–based generative adversarial network (DCGAN) was used for tempo and speed perturbations in addition to learning hidden unit contribution–based speaker adaptation (Jin et al., 2023). [76].

In summary, various data augmentation techniques have been proposed and applied to improve ASR performance for dysarthric speech. These techniques include vocal tract length, speed, and tempo perturbations; DCGAN; SD acoustic models; voice conversion–based techniques; visual speech features; raw-waveform acoustic modeling; and parametric CNNs. These techniques have shown significant improvement in ASR performance for dysarthric speech. Additionally, multistage data augmentation techniques, including traditional data augmentation, TDNN-DAE–based data augmentation, and dynamic data augmentation, have been proposed to improve ASR performance further.

2.5.4 Transfer Learning

Traditional ASR models necessitate vast quantities of training data specific to the target domain. In the case of dysarthric speech, collecting such extensive data sets proves challenging due to cost and logistical constraints. TL emerges as an effective strategy. In this approach, a model trained on one task (source domain) is leveraged to improve performance on a related task (target domain) (Pan & Yang, 2009) [143]. This proves particularly beneficial when acquiring data for the target domain is expensive, time-consuming, or limited. TL enables researchers to leverage knowledge acquired from a source domain rich in data (e.g., typical speech) and strategically apply it to the target domain with limited data (e.g., dysarthric speech). We present recent studies that leverage the TL technique to improve the ASR performance for dysarthric speech.

Xiong et al. (2020) [208] have highlighted the importance of selecting an ap-

propriate source domain for TL in dysarthric speech recognition. The proposed approach leveraged a pretrained, state-of-the-art CNN-TDNN-F ASR model, initially trained on a vast data set of typical speech (source domain). This model served as a foundation for personalized recognition. To adapt it to the specificities of dysarthric speech (target domain), the study employed neural network weight adaptation with limited data from each individual speaker. Compared to conventional SD training and data combination methods, this approach yielded an average improvement of 11.6% and a relative improvement of 7.6%, as evaluated using the UA Speech corpus. The study by Celin et al. (2023) [123] investigated data augmentation for TL in ASR for continuous dysarthric speech. To bridge vocabulary gaps, it augmented dysarthric speech with speed/volume variations, virtual microphone arrays, and multiresolution features before transferring knowledge from a pretrained normal speech model. This approach, evaluated on isolated and continuous speech data sets, tackled out-of-vocabulary challenges, with an improvement of 7.11% in WER for very low intelligibility categories. The authors proposed a TL approach for dysarthric speech recognition using Whisper, a pretrained model for various speech tasks (Rathod et al., 2023) [162]. Whisper's transformer encoder module was used to extract relevant features for dysarthric word recognition. This approach achieved an average accuracy of 59.78% on the UA Speech corpus (155 words) using a BLSTM classifier. The authors hypothesized that the Whisper encoder effectively captures necessary speech information into a fixed-length vector representation, termed "Whisper encoder features." In summary, TL is a valuable tool for improving dysarthric speech recognition systems and has been shown to significantly reduce WER. Exploring different source domains, using data augmentation techniques, and fine-tuning pretrained models are promising avenues for further research in this area.

To summarize, ASR for dysarthric speech has seen advancements through various techniques such as speaker adaptation, speech enhancement, data augmentation, and TL. These approaches address challenges like limited data and achieve significant improvements in recognition accuracy. A recurring issue in all the investigations pertaining to ASR for dysarthric speech is the limited availability of dysarthric speech data for speaker adaptation. Substantial improvements were achieved with some form of speaker adaptation in every study. However, we must acknowledge that the performance of an ASR in an SI scenario is crucial for practical purposes M. J. Kim et al. (2018) [104], demonstrate the effectiveness of convolutional LSTM recurrent neural networks (CLSTM-RNNs) for recognizing dysarthric speech in an SI manner. The authors hypothesized that CLSTM-RNNs could capture the unique characteristics of dysarthric speech due to the combination of CNNs for extracting local features and LSTM-RNN temporal dependencies. The experiments involved phoneme recognition of dysarthric speech from a data set collected from nine patients. The results showed that CLSTM-RNNs outperformed both standard CNN and LSTM–RNN–based recognizers, with time-frequency convolutional LSTM-RNNs achieving the best performance.

2.6 Discussion and Conclusions

The objective of our research is to improve the quality of life for persons with dysarthria through research into speech technology that will allow conversational interactions for people with dysarthria. A clear understanding of the characteristics of dysarthric speech is extremely important for this purpose. To make longer and faster strides, researchers typically rely on existing research and data on a global scale. Therefore, it is imperative to consolidate the existing research and present it in a form that can serve as a basis for future work. In this review article, we review the contributions of speech technologists to the area of dysarthric speech with a focus on acoustic analysis, speech features, and techniques used. It can be observed that researchers have used various low-level as well as high-level (derived) speech features such as prosodic, phonetic, articulatory, MFCC and its variants, PLP, Teager energy, glottal pulses, and neural network–based features, often in combination. They have explored several classifiers and ASR configurations to identify intelligibility as well as improve the performance of ASR. We note that with the introduction of DNNs for speech processing, there has been a tremendous effort to exploit these techniques for improving ASR and classifier performance for dysarthric speech. However, researchers are limited by the availability of dysarthric speech data, which in turn paves the way to research into data augmentation techniques. It is also worth noting that most evaluations have been done on data collected in a clinical setup. Trends from published research indicate that future work in this area will be focused on the usability of DNN and novel speech features for existing databases in languages such as English as well as low-resource languages. Another avenue worth exploring would be capturing longitudinal dysarthric speech data in real environments using non-intrusive data collection methods as conducted in the Neurospeech project (Orozco-Arroyave et al., 2018) [142] and *the homeService corpus* (Nicolao et al., 2016) [138].

Some of the challenges that are evident for research on dysarthric speech are:

- Limited data availability: This is a major hurdle across all areas of dysarthria research. Collecting data, especially longitudinal data in natural environments, is expensive, time-consuming, and requires collaboration with specialists, specifically for clinical evaluation.
- Variability of dysarthria: Dysarthria manifests differently depending on the underlying neurological condition. This variability makes it difficult to develop one-size-fits-all solutions.
- Evaluation metrics: There is a lack of standardized metrics to objectively assess dysarthria severity and treatment efficacy. This makes it challenging to compare research findings across different studies.
- Clinical versus real-world use: Many studies evaluate interventions in controlled clinical settings

However, these settings may not reflect the complexities of everyday communication.

Significant progress has been made in understanding and addressing dysarthria through advancements in speech analysis, speech recognition, and assistive technologies. However, there is still a great deal of potential for improvement. To further enhance communication accessibility and quality of life for individuals with dysarthria, future research should focus on standardization of data collection: Establishing standardized protocols for collecting dysarthric speech data across different labs and languages would facilitate collaboration and development of more generalizable models.

- Multimodal assessment: Integrating speech analysis with other modalities such as facial expressions, electromyography, and brain imaging can provide a more comprehensive understanding of dysarthria and inform treatment strategies.
- Artificial intelligence (AI)–powered speech analysis: Deep learning techniques hold promise for developing automated tools for dysarthria assessment, offering objective and efficient evaluations.
- Natural language processing (NLP) for augmentative and alternative communication (AAC): NLP can be used to create more intuitive and user-friendly AAC systems that can adapt to the specific needs of individuals with dysarthria.
- Teletherapy and remote monitoring: Teletherapy using video conferencing and remote monitoring with wearable sensors can improve access to care and enable personalized interventions for dysarthria.
- Focus on low-resource languages and cultural considerations: Research should be directed toward developing dysarthria assessment and communication tools that cater to diverse languages and cultural contexts.

In conclusion, overcoming the challenge of limited data and exploring new avenues in data collection, AI-powered analysis, and teletherapy hold immense potential for significant advancements in dysarthria research. By focusing on these future directions, researchers can develop more effective tools and interventions to improve communication, quality of life, and overall well-being for people with dysarthria.

2.7 Data Availability Statement

This is a review article; hence, we do not have ownership of the data on which the results are reported. The References section provides links to access the articles reviewed.

Table 2.1: Dysarthric Speech Corpora

Database	Ctrl Spkr	Dys Spkr	Language	Speech Material	Type of data	Cause & Type	Severity Intelligibility	Clinical Eval
Whitaker [33]	1	6	American English	• Alphabet letters, single digits, Passage • 10 control words and • 36 words from Grandfather passage	Audio	CP, Spastic	• Mild -1 • Mild to moderate -1 • Moderate -1 • Moderate to severe -1 • Severe - 2	
Nemours [126]	-	11	American English	• 74 nonsense sentences per speaker • "Grandfather" passage and • the "Rainbow" passage	Audio	CP, Spastic	• $I > 90$ - 2 • $80 < I < 90$ – 2 • $70 < I < 80$ – 2 • $60 < I < 70$ – 1 • $50 < I < 60$ – 3 • missing info - 1	FDA
Universal Access [95]	13	16	American English	• Digits (10 words X 3 reps) • Letters (26 words X 3 reps) • Computer Commands (19 words X 3 reps) • Common Words (100 words X 3 reps) • Uncommon Words (300 words X 1 rep)	Audio, Video	CP, Spastic	• $I > 85$ - 5 • $55 < I < 85$ – 3 • $25 < I < 45$ – 4 • $0 < I < 25$ – 4	
TORGO [168]	7	8	American English	• Non-words eg. repetitions of /iy-p-ah/ • Short words • Restricted sentences • Unrestricted sentences	Audio, Video, EMA	CP, ALS, Spastic, Mixed	• a – 3 • c– 2 • d/e – 3	FDA
ANH Corpus [44]	160	601	French	• Sustained vowels • Maximal phonation time • Airway interrupted sentences to estimate • Special sentences to estimate velar leakage • Text reading with several speed instructions • Spontaneous description of a picture • DDK task	Audio, Clinical, Metadata	PD, Hypokinetic	-	UPDRS

QoLT [25]	30	100	Korean	• 37 APAC words • 100 machine commands • 36 Korean phonetic alphabets • Phonetically Balanced Words	Audio	CP, Spastic	• Mild -65 • Mild to moderate - 23 • Moderate to severe - 8 • Severe - 4	APAC
Cantonese [205]	5	11	Cantonese	• 61 Word-level stimuli 23 short sentences • Phonetically rich passage • 5-minute conversation	Audio, Video	CD, Ataxic	-	-
EST Dutch [213]	-	16	Dutch	• Dutch numbers • 10 phonetically rich sentences • Plomp and Mimpen sentences • 50 most frequent utterances • from the Dutch Polyphone database • 12 semantically unpredictable sentences • 12 interrogative sentences • 5 short texts • 30 sentences with /t/, /p/ and /k/ • in the initial position and unstressed syllable • 15 sentences with /a/, /e/ and /o/ • in unstressed syllables • 3 individual vowels /a/, /e/ and /o/ • 15 bi-syllabic words	Audio	PD CVA3 TBI Congenital	• Mild – 7 • Moderate – 8 • Moderate to severe – 1	-
Tamil [2]	10	22	Tamil	• 103 isolated words • 262 sentences	Audio	CP, Spastic	• 1 – 3 • 2 – 5 • 3 – 7 • 4 – 3 • 5 – 2 • 6 – 2 .	NIEPMD [0 to 6]

Spanish [141]	50	50	Spanish	• Sustained Spanish vowels • Spanish vowels with changing tones • DDK task • Word-level stimuli	Audio	PD, Hypokinetic	-	UPDRS, Hoehn & Yahr scale
Czech [169]	23	23	Czech	• Isolated vowels • Short sentences • Short and Spontaneous monologue	Audio, Video	PD, Hypokinetic	-	UPDRS III, Hoehn & Yahr scale
German [185]	40	138	German	Complex sentences	Audio	PD, Hypokinetic	-	UPDRS III, Hoehn & Yahr scale
EasyCall [194]	24	31	Italian	• 37 commands • 30 non-commands	Audio	PD, Huntington's, ALS	• Mild – 16 • Mild to moderate – 1 • Moderate – 7 • Moderate to severe – 3 • Severe – 3 • Unknown – 1	TOM score [1-5]
IDEA corpus [122]	-	45	Italian	211 isolated words	Audio	8 pathologies	-	-

Table 2.2: Acoustic studies of dysarthric speech corpora

Task	**Speech Features**	**Outcome**
Intelligibility Prediction	$F1$, $F2 - F1$, VOT, Nasality, Burst Spectrum, Voice Quality, Articulation, Prosody, VSA, first-moment difference measures and $F2$ trajectory characteristics for diphthongs	• Identified effective features for differentiating dysarthric from normal speech with 74.8% accuracy [118] • Established that intelligibility can be predicted by a combination of these features, with articulation being the strongest contributor and 75% agreement with human judgments of intelligibility (within a 95% confidence range) [32] • Intelligibility estimates were found to be correlated with VSA, first-moment difference measures but not with $F2$ slope measures for diphthongs, and acoustic measures of supraglottal behavior [190] • Regression analysis for six words revealed that $F2$ slope for only the words 'shoot' and 'wax' with r2 values of 14.3% and 13.9% respectively, was significantly regressed against scaled speech intelligibility [106]. • Listening tests by untrained listeners indicated that 50% increments in duration and intensity are significantly sufficient for improving listener accuracy. However, a 100% increment in $F0$ is necessary to significantly improve the listener accuracy in a stress modification task to improve the intelligibility of dysarthric speech [70]
Understanding Prosody	$F0$ contour, Syllable Duration	• Revealed $F0$ contour's role in prosody for dysarthric speech [149]
Dysarthria Classification	Voice Quality, Articulation, Nasality, Prosody, Articulation Rate, Voiceless Interval Duration, Intensity Range, $F0$ Range, Vowel Space, Traditional Rhythm Metrics, Novel Combined Segment Metrics, EMS	• Identified potential acoustic measures for different dysarthria types and classification of type, severity, and cause. Accuracy of 68.6% by disease (etiology), 31.7% by type, and 54.9% by severity [107]. • Demonstrated rhythmic patterns for dysarthric classification. Most classification methods achieved 80% accuracy, and even with stricter cross-validation techniques, the accuracy remained above 70% [116]. • An automated analysis of speech EMS, which quantifies the rhythmicity of speech within specified frequency bands achieved 84%–100% classification accuracy for group membership [117].
Comparative Analysis	$F0$, $F1$, $F2$, $F3$, VOT, Spectral Variations, Duration, Intensity, Speaking Rate (Dysarthric vs Normal Speech), $F0$ (Declarative vs Interrogative Sentences), Frequency, Intensity Contrasts (Sentence Accent)	• Compared various acoustic features of dysarthric and normal speech [93]. • Found reduced difference in intonation range of 25Hz for dysarthric speech versus 83Hz for healthy speech in the last syllables of interrogative and declarative sentences. Reduced speech rate variability in dysarthric speech was also observed [35]. • Identified different strategies for sentence accent production. Both healthy and dysarthric speaker groups use $F0$ and intensity changes within the target syllable, as well as the contrast of the maximum $F0$ with the previous syllable and an intensity contrast with the rest of the sentence and the percentage of correct classification between accented and unaccented syllables can reach values above 78.8% [125]
Impact of Speech Compression	Mean HNR, Voiced/Unvoiced Frames, Pitch Statistics, VSA, Jitter, Shimmer, CPP, GoP	Analyzed effects of compression on various acoustic features of dysarthric speech [192].

Table 2.3: Dysarthria severity level Identification

Task	Database	Features	Classifier/technique	Outcome (Accuracy)
Non-reference based Approaches				
Intelligibility Prediction [68]	UA Speech	P.563, delta Energy Features, LP features, Kurtosis of spectral flatness	P.563 standard for blind speech quality assessment, Linear regression	Kurtosis of spectral flatness ranked best among all the features explored. RMSE for the baseline P.563 features was 18.3%, for proposed features SD classification 14% and SI scenario 14.3%.
Intelligibility Prediction [41]	UA Speech	Kurtosis of the LP residual, long and short-term temporal dynamics, nasality, and prosody features	RMSE of the proposed composite measure	A composite measure was developed based on linearly combining a salient subset of the proposed measures and conventional prosody-related measures. A correlation of 97% was achieved using the proposed composite measure.
Intelligibility assessment [11]	NKI CCRT	Acoustic cues at different time scales (phonetic, segmental, suprasegmental)	SVM classifiers, ensemble learning	Highest recall rate of 84.8% for ensemble learning versus the baseline of 65.1%.
Reference-Based using Non-Deep Learning				
Intelligibility Prediction [130]	Flemish pathological speech	GMM-based supervector, phonological features	SVR and late fusion	A drop of the RMSE by about 8% relative as compared to the baseline
Intelligibility assessment [96]	QoLT	Phonetic quality, prosody, voice quality features	Feature selection, SVR	RMSE of 8.1 with subjectively rated scores
Dysarthria Classification [31]	Nemours	Rhythm metrics based on vocalic/consonant intervals	Gaussian-Bayes classification	Rhythm metrics based on durational characteristics of vocalic and intervocalic intervals and Pairwise Variability Index using both their raw and normalized measures are not very promising to express the severity level of the dysarthria impairment.
Intelligibility Prediction [124]	UA Speech	i-vectors derived from PLP features	SVR	Minimum RMSE of 0.2728
Dysarthria severity Classification [80]	Nemours	Jitter, Shimmer, mean Pitch, the standard deviation of Pitch, number of Periods, standard deviation of Period, the proportion of the Vocalic duration (%V), HNR (dB), Noise to Harmonics Ratio (%), Articulation Rate, and degree of voice breaks followed by LDA	SVM	Best classification rate with LDA/SVM system of 93% that was achieved over four severity levels of dysarthria

Intelligibility Classification [97]	• NKI CCRT • TORGO	Prosodic (pitch related), voice quality (HNR), jitter and shimmer), pronunciation features (MFCC and phone duration)	LDA, SVM, k-nearest neighbor (KNN)	• NKI CCRT: 73.5% for unweighted, and 72.8% for weighted average recalls of the binary classes using SVM. • TORGO: 94.1% accuracy with pronunciation feature set with an LDA classifier.
Dysarthric Speech Classification [134]	TORGO	Glottal parameters, acoustic features using OpenSMILE	SVM	70% accuracy using glottal features and a small improvement of 0.5% by adding OpenSMILE features
Intelligible Speaking rate [202]	12 participants with ALS and 2 normal subjects	Acoustic, lip movement, and tongue movement	SVM	R^2: 0.712, RMSE: 37.562 WPM
Dysarthria Classification [62]	• TORGO • QoLT	Prosody, voice quality, MFCC features	Random Forest, SVM, Neural Network	• TORGO: An accuracy of 75.63% using a neural network classifier, a relative accuracy increase of 18.13% in comparison to baseline MFCC features. • QoLT: RF classifier trained on all acoustic features leads to the highest accuracy of 70.10%, a relative accuracy increase of 16.83% compared to baseline
Dysarthria Severity Assessment [63]	• TORGO • QoLT	Standard prosodic features, Rhythm-based features	RF, SVM, ANN	• TORGO: Increase of 4.1% and 3% in detection and severity task respectively. • QoLT: 7.5% and 15% in detection and severity assessment respectively
Dysarthria detection [73]	• PC-GITA • MoSpeeDi • UA-Speech	Spectral Subspaces, Temporal Subspaces, MFCC	SVM, GDA	Best performance with temporal GDA • PC-GITA: 82.0 ± 3.5 • MoSpeeDi: 80.5 ± 4.7 • UA-Speech: 96.3

ALS Severity Prediction[136]	Data collected between September 2020 and March 2021 in cooperation with Everything ALS and the Peter Cohen Foundation	Acoustic: rate, duration, voicing, Visual: jaw & lip statistics (higher-order)	Statistical Analysis, LASSO-LARS Regression	Mean Receiver operating characteristics (ROC) curves for the classification experiments encapsulating sensitivities and specificities: • Bulbar vs Control: 0.92 ± 0.6 • Bulbar vs pre-bulbar: 0.81 ± 0.12 • Pre-bulbar vs Control: 0.62 ± 0.14
Dysarthria severity Classification [3]	Nemours	Prosody, spectral, cepstral, and voice quality features followed by seven feature selection methods	Six classifiers: SVM, LDA, ANN, CART, NB, and RF	The classification accuracy ranges from 40.41% to 95.80%,
Reference-Based using Deep Learning				
ALS detection [201]	Unpublished	MFCC, formant centralization ratio, VSA, intonation, prosody, Articulatory (lip & tongue motion)	SVM, DNN	Leave-one-subject-out: • SVM: 80.91% • DNN: 96.57%
Severity prediction [193]	Arizona State University	Envelope modulation spectrum, The long-term average spectrum features and MFCC statistics, Dysphonia features, Correlation structure features and MFCC	DNN and an intermediate DAB (seminal work of Darley, Aronson, and Brown) representation interpretable to most clinicians that work with pathological speech	Joint training strategy yields the best performance and provides the best predictive ability
Severity Level Classification [15]	• UA Speech • TORGO	Audio descriptors, multi-tapered spectral estimation	ANN	• UA Speech: 96.44% • TORGO: 98.7%
ALS detection [6]	Unpublished	7755 acoustic features using OpenSMILE such as MFCC, and the quartile of the $F0$ contour, filter-bank for CNN representation learning	ANN with statistical features, Time-CNN, Frequency-CNN	The best actual person-level performances are • ANN:74.6% (baseline) • Frequency-CNN:84.6% • Time-CNN:80.8%
Intelligibility Estimation [23]	• UA Speech • TORGO	Joint spectro-temporal features from the Mel-scale spectrogram	ANN, CNN	Time-Frequency CNN performed best (captured spectral & temporal variations) with an accuracy of 98.3% for known and 54% for unknown speakers.

Intelligibility Classification [12]	TORGO	MFCC, log filter banks, i-vectors	BLSTM networks, TL	TL with pre-trained models yielded best results with 98.2% accuracy
Intelligibility Assessment [22]	• UA Speech • TORGO	STFT, Single frequency filtered (SFF), Perceptually enhanced spectrograms, Constant-Q spectrograms	CNN	• UA Speech: CQT representation performs best with an accuracy of 98% and 95.8% for female and male speakers respectively • TORGO: CQT-SPEC12-24-48 performed best with 72.6% average RMS for word-level and 78.5% for sentence-level
Dysarthria Severity Classification [77]	· UA Speech · TORGO	MFCCs, constant-Q cepstral coefficients, speech disorder features, i-vectors	DNN, CNN, GRU, LSTM	DNN using MFCC-based i-vectors outperformed others with a 93.97% accuracy under the SD scenario and 49.22% under the SI scenario for the UA Speech database.
Dysarthria Detection [209]	Collected at Arizona State University	Mel-spectrogram, Articulatory Precision, Consonant-Vowel (CV) Transition Precision, Hypernasality, Vocal Quality	DNN with interpretable bottleneck layer	• Sample Level: $83.97 \pm 4.4\%$ • Speaker Level: $94.71 \pm 5.15\%$
Generalizable Intelligibility Assessment [211]	Not specified	MFCC, spectral representations	Deep Learning models	Doctoral thesis

Table 2.4: ASR for dysarthric speech

Database	Features	Speech Tech	Accuracy/WER
Speaker Adaptation			
TORGO	MFCC features, lexical model adaptation	HMM-based ASR, MLLR, MAP adaptation	Significant improvement for low-moderate severity speakers with an average accuracy of mean word recognition accuracy of the speaker-adapted ASR systems 68.39% [128]
UA Speech	PLP features, MFCC	HMM-GMM based ASR, MLLR, MAP adaptation	Improved recognition for low-moderate severity with an average of best accuracy of 54.1% [27]
UA Speech	PLP coefficients	Background interpolated MAP (BI-MAP), HMM-GMM based ASR. Several BI-MAPs were investigated.	8% absolute and up to 40% relative, over the standard MAP adapted baseline [181]
QoLT	MFCC features, dysarthric speaker severity level	HMM-GMM based ASR, MLLR, MAP adaptation	For mild speakers 13.9% relative WER reduction were obtained when using 100 adaptation data. [102]
TIMIT (source model) TORGO (source model) Nemours(target model)	MFCC	HMM-GMM based ASR, MLLR, constrained MLLR (CMLLR)	The WERs of the two source models are different, with the TORGO being better for recognizing severe dysarthric speech while TIMIT is better for recognizing mild dysarthric speech. [133]
UA Speech	MFCC- 12 coefficients, their first and second derivatives, and all the acoustic features	ANN	The word recognition rate (WRR) SI ASR model: average of 68.38%. The highest WRR of speaker-dependent ASR models was 95% [177].
UA Speech	MFCC	HMM-GMM based ASR, maximum likelihood, MAP adaptation, background model	11.5% relative improvement compared to the baseline [26]
ATR Japanese	Short-term Mel spectra, Convolutive Bottleneck Network (CBN) features	HMM-GMM based ASR, CBN	Word recognition accuracy Baseline (MFCCs): 84.3% CBN: 88.0% [135]
UA Speech	Jitter, shimmer, multi-taper MFCC features	GMM-HMM-based ASR, DNN-HMM-based ASR, fMLLR speaker adaptation	Fusion of MT-MFCC, jitter and shimmer (VP.JitShim) features show a relative improvement of 8.4% in GMM-HMM-based system and 10.7% in DNN-HMM-based system over the baseline MFCC features [13]
QoLT	DNN posteriors	HMM with KL-divergence based emission probabilities, Bayesian estimation, L2 and lexical regularization	Baseline DNN-HMM: 15.6% KL-HMM: 8.8% [103]
EST Dutch	MFCC	Multi-stage DNN training, fMLLR	An absolute improvement of WER of 6.3% using 2 stage training as compared to the baseline of single stage training with dysarthric data alone [220]
Unpublished (4 speakers)	MFCC, i-vectors	Kaldi ASR, Lattice-free maximum mutual information (LF-MMI) model	Average absolute improvement in WER of 15.7% across 4 speakers [187]
UA Speech	fMLLR adaptation network for Wav2Vec2	Wav2Vec2 with adaptation network for feature fine-tuning	57.72% WER (high severity), improved across all severity levels [10]
UA Speech	Spectral and temporal features	Deep Embedding Features with DNN/TDNN & Conformer	25.05% WER (test set) [52]
Speech Enhancement			
TORGO	Filter bank	GMM based resynthesis, Acoustic space modification (temporal & frequency morphing)	The proportion of words correctly recognized increased up to 121% from 72.7% to 87.9% relative to the original speech, across various parameterizations of the recognizer[167]
UA Speech (partial)	MFCC	Temporal adaptation, GMM-HMM & DNN-HMM-based ASR, MLLT and fMLLR	DNN-HMM: Highest relative WER improvement of 48.44 % for SI scenario, 20.48% for SA and 17.65% for unseen data [14]
UA speech (partial)	MFCC, temporal adaptation, DAE bottleneck features	DNN-HMM-based ASR, DAE	An overall absolute improvement of 16% was achieved using tempo adaptation followed by autoencoder-based speech front-end representation for DNN-HMM-based dysarthric speech recognition [195].
Unpublished	• Acoustic: MFCCs with derivatives. • Articulatory: Coordinates of tongue and lip sensors (after Procrustes normalization). • CCA-transformed acoustic features.	GMM-HMM, DNN-HMM, and LSTM-HMM speech recognizers. 6-fold SI cross-validation	PER (Acoustic+ 50 dimensional CCA): • GMM-HMM: 42.1 • DNN-HMM: 36.2 • LSTM-RNN-HMM:34.8 [105]

Unpublished	Log Mel-filterbank energy and its derivatives	GMM-HMM (baseline), DNN-HMM, CNN-HMM (different types: F-CNN, T-CNN, TF-CNN, PTF-CNN), and CLSTM-RNN. Used leave-one-subject-out cross-validation for SI recognition	TF-CLSTM-RNN achieved the best overall accuracy PER: • Session wise: 30.6% • Across intelligibility: 35.4% [104]
UA speech (partial)	MFCC, temporal adaptation, bottleneck features	DNN-HMM-based ASR, Time-Delay Neural Network Denoising Autoencoder (TDNN-DAE)	Absolute improvements of 13% and 3% were observed in the ASR performance for SI and SA systems respectively as compared with unenhanced dysarthric speech recognition [16].
UA Speech	Mel-generalised cepstrum (MCEP), $F0$ features, Mel-spectrogram features from MelGAN	MaskCycleGAN, MelGAN, Fill-in-the-frame data augmentation (FIF DA)	4.8% absolute improvement in the case of the male speakers, and 10.8% absolute improvement in the case of the female speakers [157].
UA Speech, PC-GITA, EasyCall	Wav2Vec2.0, Hubert, XLSR features	Self-supervised acoustic representation learning, End-to-end ASR (ESPnet)	XLSR models performed best for all 3 databases. WER UA speech: 26.1% with SA, PC-GITA-PD: 12.9% with EasyCall-PD: 16.5% [64]
Data Augmentation			
UA Speech (partial)	MFCC	Time/tempo stretching, DNN-HMM-based ASR, fMLLR	Absolute improvement in WER · Tempo based DA: 4.24% · Speed based DA: 2% [196]
UA Speech	MFCC	Interpolation followed by downsampling for tempo adjustment, DNN-HMM-based ASR with TDNN	Best overall WER of 27.88% [207]
UA Speech	MFCC	Speed, tempo, and VTLP-based augmentation. (LHUC) based speaker adaptive and multi-task learning (MTL)-based training for DNN	2.92% absolute (9.3% relative) word error rate (WER) reduction over the baseline system without data augmentation, and gave an overall WER of 26.37% [51]
UA Speech, Tamil dysarthric speech corpus	MFCC	Virtual linear microphone array-based synthesis followed by multi-resolution feature extraction (MRFE), DNN-HMM-based ASR system	A reduced WER of up to 32.79% and 35.75% for low and very low intelligible speakers [21]
UA Speech	Voice grams	Visual data augmentation, Speech vision ASR using Spatial Convolutional Neural Network (S-CNN)	Absolute average WRAs of 64.71% with DA [176]
TORGO	Raw waveform features, data augmentation	Parametric CNNs, multi-stream acoustic modelling	WER: Parametric CNN: 36.2% (3.4% absolute error reduction) Multi-stream acoustic modelling: 33.2% [218]
UA Speech	Mel filter bank,	TDNN-DAE, speed, tempo, and loudness based DA, specialized Spec Augment. End-to-end ASR (ESPnet)	WER: An absolute improvement of 16% with a final WER of 20.6% with DA [17]
UA Speech	Wav2vec 2.0 embedding features	Variational auto-encoder generative adversarial network (VAE-GAN)-based DA, LF-MMI factored TDNN, LHUC-SAT, ESPnet toolkit	WER: 27.78%, with 57.31% on the subset of speakers with "Very Low" intelligibility [76]
Transfer Learning			
UA Speech	CNN-TDNN-F features	DNN	11.6% average improvement [208]
Various	Speed/volume variations, VM-MRFE	DNN	Improved performance for low intelligibility [123]
UA Speech (155 words)	Whisper features, BLSTM	DNN	59.78% average accuracy [162]

Chapter 3

Automatic Assessment of Intelligibility of Dysarthric Speech

This chapter is based on the following publications:

I. Bhat, C., Vachhani, B., & Kopparapu, S. K. (2017). Automatic Assessment of Dysarthria Severity Level using Audio Descriptors. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5070-5074). IEEE.

II. Bhat, C., & Strik, H. (2020). Automatic Assessment of Sentence-level Dysarthria Intelligibility using BLSTM. IEEE Journal of Selected Topics in Signal Processing, 14(2), 322-330.

3.1 Automatic Assessment Of Dysarthria Severity Level Using Audio Descriptors

Dysarthria is a motor-speech impairment, often characterized by speech that is generally indiscernible by human listeners. Assessment of the severity level of dysarthria provides an understanding of the patient's progression in the underlying cause and is essential for planning therapy, as well as improving automatic dysarthric speech recognition. In this chapter, we propose a non-linguistic manner of automatic assessment of severity levels using audio descriptors or a set of features traditionally used to define timbre of musical instruments and have been modified to suit this purpose. Multi-tapered spectral estimation-based features were computed and used for classification, in addition to the audio descriptors for timbre. An Artificial Neural Network (ANN) was trained to classify speech into various severity levels within the Universal Access dysarthric speech corpus and the TORGO database. An average classification accuracy of 96.44% and 98.7% was obtained for the UA Speech corpus and the TORGO database, respectively.

3.1.1 Introduction

Dysarthria is a motor speech impairment, often characterized by speech that is indiscernible by human listeners. Dysarthria is generally caused by neurological diseases such as amyotrophic lateral sclerosis, PD, cerebral palsy, or neurological trauma, manifesting as weakness, paralysis, or a lack of coordination of the motor-speech system, resulting in a reduction in intelligibility, audibility, naturalness, and efficiency of vocal communication. Assessment of the severity level of dysarthria could be treated as a diagnostic step and is crucial to understanding the patient's progression in the underlying cause, to make clinical decisions regarding the course of therapy or medication, as well as to plan speech therapy sessions whenever applicable. Severity assessment is generally conducted by trained speech-language pathologists, but it can be expensive and may vary between practitioners due to the use of a combination of standardized and non-standardized methods during therapy [94, 60]. On the other hand, an objective severity assessment has the advantage of being cost-effective, repeatable, and paves the way for further automation such as improved speech recognition of dysarthric speech. An understanding of severity has contributed to improved speech recognition of dysarthric speech as seen in [102, 174, 132]. In general, speech intelligibility has been used as an indicator of the severity of speech disorders [120]. Automatic intelligibility assessment of pathological speech has been carried out broadly by either (a) Automatic Speech Recognition (ASR) based methods that require reference data as well linguistic know-how [129, 120, 36] or (b) blind intelligibility assessment [11, 100, 69]. In [130], authors discuss the applicability of acoustic and phonological ASR-free features for intelligibility assessment. The authors discuss the classification of pathological speech as intelligible or non-intelligible using scores from the fusion of multiple

subsystems addressing various aspects of speech such as phonological, intonation, etc. in [96]. Literature indicates that research is trending toward moving away from language-specific ASR-based methods to language-independent automatic intelligibility assessment. While speech quality and intelligibility are closely related, their relationship is not trivial. Frenchay Dysarthria Assessment (FDA) [37] defines several parameters that need to be considered for automatic assessment of the severity level of dysarthria, of intelligibility is but one. For PD, voice quality symptoms are visible earlier than intelligibility symptoms. Hence, it is desirable to assess dysarthria severity level using the speech utterance at the voice quality level, in addition to the granular level of articulatory accuracy. In this chapter, we propose the applicability of a set of acoustic descriptors that have been used to characterize the timbre of a musical instrument [152]. Timbre is the quality of music or voice that renders each one distinct. We investigate the use of features suggested in [152] for dysarthria severity classification. Additionally, we compute the acoustic descriptors using a multi-taper-based spectral estimation [189] for improved spectral resolution. Significant improvement in severity level assessment was seen using the multi-taper-based timbre acoustic descriptors as compared to the work in literature, wherein authors reported 95% classification accuracy using feature fusion on the Universal Access (UA) Dysarthric Speech Corpus [69] and in [79] authors reported 93.2% correct classification rate of dysarthria severity levels on the TORGO and the Nemours database.

The rest of the chapter is organized as follows. Section 3.1.2 describes the audio descriptors and their role in dysarthric speech severity classification, Section 3.1.3 discusses the severity classification methodology and a description of the data used, Section 3.1.4 discusses the experimental setup used, Section 3.1.5 describes the results and analysis and we conclude in Section 3.1.6.

3.1.2 Audio Descriptors

In this chapter, we use audio descriptors that have been designed for timbre characterization of a musical instrument as a set of features for dysarthric speech severity classification. Timbre is a multidimensional attribute encompassing a set of auditory descriptors, in addition to pitch, loudness, duration, and spatial position [152]. In [152], authors define a set of audio descriptors that can be categorized into *global descriptors*, that are computed across the utterance and *time-varying descriptors*, that are extracted within each frame of the utterance. Audio descriptors are computed using various representations of the speech utterance such as (1) Temporal Energy Envelope (2) Short-term Fourier transform (STFT) (3) Equivalent rectangular Bandwidth (ERB) based auditory model and (4) Harmonics. For each audio descriptor, as shown in Table 3.1, median and interquartile range have been considered.

Table 3.1: Audio descriptors used for severity classification

Serial Number	Audio Descriptor	Serial Number	Audio Descriptor
1	Attack	17	Spectral Slope
2	Decay	18	Spectral Decrease
3	Log-Attack time	19	Spectral Rolloff
4	Attack-slope	20	Spectro-temporal variation
5	Decrease slope	21	Frame energy
6	Temporal Centroid	22	Spectral Flatness
7	Effective Duration	23	Spectral Crest
8	Frequency of Energy Modulation	24	Harmonic Energy
9	Amplitude of Energy Modulation	25	Noise Energy
10	RMS-Energy Envelope	26	Noisiness
11	Autocorrelation-12 coefficients	27	Fundamental Frequency
12	Zero Crossing Rate	28	Inharmonicity
13	Spectral Centroid	29	Tristimulus (3 coefficients)
14	Spectral Spread	30	Harmonic Spectral Deviation
15	Spectral Skewness	31	Odd to Even Harmonic Ratio
16	Spectral Kurtosis		

3.1.2.1 Multi-taper Spectral Estimation

In our work, we investigate the usage of multi-taper spectral estimation to compute STFT and Harmonic-based features. Conventional spectral estimation of speech uses a Hamming window or a single taper. Using a single taper windowing results in a significant portion of the signal being discarded and the data points at the extremes being down-weighted, giving a high variance for the direct spectral estimate [158]. Hence, a multi-taper method is used so that the statistical information lost by using just one taper is partially recovered by using multiple windows for the same duration. The multi-taper spectrum is thus a weighted sum of the several tapered periodograms. Spectral estimation of a signal S using multi-taper method is as follows,

$$S_(m,k) = \frac{1}{M}\sum_{p=0}^{M-1}\lambda(p)\sum_{j=0}^{N-1} w_p(j)s(m,j)e^{-i2\pi\frac{k}{N}j} \tag{3.1}$$

where $w_p(j)$ is the p^{th} data taper function, M is the number of tapers and $\lambda(p)$ is the weight corresponding to the p^{th} taper, N is the speech frame length, $s(m,j)$ is the j^{th} speech frame and k is the FFT points. In practice, weights are designed to compensate for increased energy loss at higher-order tapers.

10 feature sets have been used for dysarthria severity classification is as shown in Table 3.2.

Table 3.2: Feature sets for severity classification

Feature Set	Dimension	Input Representation	Acoustic Descriptors
F1	22	Temporal Energy Envelope	1-10
F2	26	Audio Signal	11-12
F3	22	STFT – Magnitude	13-23
F4	22	STFT – Power	13-23
F5	22	ERB – FFT	13-23
F6	22	ERB – Gammatone	13-23
F7	38	Harmonic	15-31
F8	22	Multi-taper Magnitude	13-23
F9	22	Multi-taper Power	13-23
F10	38	Multi-taper Harmonic	15-31

3.1.3 Severity Classification

In this chapter, an Artificial Neural Network (ANN) has been used as a classifier for dysarthria severity classification. The ANN consists of three layers: an input layer, a hidden layer, and an output layer. The input layer comprises I nodes equivalent to the dimension of the input feature set being used and the output layer comprises K nodes, the number of classes into which dysarthria severity is classified by the classifier. The number of nodes in the hidden layer J is varied based on the dimension of the input feature set being used. ANN configuration is as shown in the Figure 3.1.

3.1.3.1 Data

The proposed technique was validated using two different dysarthric databases, i.e., (a) the Universal Access (UA) Dysarthric Speech Corpus [95] and (b) the TORGO database [168].

- **The UA Dysarthric Speech Corpus:**
  The UA speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. The recording material consisted of 455 distinct words with 10 digits, 26 international radio alphabet letters, 19 computer commands, 100 common words and 300 uncommon words that were distributed into three blocks. Three blocks of data were collected for each speaker such that in each block, the speaker recorded the digits, radio alphabets, computer commands, common words and 100 of the uncommon words. Thus each speaker recorded 765 isolated words. Speech intelligibility ratings for each dysarthric speaker, as assessed by five naive listeners are also included in the corpus. Speakers were divided into four different categories based on the intelligibility, namely high, mid, low, and very low. We use this

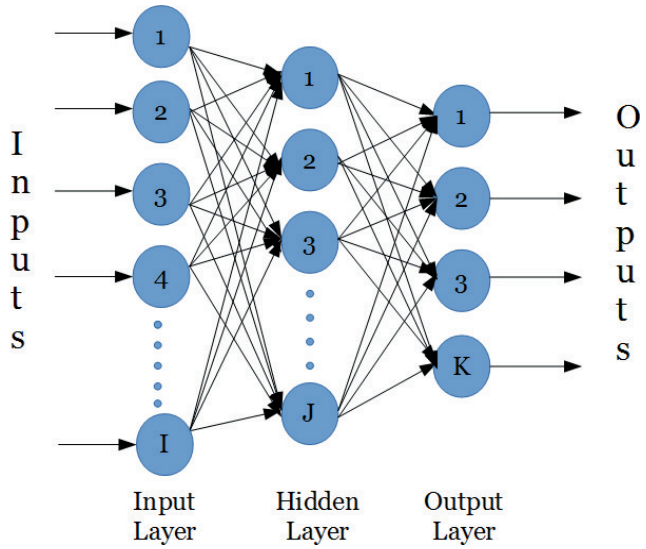

Figure 3.1: ANN configuration for severity classification

information to classify dysarthria severity levels.

- **TORGO:**
  The TORGO database of dysarthric articulation consists of aligned acoustics and measured 3D articulatory features from speakers with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). The TORGO database consists of 8 dysarthric (DYS) speakers (three females and five males) and seven non-dysarthric or healthy control (HC) speakers (three females and four males) as a control group. The acoustic data were recorded through two different microphones; an array microphone with eight recording elements placed at a distance of 61 cm facing the speaker, and a head-mounted microphone. The corpus consists of (1) non-words, (2) Short words such as digits, international radio alphabet letters, (3) Restricted sentences, (4) Unrestricted sentences. The motor functions of every subject were assessed according to the standardized Frenchay Dysarthria Assessment (FDA) [37] by a speech-language pathologist. FDA measures 28 relevant perceptual dimensions of speech grouped into eight categories, namely reflex, respiration, lips, jaw, soft palate, laryngeal, tongue, and intelligibility.

The speaker-wise severity classification for both the UA Speech and the TORGO database is shown in Table 3.3. The severity classification for the UA speech

database is based on intelligibility, whereas for the TORGO database, the overall FDA score for the dysarthric speakers as per [79] is used.

Table 3.3: Speaker-wise severity distribution for the UA Speech and the TORGO database (F** for female speakers, M** for male speakers)

Severity	**UA Speech**	**TORGO**
Very Low	F05, M08, M09, M10, M14	F03, F04, M03
Low	F04, M05, M11	F01, M05
Medium	F02, M07, M16	M01, M02, M04
High	F03, M04, M12, M01	—

3.1.4 Experimental Setup

3.1.4.1 Data

For the UA Speech corpus, a total of 2812 dysarthric utterances with utterances corresponding to 10 digits and 19 computer commands from block B1 and B2 for training and testing of the classifier has been used.

For the TORGO database, we have used a total of 1540 dysarthric utterances for experimentation.

3.1.4.2 Multi-taper Spectral Estimation

Multi-taper spectral estimation was done using Discrete Prolate Spheroidal sequences (DPSS) or Thomson or Slepian tapers [189] with 6 orthonormal tapers.

$$w_p(j) = \frac{\sin[\omega_c T(p-j)]}{(p-j)}, \qquad j = 0, 1, \ldots, N-1 \tag{3.2}$$

where N denotes the desired window length in samples, ω_c is the desired main-lobe cut-off frequency in radians per second, and T is the sampling period in seconds. Twelve-dimensional MFCC features were computed using Thomson multi-taper spectral estimation with a $30\,ms$ window and a $10\,ms$ shift rate.

3.1.4.3 ANN Configuration

The classification was carried out for 8 for different settings of hidden layer neurons. For the hidden layer, the number of neurons J or nodes is varied based on the dimension I of the input feature set, and is given as $J = I \star m$, where $m \in \{0.5, 0.66, 0.75, 0.8, 0.83, 1, 1.25, 1.5\}$. The number of output nodes $K = 4$ and 3 for the UA Speech and the the TORGO database, respectively. For both the UA Speech and TORGO data, 70% of the data was used for training the network, 15% was used for validation, and 15% was used for testing.

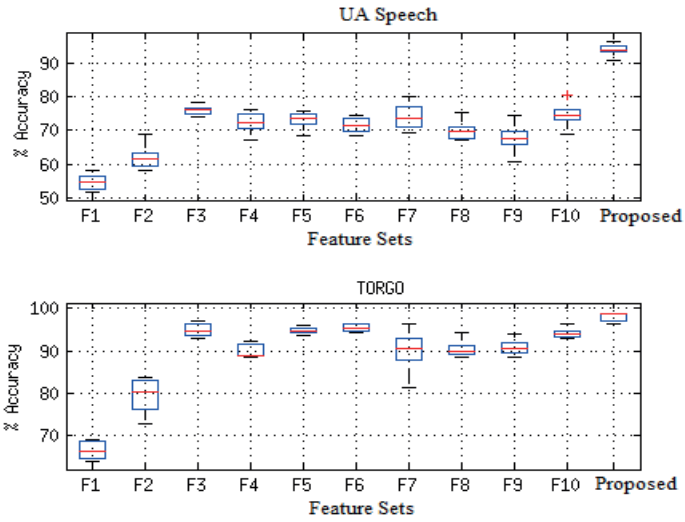

Figure 3.2: Feature-wise classification accuracy for varying hidden layer nodes

3.1.5 Results and Discussion

Severity classification was carried out using the experimental setup described in Section 3.1.4. It was observed that feature set F1, corresponding to the Temporal Energy Envelope, performed poorly compared to the other feature sets. Feature sets STF magnitude(F3), ERB FFT(F4), ERB Gammatone(F5), Multi-taper Harmonics (F10) performed well for all settings. This could be attributed to the fact that this is a global measure and hence is unable to characterize the severity adequately. Also, for each of the feature sets, similar accuracies were observed across validation and training sets, indicating that there is no overfitting or underfitting. The classification accuracy for individual feature sets F1-F10 across different numbers of hidden nodes (varied as discussed in Section 3.1.4), is as seen in Figure 3.2. Multi-taper spectral estimation(F10) outperformed the Hamming window-based Harmonics audio descriptors (F7) in the severity classification accuracy. This could be attributed to the inherent noise robustness of the multi-taper spectral estimation [108]. We obtained the best classification accuracy when the fusion of all the features from F1-F6 and F10 (Proposed) were used together to give a comprehensive feature of dimension 164. Here we replace the Harmonic timbre feature set F7 with multi-taper-based Harmonic feature set F10. Severity-wise classification accuracy for the above fusion set is given in Table 3.4.

Table 3.4: Severity-wise classification accuracy for the UA Speech and the TORGO database

Severity	**UA Speech**	**TORGO**
Very Low	96.1	99.1
Low	95.1	98.4
Medium	96.7	97.0
High	95.7	—

3.1.6 Conclusion

Dysarthria is a motor-speech impairment, often characterized by speech that is generally indiscernible by human listeners. Assessment of the severity level of dysarthria is essential for planning therapy, as well as improving automatic dysarthric speech recognition. Objective assessment of the severity level or intelligibility of dysarthric speech is essential with reliability, speed, and consistency in view. Literature suggests that automatic speech recognition of dysarthric speech can be improved if prior knowledge of the severity of dysarthria is available. In this chapter, we propose a non-linguistic technique for automatic assessment of severity levels using audio descriptors or a set of features traditionally used to define the timbre of musical instruments. Additionally, we use multi-taper-based spectral estimation to compute the spectral and harmonic features. It was observed that classification accuracies using multi-taper-based harmonics were higher than the Hamming window-based harmonic features. An Artificial Neural Network (ANN) was trained to classify speech into various severity levels within the Universal Access dysarthric speech corpus and the TORGO database. A fusion of feature sets F1-F6 and F10 (proposed) to give a comprehensive feature set of dimension 164 provided an average classification accuracy of 96.44% for the UA speech corpus 98.7% for the TORGO database, respectively. For both the UA Speech and the TORGO databases, the overall classification accuracy as well as classification accuracy at the feature level outperforms the accuracies cited in one of the most recent works [69, 79] for these dysarthric speech corpora.

3.2 Automatic Assessment of Sentence-Level Dysarthria Intelligibility using BLSTM

Dysarthria is a motor speech impairment, often characterized by slow and slurred speech that is generally incomprehensible by human listeners. An understanding of the intelligibility level of the patient's dysarthric speech can provide insight into the progression/status of the underlying cause and is essential for planning therapy. Automatic assessment of dysarthric speech intelligibility can be of immense value and serve to assist speech-language pathologists in diagnosis and therapy. However, this is a non-trivial problem due to the high intra and inter-speaker variability in dysarthric speech. In this work, we propose a machine learning-based method to automatically classify dysarthric speech into intelligible (I) and non-intelligible (NI) using Bidirectional Long-Short Term Memory (BLSTM) Networks. We explored the balancing of training data to represent both classes almost equally and its implications on binary classification. Additionally, we present a mechanism to use the available pre-trained acoustic models for transfer learning. It was observed that the transfer learning method was able to handle channel noise. This technique provided a significant improvement of roughly 6% as compared to the traditional machine learning method.

3.2.1 Introduction

Human communication relies heavily on speech intelligibility with a significant impact on quality of life, especially in the case of pathological speech. Speech production is governed by two key events: linguistic and acoustic composition that demand a seamless coordination of muscle groups driven by motor planning and motor programming to ensure intelligible speech [92]. Dysarthria is a motor speech disorder that results from neurological causes. This manifests as weakness, paralysis, or a lack of coordination and imprecise movements of the muscles of the motor-speech system, resulting in a reduction in intelligibility, audibility, naturalness, and efficiency of vocal communication. Dysarthria can be categorized as either progressive or non-progressive. Progressive dysarthrias are seen in patients with PD, Huntington's disease, multiple sclerosis, motor neuron disease, and so forth. Although the disease progression can be delayed, progressive dysarthria leads to a progressive decline in muscle functioning over time [160].

Speech intelligibility has been used as an indicator of the severity of speech disorders, in general, [120], including dysarthria [94]. Intelligibility and speaking rate can be used as quantitative measures of dysarthria severity level [31]. Assessment of the intelligibility of dysarthric speech may be considered as a diagnostic step to assess the outcome of a surgical procedure, to understand the patient's progression in the underlying cause, or to make clinical decisions regarding the course of therapy or medication and speech therapy planning. Intelligibility assessment is undertaken by a trained speech-language pathologist, which can be expensive and

time-consuming. On the other hand, speech technology-based objective assessment of intelligibility can support clinical decision-making while being accurate, cost-effective, and robust. Further, it also paves the way for further automation, such as improved speech recognition of dysarthric speech. An understanding of severity has contributed to improved speech recognition of dysarthric speech as seen in [102, 174, 132, 14].

Automatic intelligibility assessment has been carried out broadly by either (a) Model-based methods that require reference data as well as linguistic know-how [120, 36] or (b) blind intelligibility assessment [11, 100, 69]. In [130], authors discuss the applicability of acoustic and phonological ASR-free features for intelligibility assessment. Authors discuss the classification of pathological speech as intelligible or non-intelligible using scores from the fusion of multiple subsystems addressing various aspects of speech such as phonological, intonation, etc. in [96]. Literature indicates that research is trending towards moving away from language-specific ASR-based methods to language-independent automatic intelligibility assessment, as seen by some of the recent works reviewed next. Dysarthric speech from the Quality-of-Life Technology (QoLT) database in the Korean language was classified using a subset of features from phonetic quality, prosody, and voice quality features along with a linear-kernel support vector machine (SVM) classifier to yield a root mean square error of 8.1 [101]. Gaussian Bayes classification technique was used to classify dysarthric speech from the Nemours database into three classes using rhythm metrics based on acoustic measures of the duration of vocalic and consonantal intervals in continuous speech [31]. i-vectors modeled by factor analysis using perceptual linear predictors (PLP) as acoustic features were used for intelligibility prediction of the Universal Access dysarthric speech [124]. A set of prosodic features selected using Linear Discrimination Analysis (LDA) and an SVM classifier was used to classify dysarthric speech from the Nemours database into four classes [80]. Abnormal variation in the prosodic, voice quality, and pronunciation aspects at the sentence level are used as features, followed by a post-classification posterior smoothing scheme to evaluate pathological speech into binary intelligibility labels (intelligible and not-intelligible). Further, feature-level fusions and subsystem decision fusion are used for arriving at a final intelligibility decision [97]. The histograms of the pronunciation mappings are generated by aligning the phone sequence obtained from an ASR with the canonical phone sequence and used as features along with a structured sparse linear model incorporated with phonological knowledge for intelligibility prediction [98].

In this work, we explore the use of Bidirectional Long-Short Term Memory (BLSTM) type of Recurrent Neural Networks (RNN) for binary intelligibility classification of dysarthric speech. The performance of the classifier using speech parameters such as Mel Frequency Cepstral Coefficients (MFCC), log filter banks, and i-vectors has been compared. Further, a transfer learning approach is adopted wherein dysarthria intelligibility level is predicted using acoustic models pre-trained on normal speakers' speech to improve the intelligibility classification. We also

demonstrate that the transfer learning approach can be used to effectively classify the intelligibility of Dutch dysarthric speech, accommodating the differences in language, recording environment, and speaker variations.

The rest of the chapter is organized as follows. In Section 3.2.2, we briefly describe the TORGO and the Dutch dysarthric speech corpora. In Section 3.2.3, we describe the classifier design and the features used. In Section 3.2.4, we provide the details of the experimental setup and details on how the proposed system can be used for transfer learning. In Section 3.2.5, we discuss the evaluation of the proposed system and visualize the BLSTM network learning, followed by a conclusion in Section 3.2.6.

3.2.2 Databases

3.2.2.1 TORGO Dysarthric Speech Corpus

The TORGO database [168] of dysarthric articulation consists of aligned acoustics and measured 3D articulatory features from speakers with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). The TORGO database consists of 8 dysarthric (DYS) speakers (three females and five males) and seven non-dysarthric or healthy control (HC) speakers (three females and four males) as a control group. The age of the patients ranged from 16 to 50. The individuals selected for the control group were matched according to age and gender with dysarthric subjects so as to be able to compare acoustic and articulatory differences, as well as to analyze their relationships mathematically and functionally. Speaker IDs beginning with F represent female speakers, and ones that begin with M represent male speakers. The acoustic data were recorded through two different microphones: a microphone array with eight recording elements placed at a distance of 61 cm facing the speaker and a head-mounted microphone. The corpus consists of (1) non-words, (2) Short words such as digits in international radio alphabet letters, (3) Restricted sentences, and (4) Unrestricted sentences. The motor functions of each subject were assessed according to the standardized Frenchay Dysarthria Assessment (FDA) [37] by a speech-language pathologist. FDA measures 28 relevant perceptual dimensions of speech grouped into eight categories, namely reflex, respiration, lips, jaw, soft palate, laryngeal, tongue, and intelligibility. Although this database was recorded with various types of stimuli, the speech audio recordings corresponding to restricted sentences are used in this study. The prompts used for recording sentence-level speech audio comprise three pre-selected phoneme-rich sentence sets: the "Grandfather passage", 162 sentences from the sentence intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech, 460 sentences from the MOCHA database, and spontaneously elicited descriptive texts.

This database provides intelligibility labels in five categorical grades [a, b, c, d, e], which were reduced from an initial 9-point scale, where 'a' is the label corresponding to the best intelligibility and 'e' is the worst. The data was divided into

Table 3.5: Speaker-level intelligibility for the TORGO sentences

Speaker	Number of Utterances	Dysarthria Level	Category
FC01	30		
FC02	51		
FC03	84	Control	I
MC01	83		
MC02	84		
MC03	84		
MC04	84		
F03	92	a	
F04	63	a	I
M03	80	a	
F01	24	d/e	
M02	84	d/e	NI
M04	91	d/e	
M05	53	c	

two classes: intelligible (I) and non-intelligible (NI). While the speakers with grades a, b and control speakers were tagged with the category I, the speakers with grades c, d or e were tagged with NI. Speaker-wise intelligibility for sentence-level recordings is shown in Table 3.5. Considering that the sentence-level data is unbalanced with fewer utterances of dysarthric speech, only a subset of control speech sentences that are available as dysarthric speech have been used, hence the lower number of utterances.

3.2.2.2 Dutch Dysarthric Database

The speech material used in the study was selected from the recordings of dysarthric speakers prior to speech therapy [49]. To avoid speaker familiarity influencing the evaluation procedure, materials from seven different speakers were used, as shown in Table 3.6. These were all male and suffered from hypokinetic dysarthria caused by PD. To investigate the different levels of granularity in intelligibility evaluation for a broad range of speech material, four different types of recordings were used: lists of single words, declarative Semantically Unpredictable Sentences (SUS) sentences, interrogative SUS sentences, and regular sentences. All samples consisted of existing Dutch words. The word lists contained three or five words, the SUS sentences all contained six words, and the length of the regular sentences varied between five and eight words. Speech fragments with different levels of intelligibility, from low to high, were selected based on annotations by two listeners who did not participate in the current experiment. In this work, we assess the intelligibility of utterance-level speech for six speakers. Utterance level intelligibility evaluations were obtained

Table 3.6: Overview of speech material used for Dutch dysarthric data

Type of speech material	Speaker	Speech fragments
Word lists	S1	5 word lists (5 words each)
	S2	5 word lists (3 words each)
Declarative SUS sentences	S3	6 sentences
	S4	6 sentences
Interrogative SUS sentences	S5	6 sentences
	S6	6 sentences
Regular sentences	S7	8 sentences
	S1	8 sentences

using subjective rating scales, namely the Visual Analogue Scale (VAS) and the Likert scale. The VAS intelligibility score for the six speakers ranged from 39% to 89%. Speakers (4) with scores less than 75% were considered non-intelligible, and speakers (2) with scores above 75% were considered intelligible.

3.2.3 Classifier and Feature Design

3.2.3.1 Classifier Design

Traditional Artificial Neural Networks (ANN) use either time-aggregated or time-sequence features as inputs. Time-aggregated features are typically a combination of the statistics of frame-wise features of a speech utterance and some meta-information that is computed as one feature vector that represents an entire utterance. Time-sequence features are computed frame-wise and used to train an ANN over a fixed context of the right and left frames of the intended frame. This does not ensure learning of the temporal development of the speech signal by the ANN. An understanding of the temporal distance between events is essential for sequential tasks such as motor control and rhythm detection [53]. HMMs also fail to learn this information, whereas Recurrent neural networks (RNNs) can, in principle, learn to make use of it. Particularly Long Short-Term Memory (LSTM) RNN-type networks have been shown to perform better than simple RNNs on tasks involving long time lags. A deep LSTM-RNN architecture uses a combination of multiple levels of representation of the speech utterance along with the flexible use of long-range context to provide significant improvement in speech applications as shown in the end-to-end phoneme recognition task [56].

The LSTM RNN comprises special units called memory blocks. The memory blocks perform three crucial functions, namely (1) store the temporal state of the network in the memory cells (2) control the flow of information using gates or special multiplicative units, (3) appropriately handle the processing of continuous input streams using forget gates [170]. Each memory block in the original architecture contained an input gate and an output gate. The flow of input activations into

the memory cell is controlled by the input gate, while the output gate controls the output flow of cell activations into the rest of the network. The internal state of the cell is scaled by the forget gate and then added to the cell through the self-recurrent connection, whereby it adaptively resets the cell's memory. The precise timing of the outputs is learned using peephole connections from the internal cells to the gates in the same cell. Figure 3.3 shows a building block of an LSTM-RNN. Multiple LSTM layers are stacked to construct deep LSTM RNNs as seen in Figure 3.4. The input to the LSTM network at a given time step goes through multiple LSTM layers in addition to propagation through time.

In our work, the network needs to learn the aspects of speech that define the intelligibility of speech, which has possibly been afflicted with a motor-speech disorder (dysarthria). We choose to use a special type of LSTM called the Bidirectional LSTM (BLSTM) network to address the long-range bidirectional interdependencies within the speech data. BLSTM networks operate on the input sequence in both directions in order to make a decision for the current input. Deep BLSTM-HMM hybrid acoustic models have shown to perform well for TIMIT as well as Wall Street Journal corpora [55]. Online speech recognition has also been achieved using latency-controlled BLSTM acoustic models [210]. Various training schemes and parameter tuning aspects have been compared to provide an overview of the performance of deep BLSTM-based ASRs [221]. In this work, we use BLSTM acoustic models to classify dysarthric speech sequences into binary classes intelligible (I) and non-intelligible (NI).

3.2.3.2 Feature Extraction

We use three different types of feature sets for the dysarthric speech classification task. The objective is to be able to classify *unseen* speakers or speech utterances with high reliability. This drives us to make use of the features that retain the speech characteristics while masking the speaker and/or channel variations. We use log filter banks, Mel-frequency Cepstral Coefficients (MFCC), and i-vectors as features individually and in combination to learn the dysarthric speech classification.

Log filter bank
Pre-emphasis filter is first applied to a speech signal to provide a smooth spectral representation. The signal is then segmented into overlapping frames on which the Hamming window function is applied. A short-term Fourier transform is then applied to each frame to calculate the power spectrum. Subsequently, the filter banks are computed on the Mel scale to emulate human ear perception. At this point, we extract the log filter bank features.

MFCC
The output of the log filter bank is further processed by applying Discrete Cosine

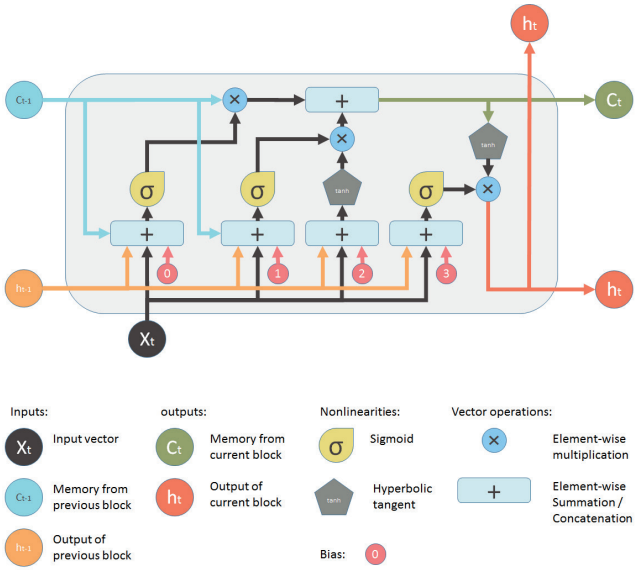

Figure 3.3: LSTM building block [182]

Transform (DCT) to decorrelate the filter bank coefficients and yield a compressed representation of the filter banks.

i-vector

The i-vector extraction is performed by a model-based cluster adaptive training (CAT) estimation where the underlying models are Gaussian mixture models (GMM) rather than Hidden Markov Models (HMMs) [84]. A GMM universal background model (UBM) with M mixture components, denoted as $\mathcal{M}$, is used to represent the intrinsic variability of phonemes within the speech data. A mean super-vector of component means ${\mu_0}^{(m)}$, diagonal component covariance matrices $\sum^{(m)}$ and mixture coefficients $\omega^{(m)}$ is used to define the UBM. The canonical model $\mathcal{M}$ is then used to generate the input acoustic feature vectors $x_t \in \mathbb{R}^D$.

All the speech data belonging to a particular speaker is used to generate an i-vector for that speaker, which is used for training the classifier. This i-vector represents all the utterances of that particular speaker. The i-vectors thus generated span the *speaker Eigenspace*, and each speaker is represented by a point in this *speaker Eigenspace*. For a Gaussian component $m \in M$, the linear dependence between the speaker-adapted means and the canonical means is computed as

$$\mu^{(sm)} = \mu_0^{(m)} + M^{(m)}\lambda^{(s)} \tag{3.3}$$

$\mu^{(sm)}$ is the m-th component of speaker-dependent super-vector, $M^{(m)}$ is the factor submatrix for component m of size $D \times P$, representing P bases spanning

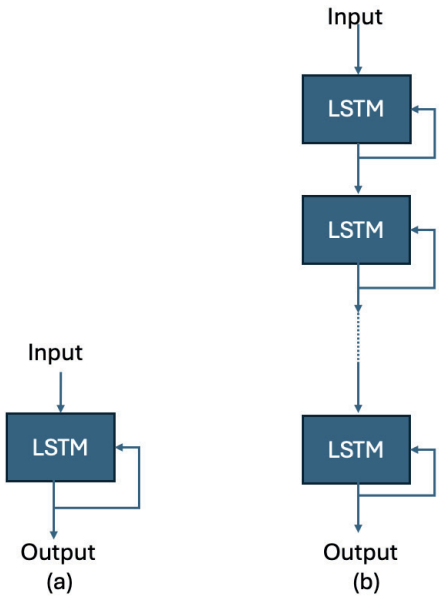

Figure 3.4: (a) LSTM network (b) Deep LSTM network

the subspaces with the highest variability in the mean super-vector space and $\lambda_{iv}^{(s)}$ is the P-dimensional *i-vector* of speaker s.

The factor matrix $M^{(m)}$, when decomposed into two components, represents two distinct acoustic factors, namely speaker and noise-dependent supervector. Each factor is represented by a subspace of the i-vector space. For each speaker s and noisy environment n, Equation 3.3 can be written as

$$\begin{aligned} \mu^{(snm)} &= \mu_0^{(m)} + M_s^{(m)}\lambda^{(s)} + M_n^{(m)}\lambda^{(n)} \\ &= \mu_0^{(m)} + M_{sn}^{(m)}\lambda^{(sn)} \end{aligned} \tag{3.4}$$

where $M_{sn}^{(m)}\lambda^{(sn)} = \begin{bmatrix} M_s^{(m)} & M_n^{(m)} \end{bmatrix} \begin{bmatrix} \lambda^{(s)} \\ \lambda^{(n)} \end{bmatrix}$ This gives i-vectors an advantage in handling the variations caused by noise and retaining the speech characteristics.

3.2.4 Experiments

Our experiments are designed to evaluate an unseen speaker for intelligibility. We follow the leave-one-out (LOO) method for training and testing, `i.e.` the speaker whose utterances are being tested is not available in the training data.

3.2.4.1 Balancing Training Data

In order to exploit the machine learning techniques fully, suitable data to build these systems is imperative. However, owing to speaker muscle weakness and fatigue, the collection of dysarthric data is tedious, especially for speakers with severe dysarthria. Additionally, since dysarthria can stem from a variety of neurological disorders, the characterization of dysarthric speech is complex, which makes the data collection process difficult. This is evident from the amount of sentence-level data available for unintelligible dysarthric speech, intelligible dysarthric speech, and healthy control speech in the TORGO database. In order for the BLSTM network to learn the patterns in dysarthric speech for intelligibility classification, it is important that the training data represents the population appropriately. It is also important to note that there are very few dysarthric unintelligible utterances for a female speaker (F01), with only 24 utterances. The parameters considered for the selection of training data for each speaker are gender, dysarthria status (dysarthric/-control), intelligibility status (I/NI), and number of utterances available. Our focus in this work is to balance the number of I and NI utterances in the training data for each test speaker. Table 3.7 shows the distribution of data used for training and testing the BLSTM. The cells in blue represent the test data, while the pink and yellow background indicate NI and I speech, respectively. This scheme of training and test data is well balanced for the task at hand, binary classification of speech utterances. Figure 3.5 shows the distribution of data used for training each test speaker as per Table 3.7, in terms of gender and dysarthria status. It can be seen from Figure 3.5 that gender-based distribution is not very balanced, with fewer utterances for female speakers. Dysarthria status-based distribution is fairly uniform for male speakers.

Table 3.7: Test speaker-wise data distribution

Test Speaker	F01	F03	F04	FC01	FC02	FC03	M01	M02	M03	M04	M05	MC01	MC02	MC03	MC04
F01	24				51	84	84	84		91	53		84	84	
F03	24	92	63	30		84	84	84	80	91	53		84		
F04	24	92	63	30	51		84	84		91	53		84	84	
FC01	24		63	30	51	84	84	84		91	53	83	84		
FC02	24		63	30	51	84	84	84		91	53	83	84		
FC03	24	92		30	51	84	84	84		91	53		84	84	
M01	24			30	51		84	84		91	53			84	84
M02	24			30	51		84	84		91	53			84	84
M03	24	92	63				84	84	80	91	53	83	84		
M04	24			30	51		84	84		91	53	83	84		
M05	24				51	84	84	84		91	53		84	84	
MC01	24			30	51		84	84		91	53	83	84	84	84
MC02	24			30	51		84	84		91	53	83	84	84	84
MC03	24			30	51		84	84		91	53	83	84	84	84
MC04	24			30	51		84	84		91	53	83	84	84	84

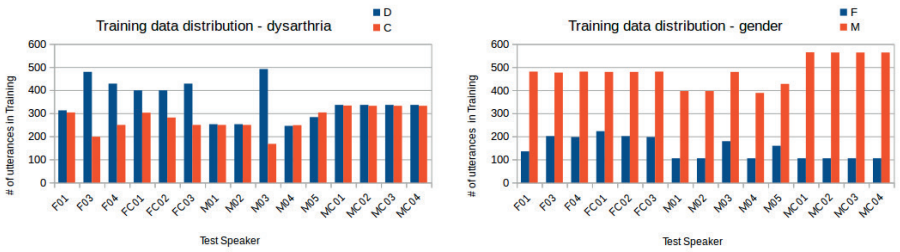

Figure 3.5: Dysarthria status and gender-based Training data distribution.

3.2.4.2 Feature Extraction and BLSTM Classifier

The input utterance is represented by a sequence of frames, with an effective frame rate of 10ms per frame. 26-dimensional log filter bank coefficients and 26-dimensional MFCCs are computed for each frame and used as inputs to the encoder-decoder framework of the BLSTM.

100-dimensional i-vectors were computed per frame and used along with the 26-dimensional MFCCs as inputs to the BLSTM. The Kaldi speech processing toolkit [156] was used to compute the i-vectors.

A BLSTM encoder-decoder model is trained for the binary classification of intelligibility. The encoder consists of a stack of 2 BLSTM layers with both layers comprising 64 LSTM units each in the forward and backward LSTMs. Max-pooling over time is used to obtain the encoder representation from the second BLSTM layer. The decoder is a fully connected dense layer whose size is adjusted as per the size of the encoder representation and the number of output labels (2 for binary classification). The encoder input dimensionality depends on the feature set being used: 26 for MFCC and log filter bank, 126 for concatenation of MFCC and i-vector, and 1024 for transfer learning experiments discussed in Section 3.2.4.3. We use an ADAM optimizer to train the model.

3.2.4.3 Transfer Learning (TL) using i-vectors

BLSTM encoder-decoder models trained using speech parameters like MFCC, log filter bank, or i-vectors require large amounts of data, which in turn results in severe degradation in performance in low resource speaker-independent scenarios such as the task at hand [159]. We hypothesize that the BLSTM is unable to handle speaker variability and learn the temporal context simultaneously with small amounts of dysarthric speech. Hence, we adopt a transfer learning approach, wherein the internal representations are learned by DNN-based ASR models. The motivation is that apart from handling speaker invariability, the pre-trained model should also give us representations that capture the temporal context in the sequence of input speech frames. We use the publicly available ASpIRE ASR DNN model [150]

which is pre-trained on the Fisher English corpus for this purpose. It consists of 6 layers of Time Delay Neural Networks (TDNNs), each with ReLU activations and re-normalization, followed by two branches: one for sequence (Chain) training and the other for cross-entropy (Xent) regularization. The input to the ASpIRE model is a sequence of frames, with each frame comprising 26-dimensional MFCCs and a 100-dimensional i-vector feature from the TORGO database. Each input frame (along with its context) is transformed into an intermediate representation by each of the layers of the ASpIRE model. In our experiments, we splice the model until layer 3 and beyond to form the pre-trained layers. We skip the final outputs because of their high dimensionality. We use the output of layer 3 of the TDNN as features to train the BLSTM for binary classification of intelligibility. Speakers from the TORGO database are evaluated for intelligibility using this BLSTM network.

Further, we use the BLSTM network and are thus trained to evaluate speakers from the Dutch dysarthric database for intelligibility. We hypothesize that the BLSTM encoder-decoder, along with the pre-trained framework outperforms the BLSTM network that uses the TORGO training data alone.

3.2.5 Results and Discussion

In this section, we discuss the performance of the BLSTM network using various different features for training the network. Table 3.8 shows the classification performance for features such as MFCC, log filter bank, the two sets of features concatenated together, and for i-vector-based features. It can be seen that the overall performance improved for balanced training data for each of the feature sets as well as for a large majority of individual speakers across the features. The classification for utterances corresponding to speaker F01 was consistently low owing to the fact that this was the only female dysarthric non-intelligible speech available and was not represented well during the training.

Table 3.8 also shows the performance of the BLSTM using the transfer learning technique. The overall classification performance improved by approximately 6% when TL was used. The improvement was significant for the utterances corresponding to speaker F01, by roughly 37%. We hypothesize that the TL technique enabled the BLSTM to learn the distinction between intelligible and non-intelligible speech better since it was extracted from pre-trained models of normal English speech.

The speech audio of the TORGO database was recorded using a head-mounted microphone and a microphone array as described in Section 3.2.2. The speech data that was recorded using a head-mounted microphone often contained considerable channel noise [97]. The results reported in Table 3.8 correspond to both head-mounted and array microphone recordings of the TORGO dysarthric speech. Table 3.9 reports our observations when we examine the difference in accuracy for the two types of recordings, namely head-mounted microphone and microphone array, using MFCC features for RNN and TL technique. It is seen that across speakers, the accuracy has significantly improved for head-mounted microphone audio when the TL

Table 3.8: Speaker-wise classification accuracy(%) for unbalanced and balanced data

Speaker	log filter		MFCC		MFCC+logf		i-Vector		TL	
	Unbal	Bal	Unbal	Bal	Unbal	Bal	Unbal	Bal	Unbal	Bal
F01	50	83.33	29.17	45.83	45.83	58.33	23.47	41.67	66.67	83.33
F03	93.48	76.09	81.52	90.22	68.48	91.3	91.3	88.04	97.83	92.39
F04	61.9	76.19	100	98.41	100	79.37	100	98.41	100	98.41
FC01	100	100	100	100	100	100	100	100	100	100
FC02	100	100	100	100	100	100	100	100	100	100
FC03	89.29	100	100	98.81	100	97.62	100	100	100	100
M01	90.48	95.24	96.43	100	91.67	94.05	100	100	100	100
M02	70.24	89.29	70.24	98.81	73.81	94.05	70.24	95.24	97.62	100
M03	100	50	85	67.5	82.5	70	87.5	75	100	100
M04	62.64	56.04	47.25	91.21	68.13	89.01	67.03	87.91	82.42	100
M05	9.43	79.25	96.23	98.11	94.34	92.45	98.11	98.11	100	100
MC01	93.98	96.39	100	96.39	92.77	92.77	100	95.18	100	100
MC02	100	100	98.81	97.62	98.81	98.81	100	98.81	100	100
MC03	98.81	77.38	98.81	98.81	98.81	97.62	98.81	98.81	100	100
MC04	94.05	98.81	92.86	95.24	92.86	79.76	100	97.62	98.81	98.81
Average accuracy	80.95	85.2	86.42	91.8	87.2	89.01	89.10	91.65	96.22	98.2

technique is used, especially for F01 and M03. F01 is the only female speaker with non-intelligible speech, and M03 is the only male dysarthric speaker with intelligible speech, making these speakers not very well represented in the training set. This indicates that using the i-vector-based TL method performs well even for speech riddled with channel noise. We would like to note that our approach yields higher classification accuracy as compared to the results reported in [97] for the TORGO database, wherein the best performance was obtained for the *Pronunciation subsystem* at 94.1% for the Linear Discriminant Analysis (LDA) classifier. Some of the key differences in our approaches are (1) In order to build a Pronunciation subsystem in [97], it is crucial to identify the vowel segments accurately, making manual phonetic transcription a necessity; in our approach, we do not need phonetic transcription of the audio. (2) Pre-processing was applied on head-mounted microphone audio in [97], compromising the speech characteristics, whereas we worked directly with the noisy audio. (3) No data balancing was applied for the classification task in [97].

We also used the i-vector-based training and TL training to classify the Dutch dysarthric speech described in Section 3.2.2. The objective was to assess if the BLSTM-based encoder trained using speech features for English, a rich-resourced language, the TORGO database of dysarthric speech can be used to decode dysarthric speech in a less-resourced language like Dutch. The Dutch database used in this study contains only 39 sentence-level utterances from 6 speakers. The TL-based training improved the classification to an overall accuracy of 44% compared to the i-vector-based classification of 28%. However, considering that the chance classifica-

Table 3.9: Speaker-wise classification based on recording method

Speaker	MFCC		TL	
	head_mic	mic_array	head_mic	mic_array
F01	8.33	83.33	74.99	91.67
F03	86.96	93.48	86.95	97.83
F04	100	97.62	96.82	100
FC01	100	100	100	100
FC02	100	100	100	100
FC03	97.62	100	100	100
M01	100	100	100	100
M02	100	97.62	100	100
M03	45	90	100	100
M04	90	92.16	100	100
M05	97.62	100	100	100
MC01	92.68	100	100	100
MC02	95.24	100	100	100
MC03	97.62	100	100	100
MC04	90.48	100	97.62	100
Average accuracy	86.77	96.95	97.09	99.3

tion accuracy is 50% for binary classification, we have to surmise that the BLSTM was unsuccessful in decoding the Dutch dysarthric data classes. Transfer learning using an existing English baseline for Dutch speech is not straightforward. This could be attributed to the difference in language of the pre-trained models used in the TL-based method. We will examine this further through visualizing the BLSTM learning.

We examine the classification performance of the BLSTM network for two different training techniques, namely, MFCC-based and TL-based by visualizing the network learning, using saliency maps extracted at the input layer of the BLSTM network. Figure 3.6 shows a visualization of the feature vector, log filter bank, saliency maps for ground truth, and predicted activations for two speech utterances. Figures (a) and (b) are for the same utterance from speaker F01 (class=NI), while (b) and (c) belong to speaker MC04 (class =I). BLSTM learning for MFCC and TL are being shown, wherein the MFCC feature vector is 26 dimensions and the TL feature vector is 1024 dimensions. The visualization was plotted using the *nipy_spectral* colormap provided in the python *matplotlib* library. It can be seen from Figure 3.6, that when TL was used, the saliency maps corresponding to network prediction and the ground truth are similar. Also, for speaker F01, the network saliency maps at the input looked significantly different when MFCC features were used as compared to when the TL-based learning was used. We note that the activations of the correctly classified utterances correspond to vowel regions. The

learning of the BLSTM is then comparable to the best-performing speech subsystem, namely the Pronunciation subsystem described in [97], which is at 94.1% for the TORGO dataset. We achieve a 4% absolute improvement as compared to this system.

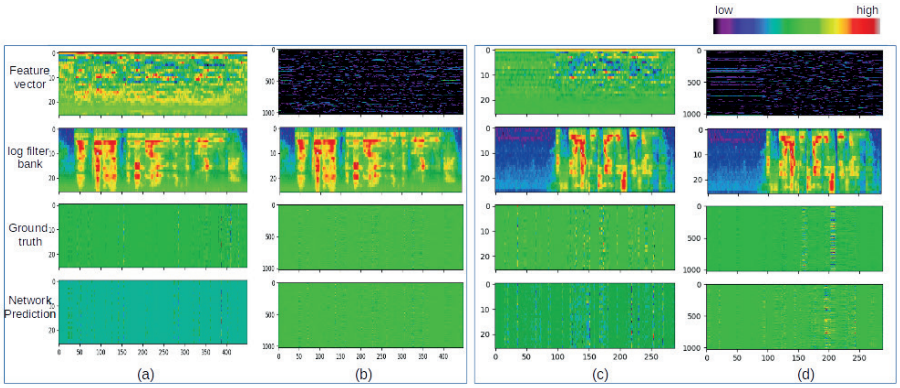

Figure 3.6: Visualization of the network learning for speakers F01 (class = NI) and MC04 (class = I) for data trained using MFCC and Transfer learning (TL). (a) F01-MFCC, (b) F01-TL, (c) MC04-MFCC, (d) MC04-TL

3.2.6 Conclusion

Dysarthria is a motor speech impairment, often characterized by slow and slurred speech with reduced intelligibility. Automatic assessment of dysarthric speech intelligibility can be of immense value and useful to assist speech-language pathologists in diagnosis and therapy. In this work, we propose a machine learning-based method to automatically classify dysarthric speech into intelligible (I) and non-intelligible (NI) using Long-Short Term Memory Neural Networks. Bidirectional LSTM type of RNNs were used to train and classify dysarthric speech. Speech parameters were chosen in such a way as to retain the speech characteristics while normalizing the speaker-specific nature of the speech. We adopt a transfer learning approach, wherein the internal representations are learned by DNN-based ASR models. The motivation is that apart from handling speaker invariability, the pre-trained model should also give us representations that capture the temporal context in the sequence of input speech frames. We explored the balancing of training data to represent both classes almost equally. The performance of the BLSTM was better for balanced data across all features. BLSTM performance was lowest for log filter bank features, while it was comparable for MFCC and i-vector and provided the best results for TL-based features. This technique provided a significant improvement of roughly 6% as compared to the traditional machine learning method. It was also

observed that the transfer learning method was able to handle channel noise. The visualization of the BLSTM network learning indicated that the network activations corresponded to vowel regions of the dysarthric speech, which has been shown in the literature. A deeper level of understanding of intelligibility will pave the way to better systems for automatic assessment and recognition of dysarthric speech. The direction of future work would be to be able to categorize dysarthric speech into multiple categories instead of binary classification. Cross-language learning and classification of dysarthric speech is another promising area that can be explored.

Acknowledgment

The authors would like to express sincere thanks to Ms. Xue Wei, who is currently pursuing her PhD at Radboud University, Netherlands, for providing details regarding Dutch dysarthric data and her assistance in extracting speech features for Dutch dysarthric data.

Chapter 4

Acoustic parameters and time domain adaptation of dysarthric speech

This chapter is based on the following publications:

I. Bhat, C., Vachhani, B., & Kopparapu, S. K. (2016, September). Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper Spectral Estimation. In Interspeech (pp. 228-232).

II. Bhat, C., Vachhani, B., & Kopparapu, S. (2016). 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 18 (pp. 370-377). Springer International Publishing.

4.1 Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper Spectral Estimation

Dysarthria is a motor speech disorder resulting from impairment in muscles responsible for speech production, often characterized by slurred or slow speech resulting in low intelligibility. With speech-based applications such as voice biometrics and personal assistants are gaining popularity, automatic recognition of dysarthric speech becomes imperative as a step towards including people with dysarthria in the mainstream. In this chapter, we examine the applicability of voice parameters that are traditionally used for pathological voice classification such as jitter, shimmer, F0, and Noise Harmonic Ratio (NHR) contour, in addition to Mel Frequency Cepstral Coefficients (MFCC) for dysarthric speech recognition. Additionally, we show that multi-taper spectral estimation for computing MFCC improves the unseen dysarthric speech recognition. A Deep neural network (DNN) - hidden Markov model (HMM) recognition system fared better than a Gaussian Mixture Model (GMM) - HMM-based system for dysarthric speech recognition. We propose a method to optimally use incremental dysarthric data to improve dysarthric speech recognition for an ASR with DNN-HMM. All evaluations were done on the Universal Access Speech Corpus.

4.1.1 Introduction

Dysarthria is a motor speech disorder resulting from impairment in muscles responsible for speech production. Neurological injury to the nervous system may result in weakness, paralysis, or a lack of coordination of the motor-speech system, resulting in a reduction in intelligibility, audibility, naturalness, and efficiency of vocal communication. For dysarthric speakers, speech is a more efficient/convenient mode of communication with electronic devices as compared to keyboard input [165]. Voice or speech as a computer interface for dysarthric speakers was implemented as early as 1985 [46]. The authors designed an assistive device to bypass the keyboard and activate the computer using voice control. Despite the early start, automatic recognition of dysarthric speech is poorer as compared to that of normal speech, owing to the inter-speaker and intra-speaker inconsistencies in the acoustic space as well as the sparseness of data. As per the literature, the work so far can be broadly classified into two types of research - (1) improving intelligibility by modifying or enhancing the dysarthric speech and (2) ASR-based speech recognition by the speaker adaptation. In [66], authors study the effect that certain modifications have on the intelligibility of dysarthric speech and report that by transforming the dysarthric speech at the short-term spectral levels, an increase in intelligibility was attained. In the study [82], authors have achieved increased intelligibility by transforming the vowels of a dysarthric speaker to more closely

match the vowel space of a normal speaker. Features that provided optimum performance were vowel duration and $F1$ - $F3$ (formant 1 - formant 3) stable points that were computed using shape-constrained isotonic regression. In another study [167], the author transforms various aspects of speech, such as the correction of pronunciation errors, adjustment of the tempo and the frequency characteristics of speech to obtain increased intelligibility. Yet another technique to increase both the perceptual quality of the speech as well as intelligibility are transformations to formant trajectories of dysarthric speech, to closely match that of a normal speaker [81].

In [57], one of the earlier works in ASR-based dysarthric speech recognition, the authors stress that the data insufficiency challenges and define confusability and consistency measures to predict recognizer performance. Several works [145, 102] discuss the merits of selection of ASR type, namely - speaker-independent (SI), speaker-dependent (SD) or speaker adapted (SA) by analyzing the correlation between the severity of dysarthria and best performing ASR type (one of SA or SD). In [26], authors have used a method of measuring similarity between dysarthric speakers and select only the most similar speaker data for training rather than the SI acoustic models, followed by maximum a posteriori (MAP) adaptation. Studies [181] also suggest an improvement in recognition by using more suitable prior model or background model for adaptation based on the dysarthric speaker's acoustic characteristics. Work pertaining to speaker-based lexical or pronunciation model adaptation in addition to acoustic model adaptation, [131, 127, 206] have shown improvement in the ASR performance. An understanding of the speech production process through the articulatory models for speech has proven beneficial in improved accuracy of the ASR, both conventional GMM-HMM and DNN-HMM [165, 166, 58]. More recently the application of neural network topologies [177, 135], feature space maximum likelihood linear regression (fMLLR) transformation [58] and a hybrid adaptation using maximum likelihood linear regression (MLLR) and MAP [174] have been used to improve dysarthric speech recognition.

We believe that speech-based applications such as voice biometrics and personal assistants can immensely benefit dysarthric speakers if designed well. Given the challenges in collecting dysarthric data, the thrust is now on recognition of unseen speech utterances, i.e. recognition of dysarthric speech that is not a part of the training set. In this chapter, we propose a method and examine a set of features to improve speech recognition of unseen dysarthric speech. We incorporate multi-taper MFCC (MT-MFCC), which has been proven to be effective in speaker verification and speech recognition [5, 4] as well as voice disorder classification [39]. Additionally, we examine the voice parameters (VP), such as jitter, shimmer, $F0$ features, and noise-to-harmonics ratio (NHR), that have traditionally been used for voice disorder classification [120, 188]. Some of these parameters have been used to automatically assess the severity level of dysarthria [80]. The main contribution of this chapter is a framework for unseen dysarthric speech recognition using a DNN-HMM SA-ASR system along with a combination of speaker-specific features.

To the best of our knowledge no other work has examined the usefulness of voice parameters such as jitter, shimmer F0 features and noise-to-harmonics ratio (NHR) in the context of dysarthric speech recognition.

The rest of the chapter is organized as follows. Section 4.1.2 describes the features and their role in dysarthric speech recognition, Section 4.1.3 discusses the various experimental setups and a description of the data used, Section 4.1.4 describes the results and analysis, and we conclude in Section 4.1.5.

4.1.2 Features for Dysarthric Speech Recognition

4.1.2.1 Multi-taper Spectral Estimation

Conventional spectral estimation of speech uses a Hamming window or a single taper. Using a single taper windowing results in a significant portion of the signal being discarded and the data points at the extremes being down-weighted, giving a high variance for the direct spectral estimate [158]. Hence, a multi-taper method is used so that the statistical information lost by using just one taper is partially recovered by using multiple windows for the same duration. The multi-taper spectrum is thus a weighted sum of the several tapered periodograms. Spectral estimation of a signal S using multi-taper method is as follows,

$$S_(m,k) = \frac{1}{M}\sum_{p=0}^{M-1}\lambda(p)\sum_{j=0}^{N-1} w_p(j)s(m,j)e^{-i2\pi\frac{k}{N}j} \tag{4.1}$$

where $w_p(j)$ is the p^{th} data taper function, M is the number of tapers and $\lambda(p)$ is the weight corresponding to the p^{th} taper, N is the speech frame length, $s(m,j)$ is the j^{th} speech frame and k is the FFT points. In practice, weights are designed so as to compensate for increased energy loss at higher-order tapers.

4.1.2.2 Jitter and Shimmer

Jitter and shimmer are characteristic of the speech of an individual and have been beneficial in speaker recognition tasks [42]. *Jitter* represents the perturbations that occur in the fundamental frequency $F0$ and can be interpreted as a modulation of the periodicity of the voice signal. Reduced control of vocal fold vibration, as is the case in dysarthria manifests as jitter. Pathological voices are generally characterized by a high degree of jitter and are perceived as hoarse. Hence, an estimation of jitter has been used in the classification of pathological speech. Absolute jitter is computed as per Equation 4.2.

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \tag{4.2}$$

where $T_i = 1\ /\ F0$ and N is the number of $F0$ periods.

Shimmer pertains to the amplitude variation of the sound wave and varies with the glottal resistance and mass lesions in the vocal folds manifesting as the presence of noise emission and breathiness in the voice [188]. Absolute shimmer is computed as per Equation 4.3 and is expressed in decibels (dB).

$$Shimmer(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20 * log\left(\frac{A_{i+1}}{A_i}\right)\right| \tag{4.3}$$

where A_i is the extracted peak-to-peak amplitude, and N is the number of $F0$ periods.

4.1.2.3 $F0$ Features

The role of fundamental frequency $F0$ in the intelligibility of speech has been studied for both normal and dysarthric speech [149]. These studies suggest that a higher variation in $F0$ contributes significantly to increased intelligibility. However, for dysarthric speakers, the precision and flexibility of the vocal folds, articulators, and other speech subsystems are lower, leading to reduced prosodic control, reflecting a reduction in intelligibility. Additionally, studies show that the slower articulatory rate tends to be associated with low values of mean, maximum, and variations of $F0$ [191]. $F0$ measurements such as mean and variation are also indicative of the vocal loudness of speech, which has a bearing on speech intelligibility.

4.1.2.4 Noise to Harmonic ratio (NHR)

Noise-to-Harmonics ratio (NHR) is indicative of the abnormal vibratory characteristics of the vocal folds, manifesting as hoarseness in dysarthric speech. NHR is measured in dB, calculated by the ratio of noise energy or the aperiodic part of a sustained vowel to the energy of the periodic part. NHR can be used as a measure of voice quality and is defined as below.

$$NHR(dB) = 10 * log\left(\frac{E_n}{E_p}\right) \tag{4.4}$$

where E_p is the energy of the periodic part, and E_n is the energy of the noise. NHR has been used as one of the discriminative features to evaluate the degree or severity of dysarthria in [80].

4.1.3 Speech Recognition Methodology

4.1.3.1 Data

Data from the Universal Access (UA) speech corpus [95] was used for both training and testing of the two ASR systems discussed in this section. The UA speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS)

speakers with cerebral palsy. The recording material consisted of 455 distinct words with 10 digits, 26 international radio alphabet letters, 19 computer commands, 100 common words, and 300 uncommon words that were distributed into three blocks. Three blocks of data were collected for each speaker such that in each block, the speaker recorded the digits, radio alphabets, computer commands, common words, and 100 of the uncommon words. Thus, each speaker recorded 765 isolated words. Data from all channels were used for this work. Speech intelligibility ratings for each dysarthric speaker, as assessed by five naive listeners are also included in the corpus. We use this information to analyze the performance of our recognition systems at the dysarthria severity level.

The objective of this work is to recognize unseen dysarthric data, and explore the applicability of voice parameters in recognition of dysarthric speech. The training and testing corpus, as described in Table 4.1 allows us to compare and contrast the performance of our recognition systems for seen and unseen testing data, i.e., DYS-computer command words (DYS-CC words).

Table 4.1: Training and testing corpus

Purpose	**Data**	**Number of Utterances**
Training	HC-digits	800
	HC-computer command words	1500
	DYS - digits	800
Testing	HC-digits	110
	HC - computer command words	229
	DYS - digits	169
	DYS - computer command words	361

4.1.3.2 ASR Systems and Experimental Setup

Feature extraction and normalization:

Multi-taper spectral estimation was done using Discrete Prolate Spheroidal sequences (DPSS) or Thomson or Slepian tapers [189] with 6 orthonormal tapers.

$$w_p(j) = \frac{\sin[\omega_c T(p-j)]}{(p-j)}, \qquad j = 0, 1, \ldots, N-1 \tag{4.5}$$

where N denotes the desired window length in samples, ω_c is the desired main-lobe cut-off frequency in radians per second, and T is the sampling period in seconds. Twelve-dimensional MFCC features were computed using Thomson multi-taper spectral estimation with a 30 ms window and a 10 ms shift rate.

All the *voice parameters*, such as the jitter, shimmer, $F0$, and NHR measures were computed using the voice analysis software, PRAAT [18], wherein a cross-correlation (cc) method was used for acoustic periodicity estimation, using a $30\,ms$ window and a $10\,ms$ shift rate. PRAAT gives various measurements for each of the above voice parameter. Based on experimental evidence and literature [42], features as shown in Table 4.2 were chosen for speech recognition.

Table 4.2: Voice parameters extracted from PRAAT [18]

Feature	**PRAAT Measurement**
Jitter	Jitter(local, relative)
Shimmer	Shimmer(local, dB)
Fundamental Frequency F0	Standard Deviation Range (Maximum - Minimum)
Noise to Harmonic ratio	Standard Deviation Mean

We have three sets of features, namely, MFCC, multi-taper MFCC (MT-MFCC), and voice parameters (VP).

Speech recognition:
We use the Kaldi toolkit [156] for both GMM-HMM-based and DNN-HMM-based dysarthric speech recognition. A 3-state HMM with a monophone or a triphone context model is used. GMM-HMM system was trained using a maximum likelihood estimation (MLE) training approach along with 100 senones and 8 Gaussian mixtures. Cepstral mean normalization (CMN) were applied to each of the above sets of features. Dimensionality reduction was done using Linear Discriminant Analysis (LDA), wherein LDA builds HMM states using feature vectors with a reduced feature space. We use the context of 6 frames (3 left and 3 right) to compute LDA. The feature vector size post LDA is set to 40.

The input layer of DNN has 360 ($40\times 9 frames$) dimensions using a left and right context of 4 frames. The output layer has a dimension of 96 (number of senones available in the data). We used 2 hidden layers with 512 nodes in each layer. Trigram language model was used, and the performance of each of the recognition systems is reported in terms of word error rate (WER).

We explore the use of our feature sets - MFCC, MT-MFCC, and MT-MFCC-VP for speech recognition with speaker adaptation(SA).

Speaker Adaptation:
Traditionally, speaker adaptation techniques such as MLLR, MAP are applied on SI acoustic models at the time of decoding. We use Maximum Likelihood Linear Transform (MLLT) for speaker normalization. MLLT derives a unique transforma-

tion for each speaker using the reduced feature space from the LDA.

An inter-speaker feature space normalization technique, known as feature space maximum likelihood linear regression (fMLLR) [48] is performed for each speaker, wherein the acoustically transformed feature vector $\hat{o}(t)$ is estimated using a transformation matrix A and a bias vector b as $\hat{o}(t) = Ao(t) + b$, where $\hat{o}(t)$ is obtained by transforming input feature vector $o(t)$ at frame t.

Speaker adaptive training (SAT) [7] is applied at the time of training the acoustic models and aims at eliminating the inter-speaker variation. fMLLR-based SAT was applied to create speaker-adapted (SA) acoustic models; further, fMLLR was applied to the features of the input utterances at the time of decoding. SAT using fMLLR remains common to both the GMM-HMM and the DNN-HMM-based systems.

Incremental training of DNN:
Considering the application of a DNN-HMM-based speech recognizer for unseen dysarthric speech, it is expected that there will be incremental data as the dysarthric user uses the system. This data can be used to improve upon the existing acoustic models and thereby improve the performance of the recognition engine. Two mechanisms of training the DNN-HMM were considered - (1) DNN weights built using the original corpus, are updated by retraining, using the incremental data alone. (2) The system is trained on the entire data (original + incremental).

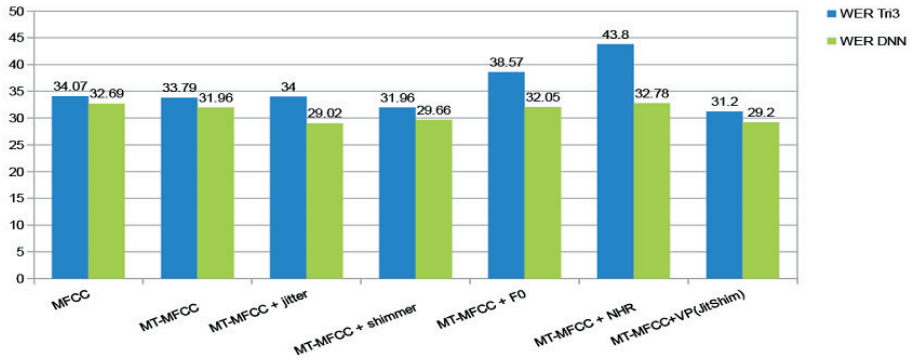

Figure 4.1: WER for GMM-HMM-based and DNN-HMM-based recognition using speaker adaptation

4.1.4 Results and Discussion

Speech recognition using the GMM-HMM system as well as the DNN-HMM system was carried out using a set of features, namely MFCC, MT-MFCC, and VP, individually as well as in fusion. Training and testing data setup was designed so as

to understand the speech recognition performance for a set of words for which no training has been done on dysarthric data such as dysarthric computer command words (DYS-CC). The train and test for all other cases such as healthy control (HC) and DYS-digits are disjoint or mutually exclusive. The word error rates (WER) for Triphone GMM-HMM and DNN-HMM systems are as shown in Figure 4.1.

It can be seen that using MT-MFCC and VP - jitter and shimmer with speaker adaptation, showed a reduction in WER for the DNN-HMM system, whereas adding $F0$ features and NHR features had an adverse impact. It has been observed that jitter and shimmer are not discernible perceptually by human listeners [113], whereas any difference in fundamental frequency $F0$ or NHR is perceptually apparent [149]. Using CMN and SAT improved speech recognition using MT-MFCC-F0 features. However, using $F0$ features in addition to MT-MFCC did not improve the overall speech recognition for any of the SA or SI systems. It was seen that features that have a clear bearing on speech perception adversely impacted the performance of the recognizers.

Fusion of MT-MFCC, jitter, and shimmer (VPJitShim) feature shows a relative improvement of 8.4% in GMM-HMM-based system and 10.7% in the DNN-HMM-based system over the MFCC features alone.

Table 4.3 shows the recognition results based on speaker type for the DNN-HMM using MT-MFCC-VPJitShim feature set. This indicates a correlation between severity of dysarthria and the accuracy of the recognition system. Similar trend was seen for MT-MFCC, MT-MFCC-Jitter and MT-MFCC-Shimmer feature-based recognition systems, wherein the WER increased with the increase in the severity of dysarthria.

Table 4.3: Dysarthria severity wise accuracy for DNN-HMM system with original training data and incremental training data for MT-MFCC-VPJitShim

Speaker type	%Accuracy-DNN-HMM **Initial**		%Accuracy-DNN-HMM **Incremental**	
	Digits	**CC words**	**Digits**	**CC words**
Healthy control	98.93	99.40	94.90	99.00
DYS – Very Low	94.66	91.47	94.70	98.66
DYS – Low	92.68	36.84	83.10	95.35
DYS – Medium	88.24	31.51	82.34	90.09
DYS – High	52.38	7.22	51.56	93.65

Experiments pertaining to incremental training were conducted for SA-based DNN-HMM recognizer, using fusion MT-MFCC and VPJitShim features. The DNN-HMM system was retrained using the initial weights from the training data mentioned in Table 4.1 and a 10% additional DYS-CC word data. This system performed poorly in comparison to the system-trained with original training, data.

This could be attributed to the updating of the neural network to a specific type of data, namely dysarthric CC word data. As expected, training the DNN-HMM system using the entire data (original data + incremental data) provided a significant improvement, especially in the recognition of DYS-CC words for dysarthric speakers, as shown in Table 4.3. Recognition of digits deteriorated for both healthy control and dysarthric data, owing to a higher number of digits being incorrectly recognized as CC words, especially the confusable pairs like the digit 'nine' and the CC word 'line'.

4.1.5 Conclusions

In this chapter, we propose a method and examine a set of features to improve speech recognition of unseen dysarthric speech. We incorporate multi-taper MFCC (MT-MFCC) and examine the applicability of voice parameters (VP) such as jitter, shimmer, F0 features and noise-to-harmonics ratio (NHR) in two types of recognition systems, namely - GMM-HMM and DNN-HMM using a speaker adaptation approach. For the MT-MFCC-VP(JitShim) fused feature set, a relative improvement of 8.4% in GMM-HMM-based system and 10.7% in the DNN-HMM-based system was seen over the MFCC features alone. This indicates that while using jitter and shimmer voice parameters was beneficial in speaker adaptation-based speech recognition, using $F0$ and NHR features added no advantage. This difference in the behavior of both recognition systems could be understood from the perspective of human listener perception of dysarthric speech. It has been observed that jitter and shimmer are not discernible perceptually by human listeners, whereas any difference in fundamental frequency $F0$ or NHR is perceptually apparent. An increment in the training data clearly increased the recognition accuracy of the DNN-HMM-based system using MT-MFCC-VPJitShim features for DYS-CC words. Our future work would involve further improving the accuracy of dysarthric speech recognition under the DNN-HMM architecture, exploring different topologies and network types that would suit the best for dysarthric speech recognition.

4.2 Improving Recognition of Dysarthric Speech Using Severity Based Tempo Adaptation

Dysarthria is a motor speech disorder, characterized by slurred or slow speech resulting in low intelligibility. Automatic recognition of dysarthric speech is beneficial to enable people with dysarthria to use speech as a mode of interaction with electronic devices. In this chapter, we propose a mechanism to adapt the tempo of the sonorant part of dysarthric speech to match that of normal speech, based on the severity of dysarthria. We show a significant improvement in recognition of tempo-adapted dysarthric speech, using a Gaussian Mixture Model (GMM) - Hidden Markov Model (HMM) recognition system, as well as a Deep neural network (DNN) - HMM-based system. All evaluations were done on the Universal Access Speech Corpus.

4.2.1 Introduction

Dysarthria is a motor speech disorder resulting from impairment in muscles responsible for speech production. Neurological injury may result in weakness, paralysis, or a lack of coordination of the motor-speech system, affecting speech subsystems, and giving rise to a reduction in intelligibility, audibility, naturalness, and efficiency of vocal communication. For dysarthric speakers, speech is a more efficient/convenient mode of communication with electronic devices as compared to keyboard input [165]. Several techniques have been proposed to improve the performance of automatic recognition of dysarthric speech, such as (1) enhancement of dysarthric speech in the acoustic domain to match that of normal speakers. (2) Automatic speech recognition (ASR) based speech recognition using speaker adaptation. Research methods to improve the intelligibility of dysarthric speech by modifying various aspects of speech, such as vowel space [82], energy, fundamental frequency, formants and tempo of dysarthric speech[167] have been proposed. In [163], the impact of manipulation of fundamental frequency on intelligibility has been studied, wherein intelligibility is reduced with reduction in variation in $F0$. Several studies have been conducted to understand the ASR performance based on severity levels. Maximum likelihood and maximum *a posteriori* (MAP) adaptation has been used for speaker adaptation in [27, 174, 102] wherein the authors analyze the performance of different types of ASR systems such as speaker-independent (SI), speaker-adapted (SA), and speaker-dependent (SD) for various severity levels. An interpolation technique along with MAP adaptation on a speaker-wise background model is used in [181] to provide improved ASR performance. In [164], the performances of speaker-dependent and speaker-adaptive models have been compared, where the speaker adaptive models performed better across various levels of severity of dysarthria.

Automatic recognition of dysarthric speech is poorer as compared to that of normal speech, owing to the inter-speaker and intra-speaker inconsistencies in the

acoustic space as well as the sparseness of data. Thus far, three popular dysarthric speech databases, namely the Universal Access (UA) speech corpus [95], Nemours [126] and TORGO [168] exist for American English. No known dysarthric speech database is available for Indian languages. The objective of our work is to build an ASR for dysarthric speakers for resource-deficient Indian languages, using zero or small amounts of dysarthric data for training the acoustic models of an automatic speech recognizer (ASR).

In this chapter, we propose a mechanism to improve the recognition of dysarthric speech using tempo adaptation of sonorants (vowels, glides, liquids, and nasals) in dysarthric speech, by using acoustic models primarily built from healthy control speakers. We show that the severity of dysarthria has a bearing on the duration of sonorants and thereby, the degree of adaptation can be selected based on the severity of dysarthria. Severity classification itself is beyond the scope of this work and can be accomplished by employing techniques available in the literature [80]. We also compare the performance of speaker-independent (SI) and speaker-adapted (SA) recognition systems when a small amount of dysarthric data is available and is used for speaker adaptation. The experimental results show that speaker-adapted dysarthric speech recognition further improved with tempo adaptation, indicating that tempo adaptation supplements the speaker-adapted dysarthric speech recognition. This improvement was seen across both Gaussian Mixture Model (GMM) - Hidden Markov Model (HMM) and Deep neural network (DNN) - HMM-based recognition system.

The rest of the chapter is organized as follows. Section 4.2.2 describes the tempo adaptation and its impact on dysarthric speech recognition, Section 4.2.3 discusses the various experimental setups and a description of the data used, in Section 4.2.4 we discuss the experimental results, and we conclude in Section 4.2.5.

4.2.2 Severity based Tempo Adaptation

Impairment of the motor nervous system impacts the articulator movements adversely, causing the articulators to move slowly. This manifests as longer durations for sonorants in dysarthric speech as compared to normal speech and tempo adaptation of the sonorants of dysarthric speech leads to improvement in the performance of ASRs [167]. Tempo adaptation involves temporal reduction of the sonorant regions of an utterance using a pre-determined adaptation parameter α.

Tempo adaptation needs to be in a manner such that it does not impact the pitch of the sonorant regions. Hence, a phase vocoder based on short-time Fourier transform (STFT) is used [155]. Magnitude spectrum and phase of the STFTs are either interpolated or decimated based on the adaptation parameter, where the magnitude spectrum is directly used from the input magnitude spectrum and phase values are chosen to ensure continuity. This ensures that the pitch of the time-warped sonorant region is intact. For the frequency band at frequency F and

frames i and $j > i$ in the modified spectrogram, the phase θ is predicted as

$$\theta'_{j^F} = \theta_{j^F} + 2\pi F \cdot (i - j)$$

The modified spectrogram is then converted into a time-domain signal using inverse Fourier transform, wherein the tempo of the sonorant regions are adapted with the pitch unchanged.

4.2.2.1 Learning the Adaptation Parameter

We propose a scheme to adapt the tempo of dysarthric automatically speech based on the severity of dysarthria. The adaptation parameter α, has been determined empirically using healthy control speech data and dysarthric speech of various severity levels. Both sets of data, healthy control and dysarthric comprise the same words. Initially, tempo adaptation is done for the sonorants at the word level, wherein the tempo of the dysarthric speech for the sonorant region in each word was adapted to match the tempo of the sonorant region in the exact same word as spoken by healthy control speakers. Consider the word W whose average sonorant duration for healthy control speakers is d_{HC} and that for a dysarthric utterance is d_{dys}. The tempo adaptation parameter for the word W is computed as

$$\alpha_{initial} = \frac{d_{HC}}{d_{dys}}$$

The sonorant region of the dysarthric utterance is adapted using α_{inital} for each dysarthric utterance. It was observed that the severity of the speakers had a clear bearing on the α_{inital} values, as shown in Figure 4.2, wherein the letters M and F in the speaker code indicate a dysarthric speaker's gender. The speaker-wise relative improvement in recognition of dysarthric speech for both GMM and DNN systems are as shown in the Figure 4.3. Also, for some speakers with high intelligibility, the word error rate (WER) increased using tempo adaptation. This factor was considered for setting the α parameter. It was also observed that the standard deviation across words was low for a particular severity class, with the highest standard deviation (0.82) being for low intelligibility.

Table 4.4: Tempo adaptation parameter α based on severity computed empirically.

Severity	Very Low	Low	Mid	High
α	1	0.6	0.5	0.4

Based on the above empirical evidence, the α parameters selected for different severity levels are as shown in Table 4.4. Figure 4.4 shows the proposed system, wherein tempo adaptation for a particular speaker is done based on the severity level. Sonorant region in a speech utterance was identified using a three-class classification technique, wherein an utterance was classified into silence, non-sonorant

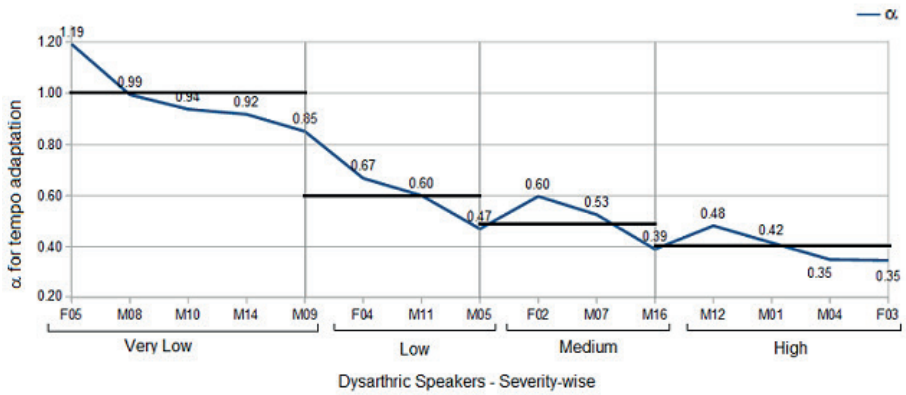

Figure 4.2: Variation in initial tempo adaptation parameter $\alpha_{initial}$ across various severity levels of dysarthria

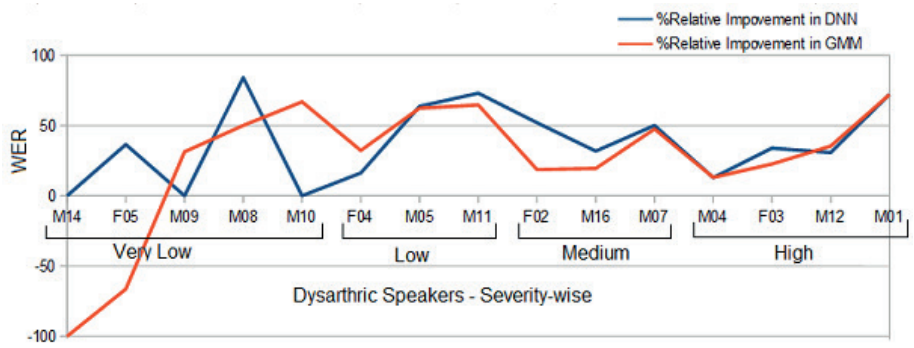

Figure 4.3: Relative improvement in WER across various severity levels of dysarthria for GMM and DNN using $\alpha_{initial}$

and sonorant regions using HTK 3.4 toolkit [216]. For this task, acoustic models corresponding to the three classes were trained using the TIMIT [50] database.

4.2.3 Experimental Setup

4.2.3.1 Data

Data from the Universal Access (UA) speech corpus [95] was used for both training and testing of the two ASR systems discussed in this section. The UA speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. The recording material consisted of 455 distinct words

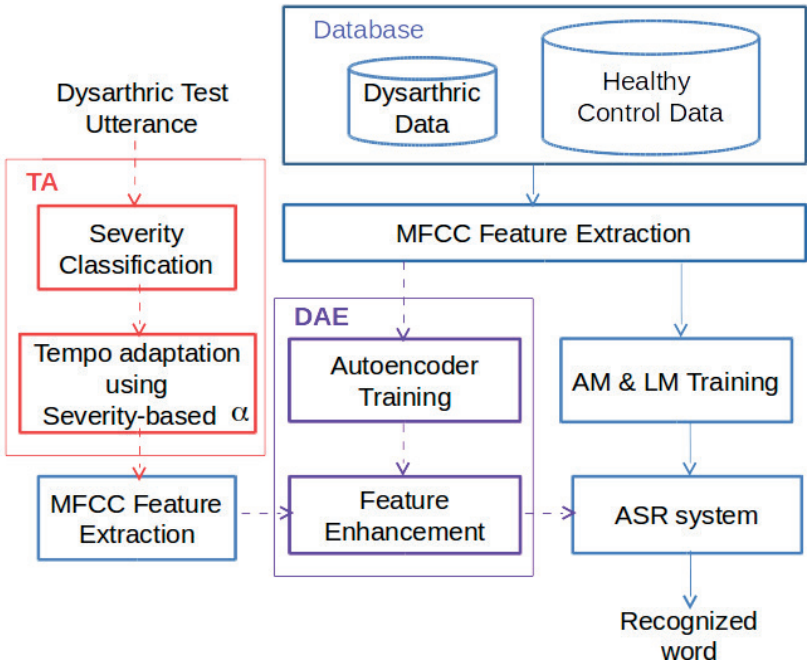

Figure 4.4: Proposed system for tempo-adapted dysarthric speech recognition

with 10 digits, 26 international radio alphabet letters, 19 computer commands, 100 common words and 300 uncommon words that were distributed into three blocks. Three blocks of data were collected for each speaker such that in each block, the speaker recorded the digits, radio alphabets, computer commands, common words and 100 of the uncommon words. Thus each speaker recorded 765 isolated words. Speech intelligibility ratings for each dysarthric speaker, as assessed by five naive listeners are also included in the corpus. Speakers were divided into four different categories based on intelligibility, namely high, mid, low, and very low. We use this information to analyze the performance of our recognition systems at the dysarthria severity level.

4.2.3.2 Speech Recognition

We use the Kaldi toolkit [156] for both GMM-HMM-based and DNN-HMM-based dysarthric speech recognition. A 3-state HMM with a monophone or a triphone context model is used. GMM-HMM system was trained using a maximum likelihood estimation (MLE) training approach along with 100 senones and 8 Gaussian mixtures. Cepstral mean and variance normalization (CMVN) was applied to each

of the above sets of features. Dimensionality reduction was done using Linear Discriminant Analysis (LDA), wherein LDA builds HMM states using feature vectors with a reduced feature space. We use the context of 6 frames (3 left and 3 right) to compute LDA. The feature vector size post LDA is set to 40.

The input layer of DNN has 360 ($40 \times nineframes$) dimensions using a left and right context of 4 frames. The output layer has a dimension of 96 (number of senones available in the data). 2 hidden layers with 512 nodes in each layer were used. The performance of each of the recognition systems is reported in terms of word error rate (WER).

We use Maximum Likelihood Linear Transform (MLLT) for speaker normalization. MLLT derives a unique transformation for each speaker using the reduced feature space from the LDA. An inter-speaker feature space normalization technique known as feature space maximum likelihood linear regression (fMLLR) [48] is performed for each speaker. Speaker adaptive training (SAT)[7] is applied at the time of training the acoustic models and aims at eliminating the inter-speaker variation. fMLLR-based SAT was applied to create speaker-adapted (SA) acoustic models; further, fMLLR was applied to the features of the input utterances at the time of decoding. SAT using fMLLR remains common to both the GMM-HMM and the DNN-HMM-based systems.

A specific combination of healthy control (HC) and dysarthric data (DYS) from each of the three blocks (B1, B2 and B3) of computer command (CC) words and digits were used for various experiments as described in Table 4.5.

Table 4.5: Training and testing corpus

System	Training	Testing	Purpose
SI-01	HC-CC	DYS-CC (B1&B3)	$\alpha_{initial}$ and α learning
SI-02	HC-CC	DYS-CC (B2)	α validation
SA-01	HC-CC, DYS-CC (B1&B3)	DYS-CC (B2)	α + Speaker adaptation
SA-02	HC-CC, HC-digits, DYS-CC (B1&B3)	DYS-digits (B2)	α validation for unseen data

The above experimental setup is used for both GMM-HMM and DNN-HMM recognizers. System SA-02 specifically shows the performance obtained for the recognition of unseen dysarthric data (which does not exist in the training set). It is expected that this would be the typical scenario, considering the challenges in collecting dysarthric data. The objective of our work is to be able to recognize dysarthric speech when no or a small amount of dysarthric data is available for training. To the best of our knowledge, no other work has reported speech recognition for this specific combination of testing and training data.

4.2.4 Evaluation Results and Discussion

The tempo adaptation parameter α was learned for each severity level, as described in Section 4.2.2. Experiments were conducted to understand the applicability of α

Table 4.6: Relative improvement in WER using tempo-adaptation

System	GMM-HMM %WER	GMM-HMM-TA %WER	%relative improvement	DNN-HMM %WER	DNN-HMM-TA %WER	%relative improvement
SI-01	77.21	40.84	47.11	75.88	39.12	48.44
SI-02	79.36	48.05	39.45	73.42	44.47	39.43
SA-01	49.59	34.12	31.2	34.67	27.57	20.48
SA-02	72.96	52.22	28.42	52.01	42.83	17.65

under various scenarios such as SI, SA, and SA with unseen data. The results indicate that the recognition accuracy improved or the WER reduced when the tempo was adapted. Acoustic models were trained using both monophone and triphone contexts. It was observed that across all experimental setups, triphone models showed higher relative improvement in recognition performance after tempo adaptation. This indicates that the tempo adaptation improves the triphone acoustic model of a phone as well.

Table 4.7: Impact of tempo adaptation on WER for SA-01 based on severity

Severity	GMM-HMM %WER	GMM-HMM-TA %WER	%relative improvement	DNN-HMM %WER	DNN-HMM-TA %WER	%relative improvement
Low	39.68	22.83	42.46	26.32	14.18	46.12
Medium	60.39	45.87	24.04	42.69	31.43	26.38
High	108.86	69.15	36.48	86.52	68.53	20.79

Further, it can be seen from Table 4.7, that the recognition performance of the best-performing system SA-01 improved across all severity levels for both GMM-based and DNN-based systems with tempo adaptation (TA). It was observed that the reduction in WER was largely due to the decrease in the number of insertions as compared to substitutions.

4.2.5 Conclusion

In this chapter, we propose a mechanism to improve the speech recognition of dysarthric speech using tempo adaptation of sonorants in dysarthric speech. We show that the severity of dysarthria has a bearing on the duration of sonorants and thereby degree of adaptation can be selected based on the severity of dysarthria. This mechanism is especially beneficial when no or less amount of dysarthric data is available in a specific language (e.g., Indian Languages), for training the acoustic models of an ASR. We compare the performance of speaker-independent (SI) and Speaker-adapted (SA) recognition systems when a small amount of dysarthric data is available and is used for speaker adaptation. The results show that speaker-adapted dysarthric speech recognition further improved with tempo adaptation, indicating that tempo adaptation supplements the speaker-adapted dysarthric speech

recognition. This improvement was seen across both Gaussian Mixture Model (GMM) - Hidden Markov Model (HMM) and Deep neural network (DNN) - HMM-based recognition system. If we consider the system wherein only healthy controls are used for training the acoustic models with no tempo adaptation as a baseline, the proposed speaker-independent and speaker-adapted systems provide an improvement of 47.11% and 55.81% respectively, for GMM-HMM-TA and 48.44% and 63.67% for DNN-HMM-TA respectively. Severity-based tempo adaptation using triphone-based acoustic models showed higher relative improvements than monophone acoustic models across all systems mentioned in the Section 4.2.3. This indicates that the tempo adaptation improves the acoustic phone model in the triphone context as well.

Chapter 5

Autoencoder-based speech enhancement of dysarthric speech

This chapter is based on the following publications:

I. Vachhani, B., Bhat, C., Das, B., & Kopparapu, S. K. (2017, August). Deep Autoencoder Based Speech Features for Improved Dysarthric Speech Recognition. In Interspeech (pp. 1854-1858).

II. Bhat, C., Das, B., Vachhani, B., & Kopparapu, S. K. (2018, September). Dysarthric Speech Recognition Using Time-delay Neural Network Based Denoising Autoencoder. In Interspeech (pp. 451-455).

5.1 Deep Autoencoder Based Speech Features for Improved Dysarthric Speech Recognition

Dysarthria is a motor speech disorder, resulting in mumbled, slurred or slow speech that is generally difficult to understand by both humans and machines. Traditional Automatic Speech Recognizers (ASR) perform poorly on dysarthric speech recognition tasks. In this chapter, we propose the use of deep autoencoders to enhance the Mel Frequency Cepstral Coefficients (MFCC) based features in order to improve dysarthric speech recognition. Speech from healthy control speakers is used to train an autoencoder which is, in turn, used to obtain improved feature representation for dysarthric speech. Additionally, we analyze the use of severity-based tempo adaptation followed by autoencoder-based speech feature enhancement. All evaluations were carried out on the Universal Access dysarthric speech corpus. An overall absolute improvement of 16% was achieved using tempo adaptation followed by autoencoder-based speech front-end representation for DNN-HMM-based dysarthric speech recognition.

5.1.1 Introduction

Neurological injury or disease such as Amyotrophic lateral sclerosis (ALS), Parkinson's disease (PD) or cerebral palsy resulting in weakness, paralysis, or a lack of coordination of the motor-speech system manifests as a speech disorder known as dysarthria. Dysarthria leads to a reduction in intelligibility, audibility, naturalness, and efficiency of vocal communication. Owing to the motor impairment, interaction with electronic devices using speech is more effective than through keyboard input [165]. Inter-speaker and intra-speaker inconsistencies in the acoustic space, as well as the sparseness of data poses a serious challenge in building automatic speech recognition engine (ASR) system for dysarthric speech. Speaker adaptation-based ASR systems and dysarthric speech enhancement to match the characteristics of normal speech are two popular techniques that have been employed to address this challenge.

In [26], a similarity measure between dysarthric speakers to select relevant speaker data for training rather than speaker-independent acoustic models, followed by maximum a posteriori (MAP) adaptation has been used. In [181] a more suitable prior model for adaptation based on the dysarthric speaker's acoustic characteristics has been used to achieve improved recognition. ASR accuracy was shown to improve by representing dysarthric speech in terms of articulatory models in [165, 166, 58]. In [177], a set of MFCC features that best represent dysarthric acoustic features was selected to be used in Artificial Neural Network (ANN)-based ASR. A hybrid adaptation using maximum likelihood linear regression (MLLR) and MAP [174] have been used to improve dysarthric speech recognition. Voice parameters such as jitter and shimmer features along with a multi-taper spectral estimation being used along with feature space maximum likelihood linear regression (fMLLR) transfor-

mation and speaker adaptation to obtain improved dysarthric speech recognition [13].

In [66], the modifications to prosody, spectral content, regions of the signal containing formants, and effects of signal processing on dysarthric speech have been studied. Transformations of dysarthric speech in both the temporal as well as spectral domain have been employed so as to match the characteristics of normal speech. In another study [167], transformations in the temporal domain by adjusting the tempo of speech using phase vocoding, spectral domain transformation using anchor-based morphing of the spectrum and phoneme level correction of pronunciation were used to give improved intelligibility and were validated both by human listeners and ASR-based recognition. In [82], vowel space transformations by manipulating vowel duration and formants $F1$ - $F3$ stable points were shown to improve the intelligibility of dysarthric speech. In their work [34], authors use speech synthesis to produce utterances with improved intelligibility corresponding to a dysarthric utterance using the dysarthric speaker characteristics. Yet another aspect that has been used to improve ASR performance is based on the severity of the Dysarthria. Traditionally, speech intelligibility has been an indicator of the severity of the speech disorder [120]. An understanding of severity has contributed to improved speech recognition of dysarthric speech as seen in [174, 132, 102].

Deep Autoencoder (DAE) based feature enhancement technique provides significant performance gain for speech recognition. A variant of basic DAE, deep denoising autoencoders (DDA), have been used to enhance speech features, especially in noisy conditions [183, 119, 43]. DDAs are also efficiently used for reverberant speech recognition [71]. In [43], a DDA is pre-trained as restricted Boltzmann machines (RBMs) and then a nonlinear mapping from noisy to clean features is learned from the corresponding clean speech features. Generally, a DDA learns a stochastic mapping from noisy to clean by using clean features for fine-tuning.

In this chapter, we train the deep autoencoder network using healthy control speech, which is in turn used to enhance the speech features of dysarthric speech. We propose a method to improve the recognition of dysarthric speech using enhanced speech features that have been extracted using a Deep Autoencoder (DAE). Additionally, we extend our earlier work [14], wherein we transform the dysarthric speech in the temporal domain using severity-based tempo adaptation (TA) and use the tempo-adapted dysarthric speech prior to feature enhancement using a DAE. We analyze the contribution of the individual techniques towards improvement in speech recognition, as well as tempo adaptation and DAE-based feature enhancement in tandem. To the best of our knowledge, autoencoder-based speech feature enhancement for dysarthric speech has not been attempted so far and is the main contribution of this chapter.

The rest of the chapter is organized as follows. Section 5.1.2 describes the methodology employed to enhance speech features for dysarthric speech recognition, Section 5.1.3 discusses the various experimental setups and a description of the data used, Section 5.1.4 describes the results and analysis, and we conclude in Section

5.1.5.

5.1.2 Speech Feature Enhancement

In this chapter, we propose (a) an improved front-end speech processing through enhanced speech features using deep autoencoders (DAE) and (b) a combination of dysarthric speech transformation in the temporal domain followed by feature enhancement using DAE. Figure 5.1 shows an overview of the proposed setup for improved dysarthric speech recognition.

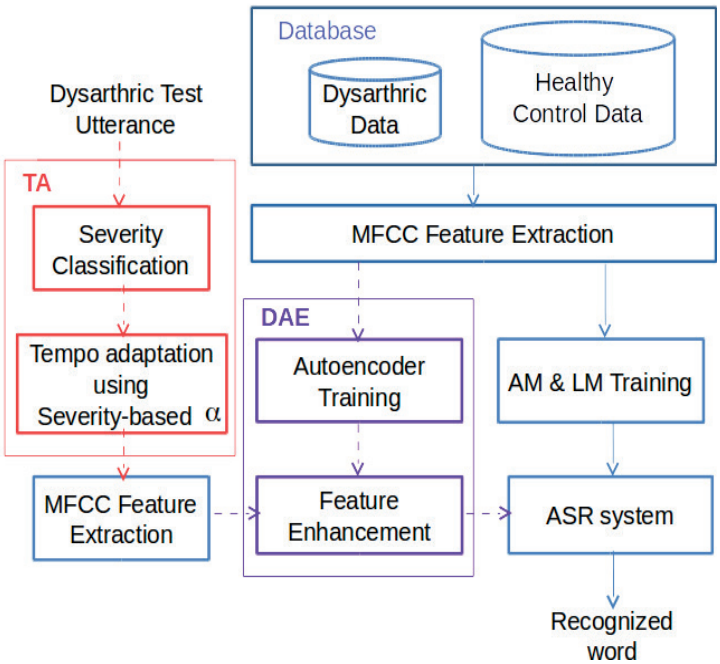

Figure 5.1: Proposed setup for improved dysarthric speech recognition

5.1.2.1 Deep Autoencoder (DAE)

Traditionally, an autoencoder is a fully connected artificial neural network system with a bottleneck layer as shown in Figure 5.2. In this chapter, we use deep autoencoder to enhance the Mel Frequency Cepstral Coefficients (MFCC) based features of dysarthric speech.

An autoencoder comprises two blocks: the encoder and the decoder. The objective of the encoder is to transform a higher dimensional input feature vector into a

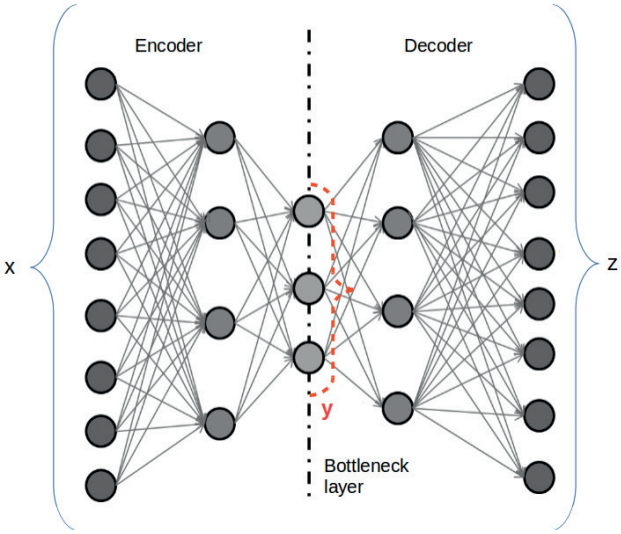

Figure 5.2: Deep Autoencoder (DAE)

lower dimensional representation at the bottleneck layer. The bottleneck features are then transformed into higher dimensional representation at the decoder end of the autoencoder, the input and output features drive the learning of the autoencoder to ensure that the bottleneck layer presents a lower dimensional representation of the input features. Encoding operation can be represented as

$$\vec{y} = f(\theta; \vec{x}) = s(W\vec{x} + \vec{b}) \tag{5.1}$$

where

- $\vec{y}$ is the bottleneck feature vector representation of the input feature vector x, which propagates through hidden layers.
- $\theta = \{W, \vec{b}\}$, where W and $\vec{b}$ are the weights and biases of the network, respectively.
- s is an activation function, linear or non-linear.

At the decoder, the bottleneck feature vector $\vec{y}$ which propagates through hidden layers is mapped to the higher dimensional representation $\vec{z}$ at the output stage as

$$z = g(\theta'; \vec{y}) = s(W'\vec{y} + \vec{b'}) \quad \text{where} \quad \theta' = \{W', \vec{b'}\} \tag{5.2}$$

Thus, the output of the DAE can be represented as a function of the weights and biases of the encoder and decoder stages, namely $\{\theta, \theta'\}$ and written as $\vec{z} = g(\theta'; (f(\theta; \vec{x}))$. DAE parameters θ and θ' are optimized such that $\vec{z}$ is as close as

possible to input/target $\vec{x}$ and maximizes $P(\vec{x}|\vec{z})$. The autoencoder parameters are optimized using mean square error (MSE) back-propagation between target $\vec{x}$ and network output $\vec{z}$.

5.1.2.2 Unsupervised Feature Extraction using Modified DAE

Unsupervised feature learning is currently being used as an alternative to the conventional MFCC features. In this chapter, we modify the DAE architecture to suit the purpose of enhancing dysarthric speech features as shown in Figure 5.3. The DAE parameters ($\theta 1$, $\theta 2$ and θ') are learned from healthy control speech. We have used MFCC features from healthy control speech as input and target, as shown in Figure 5.3(a). Learned parameters ($\theta 1$, $\theta 2$ and θ') represent the weights and biases of the DAE, which provides the minimum MSE between the input and target at the time of autoencoder training. $\theta 1$, $\theta 2$ are encoder parameters and θ' is the decoder parameter of the network where $\theta 1 = \{W1, b1\}$, $\theta 2 = \{W2, b2\}$ and $\theta' = \{W', b'\}$. We extract enhanced features from dysarthric speech using the trained autoencoder parameters. We use these enhanced features as input to the decoding process.

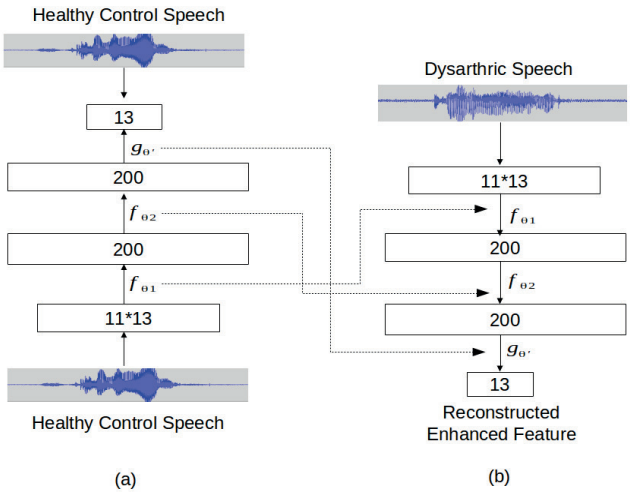

Figure 5.3: (a) Modified DAE architecture for Training (learning parameter $\theta 1$, $\theta 2$and θ') (b) Feature extraction for dysarthric speech using learned DAE parameters

5.1.2.3 Severity-based Tempo Adaptation (TA)

We examined the improvement in dysarthric speech recognition using severity-based tempo adaptation in one of our earlier works [14]. Dysarthric speech severity level classification was carried out using techniques mentioned in [15]. Malfunctioning of the motor nervous system impacts the precision and flexibility of the vocal folds, articulators, and other speech subsystems, leading to reduced prosodic control. This manifests as the longer duration for sonorants in dysarthric speech as compared to normal healthy speech [167]. Temporal reduction of sonorant regions emerges as a possible enhancement to dysarthric speech to provide improved intelligibility, both to human listeners as well as the ASR systems. This process is referred to as tempo adaptation. Tempo adaptation based on the knowledge of the severity of the dysarthric speech was found to be beneficial since the adaptation parameter α could be learned for a specific severity level, empirically using healthy control speech data and dysarthric speech of various severity levels, where exactly the same words are spoken by both healthy control speakers and dysarthric speakers. Consider a spoken word whose average sonorant duration for healthy control speakers is d_{hc} and that for a dysarthric utterance is d_{dys}. The tempo adaptation parameter for each word is computed as

$$\alpha = \frac{d_{hc}}{d_{dys}} \tag{5.3}$$

An average tempo adaptation parameter was computed for each speaker and it was found that tempo adaptation can be carried out by selecting an α value that would suit all the speakers at a certain severity level. Tempo adaptation needs to be carried out in a manner so as not to affect the pitch of the sonorant regions of dysarthric speech. A phase vocoder based on short-time Fourier transform (STFT) is used [155].

Let $X(F)$ be the Fourier transform of a speech signal $x(t)$, $x(t) \overset{\mathscr{F}}{\longleftrightarrow} |X(F)| \cdot \Theta$, where $|X(F)|$ is the magnitude and $\Theta = \angle X(F)$ is the phase. Magnitude spectrum and phase of the STFTs are either interpolated or decimated based on the adaptation parameter (α), where the magnitude spectrum is directly used from the input magnitude spectrum and phase values are chosen to ensure continuity. This ensures that the pitch of the time-warped sonorant region is intact. For the frequency band at frequency f and frames i and $j > i$ in the modified spectrogram, the phase Θ is predicted as

$$\Theta'_{jf} = \Theta^f_j + 2\pi f \cdot (i - j) \tag{5.4}$$

If the modified magnitude and phase spectrum are represented as $|X'(F)|$ and $\angle\Theta'$, the spectrogram is then converted into a time-domain signal using inverse Fourier transform, wherein the tempo of the sonorant regions are adapted with the pitch unchanged as $|X'(F)| \cdot \Theta' \overset{\mathscr{F}^{-1}}{\longleftrightarrow} x'(t)$

Additionally, we explore the possibility of using severity-based tempo adaptation in tandem with DAE-based feature enhancement as shown in Figure 5.1.

5.1.3 Experimental Setup

Data from the Universal Access (UA) speech corpus [95] was used for both training and testing. The UA dysarthric speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. Three blocks of data were collected for each speaker such that in each block, a speaker recorded 10 digits, 26 international radio alphabets, 19 computer commands, 100 common words and 100 uncommon words such that each speaker recorded 455 distinct words and a total of 765 isolated words. Speech intelligibility ratings for each dysarthric speaker, as assessed by five naive listeners are also included in the corpus. Based on this evaluation, speakers were divided into four different categories. We have used this information to analyze the performance of our recognition systems at different dysarthria severity levels.

Tempo adaptation parameters as shown in Table 5.1, were empirically determined for different severity levels in the UA speech corpus as described in [14].

Table 5.1: Tempo adaptation parameter α based on severity

Severity	Very Low	Low	Mid	High
α	1.0	0.6	0.5	0.4

We use the Kaldi [156] toolkit-based deep autoencoder for our experiments. The architecture of deep autoencoder (DAE) was 143-200- 200-13, with 143 nodes in the input layer, where 13-dimensional MFCC with a splicing of 11 contextual frames, 200 neurons in each hidden layer, and 13 nodes in the output layer. All neurons had sigmoid activation in all the layers. To demonstrate the ability of the autoencoder to capture general spectral information, the autoencoder was trained using training data as mentioned in Table 5.2 for each of the four configurations.

5.1.3.1 Speech Recognition

The Kaldi toolkit [156] was used for DNN-HMM-based dysarthric speech recognition. The system was trained using a maximum likelihood estimation (MLE) training approach along with 100 senones and 8 Gaussian mixtures. Cepstral mean and variance normalization (CMVN) was applied to each of the above sets of features. Dimensionality reduction was done using Linear Discriminant Analysis (LDA), wherein LDA builds HMM states using feature vectors with a reduced feature space. We use the context of 6 frames (3 left and 3 right) to compute LDA. The feature vector size post LDA is set to 40.

The input layer of DNN has 360 (40×9 frames) dimensions using a left and right context of 4 frames. The output layer has a dimension of 96 (number of senones available in the data). Two hidden layers with 512 nodes in each layer were used. Dysarthric speech recognition was carried out using a constrained language model (LM), wherein we restricted the recognizer to give one word as output per

utterance. The performance of each of the recognition systems is reported in terms of word error rate (WER).

A specific combination of healthy control (HC) and dysarthric data (DYS) from each of the three blocks (B1, B2 and B3) of computer command (CC) words, were used for various experiments as described in Table 5.2 to prove the feasibility of the proposed method.

Table 5.2: Training and testing setup

System	Training	Testing
S-1	HC-CC (B1, B3)	HC-CC (B2)
S-2	HC-CC (B1, B3)	DYS-CC (B2)
S-3	DYS-CC(B1, B3)	DYS-CC (B2)
S-4	HC-CC(B1, B3) + DYS-CC(B1, B3)	DYS-CC (B2)

5.1.4 Experimental Results

We examine the effectiveness of two types of enhancements to dysarthric speech for automatic speech recognition purposes, namely (1) Tempo adaptation carried out in the temporal domain (2) DAE-based MFCC feature enhancement. DNN-HMM-based speech recognition was carried out for both the above scenarios individually and in tandem. The DAE and DNN-HMM systems were configured and trained as described in Section 5.1.3. ASR performance is reported in terms of word error rates (WERs). The following four different front-end scenarios were considered for our experiments :

- MFCC features
- Tempo adaptation followed by MFCC feature extraction.
- DAE enhanced MFCC features.
- Tempo adaptation followed by DAE enhanced MFCC features.

WERs for each configuration in Table 5.2 for the relevant front-end scenarios described above can be seen in Table 5.3. The purpose of S-1 is to examine the impact of DAE on clean or healthy control speech. The WERs for MFCC and MFCC-DAE indicate that DAE-based speech feature enhancement has improved the recognition performance even for healthy-control or clean speech. Significant improvements were seen for all four configurations over the baseline MFCC-based ASR system when enhancements were applied. Although the tandem system showed significant improvement over the baseline (of the order of 16% for S-2) for all configurations, for S-4 the MFCC-DAE seemed to perform the best. When additional dysarthric data was included in the S-2 configuration for training the DAE and DNN-HMM

systems, the performance (of S-4) significantly improved across all front-end scenarios. However, the individual front-ends performed on par or slightly better than the tandem front-end. In order to understand this better, we analyze the performances of S-2 and S-4 by looking at the performances of individual and tandem scenarios at dysarthria severity levels as shown in Table 5.4.

Table 5.3: WER for different Experimental setups

System	**MFCC** (Baseline)	**TA-MFCC**	**MFCC-DAE**	**TA-MFCC + DAE**
S-1	2.26	-	0.00	-
S-2	46.89	44.25	34.51	30.71
S-3	32.80	-	27.85	-
S-4	31.59	21.30	20.14	20.69

Table 5.4: WER analysis at severity level

System	**Severity**	**MFCC** (Baseline)	**TA-MFCC**	**MFCC-DAE**	**TA-MFCC + DAE**
S-2	Very-low	14.59	-	2.86	-
	Low	43.79	39.27	14.41	15.54
	Mid	67.63	60.53	60.00	48.16
	High	82.06	80.38	78.71	71.29
S-4	Very-low	12.93	-	1.65	-
	Low	22.60	16.95	13.56	17.23
	Mid	34.47	15.79	14.47	15.79
	High	66.27	61.24	60.29	58.61

The tempo adaptation parameter used for very low severity was 1, indicating no adaptation is performed on this set of dysarthric speech. Hence we only report the MFCC-DAE performance. The ASR performance across all front-end scenarios reduces with the increase in severity. In the majority of the cases, MFCC-DAE provided the best performance or the least WER. The addition of dysarthric speech to the training data has given tremendous improvement in the overall performance of S-2 configuration. However, the majority of the contribution to this spike in performance comes from the performance improvement for mid and high-severity dysarthric speech. Based on the severity level assessment, the tandem system performs best for mid and high-severity dysarthric speech while MFCC-DAE gives significant performance gains in cases of very low and low-severity dysarthric speech. Several iterations with various combinations of data need to be conducted to arrive at an exact recommendation regarding the choice of the front end. However, the tandem system (TA-MFCC+DAE) performed the best or on par with MFCC-DAE in most cases.

5.1.5 Conclusions

The objective of this chapter was to improve dysarthric speech recognition by enhancing the MFCC-based speech front end. We used deep autoencoders to enhance the Mel Frequency Cepstral Coefficients (MFCC) based features in order to improve dysarthric speech recognition. Additionally, we analyzed the use of severity-based tempo adaptation followed by autoencoder-based speech feature enhancement. Tempo adaptation was done in the temporal domain using a severity-based parameter to match the dysarthric speech to healthy-control speech. The performance of a DNN-HMM speech recognizer for both the enhancement techniques individually as well as in tandem was analyzed. It was observed that each technique provided significant improvement over the baseline recognition. All evaluations were carried out on the Universal Access dysarthric speech corpus. An overall absolute improvement of 16% was achieved using tempo adaptation followed by autoencoder-based speech front-end representation. Further, severity level analysis of the dysarthric recognition provided insights into the choice of front-end for each severity level, wherein the tandem system (TA-MFCC+DAE) performed exceptionally well for mid and high-severity levels of dysarthria. Future work could entail optimizations of the DAE network to further improve dysarthric speech recognition.

5.2 Dysarthric Speech Recognition using Time-delay Neural Network based Denoising Autoencoder

Dysarthria is a manifestation of the disruption in the neuro-muscular physiology resulting in uneven, slow, slurred, harsh, or quiet speech. Dysarthric speech poses serious challenges to automatic speech recognition, considering this speech is difficult to decipher for both humans and machines. The objective of this work is to enhance dysarthric speech features to match that of healthy control speech. We use a Time-Delay Neural Network based Denoising Autoencoder (TDNN-DAE) to enhance the dysarthric speech features. The dysarthric speech thus enhanced is recognized using a DNN-HMM-based Automatic Speech Recognition (ASR) engine. This methodology was evaluated for speaker-independent (SI) and speaker-adapted (SA) systems. Absolute improvements of 13% and 3% were observed in the ASR performance for SI and SA systems, respectively, as compared with unenhanced dysarthric speech recognition.

5.2.1 Introduction

The speech production process comprises acoustic and linguistic events that occur through the coordination of muscle groups and the neurological programming of muscle activities to ensure fluent and accurate articulation. Acquired or developmental dysarthria results from the impairment of the motor execution function and affects the speech intelligibility of a person. Voice input-based interactions with smart devices perform poorly for dysarthric speech. Research into automatic recognition of dysarthric speech has garnered much interest due to the rising popularity and possibility of voice inputs, especially since speech-based interaction is easier for persons with neuro-motor disorders as compared to keypad inputs [165].

Several techniques are employed to improve ASR performance for dysarthric speech: acoustic space enhancement, feature engineering, Deep Neural Networks (DNN), speaker adaptation, and lexical model adaptation- individually or as a combination thereof. Formant re-synthesis preceded by modifications of formant trajectories and energy for dysarthric speech, vowels showed significant improvement in perceptual evaluation of intelligibility of CVC utterances [82]. Acoustic space modification carried out through temporal and frequency morphing improved automatic dysarthric speech recognition, as well as subjective evaluation in [167]. It can be seen that temporal adaptation based on dysarthria severity level improved the ASR performance for dysarthric speech recognition at each severity level, [14]. A Convolutive Bottleneck Network (CBN) was used for dysarthric speech feature extraction wherein the pooling operations of the CBN resulted in features that were more robust toward the small local fluctuations in dysarthric speech and outperformed the traditional MFCC feature-based recognition [135]. A comparative study

of several types of ASR systems, including maximum likelihood and maximum a posteriori (MAP) adaptation showed a significant improvement in dysarthric speech recognition when speaker adaptation using MAP adaptation was applied [27]. The word error rate for dysarthric speech was reduced using voice parameters such as jitter and shimmer along with multi-taper Mel-frequency Cepstral Coefficients (MFCC) followed by speaker adaptation [13], and using Elman back-propagation network (EBN) which is a recurrent, self-supervised neural network along with glottal features and MFCC in [175]. A multi-stage deep neural network (DNN) training scheme is used to better model dysarthric speech, wherein only a small amount of in-domain training data showed considerable improvement in the recognition of dysarthric speech [219]. In [193], authors propose a DNN-based interpretable model for objective assessment of dysarthric speech that provides users with an estimate of the severity as well as a set of explanatory features. Speaker selection and speaker adaptation techniques have been employed to improve ASR performance for dysarthric speech in [26, 181]. ASR configurations have been designed and optimized using dysarthria severity level cues in [174, 132, 102].

It has been observed that the subjective perception-based intelligibility performance for noisy and dysarthric speech is correlated, indicating that there exists similarity in the information processing of these two types of speech [214]. Extrapolating this to the objective assessment domain, we hypothesize that techniques used for noisy speech may support dysarthric speech processing as well. In this chapter, we explore the possibility of using a Time-Delay Neural Network Denoising Autoencoder (DAE) for dysarthric speech feature enhancement. DAEs have been used to enhance speech features, especially in noisy conditions [183, 119, 43]. The objective is for the network to learn a mapping between dysarthric speech features and healthy control speech features. This network is then used to enhance the dysarthric speech features that are used in a DNN-HMM-based ASR for improved dysarthric speech recognition. ASR performance indicates that the enhanced dysarthric speech features are closer to healthy control speech features rather than dysarthric speech features. Evaluation of our work is carried out on the Universal Access Dysarthric Speech corpus [95]. In our earlier work [195], we used a Deep Autoencoder to enhance dysarthric test speech features, wherein the DAE was trained using only healthy control speech. This is different from our current work in the DAE configuration and the training protocol followed.

The rest of the chapter is organized as follows. Section 5.2.2 describes the methodology employed to enhance speech features for dysarthric speech recognition, Section 5.2.3 discusses the experimental setup. In Section 5.2.4 we discuss the results of our experiments, we conclude in Section 5.2.5.

5.2.2 Dysarthric Speech Feature Enhancement

The process and techniques used to enhance dysarthric speech features is described in this Section.

5.2.2.1 Time-Delay Neural Network

TDNN architecture is capable of representing relationships between events in time using a feature space representation of these events [199]. Computation of the relationship between current and past inputs is made possible by introducing delays to the basic units of a traditional neural network as shown in Figure 5.4.

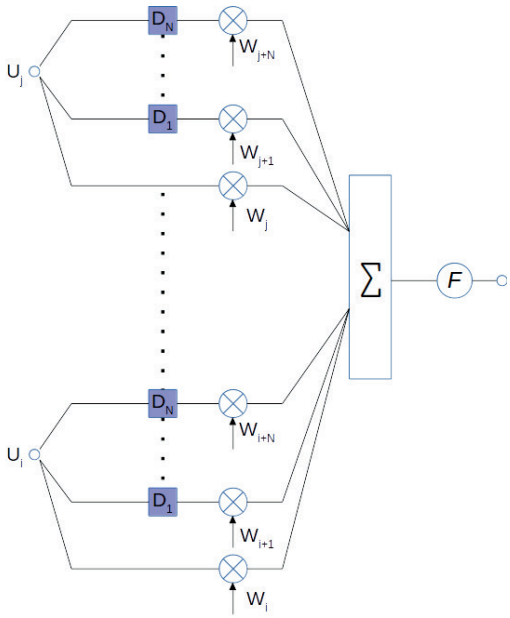

Figure 5.4: Time delay neural network unit [199]

The discovery of the acoustic features and the temporal relationship between them independent of the position of time ensures that the dysarthric speech features are not blurred by the inherent small local fluctuations. Shorter temporal contexts are used to learn the initial transforms, whereas the hidden activations from longer contexts are used to train the deeper layers. This enables the higher layers to learn longer temporal relationships [151].

Back-propagation learning is used to train TDNN-DAE, wherein the input features are extracted from noisy speech, and the target features are extracted from the corresponding clean speech.

5.2.2.2 Methodology

In traditional DAE training, the number of frames in the input utterance must necessarily be equal to the number of frames in the target utterance. This works well for scenarios wherein noise-added clean speech is the input and the corresponding

clean speech is the target. In this work, we intend to use dysarthric speech as input and its healthy control counterpart as the target speech since the objective is for the TDNN-DAE network to learn the mapping between the two. Typically, dysarthric speech is slower than healthy control speech and, hence, has a longer duration. One mechanism to match the number of frames is by using varying frame lengths and frameshifts for dysarthric utterances to match the number of frames in the corresponding healthy control utterance. However, the difference in the durations between dysarthric utterances and healthy control utterances was too high to achieve meaningful frame lengths and frameshifts.

Matching of the number of frames was done using the following two steps as depicted in Figure 5.5.

- Majority of the silence portion at the beginning and end of both dysarthric and healthy control utterances were eliminated retaining roughly 200 ms of silence.

- In order to match the durations of the input dysarthric utterance and target healthy control utterance, the dysarthric utterance was temporally adapted using phase vocoder as described in [167]. Tempo adaptation is carried out according to the adaptation parameter α given as $\alpha = \frac{d_H}{d_D}$ where d_D is the duration of the dysarthric utterance and d_H is the duration of healthy control utterance. Tempo adaptation using phase vocoder based on short-time Fourier transform (STFT) ensures that the pitch of the sonorant regions of dysarthric speech is unaffected [155]. The magnitude spectrum and phase of the STFT are either interpolated or decimated based on the adaptation parameter (α), where the magnitude spectrum is directly used from the input magnitude spectrum, and phase values are chosen to ensure continuity. This ensures that the pitch of the time-warped sonorant region is intact. For the frequency band at frequency f and frames i and $j > i$ in the modified spectrogram, the phase Θ is predicted as

$$\Theta_j^f = \Theta_i^f + 2\pi f \cdot (i - j) \tag{5.5}$$

  The modified magnitude and phase spectrum are then converted into a time-domain signal using inverse Fourier transform.

Figure 5.6 shows the proposed methodology for a TDNN-DAE-based dysarthric speech feature enhancement and recognition.

5.2.3 Experimental Setup

TDNN-DAE as well as DNN-HMM based ASR were implemented using the Kaldi speech recognition toolkit [156].

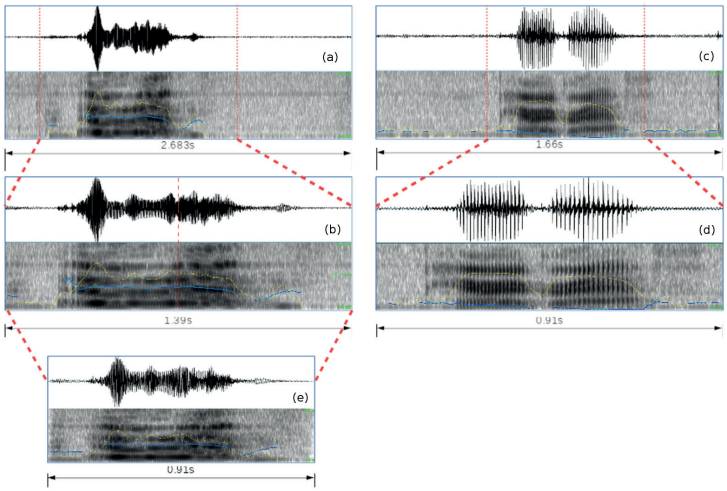

Figure 5.5: Data Preparation for TDNN-DAE training for the word 'Paragraph'- (a) Original dysarthric utterance (2.68s) (b) Dysarthric utterance after endpoint silence removal (1.39s) (c) Original healthy control utterance of duration (1.66s) (d) Healthy Control utterance after endpoint silence removal (0.91s) (e) Dysarthric utterance after tempo adaptation (0.91s) to match (d)

5.2.3.1 Dysarthric Speech Corpus

Data from the Universal Access (UA) speech corpus [95] was used for training the TDNN-DAE and DNN-HMM-based ASR systems. The UA dysarthric speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. Data was collected in three separate sessions for each speaker and categorized into three blocks B1, B2, and B3. In each block, a speaker recorded 10 digits, 26 international radio alphabet letters, 19 computer commands, 100 common words and 100 uncommon words such that each speaker recorded 455 distinct words and a total of 765 isolated words. The corpus also includes speech intelligibility rating for each dysarthric speaker, as assessed by five naive listeners.

5.2.3.2 TDNN-DAE

23-dimensional Mel-frequency cepstral coefficients (MFCC) were used as input features for all the experiments. TDNN-DAE architecture described in [151] was followed. Contexts for the DAE network with 4 hidden layers are organized as (-2,-1,0,1,2) (-1,2) (-3,3) (-7,2) (0) which is asymmetric in nature. Input temporal context for the network is set to [-13,9]. It can be observed that a narrow context is selected for the initial hidden layers, whereas higher contexts are for deeper layers.

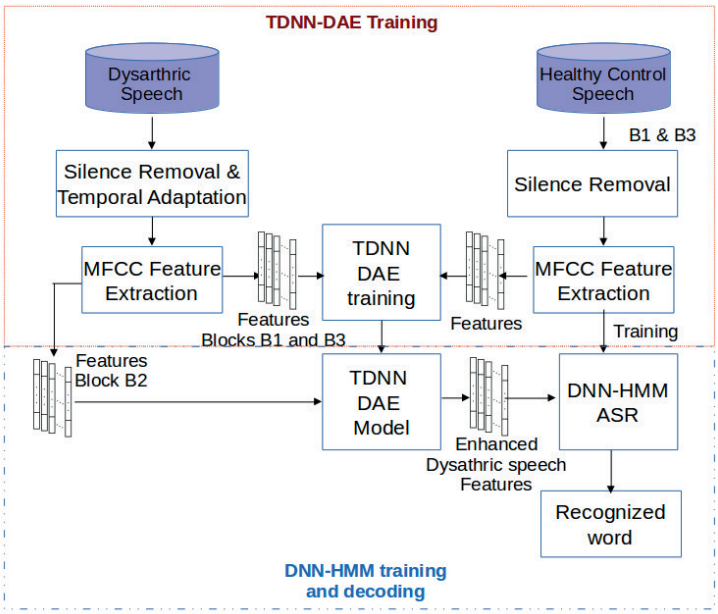

Figure 5.6: TDNN-DAE based dysarthric speech feature enhancement and recognition

Each hidden layer comprises 1024 ReLU activation nodes. TDNN-DAE was trained using training data described in Section 5.2.3.1.

Training data

In this work, we use 19 computer command (CC) words from blocks B1 and B3 of dysarthric speech and healthy control speech for TDNN-DAE training. Each dysarthric utterance was temporally adapted with each of its corresponding healthy control utterances. For example the dysarthric utterance F05_B1_C12_M2.wav (spoken by speaker F05 recorded as block B1 on channel M2) corresponding to CC word *C12:Sentence* was temporally adapted to match the duration of each of the healthy control utterances corresponding to the CC word *C12:Sentence*. Thus generating multiple dysarthric utterances from one single dysarthric utterance, as shown in the equation given below.

$$D_{u_ij} = f(D_{u_ij}, \forall_{H_{u_i}} H_{u_i}) \tag{5.6}$$

where $u_i \rightarrow$ CC utterances with $i = 1 \cdots 19$
$D_{u_ij} \rightarrow$ dysarthric utterance where $j = 1 \cdots 3511$
$H_{u_i} \rightarrow$ healthy control CC utterances with $i = 1 \cdots 19$
$f \rightarrow$ temporal adaptation(TA) function [14]

A total of 3511 dysarthric utterances were temporally adapted against their

healthy control counterparts generating around 0.6 million temporally adapted dysarthric utterances. The TDNN-DAE was trained using the temporally adapted dysarthric speech utterances as input speech while their corresponding healthy control utterances comprised the target speech.

Testing data
TDNN-DAE trained as above was used to enhance the dysarthric speech features corresponding to 1791 utterances `i.e.` computer command words from block B2. These utterances were first temporally adapted, followed by enhancement of the corresponding MFCC features using TDNN-DAE. These enhanced speech features for dysarthric speech were used to evaluate ASR recognition performance.

5.2.3.3 DNN-HMM based ASR

The dysarthric speech was recognized using the same configuration of DNN-HMM as in our previous work [195]. A maximum likelihood estimation (MLE) training approach with 100 senones and 8 Gaussian mixtures was adopted. Cepstral mean and variance normalization (CMVN) was followed by dimensionality reduction using Linear Discriminant Analysis (LDA) with a context of 6 frames (3 left and 3 right) to give a feature vector of size 40. The input layer of DNN has 360 (40×9 frames) dimensions. Two hidden layers with 512 nodes in each layer and an output layer of dimension 96 were used. A constrained Language Model (LM), wherein we restrict the recognizer to give one word as output per utterance, was used.

Healthy control (HC) and dysarthric (DYS) speech utterances from blocks B1 and B3 of computer command (CC) words were used for training the DNN-HMM based ASR as shown in Table 5.5. Training configuration S-1 comprises only healthy control (HC) speech. In the second training configuration S-2, we use dysarthric (DYS) speech from blocks B1 and B3 in addition to HC speech. In S-3, ASR was trained using HC speech and dysarthric speech from blocks B1 and B3 that were enhanced using the TDNN-DAE models. Each training configuration was evaluated using dysarthric speech features for computer command words (DYS) from block B2. In Testing Configuration 1, the dysarthric speech features were temporally adapted. In our earlier work [14], we show that temporal adaptation of the test dysarthric speech significantly reduced the ASR word error rate (WER). Hence, this chapter uses the WER corresponding to temporally adapted dysarthric speech as the baseline. In Testing Configuration 2, the temporally adapted dysarthric speech features were enhanced using the TDNN-DAE model and then evaluated. There is no overlap in the training and testing data.

5.2.4 Results and Analysis

DNN-HMM ASR recognition is evaluated for speaker adaptation (SA) and speaker-independent (SI) scenarios for the training and test configurations mentioned in

Table 5.5: ASR Training and testing configurations

System	**Training configuration** (B1, B3)	**Testing configuration 1** (B2)	**Testing configuration 2** (B2)
S-1	HC	Temporally adapted DYS (MFCC-TA)	Temporally adapted + TDNN-DAE enhanced DYS (MFCC-TA+TDNN-DAE)
S-2	HC + DYS		
S-3	HC + TDNN-DAE enhanced-DYS		

Table 5.5. Word error rates produced for the above scenarios are reported in Table 5.6. System S-1 does not use any dysarthric speech data for ASR training. An absolute improvement of 13% was observed when the test dysarthric speech data was enhanced using the TDNN-DAE. This indicates that the TDNN-DAE-based enhancement of dysarthric speech features results in these features being closely matched to healthy control speech features. Also, the drastic reduction in the ASR performance for S-2 for TDNN-DAE enhanced data, specifically in the SA scenario serves as additional confirmation that the enhanced dysarthric speech features match more closely to healthy control than to dysarthric speech data. Training configuration S-3 comprises healthy control and TDNN-DAE enhanced dysarthric data (B1 and B3). Speaker adaptation-based ASR performance is higher by 3% for TDNN-DAE enhanced dysarthric speech (B2) than SA recognition performance for S-2. Both S-2 and S-3 contain the same amount of healthy control and dysarthric speech data in the training process, except that the dysarthric speech used in S-3 is enhanced using TDNN-DAE. ASR performance for the three different training configurations indicates that using TDNN-DAE to enhance dysarthric speech features results in dysarthric speech features matching closely to healthy control speech.

Table 5.6: WER for TDNN-DAE

Training configuration	Testing configuration 1		Testing configuration 2	
	SA	SI	SA	SI
S-1	-	37.86	-	24.73
S-2	21.44	33.67	60.8	29.7
S-3	82.69	72.47	18.54	34.39

An analysis of ASR performance at dysarthria severity levels was done for the two configurations that provide the best recognition, namely S-2-SA using unenhanced dysarthric training and test data and S-3-SA using enhanced dysarthric training and test data. An improvement was seen across all dysarthria severity levels as shown in 5.7.

Table 5.7: Severity level analysis of WER

Severity	S-2-SA Testing configuration 1	S-3-SA Testing configuration 2	Absolute Improvement
Very Low	5.71	1.35	4.4
Low	11.39	9.4	1.99
Medium	22.67	19.46	3.2
High	57	52.5	4.5

5.2.5 Conclusion

In this chapter, we explain the process of enhancing dysarthric speech features using a TDNN-DAE. The objective is to enhance the dysarthric speech features to match that of healthy control speech. TDNN-DAE is trained using temporally adapted dysarthric speech as input and healthy control speech as target speech. The training process and the data used for TDNN-DAE need careful consideration to obtain optimal ASR performance. The dysarthric speech thus enhanced is recognized using a DNN-HMM-based Automatic Speech Recognition (ASR). Speaker-independent and speaker adaptation-based ASR configurations were evaluated using both unenhanced and enhanced dysarthric. An absolute improvement of 13% and 3% was observed in ASR performance for SI and SA configurations, respectively when enhanced dysarthric speech features were used. ASR performance for each of the training and testing configurations confirms that the dysarthric speech enhanced using TDNN-DAE is matched more closely to healthy speech than to dysarthric speech for the same speaker. An analysis of the two best-performing configurations clearly indicate that the ASR performance significantly improves at all severity levels of dysarthria.

Chapter 6

Data augmentation of dysarthric speech

This chapter is based on the following publications:

I. Vachhani, B., Bhat, C., & Kopparapu, S. K. (2018). Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. In Interspeech (pp. 471-475).

II. Bhat, C., & Strik, H. (2024). Two-stage Data Augmentation for Improved ASR Performance for Dysarthric Speech - submitted to Computers in Biology and Medicine

6.1 Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition

Dysarthria refers to a speech disorder caused by trauma to the brain areas concerned with motor aspects of speech giving rise to effortful, slow, slurred or prosodically abnormal speech. Traditional Automatic Speech Recognizers (ASR) perform poorly on dysarthric speech recognition tasks, owing mostly to insufficient dysarthric speech data. Speaker-related challenges complicate the data collection process for dysarthric speech. In this chapter, we explore data augmentation using temporal and speed modifications to healthy speech to simulate dysarthric speech. DNN-HMM-based Automatic Speech Recognition (ASR) and Random Forest-based classification was used for the evaluation of the proposed method. Dysarthric speech, generated synthetically, is classified for severity level using a Random Forest classifier that is trained on actual dysarthric speech. ASR trained on healthy speech, augmented with simulated dysarthric speech, is evaluated for dysarthric speech recognition. All evaluations were carried out using the Universal Access dysarthric speech corpus. An absolute improvement of 4.24% and 2% were achieved using tempo-based and speed-based data augmentation, respectively, as compared to ASR performance using healthy speech alone for training.

6.1.1 Introduction

Dysarthria is a speech disorder resulting from disruption in the execution of speech movements due to neuro-muscular disturbances to muscle tone, reflexes, and kinematic aspects of movement. It could be either acquired or developmental. Dysarthric speech is characterized by being slow, slurred, harsh or quiet, or uneven, depending on the type of dysarthria. Speech-enabled interfaces are gaining popularity, especially in the assisted and smart living domains. Also, speech is a convenient alternative to other machine interfaces such as remote controls, keyboards, or PC mice given that persons with dysarthria are often faced with physical inabilities as well [165]. While traditional, *off-the-shelf* Automatic Speech Recognition (ASR) systems perform well for normal speech; this is not the case with atypical dysarthric speech owing to the inter-speaker and intra-speaker inconsistencies in the acoustic space as well as the sparseness of data. Several techniques are employed to improve ASR performance for dysarthric speech: acoustic space enhancement, feature engineering, Deep Neural Networks (DNN), speaker adaptation, and lexical model adaptation- individually or as a combination thereof [167, 26, 135, 219, 195]. In order to exploit the machine learning techniques for ASR fully, suitable data to build these systems is imperative. However, owing to speaker muscle weakness and fatigue, collecting dysarthric data is tedious, especially for speakers with severe dysarthria. Additionally, since dysarthria can stem from a variety of neurological disorders, the characterization of dysarthric speech is complex; this makes the designing of a data collection process difficult. Thus far, three popular dysarthric

speech databases, namely the Universal Access (UA) speech corpus [95], Nemours [126] and TORGO [168] exist for American English. Two French corpora, namely the CCM corpus collected by Dr Claude Chevrie-Muller and her team and the Aix-Neurology-Hospital corpus (ANH) has been described in [44]. The authors describe a Dutch dysarthric speech database containing mildly to moderately dysarthric speech from patients with PD, traumatic brain injury, and cerebrovascular accident [213]. A Korean dysarthric speech corpus was built as a part of the Quality-of-Life technology (QoLT) project that focuses on the development of speech technologies for people with articulation disabilities [25]. A Cantonese corpus with a focus on the investigation of articulatory and prosodic characteristics of Cantonese dysarthric speech is discussed in [205]. German [185], Spanish [141] and Czech [169] corpora were collected with the intent of studying dysarthric speech in patients suffering from PD. While most of the corpora comprise data collected under clinical settings, [138] describes *the homeService corpus*, a British English corpus of realistic dysarthric data collected in the home environment. Each of the above databases was designed for a specific purpose with a broad perspective of improving the lives of people with dysarthria. However, the amount of data is substantially lower than in a speech corpus of normal speech used in training the state-of-the-art ASR systems that use machine learning. To overcome this issue of unavailability of suitable speech data, we adopt data augmentation techniques.

Data augmentation is the process by which we create new synthetic training samples by adding small perturbations to our initial training set. The objective is to make the model invariant to perturbations and enhance its ability to generalize. In [109], audio speed was modified using three-speed factors, and the effectiveness was reported for large vocabulary continuous speech recognition (LVCSR). Different audio data augmentation techniques such as time stretching, pitch shifting, dynamic range compression, and mixing with background noise were used for environmental sound classification in a Convolutional Neural Network (CNN) based architecture to significantly improve the classification accuracy [171]. Data augmentation techniques have been used to improve classification tasks such as real-life sound classification [153, 144]. In [139] Alzheimer's disease (AD) data was augmented using two normative data sets, through minority class oversampling with Adaptive Synthetic sampling (ADASYN), wherein the proposed technique outperformed state-of-the-art results in the binary classification of speech with and without AD.

In this chapter, we explore how an understanding of the deficits in speech production caused by dysarthria may be used to augment existing data. We present an analysis of phone durations in dysarthric data with a bearing on dysarthria severity level. Based on this information, we proceed with data augmentation using temporal and speed modifications to healthy speech to generate synthetic speech that matches the characteristics of dysarthric speech. Further, we classify this synthetic dysarthric speech into four severity levels using a Random Forest classifier that is trained on actual dysarthric speech so as to validate our understanding of the impact of these modifications on healthy speech and how it simulates dysarthric speech. A

DNN-HMM-based Automatic Speech Recognition (ASR) is trained using healthy speech augmented with simulated dysarthric speech. This ASR system is evaluated for dysarthric speech recognition using the Universal Access (UA) dysarthric speech corpus.

The rest of the chapter is organized as follows. Section 6.1.2 presents an analysis of phone durations in dysarthric speech, motivates the data augmentation process, and discusses the augmentation techniques used, Section 6.1.3 describes the experimental setup, In Section 6.1.4, we present the results and analysis, and we conclude in Section 6.1.5.

6.1.2 Methodology

6.1.2.1 Phoneme duration analysis

In order to modify healthy control speech to emulate dysarthric speech characteristics, we need to first understand the dysarthric speech itself. In our earlier work [14], we modified the tempo of dysarthric speech based on severity to improve ASR recognition. It was observed that the sonorant regions of dysarthric speech are of longer durations as compared to that of healthy speech. In this work, we further examine the relationship between phone durations of dysarthric speech and dysarthria severity levels. The UA Speech corpus comprises dysarthric speech of 4 severity levels, namely S1, S2, S3, S4 in the increasing order of severity. A total of 3534 utterances of dysarthric speech corresponding computer command words were force-aligned at phone level using Sphinx3 toolkit [28], using Voxforge English acoustic models trained on approximately 35 hours of speech data [198]. The alignment was then manually inspected and corrected for extraction of phone duration. A similar exercise was carried out on the TORGO dysarthric speech corpus [168]. The TORGO dysarthric speech corpus comprises dysarthric speech of 3 severity levels, namely S1, S2, and S3. A more accurate representation of the relationship between phone durations and severity can be seen for this corpus since it comprises manual annotation of utterances at the phone level. We observed that there is a strong correlation between dysarthria severity and the average duration of a phone as shown in Figure 6.1. It was found that the average phone duration is proportional to the severity of dysarthric speech; the higher the severity, the longer the phone duration.

Based on this analysis, we modify the phone durations of healthy control speech to generate synthetic dysarthric speech data. We use this modified speech along with the healthy control speech to augment the ASR training data.

6.1.2.2 Synthetic dysarthric data generation

Healthy control speech was modified using two different time domain perturbations, namely (1) Time (Speed) perturbation and (2) Tempo perturbation. *Rubberband – an audio time-stretching and pitch-shifting utility program* was used for this purpose

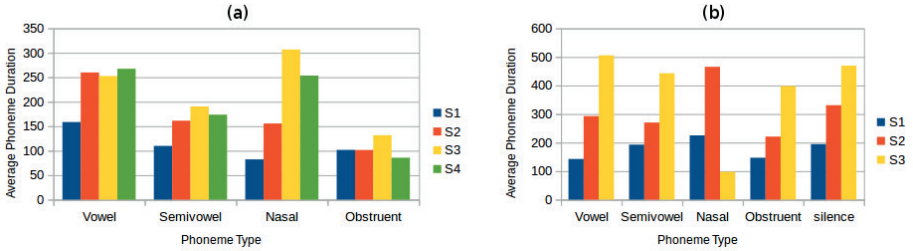

Figure 6.1: (a) Average phone duration (ms) for the UA dysarthric speech Corpus (b) Average phone duration (ms) for the TORGO dysarthric speech corpus

and is described below [20]. Healthy control speech modified in this manner amounts to synthetically generated dysarthric speech data. To the best of our knowledge, data augmentation in the context of dysarthric speech recognition has not been reported in the literature previously.

6.1.2.3 Time (Speed) Perturbation-based Data Augmentation

Speed perturbation is achieved by re-sampling the input signal by a factor $R1$. If $R1 < 1$, signal duration is reduced, and for $R1 > 1$, signal duration is increased. In this work, we use different values of $R1$ as $R1 \in \{1.2, 1.4, 1.6, 1.8, 2.0, 2.2\}$ to modify the durations of healthy control speech. The below command will stretch the given input signal duration to $R1$ times the original duration in the Rubberband toolkit.

$$rubberband \quad -t \quad R1 < infile.wav >< outfile.wav >$$

6.1.2.4 Tempo Perturbation-based Data Augmentation

The tempo of the signal is modified by factor $R2$ while ensuring that the pitch and spectral envelope of the signal do not change. If $R2 > 1$, signal duration reduces and $R2 < 1$ signal duration increases, making the healthy control speech slower. In this work we use $R2$ as $R2 \in \{0.4, 0.6, 0.8\}$ to modify the healthy control speech. The below command will modify the given input signal duration to $R2$ times the original duration.

$$rubberband \quad -T \quad R2 < infile.wav >< outfile.wav >$$

The parameters $R1$ and $R2$ were selected empirically based on the severity classification provided by the Random Forest classifier for various values of $R1$ and $R2$ as discussed in Section 6.1.3.2.

6.1.3 Experimental setup

6.1.3.1 Database

Data from the Universal Access (UA) speech corpus [95] was used for both training and testing. The UA speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. The recording material consisted of 455 distinct words with 10 digits, 26 international radio alphabet letters, 19 computer commands, 100 common words and 300 uncommon words that were distributed into three blocks. Audio data was recorded using a 7-channel microphone array, fitted to the top of a computer monitor. Speech intelligibility ratings for each dysarthric speaker, as assessed by five naive listeners are also included in the corpus. Speakers were divided into four different categories based on the intelligibility. We use this information to analyze the performance of our recognition systems at different dysarthria severity levels. In this chapter, we have used 19 computer command words from 13 healthy control (HC-CC) and 15 dysarthric (DYS-CC) speakers.

Long silence regions at the start and end of both the healthy control (HC) data used for training and dysarthric speech (DYS) used for testing of the ASR are trimmed using an energy-based method using the PRAAT tool [18]. Initial experiments were conducted to understand the effect of silence removal at the start and end of the HC and DYS speech. Traditional DNN-HMM-based system using standard MFCC features as discussed in Section 6.1.3.3. Table 6.1 shows the ASR performance in terms of word error rate (WER) for training and testing data with and without silence pre-processing. An absolute improvement of 15% (48.47% to 33.11%) was achieved by using the fMLLR transform, and further improvement of 4% (33.11% to 29.06%) was achieved using silence pre-processing. We use the best WER, wherein both training and test data were pre-processed with fMLLR-based ASR configuration as the baseline for reporting our current work.

Table 6.1: Effect of data Pre-processing on WER

Training	**Testing**	**WER**	
Total 3458 utt	**Total 3534 utt**	**w/o fMLLR**	**with fMLLR**
HC-CC	DYS-CC	48.47	33.11
SIL trimmed HC-CC	SIL trimmed DYS-CC	37.32	29.06

6.1.3.2 Dysarthria Severity Classification on Augmented Data

The validity of using synthetically generated dysarthric speech to augment the ASR training data for dysarthric speech recognition needs to be ascertained. Synthetically generated dysarthric speech is automatically classified using a Random Forest classifier trained on actual dysarthric speech. Classifier was trained using the feature set suggested by *Interspeech 2009 emotion challenge*, extracted using openSMILE

toolkit [40]. A total of 3534 dysarthric utterances were used for training the classifier using 5-fold cross-validation using WEKA toolkit [45]. An accuracy of 96% was achieved for dysarthric speech classification into 4 classes based on the intelligibility score provided in the UA Speech corpus. A total of 3458 healthy control (HC) utterances were modified using various tempo and speed perturbation parameters described in Section 6.1.2 were classified using this framework into four severity classes.

6.1.3.3 DNN-HMM-based ASR Framework

The Kaldi toolkit [156] was used for DNN-HMM-based dysarthric speech recognition. GMM-HMM system was trained using a maximum likelihood estimation (MLE) training approach along with 100 senones and 8 Gaussian mixtures. Cepstral mean and variance normalization (CMVN) was applied on 23 dimensions of MFCC features. Dimensionality reduction was done using Linear Discriminant Analysis (LDA), wherein LDA builds HMM states using feature vectors with a reduced feature space. We use the context of 6 frames (3 left and 3 right) to compute LDA. The feature vector size post LDA is set to 40.

The input layer of DNN has 360 (40×9 frames) dimensions using a left and right context of 4 frames. The output layer has a dimension of 96 (number of senones available in the data). Two hidden layers with 512 nodes in each layer were used. Feature-space Maximum Likelihood Linear Regression (fMLLR) transformed features are used as input to the DNN training, making it a feature normalization technique. In the decoding process, we use configurations with and without fMLLR transformed features as input [147]. DNN training was carried out using 15 epochs for all experiments. Dysarthric speech recognition was carried out using a constrained language model (LM), wherein we restricted the recognizer to give one word as output per utterance. The performance of each of the recognition systems is reported in terms of word error rate (WER).

Training configurations for the DNN-HMM-based ASR are shown in Table 6.2. A total of 3534 dysarthric speech utterances corresponding to 19 computer command words from blocks B1 and B3 have been used for testing purposes.

Table 6.2: Training data for different systems

Training Set	**Info**	**Total no.utterances**
A	No augmentation	3458
B	Time stretching $R1 \in \{1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2\}$	24206
C	Tempo stretching $R2 \in \{0.4, 0.6, 0.8, 1.0\}$	13832

6.1.4 Experimental Results and Analysis

The synthetically generated dysarthric speech was classified into four classes as discussed in Section 6.1.3. Table 6.3 shows the classification of 3458 healthy control utterances modified into dysarthric utterances using various augmentation parameters. Synthetically generated dysarthric speech is classified into four classes, namely S1, S2, S3, and S4, in increasing order of severity. It can be seen for both speed and tempo modifications that the synthetically generated dysarthric utterance classification is closely correlated to the duration of the utterance. Table 6.4 shows the performance of the ASR for training configurations mentioned in Table 6.2, examined at individual severity levels.

Table 6.3: Severity classification of synthetically generated dysarthric data - %Accuracy

Classifier System	**Augmentation parameter**	**S1**	**S2**	**S3**	**S4**
A	None	93.97	4.29	1.74	0.00
B1	R1 = 1.2	88.35	8.00	3.30	0.35
B2	R1 = 1.4	81.05	12.51	4.98	1.45
B3	R1 = 1.6	65.24	21.55	9.79	3.42
B4	R1 = 1.8	46.87	33.43	14.83	4.87
B5	R1 = 2.0	33.72	39.86	21.15	5.27
C1	R2 = 0.4	4.06	39.98	38.18	17.79
C2	R2 = 0.6	59.62	23.99	12.05	4.35
C3	R2 = 0.8	86.96	8.98	3.48	0.58

Tempo-based and speed-based augmentation techniques give an absolute improvement of 4.24% and 2%, respectively. Higher improvement was observed for higher severity (S4), approximately 3% and 12% absolute improvement over baseline for systems speed and tempo augmentation, respectively. Table 6.5 shows the effects of each data augmentation parameter on four different severity levels. It can be seen that the proposed method gives improvement at all severity levels.

Table 6.4: Severity wise WER for testing data

Training Set	**S1**	**S2**	**S3**	**S4**	**Overall WER**
A	1.05	17.89	44.73	78.51	**29.06**
B	0.98	19.73	36.44	75.43	**27.05**
C	1.28	15.52	37.36	66.96	**24.82**

In order to attribute the improvement in the ASR performance to the synthetically generated dysarthric speech data, we look into the ASR performance for data augmentation parameters R1 and R2 separately. 8 separate ASR systems were trained as seen in Table 6.5, each with 3458 synthetically generated dysarthric utterances. Table 6.5 shows the effect of individual augmentation parameters on ASR

performance. No healthy control data was used in the ASR training. Correlation between the ASR performance for actual dysarthric speech and the duration of the synthetic dysarthric speech data is seen for both speed and tempo perturbations. From Table 6.5 and Table 6.3, it is seen that increasing the phone durations using augmentation degrades the ASR performance for low-severity dysarthric speech (S1 and S2).

Table 6.5: Effect of data augmentation on WER for individual severity level

Training Set	**Augementation parameter**	**S1**	**S2**	**S3**	**S4**	**Overall WER**
A	None	1.05	17.89	44.74	78.51	29.06
B1	R1=1.2	0.98	17.63	42.11	72.95	27.33
B2	R1=1.4	0.90	18.82	38.42	73.68	26.91
B3	R1=1.6	1.35	21.32	36.45	69.30	26.34
B4	R1=1.8	1.20	22.63	37.76	67.25	26.46
B5	R1=2.0	1.58	21.18	34.87	69.01	26.00
C1	R2=0.4	2.33	18.95	39.87	70.47	27.16
C2	R2=0.6	1.05	20.79	37.37	77.34	27.87
C3	R2=0.8	0.75	18.55	34.61	74.56	26.15

Based on the ASR performance for synthetically generated dysarthric data, optimal values of augmentation parameters R1 and R2 to generate dysarthric data of different severity levels is as shown in Table 6.6.

Table 6.6: R1 and R2 recommendation for optimal ASR recognition

Severity	**R1**	**R2**
S1	1.4	0.8
S2	1.2	0.8
S3	2	0.4
S4	1.8	0.4

6.1.5 Conclusions

Given that speech is an attractive interface to control the devices used in assisted living and smart homes, it is imperative that we look into improving the ASR performance for pathological speech. Due to a lack of suitable data to train the ASRs, machine learning techniques are not fully exploited for pathological speech recognition. In this chapter, we address the data challenge for dysarthric speech using data augmentation to synthetically generate dysarthric speech data using healthy control speech. An understanding of the deficits in speech production caused by speech pathology has been used to augment existing data using speed and tempo modifications to the healthy control speech. A DNN-HMM-based Automatic Speech Recognition (ASR) system and Random Forest-based classification system have been

used for the evaluation of the proposed method. Synthetically generated dysarthric speech is classified into four different severity levels using a Random Forest classifier trained on actual dysarthric speech. ASR system trained using healthy control speech augmented using synthetically generated dysarthric speech is evaluated for dysarthric speech utterances. All evaluations were carried out on the Universal Access dysarthric speech corpus computer command words. An absolute improvement of 15% was achieved by using fMLLR transform as compared to our previous work [195]. Additionally, ASR performance improved by 4% using silence pre-processing. We use this WER (29.06%) as a baseline to report our current work. An absolute improvement of 4.24% and 2% was achieved using tempo-based and speed-based data augmentation systems over the baseline system.

6.2 Two-stage Data Augmentation for Improved ASR Performance for Dysarthric Speech

Machine learning (ML) and Deep Neural Networks (DNN) have greatly aided the problem of Automatic Speech Recognition (ASR). However, accurate ASR for dysarthric speech remains a serious challenge. The dearth of usable data remains a problem in applying ML and DNN techniques for dysarthric speech recognition. In the current research, we address this challenge using a novel two-stage data augmentation scheme, a combination of static and dynamic data augmentation techniques, designed by leveraging an understanding of the characteristics of dysarthric speech. We explore speaker-independent ASR using modifications to healthy speech using various perturbations, devoicing of consonants, and voice conversion, comprising stage one or static augmentations. Subsequent to the first stage, a modified SpecAugment algorithm tailored for dysarthric speech is employed. This variant, termed Dysarthric SpecAugment, leverages the characteristics of dysarthric speech and forms the second stage of the two-stage augmentation approach. This acoustic model is used to pre-train a speaker-dependent ASR using dysarthric speech. The objective of this work is to improve the ASR performance for dysarthric speech using the two-stage data augmentation scheme. An end-to-end ASR using a Transformer acoustic model is used to evaluate the data augmentation scheme on speech from the UA dysarthric speech corpus. We achieve an absolute improvement of 10.7% and a relative improvement of 29.2% in word error rate (WER) over a baseline with no augmentation, with a final WER of 25.9% for the speaker-dependent system.

6.2.1 Introduction

Speech production is one of the most complex human motor skills and involves both linguistic units and acoustic events. Motor speech problems caused by neurological difficulties can be congenital or acquired, impacting one or several speech subsystems, namely respiratory, phonatory, and articulatory. Congenital dysarthria can be attributed to an inherited condition, such as Cerebellar Palsy, which affects the muscles of speech production. Dysarthria acquired later in life may result from stroke, traumatic brain injury (TBI), tumors, infection, or progressive neurological diseases such as amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS), or Parkinson's disease (PD). Manifestations of dysarthria may include restricted movement of the lip, tongue, and jaw; slurred speech; slow speech; rapid mumbled speech; soft or inaudible speech; breathiness; hoarseness; drooling; and difficulty swallowing. For intelligible speech production, the muscles and muscle groups in the speech subsystems must be well-coordinated in time and space, rendering dysarthric speech generally unintelligible. The higher the severity of dysarthria, the lower the intelligibility of dysarthric speech. The neurological damage that affects speech-motor function impacts physical activities associated with the motor neurons as well. Typical human interface with gadgets and devices involves typing

into a keyboard. Keyboard input using hand movements is slowed down by a factor of 150 to 300 in severe cases of dysarthria in comparison with regular users [66]. However, dysarthric speech is slow by a factor of 10 to 17 as compared to regular speech, at about 15 words per minute in the most severe cases [165]. Also, it has been found that dysarthric speakers exhibit good prosodic control which in turn aids communication efficiency [148]. Hence, persons with dysarthria may benefit immensely from Automatic Speech Recognition (ASR)-based personal assistants. ASR utilizes a combination of signal processing and machine learning techniques to convert spoken language into text. Performance of ASR systems and personal assistants has made great strides owing to the recent Machine learning (ML) and Deep Neural Networks (DNN) techniques, albeit this is not the case with the atypical dysarthric speech due to the inter-speaker and intra-speaker inconsistencies in the acoustic space as well as the sparseness of data. In order to capitalize on the current research on ML techniques for ASRs, such as the End-to-End (E2E) ASR systems, suitable and abundant data to build these systems is imperative. However, the collection of dysarthric data is tedious, especially for speakers with severe dysarthria, on account of speaker muscle weakness and fatigue. Data augmentation policies designed specifically for dysarthric speech can act as a key factor in improving dysarthric ASR with limited intrusion on the speakers for data collection.

Data augmentation (DA) is a common approach employed to increase the amount of training data, especially for DNN training in order to avoid overfitting while generating robust DNN models. Several types of DA techniques have been employed to vastly increase the quantity of matching training data [109, 137, 30], especially in atypical scenarios such as reverberant speech [110], child speech [24] and dysarthric speech [51, 74, 196]. Simple audio-level augmentations such as speed, pitch, tempo, and volume perturbations applied directly on raw speech to increase the training data multi-fold [109, 24] have proven to be extremely effective. SpecAugment [146] and DA techniques inspired by SpecAugment such as frame level SpecAugment [115] and usage of semantic masks [200] have been successful in improving ASR performance through dynamic DA.

Research on DA in the context of dysarthric speech is sparse since this involves a clear understanding of dysarthric speech patterns. Time and tempo-stretching of healthy speech-based DA for improving speech recognition has been investigated in [196]. In order to address ASR performance on severe dysarthric speech, speaker-dependent acoustic models based on phoneme-level speech tempo ratio between typical and speaker-specific dysarthric speech have been created to augment existing dysarthric speech [207]. Two separate augmentation policies involving speed, tempo, and vocal tract length perturbation (VTLP) applied on healthy and dysarthric speech showed significant improvement in the ASR performance [51]. A transformation of healthy speech to dysarthric speech using voice conversion-based techniques involving speaking rate modification, pitch modification, and spectral feature transformation using adversarial training, have been employed to simulate training data using healthy speech [21]. Visual DA techniques are applied to speech

features that are extracted visually [176]. A Deep convolution-based generative adversarial network (DCGAN) was used for tempo and speed perturbations in addition to learning hidden unit contributions (LHUC) based speaker adaptation in [75]. The study [123] investigates DA for transfer learning in ASR for continuous dysarthric speech. It augments dysarthric speech with speed/volume variations, virtual microphone arrays, and multi-resolution features (VM-MRFE) to bridge vocabulary gaps before transferring knowledge from a pre-trained normal speech model. This approach, evaluated on isolated and continuous speech datasets, tackles out-of-vocabulary challenges. Table 6.7 summarizes the research work on DA for dysarthric speech.

Through this chapter we extend the work presented in [17], which explored various DA techniques using hand-crafted explainable features as well DNN-based augmentations in order to achieve improved ASR performance for dysarthric speech in terms of word error rate (WER). A two-stage DA process that involves a static and dynamic augmentation of dysarthric speech data was used. We refer to the process of DA prior to DNN training as static and DA as a part of the DNN training as dynamic. The features used for static DA are handcrafted and explainable, whereas the dynamic augmentation is done in real-time, using features that are generated as part of the neural network training ensuring that the DNN system benefits from viewing diverse data at every epoch, which in turn strengthens the training process. Since the features generated during the DNN training are in real-time, we consider these features non-explainable. Our main contribution is the use of various static augmentation policies and SpecAugment [146] in novel ways to achieve dysarthric speech DA. In this chapter, we introduce two new concepts that improve the performance of ASR for dysarthric speech. We leverage our knowledge of dysarthric speech to design a dynamic DA method akin to SpecAugment, which we call the Dysarthric SpecAugment (DSA) scheme. In DSA, we dynamically introduce breathiness, stuttering, and hypernasality to healthy speech, thereby increasing the diversity of the training data. We analyze the performance of an End-to-End Transformer-based ASR in terms of WER for scenarios with no dysarthric data being available for training, as well as when some dysarthric speech data is available. Transformer-based neural network training was used because transformers offer significant advantages in capturing long-range dependencies, faster training, and improved modeling of speech features. These advancements lead to more robust, accurate, and versatile speech recognition systems that can power a wide range of applications [85, 203]. We use the ESPnet toolkit [204] for all our ASR experiments. ASR performance is evaluated on the UA dysarthric speech corpus.

We describe static and dynamic DA techniques in Section 6.2.2. We discuss the experimental set-up in Section 6.2.3 and present the contributions of various combinations of DA techniques on the WER in Section 6.2.4. Finally, we conclude with our observations and recommendations in Section 6.2.5.

Table 6.7: Data Augmentation for dysarthric speech ASR in Literature

Study	Database	Features	Speech Tech	Accuracy/WER
Vachhani et al., 2018 [196]	UA Speech (partial)	MFCC	Time/tempo stretching, DNN-HMM-based ASR, fMLLR	Tempo-based: 24.82% Speed-based: 27.05%
Xiong et al., 2019 [208]	UA Speech	MFCC	Interpolation followed by downsampling for tempo adjustment, DNN-HMM based ASR with TDNN	27.88%.
Geng et al., 2020 [51]	UA Speech	MFCC	Speed, tempo, and VTLP-based augmentation. (LHUC) based speaker adaptive and multi-task learning (MTL)-based training for DNN	26.37%
Celin et al., 2020 [21]	UA Speech Tamil dysarthric speech corpus	MFCC	Virtual linear microphone array-based synthesis followed by multi-resolution feature extraction (MRFE), DNN-HMM-based ASR system	32.79% - low 35.75% - very low
Shahamiri, 2021 [176]	UA Speech	Voice grams	Visual DA, Speech vision ASR using Spatial Convolutional Neural Network (S-CNN)	Absolute average WRAs of 64.71%
Yue et al., 2022 [218]	TORGO	Raw waveform features, DA	Parametric CNNs, multi-stream acoustic modelling	Parametric CNN: 36.2% Multi-stream acoustic modelling: 33.2%
Bhat et al., 2022 [17]	UA Speech	Mel filter bank,	TDNN-DAE, speed, tempo, and loudness based DA, specialized Spec Augment. End-to-end ASR (ESPnet)	20.6%
Jin et al., 2023 [76]	UA Speech	Wav2vec 2.0 embedding features	Variational auto-encoder generative adversarial network (VAE-GAN)-based DA, LF-MMI factored TDNN, LHUC-SAT, ESPnet toolkit	27.78%
Celin et al., 2023 [123]	UA Speech Tamil dysarthric speech corpus	Speed/volume variations, VM-MRFE	DNN	32.97% -low 63.38% - very low category

6.2.2 Two-stage Data Augmentation

Building a dysarthric speech corpus is challenging owing to the difficulties faced by the patients. Our approach offers a significant advantage: it leverages readily available healthy speech data and transforms it to address the scarcity of dysarthric speech data. The combined healthy speech, augmented speech, and dysarthric speech data help ASR systems learn the nuances of dysarthric speech patterns, leading to better recognition accuracy. Two-stage data augmentation involves DA done in two steps. First, static DA (SDA) techniques described in Section 6.2.2.1 are applied, followed by dynamic DA (Section 6.2.2.2) as the next step. The term *Dynamic DA* refers to a special technique applied during neural network training. It is a variant of SpecAugment, which we call Dysarthric SpecAugment (DSA). DSA algorithms aim to transform healthy speech features through specific manipulations to create speech that acoustically resembles dysarthric speech. Therefore, both SDA, using explainable speech features, and DSA which uses speech features generated in real-time and hence non-explainable, are applied to healthy speech data. The ASR models built using healthy speech and augmented speech are adapted using dysarthric speech data to significantly improve the ASR performance in terms of WER.

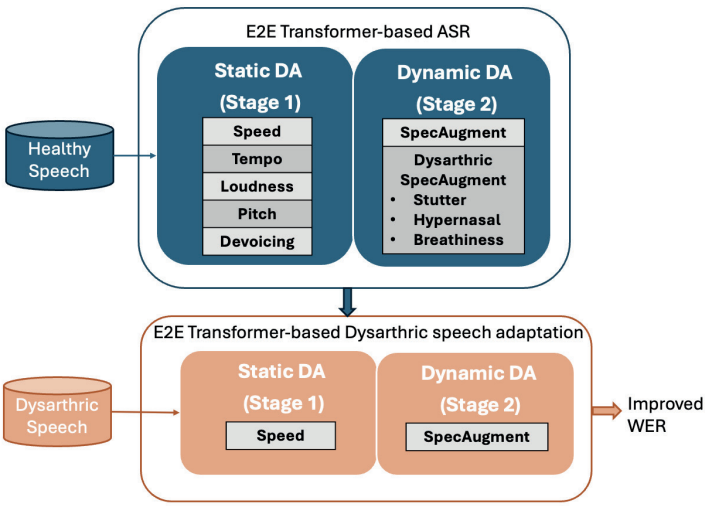

Figure 6.2: Two-stage DA for dysarthric speech recognition

6.2.2.1 Static Data Augmentation (SDA)

The DA techniques applied prior to DNN training augmented the dysarthric speech data in a static manner. We have used two types of static augmentation as described

in the rest of this section.

Speed, tempo, volume, and pitch perturbation

A. Speed perturbation is a recommended method for DA since it has been known to improve speech recognition as well as has a low implementation cost [109]. However, it is to be noted that speed perturbation affects both pitch and tempo of the original speech since it involves resampling of the original speech signal. In the current work, we have applied speed modifications to both healthy speech as well as dysarthric speech to provide three different versions, resulting in DA by a factor of three.

B. The relationship between the tempo of healthy speech and dysarthric speech has been investigated and leveraged to improve dysarthric speech recognition [14, 207, 51]. Typically, dysarthric speech is slow and slurred, indicating a slower tempo as compared to healthy speech. We have applied three different tempo modifications on healthy speech to match the severity levels in the dysarthric speech corpus. Tempo perturbation does not alter the pitch of the speech being modified.

C. Training dysarthric speakers to increase the loudness of speech to improve intelligibility is a known therapy technique that results in a higher articulatory-acoustic working space as well as improved acoustic contrast for dysarthric speakers[190]. In order to match the characteristics of dysarthric speech, we have applied loudness modifications to healthy speech by reducing the loudness. Two different loudness factors have been used to generate two distinct versions of healthy speech.

D. It has been observed that the three popularly used English dysarthric speech corpora, the UA Speech corpus, the TORGO, and the Nemours database have fewer female speakers as compared to male speakers. The skewed representation of female speakers in the training data impacts the ASR performance for female dysarthric speech. This article explores several techniques for augmenting female dysarthric speech.

    - Pitch shifting-based DA has been applied to improve the ASR performance [178] as well as in case of classification of noise [217]
    - Vocal tract length perturbation (VTLP) has been applied in different speech recognition tasks, especially with the limitations of low-resource languages. Combined with Deep Neural Networks (DNN), VTLP-based augmentation has proved beneficial in improving ASR performance [83, 109]
    - Spectral warping: applied to the linear prediction (LP) spectrum of children's speech data, aims to increase spectral variability. The resulting

modified features are expected to improve ASR performance through enhanced spectral diversity [88].

- Voice conversion swaps a voice fingerprint, preserving words but changing speaker, accent, or even emotion. It's a voice mask for new applications like speaking aids, personalized narration, or anonymizing recordings. We use voice conversion to generate new female voices and use this data for augmenting the UA Speech corpus[61, 222]

Devoicing of consonants

Devoicing of consonants is one of the possible characteristics of dysarthric speech. Devoicing refers to the lack of vocal fold vibration during the production of voiced consonant sounds. Normally, when voicing a consonant, such as b, d, or g, the vocal folds vibrate, creating a voiced sound. However, in dysarthria, the coordination and control of the muscles involved in vocal fold vibration may be impaired, leading to inconsistent or absent voicing. In dysarthric speech, devoicing of consonants can manifest in different ways [121]:

- Partial devoicing: The consonant may have reduced voicing, meaning that the vocal folds do not fully vibrate, resulting in a weak or breathy sound. For example, the b sound may be pronounced as a partially voiced p sound.
- Complete devoicing: The vocal folds may fail to vibrate altogether, resulting in a voiceless sound. For instance, the d sound may be produced as a voiceless t sound.
- Variable devoicing: The presence or absence of voicing may be inconsistent and unpredictable. In some instances, the consonant may be voiced, while in others, it may be devoiced. This variability can make speech sound unclear or difficult to understand.

It is important to note that dysarthria can present in different forms and severity levels, depending on the underlying cause and individual factors. Therefore, specific characteristics of dysarthric speech, including the devoicing of consonants, can vary from person to person. We, therefore, modify healthy speech data by introducing the devoicing of consonants to different degrees, as mentioned above. We observe the changes in the ASR performance for dysarthric speech at different intelligibility levels corresponding to the devoicing method used.

6.2.2.2 Dynamic Data Augmentation

The term *Dynamic DA* refers to a DA technique applied during neural network training. SpecAugment and variants of this algorithm, specifically designed to augment dysarthric speech (DSA), are applied to healthy speech data post the static

DA process. DSA algorithms are designed to embed speech characteristics typical of dysarthric speech into healthy speech data, transforming healthy speech features to create speech that acoustically resembles dysarthric speech.

SpecAugment
SpecAugment is a DA technique that is directly applied to the spectral speech features used for DNN training. The augmentation policy is designed to build a robust ASR by allowing for the prediction of changes to data in the time direction, partial loss of information in the frequency direction, as well as due to loss of small segments of speech [146]. Towards this end, masks are constructed to dynamically mask or modify the information in the time and frequency directions. The width and location of the masks are determined randomly, ensuring that the DNN is exposed to a different version of the input speech at every epoch of the training process.

Dysarthric SpecAugment (DSA)
We have leveraged our understanding of dysarthric speech to design three DSA policies specific to dysarthric speech. These masks have been applied only on healthy speech in a manner similar to the SpecAugment process described above.

A. **Stutter mask**
Stuttering is a speech disorder that manifests as either the arrested articulation of a syllable or clonic repetition of the same syllable [161]. Dysarthria is a motor-speech disorder, and the resulting speech has some of the characteristics of a stutter.

A mask was constructed along the time direction, wherein random and small segments of speech were repeated to emulate stuttering. Stutter mask was applied so that t consecutive time steps $[t0,\ t0 + t)$ were repeated, where the mask width t was randomly chosen from a uniform distribution such that $t \sim U(0, T)$ where T is the stutter mask parameter and $t0$ is chosen from $[0,\ \tau - t)$, where τ is the length of the utterance.

B. **Hypernasal mask**
Hypernasality is a consequence of velopharyngeal dysfunction (VPD) or velopharyngeal incompetence (VPI), which manifests as excessive nasal resonance in speech. It is the outcome of improper closure of the soft palate that regulates the airflow between the oral and nasal cavities. Hypernasality is a common occurrence in motor-speech disorders such as dysarthria [172]. Hypernasal speech exhibits significantly higher energy levels at frequency bands centered at 630, 800, and 1000 Hz and significantly lower amplitude for the band centered at 2500 Hz as compared to healthy speech [114, 112].
In order to simulate hypernasality in healthy speech, the SpecAugment was modified along the frequency direction. Channels of the Mel filter bank corresponding to the frequency regions from 600 to 1600 Hz were identified as the first region and the second region corresponding to the frequencies around

2500 Hz. The energy level in f consecutive mel frequency channels around the first region was increased by three times and the amplitude in the second region was reduced by half. f consecutive Mel frequency channels for modification [*f0, f0 + f*) were randomly chosen from a uniform distribution such that $f \sim$ *U(0,F)* where F is the hypernasal mask parameter and $f0$ is chosen from $[\nu 1, \nu 2 - f)$, where $\nu 1$ and $\nu 2$ correspond to the first and last Mel channels of the corresponding region.

C. **Breathiness (Noise) mask**
Dysarthria is often associated with disturbances of respiration, laryngeal function, airflow direction, and articulation, resulting in breathy speech quality and reduced intelligibility. Breathiness is typically caused by glottal air leakage, and acoustic measures related to breathiness are often used to distinguish between different physiological phonation conditions for pathological speech [47].
A scaled white noise mask was applied to healthy speech in order to replicate the presence of breathiness along both the time and frequency directions. The initial point and the width of the mask were chosen randomly as described in the stutter and hypernasal sections.

6.2.2.3 Transformer Models for ASR

Researchers are increasingly turning to Transformer models for ASR tasks. These models offer two key benefits: the ability to process information in parallel (parallelization) and an internal mechanism for focusing on important parts of the input, i.e., the attention mechanism. Unlike recurrent neural networks (RNNs), which process information sequentially, Transformers can learn faster. End-to-end (E2E) automatic speech recognition is a recent advancement in ASR that leverages the power of neural networks. It utilizes a single, unified model trained at a lower audio frame rate, significantly simplifying the training process. This translates to faster learning times, quicker decoding, and the possibility of jointly optimizing the system for tasks like natural language understanding after recognizing the speech. Conventional E2E models for speech recognition rely on a simple setup: one encoder, one decoder, and an attention mechanism. The encoder transforms the raw audio features (vectors) into a different format. The decoder then predicts the word sequence based on this new representation. The attention mechanism helps the decoder focus on crucial parts of the encoded information for each word prediction. In contrast, Transformer models are more complex. They can have multiple encoders and decoders, each equipped with an internal attention mechanism. This allows the model to capture relationships between different parts of the speech signal more effectively, leading to potentially better recognition accuracy [140] as shown in Figure 6.3.

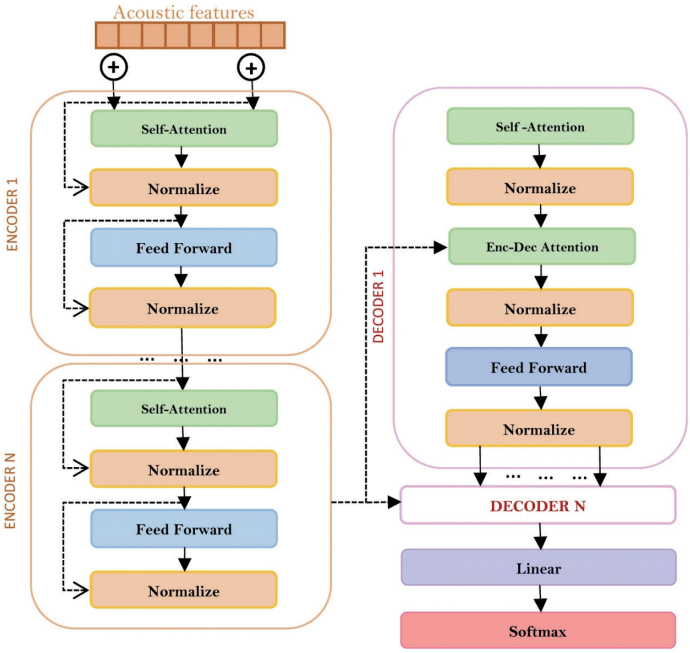

Figure 6.3: Transformer architecture [140]

6.2.3 Experimental setup

6.2.3.1 Data

Data from the Universal Access (UA) speech corpus [95] was used for training the End-to-End(E2E) Transformer-based ASR. The UA dysarthric speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. Data was collected in three separate sessions for each speaker and categorized into three blocks B1, B2, and B3. The speech material contains 155 words that are common to all three blocks and 100 words that are distinct for each block. Blocks B1, B2, and B3 from healthy speakers and blocks B1 and B3 from dysarthric speakers were treated as training sets, and block B2 from dysarthric speakers was treated as the test set. We have not included uncommon words in the test setup. The corpus also includes speech intelligibility ratings for each dysarthric speaker, as assessed by five naive listeners.

6.2.3.2 Hardware and Training Configuration

The configuration used for ESPnet training is as shown in Figure 6.4. We use 12 encoder and 2 decoder layers. Details pertaining to attention, CTC as well

optimization used are as seen in Figure 6.4.

```
# network architecture
# encoder related
elayers: 12
eunits: 2048
# decoder related
dlayers: 2
dunits: 1024
# attention related
adim: 256
aheads: 4

# hybrid CTC/attention
mtlalpha: 0.3

# label smoothing
lsm-weight: 0.1

# minibatch related
batch-size: 32
maxlen-in: 512  # if input length  > maxlen-in, batchsize is automat
maxlen-out: 150 # if output length > maxlen-out, batchsize is autom

# optimization related
sortagrad: 0 # Feed samples from shortest to longest ;
opt: noam
accum-grad: 2
grad-clip: 5
patience: 0
epochs: 20
dropout-rate: 0.1

# transformer specific setting
backend: pytorch
model-module: "espnet.nets.pytorch_backend.e2e_asr_transformer:E2E"
transformer-input-layer: conv2d     # encoder architecture type
transformer-lr: 10.0
transformer-warmup-steps: 25000
transformer-attn-dropout-rate: 0.0
transformer-length-normalized-loss: false
transformer-init: pytorch
```

Figure 6.4: ESPnet E2E training configuration

The experiments were conducted using a single NVIDIA GPU. CUDA, driver, and other details regarding the GPU are provided in Figure 6.5.

Noam optimizer introduced in the article [197] was used during the training phase as shown in the training configuration (Figure 6.4). Noam optimizer has a warm-up period and then an exponentially decaying learning rate as shown in Figure 6.6 [1].

[1] https://nn.labml.ai/optimizers/noam.html

```
| NVIDIA-SMI 460.91.03    Driver Version: 460.91.03    CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  GeForce RTX 2070    Off  | 00000000:01:00.0 Off |                  N/A |
| N/A   51C    P8     2W /  N/A |    696MiB /  7982MiB |      5%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
```

Figure 6.5: NVIDIA GPU details

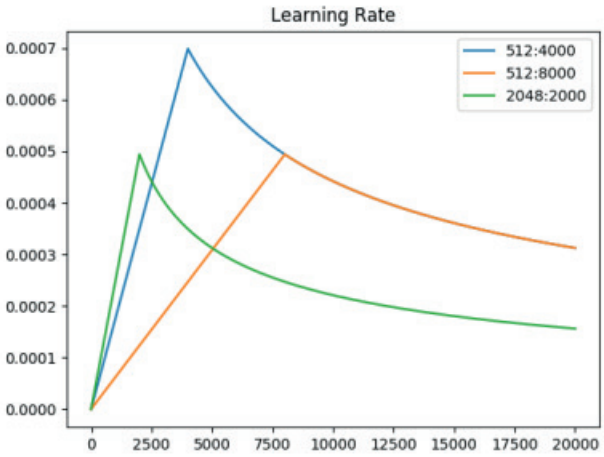

Figure 6.6: noam optimizer learning rate

6.2.3.3 Static Augmentation

ESPnet toolkit [204] was used as the E2E-Transformer-based [86] system to evaluate the ASR performance for DA of dysarthric speech, along with a word-based language model. The training was conducted for **20 epochs**.

SDA - SoX
SoX was used for speed, tempo, and loudness perturbations. The options and factors are as mentioned below:

- *speed* option for speed perturbation with the factors 0.9, 1.0, and 1.1
- *tempo* option for tempo perturbation with the factors 0.7, 0.5, and 0.4 based on the factors mentioned in [14]
- *vol* option for loudness perturbation with the factors 0.7, 0.5

SDA-female speech
It has been observed that the UA Speech corpus has fewer female speakers as

compared to male speakers across intelligibility levels in both healthy control and dysarthric speaker categories, as shown by Table 6.8. This skewness in data impacts the ASR performance of female speakers. The DA techniques described in Section 6.2.2.1 have been applied to male speaker data to address this skewness.

Table 6.8: Gender-wise data distribution in the UA Speech Corpus

Category	**Female**		**Male**		**% Female**
	Speakers	**Utterances**	**Speakers**	**Utterances**	**Speakers**
Healthy	4	21440	9	48240	30.77
Dysarthric	4	21440	11	52585	26.66

- Pitch shifting: Pitch shifting of male speakers' speech data was done, by a factor of 2 and 4 semitones using the techniques mentioned in [87].
- VTLP: One way of perturbing data is to transform spectrograms, using a randomly generated warp factor α to warp the frequency dimension, such that a frequency f is mapped to a new frequency f$\prime$ via a function of α [72]. VTL perturbed data was created using the python toolkit **nlpaug** random warp factor in the range of 0.9,1 was used.
- Spectral warping was applied using three parameters, 0.1, 0.15, and 02, based on the study [88].
- Voice conversion of healthy female speakers was carried out using ESPnet-TTS toolkit [61], which is an end-to-end text-to-speech (E2E-TTS) toolkit, which is an extension of the open-source speech processing toolkit ESPnet [204]. The toolkit supports state-of-the-art E2E-TTS models, including Tacotron 2, Transformer TTS, and FastSpeech. We have x-vector-based TTS for voice conversion, using the *libritts* pre-trained model with the model tag *kan-bayashi/libritts_xvector_vits* from the ESPnet model zoo. Voice conversion was limited to two female speakers from the pre-trained models as target speakers.

While the first three methods are algorithm-based, voice conversion uses DNNs and pre-trained models of the target speakers and hence can be categorized as non-explainable feature-based augmentation.

Devoicing of consonants

The distribution of each target voiced stop consonant as a percentage of the total target consonants in the UA speech corpus is as shown in Table 6.9. The percentage of /b/ is lower as compared to the rest of the phonemes. Also, there are no isolated words present in the database that comprise both /b/ and any of the other consonants in Table 6.9. These two factors indicate that the majority of the audio generated using mixed-devoicing will be the same as full-devoicing audio. Three

Table 6.9: Distribution of target voiced stop consonants for devoicing

Phoneme	% Occurrence
/b/	21
/d/	43
/dh/	23
/g/	11
/jh/	2

different types of devoicing parameters, namely partial, full, and mixed, were applied to healthy control speaker data. The location of voiced stop consonants such as /g/, /jh/, /d/, /D/, and /b/ within the audio was determined using AS R-based forced alignment. The ASR used was trained on librispeech corpus.

- Full devoicing: Applied on all stop consonants by replacing the voicing prior to the burst location of the voiced consonant with noise
- Partial devoicing: Applied on all stop consonants by replacing the reducing the voicing prior to the burst location of the voiced consonant by a degree of 0.5
- Mixed devoicing: Partial devoicing was applied on /b/, whereas full devoicing was applied on the rest of the voiced stops.

It was found that the UA speech corpus data considered in our experiments did not comprise any isolated word that had the labial voiced consonant /b/ and another voiced stop in the same word.

6.2.3.4 Dynamic Augmentation

SpecAugment is the baseline SpecAugment method as discussed in [146]. SpecAugment was applied on both healthy as well as dysarthric speech. Three different dysarthric SpecAugment (DSA) techniques were devised as discussed in Section 6.2.2.2. The details of the configurations for the DSA masks are provided in Table 6.10. Stutter, Hypernasal, and Breathiness masks were designed to augment healthy speech to match dysarthric speech. Hence, they were applied only to healthy speech as a part of the ASR training. This model was then used as a pre-trained ASR model for adaptation using dysarthric speech to arrive at a final WER.

The DA sequence and manner of application can be visualized as shown in Figure 6.2. In order to demonstrate the benefits of SDA, and DSA to DA of dysarthric speech, we have examined the performance of E2E ASR using each of the techniques separately based on the set-ups discussed below:

- Healthy speech for scenarios of no augmentation, with SDA and DSA.

Table 6.10: Masks used in Dynamic Data augmentation

System	Mask type
SpecAugment	time warp, time mask, freq mask
DSA-Stutter	time warp, noise mask, stutter mask
DSA-Hypernasal	time warp, time mask, hypernasal mask
DSA-Breathiness	time warp, frequency mask, noise mask
DSA-All	time warp, time mask, noise mask, stutter mask, hypernasal mask

- SDA and SpecAugment applied to dysarthric speech using the pre-trained acoustic model built using augmented healthy speech.
- Effect of different combinations of augmentation procedures along with model adaptation using dysarthric speech.

6.2.4 Results and Discussion

6.2.4.1 Static Augmentation

The E2E-Transformer-based ASR was trained on a combination of augmented healthy speech and dysarthric speech and evaluated on dysarthric speech. We present the baseline WER wherein no augmentation techniques were applied. We then proceed to present the ASR performance for static augmentation and dynamic augmentations separately. Finally, the ASR is evaluated for the two-stage augmentation.

Data augmentation-female speech
Several DA methods were looked into for augmenting female-speaker data as described in Section 6.2.2.1. A comparison of the performance of each of these methods is shown in Table 6.11. It can be seen from the table that each augmentation method leads to improved ASR performance. However, across all methods, the best improvement in performance was obtained for speech with high intelligibility. This can be attributed to the fact that only healthy control speech has been used for augmentation at this stage. Also, the greatest improvement was achieved with pitch shifting, across speakers with varying degrees of dysarthria severity, with an absolute overall improvement of 10.7%. The VTLP-based method showed the least improvement. It's important to note that VTLP alone might not be sufficient to convincingly convert a male voice to a female voice, as there are other factors influencing the perception of gender in speech. The pitch, or fundamental frequency ($F0$), is a crucial factor in gender perception. In general, female voices tend to have a higher pitch than male voices. This indicates that a combination of pitch-shifting techniques with other voice transformation methods will be beneficial to create a more convincing male-to-female voice conversion. Based on the ASR performance,

we proceed with pitch-shifting-based DA.

Devoicing of consonants

Devoicing of voiced stop consonants, such as b, d, or g was carried out as explained in Section 6.2.2.1. The results indicate that mixed devoicing works best amongst all three variations of devoicing. It can be noted that the WER for mixed devoicing is not significantly better than full devoicing. This can be attributed to the fact that the majority of the mixed-devoiced utterances are the same as fully-devoiced, as explained in Section 6.2.3 A small improvement of 1.7% was achieved when this method was used for augmentation as shown in Table 6.12.

Table 6.11: Impact of DA on the WER for female dysarthric speech

Speaker	Intelligibility	No Augmentation	Pitch Shifting	VTLP	Spectral Warping	Voice Conversion
F02	Low	78.1	74.8	79.4	77.4	77.7
F03	Very low	97.4	96.8	96.8	95.8	96.1
F04	Mid	69.2	56.6	62.9	62.3	61.9
F05	High	48.1	21.9	49.7	23.2	24.8
Average		73.2	62.53	72.2	64.68	65.13

Table 6.12: Impact of Devoicing-Based DA on WER

Devoicing	WER
No Augmentation	68.0
Partial	68.3
Full	66.7
Mixed	66.3

We summarize our findings for static DA (SDA) in Table 6.13. It is clearly visible from the table that both pitch-based and devoicing-based static augmentation techniques contribute significantly to the reduction of the WER of the ASR for dysarthric speech.

Table 6.13: Impact of Static DA on WER

Augmentation	Overall WER	Intelligibility			
		Very Low	Low	Mid	High
No Augmentation	68.0	97.03	85.42	71.43	43.03
Speed, Tempo Volume	63.0	95.61	81.16	64.54	36.90
+Pitch	61.4	94.19	80.00	64.15	34.32
+Devoicing	60.1	94.06	78.58	62.32	32.84
Relative Improvement	11.6	3.06	8.0	12.8	23.7

6.2.4.2 Dynamic Augmentation

Further, as a part of the proposed approach of two-stage DA, DSA policies are applied to the healthy speech data. Table 6.14 demonstrates the impact of DSA on the performance of the ASR, wherein the highest reduction in WER can be seen when all the DSA policies are applied.

Table 6.14: Impact of dysarthric SpecAugment on WER

System	**Speaker Ind.**	**Speaker dependent**	**Intelligibility**			
			Very Low	**Low**	**Mid**	**High**
No Augmentation	68.0	36.6	78.9	40.8	36.1	16.1
SpecAugment	65.5	35.0	72.0	36.0	33.8	16.8
DSA-Stutter	64.6	32.9	75.8	35.5	29.9	14.0
DSA-Hypernasal	64.2	30.8	75.8	29.1	27.4	13.1
DSA-Breathiness	63.1	30.4	75.2	28.0	27.0	13.0
DSA-All	62.1	**29.0**	73.6	25.1	25.1	12.7
Relative Improvement	8.7	**20.8**	6.7	38.5	30.5	21.1

As mentioned in Section 6.2.3, dynamic augmentations (DSA) specific to dysarthric speech have been applied only on healthy speech data, in order to achieve maximum matching between training and test data. The E2E-Transformer models are trained using DSA-applied healthy speech followed by model adaptation using dysarthric speech data. Tables 6.14 and 6.16 demonstrate that applying DSA improves the ASR performance across all the DSA techniques. We achieve the lowest WER when all the DSAs are applied together in the final DSA-All system. An absolute improvement of 5% is achieved for healthy data and 7.6% post-model adaptation. We also examine the improvement in WER at dysarthria intelligibility levels, that the UA dysarthric speech corpus has provided. While both SDA and DSA improve the WER across intelligibility levels, we note that SDA gives higher improvement for dysarthric speech. It can be observed from Tables 6.14 and 6.16 that each of the augmentation techniques applied contributes to the matching of augmented healthy speech to dysarthric speech.

The overall system comprises both static and dynamic augmentations in cascade, as shown in the figure 6.2. For the final system, a combination of (1) as step 1, healthy speech SDA followed by DSA-All and (2) adaptation of the model from step 1 using dysarthric speech SDA followed by SpecAugment has been used to improve the overall WER of the E2E-Transformer ASR. Table 6.15 highlights the reduction WER for both speaker-independent and speaker-dependent systems when DA policies are applied. It can also be observed from Tables 6.13, 6.14, and 6.16 that the improvement is across all intelligibility levels. Highest WER improvement using SDA is achieved for high intelligibility dysarthric speech while DSA achieves better ASR performance for low and mid intelligibility levels, the cascading effect

of which can be seen in 6.16. Please note that all the experimental results reported in Tables 6.11, 6.9, 6.13, 6.14, 6.15 and 6.16 correspond to a training set-up of 20 epochs. While DA plays a key role in the ASR outcome, significant gains were achieved using model adaptation using dysarthric speech. It can be seen from Table 6.16 that WER has improved at each of the four intelligibility levels provided by the UA Speech corpus We achieve an absolute improvement of **10.7%** in word error rate (WER) over a baseline with no augmentation, with a final WER of **25.9%**. These results can be benchmarked against some of the recent studies that have used data augmentation techniques to improve ASR performance and evaluated them on the UA Speech corpus. Time and tempo-stretching of healthy speech was used for data augmentation by [196] and achieved the best WER of 24.82% when evaluated on a part of the UA speech corpus. Authors used speaker-dependent acoustic models based on phoneme-level speech tempo ratio between typical and speaker-specific dysarthric speech to augment existing dysarthric speech and achieved a WER of 27.88% [207]. Speed, tempo, and VTLP-based augmentation followed by LHUC-based speaker adaptive and multi-task learning (MTL)-based training for DNN carried out in [51] reported a WER of 26.37%. [75] achieve a WER of 27.78% using Wav2vec 2.0 embedding features along with VAE-GAN and other techniques.

Table 6.15: Impact of two-stage DA on WER

System	**Speaker Ind.**	**Speaker Dependent**
No Augmentation	68.0	36.6
SDA	60.1	28.0
DSA-All	62.1	29.0
SDA + DSA-all	55.4	25.9
Relative Improvement	18.5	29.2

Table 6.16: Impact of two-stage DA on WER at 4 Intelligibility levels

System	**Overall WER**	**Intelligibility**			
		Very low	**Low**	**Mid**	**High**
No Augmentation	36.6	78.9	40.8	36.1	16.1
SDA	28.0	67.0	28.5	21.5	11.3
DSA-All	29.0	73.6	25.1	25.1	12.7
SDA + DSA-all	25.9	60.5	24.4	21.0	11.5
Relative Improvement	29.2	23.3	40.2	41.8	28.6

t-SNE, or t-Distributed Stochastic Neighbor Embedding, is a powerful technique for visualizing high-dimensional data in a lower-dimensional space. By analyzing

the t-SNE plots of transformer decoder output features at different epochs and with/without data augmentation, we can gain insights into the model's learning process and the impact of data augmentation. Figure 6.7 t-SNE plots across 20 epochs for two different training configurations namely, (a) Only healthy data with no augmentation (68% accuracy) and (b) Augmented healthy and dysarthric data (accuracy 25.9%). The formation of distinct clusters in the t-SNE plots indicates that the model has learned to distinguish between different categories or classes in the data. The density of points in the feature space provides insights into the distribution of data points and the model's ability to capture fine-grained distinctions. Key Observations from the plots in Figure 6.7 are as below:

- Evolution of Feature Space with Epochs:
  - Early Epochs (1): At the beginning of training, the feature space is more dispersed and less structured. This indicates that the model has not yet learned to represent the underlying data distribution effectively.
  - Intermediate Epochs (10, 15): As the model trains, the feature space becomes more compact and clusters start to emerge. This suggests that the model is learning meaningful representations of the data.
  - Later Epochs (20): In the later stages of training, the feature space becomes even more structured, with distinct clusters forming, indicating that the model has converged to a solution that captures the underlying patterns in the data.
- Impact of Data Augmentation: Without data augmentation, the model struggles to generalize to unseen data, leading to overfitting. This can be observed in the t-SNE plots as a more tightly clustered feature space, indicating that the model has learned specific patterns in the training data but may not be able to recognize variations. Data augmentation introduces artificial diversity into the training data, helping the model to learn more robust and generalizable representations. In the t-SNE plots, this can be seen as a more dispersed feature space with less distinct clusters. This suggests that the model has learned to capture a wider range of variations in the data.

Figure 6.8 depicts the neural network training loss for the speaker-independent (SI) and speaker-dependent (SD) systems with no augmentation policies applied and after the application of both static and dynamic augmentation. The loss diagram for each of the training cycles indicates a reduction in the training loss over the subsequent epochs. It can be seen that the reduction in the training loss for SI systems is slower as compared to the SD systems.

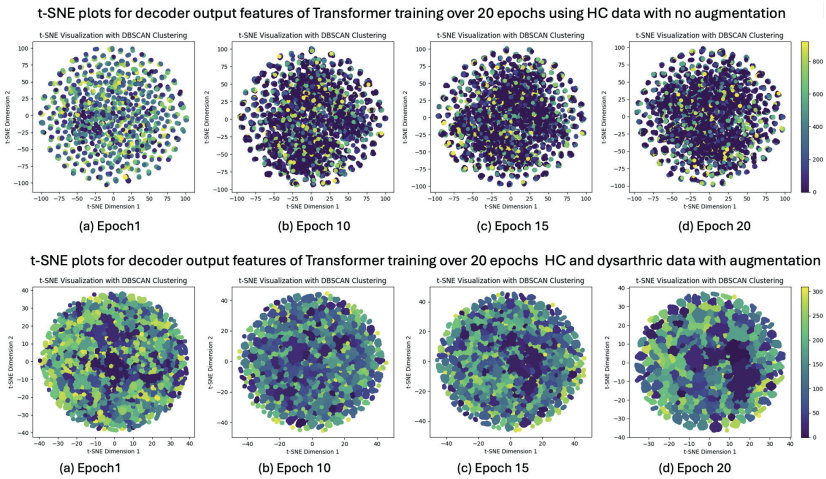

Figure 6.7: t-SNE plots for visualizing transformer-decoder outputs

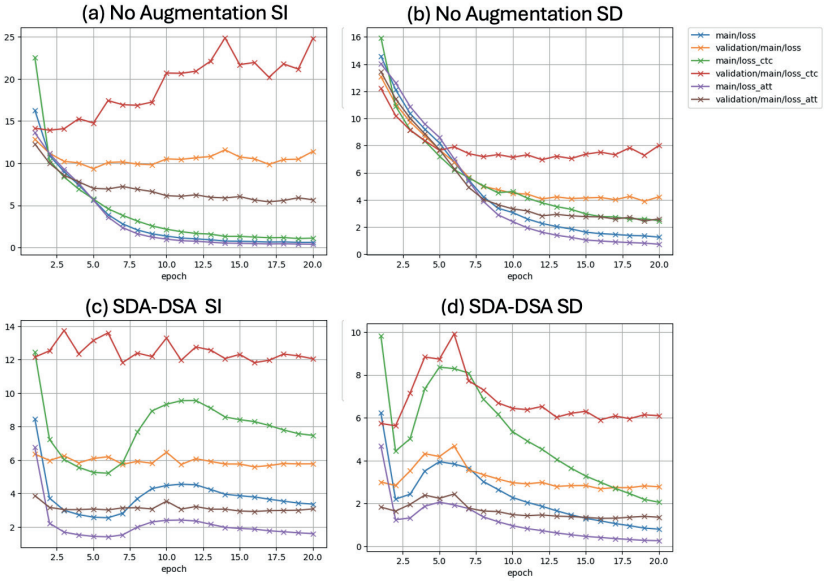

Figure 6.8: Training loss (a) No Augmentation SI (b) No Augmentation SD (c) SDA-DSA SI (d) SDA-DSA SD

6.2.5 Conclusion and Future work

This work proposes a novel two-stage DA scheme to improve ASR performance in terms of WER. This approach tackles the challenge of limited dysarthric speech data, significantly improving ASR performance for individuals with speech impairments. The objective is to leverage healthy speech data, applying static augmentations like speed, tempo, and pitch modifications alongside voice conversion techniques to address the gender imbalance often present in dysarthric corpora. Additionally, dynamic augmentations are introduced through dysarthric-specific masks like stutter and breathiness, incorporated during training with SpecAugment as the baseline. This two-stage approach, coupled with efficient Transformer-based E2E ASR models, significantly reduces word error rate compared to using unaugmented data. Notably, the method demonstrates substantial improvement when further fine-tuned on actual dysarthric speech. By augmenting training data with modified healthy speech to match the dysarthric speech characteristics, this technique effectively addresses the data scarcity issue, paving the way for more accurate ASR systems specifically designed for individuals with dysarthric speech. We observe that augmenting female speech improves the ASR performance significantly while devoicing of consonants provided marginal improvement. Both SDA and DSA contribute to the improvement of ASR performance individually and when used in a two-stage sequence across all intelligibility levels, for both speaker-independent and speaker-dependent configurations with an absolute and relative improvement of 10.7% and 29.2% respectively for a speaker-dependent scenario. It may be worth exploring the concepts of *Sequence-to-sequence learning (Seq2Seq)* and *Generative adversarial networks (GAN)* for DA, and thereby a robust ASR for dysarthric speech.

Chapter 7

Conclusion and Discussion

This thesis has explored the application of speech technology for dysarthric speech, focusing on two key aspects: automatic intelligibility assessment and automatic speech recognition. These advancements have the potential to significantly improve assisted speech therapy and develop assistive living technologies for individuals with dysarthria.

The proposed system, as illustrated in Figure 1.2, integrates modules for Automatic Intelligibility Assessment (AIA) and Automatic Speech Recognition (ASR) for dysarthric speech. The intelligibility information obtained from the AIA module serves as a crucial input for designing more efficient ASR systems.

7.1 Answering the Research Questions

We address the three key research questions raised in Chapter 1, section 1.3 through our contributions to dysarthric speech analysis and recognition outlined in this thesis.

7.1.1 State of the art (SOTA)

RQ1: What is the status of research into the interdisciplinary area of dysarthric speech and speech technology and which knowledge gaps should be addressed?

In Chapter 2, we have presented an overview of the contributions of speech technologists to the area of dysarthric speech, focusing on acoustic analysis, speech features, and techniques used. It also discusses the challenges of limited data availability, variability of dysarthria, and lack of standardized metrics. While there has been significant progress in understanding and addressing dysarthria through advancements in speech analysis, speech recognition, and assistive technologies, there is still a great deal of potential for improvement. We recommend that future research should focus on standardization of data collection to facilitate collaboration and development of more generalizable models, and multi-modal assessment to

provide a more comprehensive understanding of dysarthria and informed treatment strategies. Additionally, some of the key contributions of the studies that have been discussed have the potential to translate into AI-powered speech analysis for developing automated tools for dysarthria assessment, natural language processing (NLP) for augmentative and alternative communication (AAC) to create more intuitive and user-friendly AAC systems, teletherapy and remote monitoring to improve access to care and enable personalized interventions for persons with dysarthria. To make longer and faster strides, researchers typically rely on existing research and data on a global scale. Therefore, it is imperative to consolidate the existing research and present it in a form that can serve as a basis for future work. In this chapter, we have reviewed the contributions of speech technologists to the area of dysarthric speech with a focus on acoustic analysis, speech features, and techniques used. By focusing on the existing research and future directions, researchers can develop more effective tools and interventions to improve communication, quality of life, and overall well-being for people with dysarthria

7.1.2 Automatic Intelligibility Assessment

RQ2: How can we efficiently and automatically assess the intelligibility of dysarthric speech?

Chapter 3 comprises two parts, in which we have presented our research work that delves into the automatic intelligibility assessment of dysarthric speech. In the first part, we examined methods to identify the intelligibility of dysarthric speech. We evaluated multiple feature sets and classification techniques and have highlighted the effectiveness of multi-taper spectral estimation and neural network-based approaches in improving classification accuracy. Unlike traditional methods that rely on linguistic features, we use audio descriptors, which are more objective and less language-dependent.

In the second part, we further explore efficient identification of intelligibility, focusing on BLSTM networks trained with various features and transfer learning techniques. We address challenges such as noise in recordings and gender representation in datasets. In this work, we have developed a robust model for the intelligibility assessment of dysarthric speech, provided insights into the features that are important for distinguishing intelligible and non-intelligible speech, and demonstrated the potential of transfer learning for improving the performance of speech recognition systems for dysarthric speakers. These studies demonstrate the effectiveness of using Artificial Neural Networks (ANN) and Bidirectional Long Short-Term Memory (BLSTM) networks with transfer learning for classifying dysarthric speech intelligibility. By accurately predicting intelligibility, researchers and clinicians can develop targeted interventions to improve communication for individuals with dysarthria.

7.1.3 Improved ASR Performance for Dysarthric Speech

RQ3: How can we improve the automatic recognition of dysarthric data in terms of word error rate?
In Chapters 4, 5, and 6 we have presented various techniques to improve ASR performance for dysarthric speech. These techniques address the challenges posed by the unique acoustic characteristics of dysarthric speech. Each of the aforementioned chapters consists of two peer-reviewed articles.

7.1.3.1 Acoustic Feature Exploitation

In Section 4.1, the impact of different feature sets (MFCC, MT-MFCC, jitter, shimmer) on speech recognition performance was investigated. GMM-HMM and DNN-HMM systems were compared to evaluate their effectiveness for dysarthric speech. Furthermore, we analyzed the relationship between feature sets, speech recognition systems, and dysarthria severity. Key outcomes of this work were (1) Providing insights into the optimal feature set for dysarthric speech recognition. (2) Comparing the performance of different speech recognition systems for this challenging task. (3) Identifying the impact of dysarthria severity on speech recognition accuracy.

In Section 4.2, we have described an innovative approach to address the challenges posed by varying speech rates in dysarthric individuals. We explored the concept of adjusting the tempo of speech to improve recognition accuracy and investigated the impact of tempo adaptation on different speech recognition systems such as GMM-HMM and DNN-HMM. We also analyzed the ASR performance at dysarthria severity levels. We demonstrated how tempo adaptation can enhance the performance of speech recognition systems, particularly for triphone models. By focusing on tempo adaptation, we have introduced a promising strategy for enhancing communication capabilities for individuals with dysarthria. Collectively, the two articles presented in Chapter 4 highlight the importance of both feature engineering and speech processing techniques in developing effective speech recognition systems for dysarthric individuals. The findings contribute to a deeper understanding of the challenges and potential solutions for improving communication for people with dysarthria.

7.1.3.2 Speech Feature Enhancement

Chapter 5 investigates the feature-domain enhancement of dysarthric speech using autoencoders. Section 5.1 focuses on enhancing dysarthric speech through tempo adaptation and DAE-based feature enhancement. This research work is significant because we have addressed the challenge of improving speech intelligibility for individuals with dysarthria by transforming the speech signal. By modifying speech characteristics of dysarthric speech, we aim to improve the raw input for speech recognition systems. A comparison of the effectiveness of both techniques offers valuable insights into their strengths and weaknesses, providing a foundation for

further research into speech enhancement for individuals with dysarthria.

In section 5.2 the significance of using TDNN-DAE to enhance dysarthric speech features, leading to substantial improvements in ASR accuracy, especially in recognizing dysarthric speech more akin to healthy control speech patterns has been investigated. We focus on evaluating DNN-HMM ASR systems in speaker adaptation (SA) and speaker-independent (SI) scenarios. The impact of TDNN-DAE-based enhancement of dysarthric speech features for ASR has been explored. A comparison of the recognition performance, when dysarthric speech data is enhanced using TDNN-DAE versus unenhanced data, has been presented. It was demonstrated that TDNN-DAE enhancement aligns dysarthric speech features more closely with healthy control speech features. This alignment resulted in improved ASR performance. Different training configurations, including those using only healthy control data, only dysarthric data, and a combination of both enhanced dysarthric data and healthy control data were also investigated. An analysis of the ASR performance across different severity levels of dysarthria showed consistent improvements with TDNN-DAE enhancement. This analysis helps understand the effectiveness of the enhancement technique across varying degrees of speech impairment. Based on these findings, we suggest further exploration into optimizing front-end configurations for ASR systems tailored to dysarthric speech, potentially leveraging TDNN-DAE and other enhancement techniques.

7.1.3.3 Data Augmentation

In Chapter 6 we address the limited availability of dysarthric speech data by proposing data augmentation techniques. These techniques involve transforming healthy speech data in the time domain and frequency domain to mimic dysarthric speech characteristics. The effectiveness of data augmentation is evaluated in both speaker-dependent and speaker-independent scenarios. In Section 6.1, we have focused on synthetic data generation for dysarthric speech using healthy speech. This was a novel approach to address the critical issue of limited available dysarthric speech data. We have explored tempo and speed modification techniques to generate artificial dysarthric speech samples that can be used to augment existing training data, potentially improving model performance. The synthetically generated dysarthric speech was classified into four severity classes based on modifications applied to healthy control utterances using different augmentation parameters and demonstrated that the severity of dysarthric speech correlates closely with the duration of the utterance. We show that alterations in speed and tempo impact the severity classification, with longer durations generally correlating with higher severity levels. The results indicated improvements in ASR accuracy when using tempo-based and speed-based augmentation techniques. Notably, higher severity levels showed significant absolute improvements over baseline performance.

In Section 6.2 we introduce a novel two-stage data augmentation (DA) scheme aimed at enhancing ASR performance, specifically leveraging the characteristics of

dysarthric speech. The approach involves applying static augmentations like speed, tempo, and pitch modifications and dynamic augmentations using dysarthric-specific masks such as hypernasality, stutter, and breathiness. Dynamic augmentations are incorporated during training using the SpecAugment baseline. We have applied techniques to handle gender imbalances in dysarthric corpora since this is crucial as dysarthric speech datasets often lack gender diversity. The two-stage DA approach, combined with efficient Transformer-based End-to-End (E2E) ASR models, resulted in a significant reduction in Word Error Rate (WER) compared to unaugmented data. The method showed substantial gains when fine-tuned on actual dysarthric speech, effectively overcoming data scarcity issues. Augmenting female speech and introducing devoicing of consonants were noted to improve ASR performance. Static and dynamic augmentations (SDA and DSA) individually contribute to performance improvements, with a substantial absolute and relative improvement observed in speaker-dependent scenarios.

Overall, in Chapter 6, we have presented two approaches to augmenting dysarthric training data to better match dysarthric speech characteristics, thereby significantly enhancing ASR accuracy and performance for individuals with dysarthria.

7.2 Limitations

While this research has made significant strides in improving speech technology for dysarthric speech, there are limitations to consider. These limitations not only acknowledge the current boundaries of this work but also pave the way for future research directions.

7.2.1 Data Availability

7.2.1.1 Limited Dysarthric Speech Data

A major challenge is the limited availability of dysarthric speech data. This restricts the ability to train and validate complex deep learning models, potentially hindering performance compared to models trained on abundant healthy speech data. We have presented the available dysarthric corpora in detail in Section 2.2. We discuss the limitations and how it has been addressed.

- Limited Number of Speakers: Dysarthric speech corpora often include data from a relatively small number of speakers, making it difficult to generalize findings across a broader population. Efforts have been made to include diverse speakers from various backgrounds and severities of dysarthria. For instance, the Universal Access (UA) dysarthric speech corpus includes data from multiple speakers with different severity levels, and the IDEA database includes 45 speakers affected by various pathologies.

- Skewed Data Distribution: Data tends to be skewed towards speakers with lower severity levels of dysarthria, with less representation of severe cases. There is a call for more inclusive speech corpora that cover a broader range of severity levels. Some corpora, like the *homeService* corpus, focus on collecting realistic dysarthric data from speakers with severe dysarthria over time.

- Time-Consuming and Tedious Data Collection: Collecting speech from dysarthric speakers, particularly those with severe dysarthria, is challenging and time-intensive due to muscle weakness and fatigue. Longitudinal studies such as the Neurospeech project [142] and *the homeService corpus* [138] are suggested to observe changes in speech quality and intelligibility over time. This approach helps in understanding the progression of the disorder and the impact of therapy and medication, which can inform the development of assistive technologies.

- Complex Characterization of Dysarthric Speech: Dysarthria results from various neurological disorders, making it complex to characterize and collect consistent data. Detailed databases like the TORGO include aligned acoustics and 3D articulatory features, providing comprehensive data that aids in the explicit learning of hidden articulatory parameters.

7.2.1.2 Speaker Variability

Dysarthria manifests differently depending on the underlying neurological condition and disease severity. The data used in our research might not fully capture this variability, potentially limiting the generalizability of our findings to a broader dysarthric population. Dysarthria stems from a variety of neurological disorders, including conditions like cerebral palsy, Parkinson's disease, ALS, MS, and TBI. This variability influences the speech patterns and characteristics observed in different speakers. The type of speech task used in data collection (e.g., reading, monologue, diadochokinetic evaluation) can also significantly affect the speech data. Simple reading tasks might not fully capture the variability and challenges faced by dysarthric speakers compared to more complex tasks like conversations or monologues. While dysarthric speech corpora have been developed for multiple languages, including French, Korean, Cantonese, Dutch, Tamil, Spanish, Czech, and German, the variability in dysarthric speech across different linguistic and cultural contexts poses challenges in designing tools for automatic analysis of dysarthric speech that can be applicable across these contexts.

7.2.2 Technical Limitations

- Accuracy of Automatic Techniques: While our research proposes methods to improve automatic intelligibility assessment and ASR for dysarthric speech, these techniques might not achieve perfect accuracy. Errors in these systems

could lead to misinterpretations of dysarthric speech, hindering the effectiveness of communication aids, and making it difficult for dysarthric speakers to convey their intended messages. This can affect their ability to engage in everyday activities and interactions, impacting their quality of life. Continuous improvement in machine learning models, particularly in handling variations in speech due to different types and severities of dysarthria, is necessary. Techniques such as transfer learning, where models trained on typical speech are fine-tuned on dysarthric speech, and incorporating more robust feature extraction methods can help enhance accuracy.

The limited amount of dysarthric speech data available for training ASR systems can constrain the performance of these models. Insufficient and imbalanced training data can result in models that are not generalizable, performing poorly when encountering speech patterns that differ from those seen during training. Increasing the size and diversity of dysarthric speech corpora, and employing data augmentation techniques to create varied training examples, can help mitigate this limitation.

- Real-world applicability: The controlled settings used in our research might not fully translate to real-world scenarios, where background noise, varying communication contexts, and environmental factors are present. Systems that perform well in controlled settings may fail to maintain their performance in real-world scenarios, reducing their practical utility for dysarthric speakers. Testing and training systems in more diverse and realistic environments, including home settings and public spaces, can help improve their robustness. Incorporating noise reduction techniques and context-aware models can also enhance real-world performance.

  The high variability in dysarthric speech due to differences in the underlying neurological conditions, severity levels, and individual speaker characteristics poses a challenge for automatic systems. A system optimized for one type or severity of dysarthria may not perform well for another, limiting its effectiveness for a broader range of users. Developing adaptive models that can dynamically adjust to different speakers and conditions, and using personalized models trained on individual users' speech data, can help address speaker variability.

  Further, real-time processing requirements for ASR and intelligibility assessment systems necessitate low-latency responses, which can be computationally intensive. High latency or computationally demanding systems may not be feasible for real-time use, especially on resource-constrained devices like mobile phones or wearable technologies. Optimizing algorithms for efficiency, using lightweight models, and leveraging edge computing to distribute processing tasks can help reduce latency and computational demands.

  Lastly, the metrics used to evaluate the performance of ASR and intelligibility

assessment systems may not fully capture the nuances of dysarthric speech. Traditional metrics like word error rate (WER) may not adequately reflect the intelligibility improvements needed for effective communication by dysarthric speakers. Developing more comprehensive evaluation metrics that consider the intelligibility, naturalness, and usability of the output, as well as user satisfaction, can provide a better assessment of system performance.

Despite these limitations, the presented research offers a valuable contribution to the field of speech technology for dysarthric speech. By addressing the research questions and proposing innovative techniques, this work paves the way for further advancements. As can be seen from Chapter 2, our article published in January 2025, very little has been done for dysarthric speech. Starting from 2016, our attempt has been to apply the latest techniques in speech processing including the leaps in AI to recognize and analyse dysarthric speech. We have explored multiple strategies, including deep neural networks (DNN-HMM), time-delay neural networks (TDNN-DAE), and advanced data augmentation techniques to enhance ASR performance for dysarthric speech. These approaches have demonstrated consistent improvements in both AIA and ASR for dysarthric speech. While the foundational studies in this thesis were conducted between 2016 and 2020, their relevance remains strong because our findings provide a critical benchmark for future developments, offering methodologies that continue to be applicable as AI models evolve. The research bridges the gap between conventional ASR and the specialized needs of dysarthric speakers, ensuring that advancements in deep learning and speech technology are inclusively designed to support those with speech impairments. While AI research progresses rapidly, the novel contributions of this work remain highly pertinent. This is borne out by the results presented in our 2025 overview (Chapter 2). Future studies can build upon the methodologies presented here, including the development of a unified system, incorporating additional information, exploring advanced deep learning architectures, and pursuing real-world applications as described in the following Section 7.3. Addressing the evolving landscape of AI while maintaining a focus on dysarthric speech is crucial for ensuring that these technological advancements translate into tangible societal benefits.

7.3 Future Directions

The advancements presented in this thesis lay the groundwork for a future where speech technology empowers individuals with dysarthria to communicate effectively and confidently. However, there remains significant potential for further exploration and development. This Section outlines several avenues for future research that can build upon this foundation and push the boundaries of speech technology for dysarthric speech analysis and recognition.

7.3.1 Development of a Unified System

- User-friendly Interface: The system envisioned in Figure 1.2 can be built into a software application with an intuitive interface. This would allow speech pathologists to easily access features like intelligibility assessment and ASR for dysarthric speech. Additionally, the system could be adapted for use by individuals with dysarthria, providing features like text-to-speech conversion or speech output correction.

- Modularity and Customization: The system could be designed with modular components, allowing clinicians to choose the functionalities they need (e.g., intelligibility assessment only, ASR with a specific output format). Customization options could also be included, enabling adjustments for different user preferences or dysarthria severities.

7.3.2 Incorporation of Additional Information

- Speaker Characteristics and meta information: By incorporating data on speaker demographics (age, gender), the system could potentially account for natural variations in speech patterns and improve accuracy. Additionally, factoring in the type of dysarthria (spastic, flaccid, etc.) could allow for more targeted analysis and recognition techniques.

- Emotional State Recognition: Integrating emotional recognition capabilities would enhance the system's understanding of the speaker's intent and communication style. This could be particularly valuable for individuals with dysarthria who might struggle to convey emotions effectively due to their speech impairment.

7.3.3 Exploration of Deep Learning Architectures

- Advanced Network Designs: As the amount of dysarthric speech data grows, researchers can explore more sophisticated deep learning architectures specifically tailored for this domain. These architectures could leverage techniques like recurrent neural networks (RNNs) or transformers to capture the complex temporal and sequential nature of dysarthric speech.

- Transfer Learning and Multi-modal Learning: Transfer learning from pretrained models on healthy speech data can be a powerful approach for dysarthric speech analysis when limited data is available. Multi-modal learning, which incorporates visual information alongside audio, could also be explored to improve intelligibility assessment and recognition accuracy.

7.3.4 Real-World Applications

While real-world applications are the end objective of the research work presented in the current thesis, ethical and legal concerns regarding assistive technology (AT) for dysarthric speakers need to be addressed. The study by Shanmugam & Marimuthu (2021) [179] investigated the evolution of AT services, relevant laws, and their influence on the well-being of individuals with dysarthria. A comparative analysis of legal frameworks and professional organizations in the United States, India, and Europe has been presented. Additionally, the study also assessed the impact of AT on the quality of life of dysarthric patients.

- Communication Aids: The developed techniques can be integrated into communication aids for individuals with dysarthria. These aids could offer features like real-time speech intelligibility feedback, word prediction, or alternative communication methods for situations where speech recognition fails.
- Speech Rehabilitation Systems: The system's intelligibility assessment capabilities could be used as a feedback mechanism in speech rehabilitation programs. By providing objective data on speech intelligibility, the system could help individuals with dysarthria track their progress and tailor their therapy exercises.
- Dysarthria Screening Tools: The automatic intelligibility assessment module could be adapted for use as a dysarthria screening tool. This could allow for early detection of dysarthria, particularly in populations where access to speech pathologists might be limited.

By pursuing these future directions, speech technology has the potential to become a transformative tool for individuals with dysarthria. Advancements in speech technology hold immense promise for achieving a better quality of life for people with dysarthria. Improved automatic speech recognition (ASR) will ensure their words are understood, reducing frustration and fostering stronger social connections. Tools like real-time intelligibility feedback and alternative communication methods will lessen the communication effort, empowering them to participate actively in all aspects of life. This newfound ability to communicate effectively will translate to greater independence and participation. Individuals with dysarthria will have access to a wider range of employment opportunities, engage more fully in education and social interactions, and experience a significant improvement in their overall quality of life. Speech technology has the power to break down communication barriers and empower individuals with dysarthria to live life to the fullest.

Appendix A

Author contributions in publications

A.1 First Author

The author of this thesis is the first author of the following articles included in the current thesis

1. Bhat, C., Vachhani, B., & Kopparapu, S. K. (2016). Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper Spectral Estimation. In Interspeech (pp. 228-232). (Section 4.1)

2. Bhat, C., Vachhani, B., & Kopparapu, S. (2016). 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 18 (pp. 370-377). Springer International Publishing. (Section 4.2)

3. Bhat, C., Vachhani, B., & Kopparapu, S. K. (2017). Automatic Assessment of Dysarthria Severity Level using Audio Descriptors. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5070-5074). IEEE. (Section 3.1)

4. Bhat, C., Das, B., Vachhani, B., & Kopparapu, S. K. (2018). Dysarthric Speech Recognition Using Time-delay Neural Network Based Denoising Autoencoder. In Interspeech (pp. 451-455). (Section 5.2)

5. Bhat, C., & Strik, H. (2020). Automatic Assessment of Sentence-level Dysarthria Intelligibility using BLSTM. IEEE Journal of Selected Topics in Signal Processing, 14(2), 322-330. (Section 3.2)

6. Bhat, C., & Strik, H. (2024). Two-stage Data Augmentation for Improved ASR Performance for Dysarthric Speech - submitted to Computers in Biology and Medicine (Section 6.2)

7. Bhat, C., & Strik, H. (2024). Speech Technology for Automatic Recognition and Assessment of Dysarthric Speech: An Overview, Submitted to ASHA Journal of Speech, Language and Hearing Research (Chapter 2)

A.2 Second Author

The author of this thesis is the second author of the following articles included in the current thesis. The problem statement was designed by the author. She was directly involved in daily discussions in implementing the models, analyzing the results, and writing the paper.

1. Vachhani, B., Bhat, C., Das, B., & Kopparapu, S. K. (2017, August). Deep Autoencoder Based Speech Features for Improved Dysarthric Speech Recognition. In Interspeech (pp. 1854-1858). (Section 5.1)

2. Vachhani, B., Bhat, C., & Kopparapu, S. K. (2018, September). Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. In Interspeech (pp. 471-475). (Section 6.1)

Appendix B

Research Data Management

B.1 Datasets Overview

B.1.1 The Universal Access Dysarthric Speech Corpus

The UA Speech Corpus is a publicly available dataset designed for research on dysarthric speech recognition. It contains recordings of speakers with varying levels of dysarthria, providing word-level and sentence-level speech data. This corpus has been widely used for Automatic Speech Recognition (ASR) research.

B.1.1.1 Access:

The UA Speech Corpus can be accessed from the University of Illinois at Urbana-Champaign's official repository. Access typically requires a request and approval for research purposes.

B.1.1.2 Contact Information:

For access, visit the official website: `http://www.isle.illinois.edu/sst/data/ua-speech/`
Alternatively, access can be requested via email to Prof. Hasegawa-Johnson (`https://ece.illinois.edu/about/directory/faculty/jhasegaw`)

B.1.2 The TORGO Dysarthric Speech Database

The TORGO database contains recordings of dysarthric and healthy control speakers, offering detailed articulatory data alongside audio recordings. This corpus is useful for both ASR and speech therapy research.

B.1.2.1 Access:

TORGO database access requires submitting a formal request to the University of Toronto. Approved researchers are granted access for non-commercial purposes.

B.1.2.2 Contact Information:

For access, visit: `https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html`

B.2 Modifications Made for Research

The following modifications and enhancements have been applied to the datasets for experimental purposes in published studies:

B.2.1 Tempo-based Modification

- Reference: Bhat, C., Vachhani, B., & Kopparapu, S. (2016). 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 18 (pp. 370-377). Springer International Publishing.
- Description: Tempo adaptation techniques were applied to dysarthric speech to improve ASR performance.
- Tempo-based Modification Code: `https://www.mathworks.com/matlabcentral/fileexchange/45441-phase-vocoder`

B.2.2 Feature-Level Enhancements

1.
   - Reference: Vachhani, B., Bhat, C., Das, B., & Kopparapu, S. K. (2017, August). Deep Autoencoder Based Speech Features for Improved Dysarthric Speech Recognition. In Interspeech (pp. 1854-1858)
   - Description: Deep Autoencoders were used for feature enhancement, and the Kaldi recipes were modified to integrate these enhancements.
2.
   - Reference: Bhat, C., Das, B., Vachhani, B., & Kopparapu, S. K. (2018). Dysarthric Speech Recognition Using Time-delay Neural Network Based Denoising Autoencoder. In Interspeech (pp. 451-455).
   - Description: TDNN-based Deep Autoencoders were used for feature enhancement, and the Kaldi recipes were modified to integrate these enhancements.

B.2.3 Data Augmentation

1. • Reference: Vachhani, B., Bhat, C., & Kopparapu, S. K. (2018, September). Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. In Interspeech (pp. 471-475).
   • Description: Healthy speech data was augmented by applying tempo and time-stretching methods to simulate dysarthric speech patterns.
   • Rubberband toolkit: `https://breakfastquay.com/rubberband/#:~:text=Rubber%20Band%20Library&text=It%20permits%20you%20to%20change,any%20desktop%20or%20mobile%20platform`
2. • Reference: Bhat, C., & Strik, H. (2024). Two-stage Data Augmentation for Improved ASR Performance for Dysarthric Speech - submitted to Computers in Biology and Medicine.
   • Description: ESPNet-based code was used to implement a two-stage augmentation pipeline.

B.3 Data Availability

The modified versions of the datasets, along with scripts and configurations used for experimentation, are not publicly distributed due to restrictions from the original dataset agreements. However, researchers can replicate the experiments by accessing the original datasets and applying the methods described in the respective publications.

B.4 Ethical Considerations

All modifications and experiments comply with the ethical guidelines and licensing agreements of the respective datasets. Data use is strictly for research purposes, and no commercial applications have been developed.

Contact for Further Details
For queries regarding the modifications or access to experiment scripts, please contact:
- Chitralekha Bhat: chitralekha.bhat@gmail.com

Bibliography

[1] https://www.mndassociation.org/professionals/management-of-mnd/dysarthria. Viewed July 2024. 8

[2] Mariya Celin T A, Nagarajan T, and Vijayalakshmi P. Dysarthric speech corpus in Tamil for rehabilitation research. In *2016 IEEE Region 10 Conference (TENCON)*, pages 2610–2613, 2016. URL https://doi.org/10.1109/TENCON.2016.7848510. 17, 20, 40

[3] Bassam Ali Al-Qatab and Mumtaz Begum Mustafa. Classification of Dysarthric Speech According to the Severity of Impairment: an Analysis of Acoustic Features. *IEEE Access*, 9:18183–18194, 2021. URL https://doi.org/10.1109/ACCESS.2021.3053335. 28, 45

[4] Md Jahangir Alam, Patrick Kenny, and Douglas O'Shaughnessy. A Study of Low-variance Multi-taper Features for Distributed Speech Recognition. In Carlos M. Travieso-González and Jesús B. Alonso-Hernández, editors, *Advances in Nonlinear Speech Processing*, pages 239–245, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25020-0. URL https://doi.org/10.1007/978-3-642-25020-0_31. 75

[5] Md. Jahangir Alam, Patrick Kenny, and Themos Stafylakis. Combining amplitude and phase-based features for speaker verification with short duration utterances. In *Proc. Interspeech 2015*, pages 249–253, 2015. doi: 10.21437/Interspeech.2015-94. URL https://doi.org/10.21437/Interspeech.2015-94. 75

[6] Kwanghoon An, Myungjong Kim, Kristin Teplansky, Jordan Green, Thomas Campbell, Yana Yunusova, Daragh Heitzman, and Jun Wang. Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks. In *Proc. Interspeech 2018*, pages 1913–1917, 2018. URL https://doi.org/10.21437/Interspeech.2018-2496. 28, 45

[7] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 2,

pages 1137–1140 vol.2, 1996. URL https://doi.org/10.1109/ICSLP.1996.607807. 80, 88

[8] Ahmet S. Asan, James R. McIntosh, and Jason B. Carmel. Targeting sensory and motor integration for recovery of movement after CNS injury. *Frontiers in Neuroscience*, 15, 2022. ISSN 1662-453X. URL https://doi.org/10.3389/fnins.2021.791824. 6

[9] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. URL https://dl.acm.org/doi/abs/10.5555/3495724.3496768. 34

[10] Murali Karthick Baskar, Tim Herzig, Diana Nguyen, Mireia Diez, Tim Polzehl, Lukas Burget, and Jan Černocký. Speaker adaptation for Wav2vec2 based dysarthric ASR. In *Proc. Interspeech 2022*, pages 3403–3407, 2022. URL https://doi.org/10.21437/Interspeech.2022-10896. 32, 47

[11] Visar Berisha, Rene Utianski, and Julie Liss. Towards a clinical tool for automatic intelligibility assessment. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2825–2828, 2013. URL https://doi.org/10.1109/ICASSP.2013.6638172. 25, 43, 50, 59

[12] Chitralekha Bhat and Helmer Strik. Automatic assessment of sentence-level dysarthria intelligibility using blstm. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):322–330, 2020. URL https://doi.org/10.1109/JSTSP.2020.2967652. 10, 29, 46

[13] Chitralekha Bhat, Bhavik Vachhani, and Sunil Kopparapu. Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper Spectral Estimation. In *Proc. Interspeech 2016*, pages 228–232, 2016. URL https://doi.org/10.21437/Interspeech.2016-1085. 11, 32, 47, 93, 103

[14] Chitralekha Bhat, Bhavik Vachhani, and Sunil Kopparapu. Improving Recognition of Dysarthric Speech Using Severity Based Tempo Adaptation. In Andrey Ronzhin, Rodmonga Potapova, and Géza Németh, editors, *Speech and Computer*, pages 370–377, Cham, 2016. Springer International Publishing. ISBN 978-3-319-43958-7. URL https://doi.org/10.1007/978-3-319-43958-7_44. 11, 33, 47, 59, 93, 97, 98, 102, 107, 108, 114, 126, 132

[15] Chitralekha Bhat, Bhavik Vachhani, and Sunil Kumar Kopparapu. Automatic assessment of dysarthria severity level using audio descriptors. In

2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5070–5074, 2017. URL https://doi.org/10.1109/ICASSP.2017.7953122. 10, 45, 97

[16] Chitralekha Bhat, Biswajit Das, Bhavik Vachhani, and Sunil Kumar Kopparapu. Dysarthric Speech Recognition Using Time-delay Neural Network Based Denoising Autoencoder. In *Proc. Interspeech 2018*, pages 451–455, 2018. URL https://doi.org/10.21437/Interspeech.2018-1754. 11, 33, 48

[17] Chitralekha Bhat, Ashish Panda, and Helmer Strik. Improved ASR Performance for Dysarthric Speech Using Two-stage DataAugmentation. In *Proc. Interspeech 2022*, pages 46–50, 2022. URL https://doi.org/10.21437/Interspeech.2022-10335. 11, 35, 48, 123, 124

[18] Paul Boersma. Praat, a system for doing phonetics by computer. In *Glot International*, 2002. URL https://api.semanticscholar.org/CorpusID:60531168. 79, 116

[19] Kate Bunton, Ray D. Kent, Jane F. Kent, and Joseph R. Duffy. The effects of flattening fundamental frequency contours on sentence intelligibility in speakers with dysarthria. *Clinical Linguistics & Phonetics*, 15(3):181–193, 2001. doi: 10.1080/02699200010003378. URL https://doi.org/10.1080/02699200010003378. 21

[20] Chris Cannam. Rubber band: A library and utility program for changing tempo and pitch of an audio recording. URL http://breakfastquay.com/rubberband/. 115

[21] T. A. Mariya Celin, T. Nagarajan, and P. Vijayalakshmi. Data Augmentation Using Virtual Microphone Array Synthesis and Multi-Resolution Feature Extraction for Isolated Word Dysarthric Speech Recognition. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):346–354, 2020. URL https://doi.org/10.1109/JSTSP.2020.2972161. 35, 48, 122, 124

[22] H. M. Chandrashekar, Veena Karjigi, and N Sreedevi. Investigation of Different Time-Frequency Representations for Intelligibility Assessment of Dysarthric Speech. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):2880–2889, 2020. URL https://doi.org/10.1109/TNSRE.2020.3035392. 29, 46

[23] H. M. Chandrashekar, Veena Karjigi, and N. Sreedevi. Spectro-Temporal Representation of Speech for Intelligibility Assessment of Dysarthria. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):390–399, 2020. URL https://doi.org/10.1109/JSTSP.2019.2949912. 28, 45

[24] Guoguo Chen, Xingyu Na, Yongqing Wang, Zhiyong Yan, Junbo Zhang, Sifan Ma, and Yujun Wang. Data augmentation for children's speech recognition - the "Ethiopian" system for the SLT 2021 children speech recognition challenge. *CoRR*, abs/2011.04547, 2020. URL `https://arxiv.org/abs/2011.04547`. 122

[25] D. L. Choi, B. W. Kim, Y. J. Lee, Y. Um, and M. Chung. Design and creation of Dysarthric Speech Database for development of QoLT software technology. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 47–50, Oct 2011. URL `https://doi.org/10.1109/ICSDA.2011.6085978`. 17, 40, 113

[26] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain. Automatic selection of speakers for improved acoustic modelling: recognition of disordered speech with sparse data. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 254–259, Dec 2014. URL `https://doi.org/10.1109/SLT.2014.7078583`. 32, 47, 75, 92, 103, 112

[27] Heidi Christensen, Stuart Cunningham, Charles Fox, Phil Green, and Thomas Hain. A comparative study of adaptive, automatic recognition of disordered speech. In *Proc. Interspeech 2012*, pages 1776–1779, 2012. URL `https://doi.org/10.21437/Interspeech.2012-484`. 31, 47, 83, 103

[28] CMU Sphinx:. The Carnegie Mellon Sphinx Project. URL `http://cmusphinx.sourceforge.net/`. 114

[29] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430, 2021. URL `https://doi.org/10.21437/Interspeech.2021-329`. 34

[30] Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury. Data Augmentation for deep neural network acoustic modeling. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5582–5586, 2014. URL `https://doi.or/10.1109/ICASSP.2014.6854671`. 122

[31] H. Dahmani, S.-A. Selouani, D. O'shaughnessy, M. Chetouani, and N. Doghmane. Assessment of dysarthric speech through rhythm metrics. *Journal of King Saud University - Computer and Information Sciences*, 25 (1):43–49, 2013. ISSN 1319-1578. doi: https://doi.org/10.1016/j.jksuci.2012.05.005. URL `https://www.sciencedirect.com/science/article/pii/S1319157812000304`. 24, 26, 43, 58, 59

[32] Marc S. De Bodt, Maria E. Hernández-Díaz Huici, and Paul H. Van De Heyning. Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders*, 35(3):283–292, 2002. ISSN 0021-9924.

doi: https://doi.org/10.1016/S0021-9924(02)00065-5. URL https://www.sciencedirect.com/science/article/pii/S0021992402000655. 22, 42

[33] Jr. Deller, J. R., M. S. Liu, L. J. Ferrier, and P. Robichaud. The Whitaker database of dysarthric (cerebral palsy) speech. *The Journal of the Acoustical Society of America*, 93(6):3516–3518, 06 1993. ISSN 0001-4966. doi: 10.1121/1.405684. URL https://doi.org/10.1121/1.405684. 17, 39

[34] M. Dhanalakshmi and P. Vijayalakshmi. Intelligibility modification of dysarthric speech using hmm-based adaptive synthesis system. In *2015 2nd International Conference on Biomedical Engineering (ICoBE)*, pages 1–5, 2015. URL https://doi.org/10.1109/ICoBE.2015.7235130. 93

[35] Guylaine Le Dorze, Lisa Ouellet, and John Ryalls. Intonation and speech rate in dysarthric speech. *Journal of Communication Disorders*, 27(1):1 – 18, 1994. ISSN 0021-9924. doi: https://doi.org/10.1016/0021-9924(94)90007-8. URL http://www.sciencedirect.com/science/article/pii/0021992494900078. 23, 42

[36] Philip C Doyle, Herbert A Leeper, Ava-Lee Kotler, Nancy Thomas-Stonell, Charlene O'Neill, Marie-Claire Dylke, and Katherine Rolls. Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of rehabilitation research and development*, 34:309–316, 1997. URL https://www.rehab.research.va.gov/jour/97/34/3/pdf/doyle.pdf. 50, 59

[37] P. Enderby. Frenchay Dysarthria Assessment. *Int J Lang Commun Disord*, 15(3):165–173, December 2010. doi: 10.3109/13682828009112541. URL http://dx.doi.org/10.3109/13682828009112541. 17, 19, 51, 54, 60

[38] Pamela M Enderby and Alexandra John. Therapy outcome measures in speech and language therapy: Comparing performance between different providers. *International Journal of Language & Communication Disorders*, 34(4):417–429, 1999. doi: https://doi.org/10.1080/136828299247360. URL https://onlinelibrary.wiley.com/doi/abs/10.1080/136828299247360. 20

[39] Ömer Eskidere and Ahmet Gürhanlı. Voice Disorder Classification Based on Multitaper Mel Frequency Cepstral Coefficients Features. *Computational and Mathematical Methods in Medicine*, 2015(1):956249, 2015. doi: https://doi.org/10.1155/2015/956249. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2015/956249. 75

[40] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 835–838, New York, NY, USA, 2013. Association for

Computing Machinery. ISBN 9781450324045. doi: 10.1145/2502081.2502224. URL https://doi.org/10.1145/2502081.2502224. 117

[41] Tiago H. Falk, Wai-Yip Chan, and Fraser Shein. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication*, 54 (5):622–631, 2012. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2011.03.007. URL https://www.sciencedirect.com/science/article/pii/S0167639311000513. Advanced Voice Function Assessment. 25, 43

[42] Mireia Farrús, Javier Hernando, and Pascual Ejarque. Jitter and shimmer measurements for speaker recognition. In *Interspeech*, 2007. URL https://api.semanticscholar.org/CorpusID:16303874. 76, 79

[43] X. Feng, Y. Zhang, and J. Glass. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1759–1763, May 2014. URL https://doi.org/10.1109/ICASSP.2014.6853900. 93, 103

[44] Cécile Fougeron, Lise Crevier-Buchman, Corinne Fredouille, Alain Ghio, Christine Meunier, Claude Chevrie-Muller, Jean-Francois Bonastre, Antonia Colazo Simon, Céline Delooze, Danielle Duez, Cédric Gendrot, Thierry Legou, Nathalie Levèque, Claire Pillot-Loiseau, Serge Pinto, Gilles Pouchoulin, Danièle Robert, Jacqueline Vaissiere, François Viallet, and Coralie Vincent. The DesPho-APaDy project: Developing an acoustic-phonetic characterization of dysarthric speech in French. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/626_Paper.pdf. 17, 39, 113

[45] Eibe Frank, Mark A. Hall, and Ian H. Witten. The weka workbench. online appendix for data mining: Practical machine learning tools and techniques. Morgan Kaufmann, Fourth Edition, 2016. URL https://www.goodreads.com/book/show/36012530-the-weka-workbench-online-appendix-for-data-mining. 117

[46] Melanie Fried-Oken. Voice recognition device as a computer interface for motor and speech impaired people. *Archives of physical medicine and rehabilitation*, 66 10:678–81, 1985. URL https://api.semanticscholar.org/CorpusID:10711427. 74

[47] M. Frohlich, D. Michaelis, and H. Werner Strube. Acoustic "breathiness measures" in the description of pathologic voices. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 2, pages 937–940 vol.2, 1998. URL https://doi.org/10.1109/ICASSP.1998.675420. 129

[48] M.J.F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech & Language*, 12(2):75–98, 1998. ISSN 0885-2308. doi: https://doi.org/10.1006/csla.1998.0043. URL https://www.sciencedirect.com/science/article/pii/S0885230898900432. 80, 88

[49] Mario Ganzeboom, Marjoke Bakker, Catia Cucchiarini, and Helmer Strik. Intelligibility of Disordered Speech: Global and Detailed Scores. In *Proc. Interspeech 2016*, pages 2503–2507, 2016. URL https://doi.org/10.21437/Interspeech.2016-1448. 61

[50] J. S. Garofolo. Getting started with the darpa TIMIT cd-rom: An acoustic phonetic continuous speech database. *NIST*, 1988. URL https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir4930.pdf. 86

[51] Mengzhe Geng, Xurong Xie, Shansong Liu, Jianwei Yu, Shoukang Hu, Xunying Liu, and Helen Meng. Investigation of Data Augmentation Techniques for Disordered Speech Recognition. In *Proc. Interspeech 2020*, pages 696–700, 2020. URL https://doi.org/10.21437/Interspeech.2020-1161. 34, 48, 122, 124, 126, 138

[52] Mengzhe Geng, Xurong Xie, Zi Ye, Tianzi Wang, Guinan Li, Shujie Hu, Xunying Liu, and Helen Meng. Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2597–2611, 2022. URL https://doi.org/10.1109/TASLP.2022.3195113. 33, 47

[53] Felix A. Gers, Nicol N. Schraudolph, and Jürgen Schmidhuber. Learning precise timing with lstm recurrent networks. *J. Mach. Learn. Res.*, 3(null): 115–143, mar 2003. ISSN 1532-4435. doi: 10.1162/153244303768966139. URL https://doi.org/10.1162/153244303768966139. 62

[54] Alexander M. Goberman, Carl A. Coelho, and Michael P. Robb. Prosodic characteristics of parkinsonian speech: The effect of levodopa-based medication. *Journal of Medical Speech-Language Pathology*, 13(1):51 – 68, 2005. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-15944419989&partnerID=40&md5=d9949ad00d04a67941c081f01c09c23f. Cited by: 46. 22

[55] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with Deep Bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech*

Recognition and Understanding, pages 273–278, Dec 2013. URL https://doi.org/10.1109/ASRU.2013.6707742. 63

[56] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, May 2013. URL https://doi.org/10.1109/ICASSP.2013.6638947. 62

[57] Phil Green, James Carmichael, Athanassios Hatzis, Pam Enderby, Mark Hawley, and Mark Parker. Automatic speech recognition with sparse training data for dysarthric speakers. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 1189–1192, 2003. URL https://doi.org/10.21437/Eurospeech.2003-384. 75

[58] Seongjun Hahm, Daragh Heitzman, and Jun Wang. Recognizing Dysarthric Speech due to Amyotrophic Lateral Sclerosis with Across-Speaker Articulatory Normalization. In Jan Alexandersson, Ercan Altinsoy, Heidi Christensen, Peter Ljunglöf, François Portet, and Frank Rudzicz, editors, *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 47–54, Dresden, Germany, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-5109. URL https://aclanthology.org/W15-5109. 75, 92

[59] Kyle Hall, Andy Huang, and Seyed Reza Shahamiri. An Investigation to Identify Optimal Setup for Automated Assessment of Dysarthric Intelligibility using Deep Learning Technologies. *Cognitive Computation*, 2022. URL https://doi.org/10.1007/s12559-022-10041-3. 29

[60] Anna Eva Hallin and Petri Partanen. Factors affecting speech-language pathologists' language assessment procedures and tools – challenges and future directions in sweden. *Logopedics Phoniatrics Vocology*, 49(3):104–113, 2024. doi: 10.1080/14015439.2022.2158218. URL https://doi.org/10.1080/14015439.2022.2158218. PMID: 36576225. 50

[61] Tomoki Hayashi, Ryuichi Yamamoto, Takenori Yoshimura, Peter Wu, Jiatong Shi, Takaaki Saeki, Yooncheol Ju, Yusuke Yasuda, Shinnosuke Takamichi, and Shinji Watanabe. ESPnet2-TTS: Extending the Edge of TTS Research. *ArXiv*, abs/2110.07840, 2021. URL https://api.semanticscholar.org/CorpusID:239009547. 127, 133

[62] Abner Hernandez, Sunhee Kim, and Minhwa Chung. Prosody-Based Measures for Automatic Severity Assessment of Dysarthric Speech. *Applied Sciences*, 10(19), 2020. ISSN 2076-3417. doi: 10.3390/app10196999. URL https://www.mdpi.com/2076-3417/10/19/6999. 27, 44

[63] Abner Hernandez, Eun Jung Yeo, Sunhee Kim, and Minhwa Chung. Dysarthria Detection and Severity Assessment Using Rhythm-Based Metrics. In *Proc. Interspeech 2020*, pages 2897–2901, 2020. URL https://doi.org/10.21437/Interspeech.2020-2354. 27, 44

[64] Abner Hernandez, Paula Andrea Pérez-Toro, Elmar Nöth, Juan Rafael Orozco-Arroyave, Andreas Maier, and Seung Hee Yang. Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition. In *Proc. Interspeech 2022*, pages 51–55, 2022. URL https://doi.org/10.21437/Interspeech.2022-10674. 34, 48

[65] Margaret M. Hoehn and Melvin D. Yahr. Parkinsonism: onset, progression, and mortality. *Neurology*, 17(5):427–427, 1967. doi: 10.1212/WNL.17.5.427. URL https://www.neurology.org/doi/abs/10.1212/WNL.17.5.427. 19

[66] J.-P. Hosom, A.B. Kain, T. Mishra, J.P.H. van Santen, M. Fried-Oken, and J. Staehely. Intelligibility of modifications to dysarthric speech. In *Proc. ICASSP 2003*, 2003. URL https://doi.org/10.1109/ICASSP.2003.1198933. 30, 74, 93, 122

[67] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537, 2021. URL https://doi.org/10.1109/ICASSP39728.2021.9414460. 34

[68] Richard Hummel. Objective Estimation of Dysarthric Speech Intelligibility. *Thesis, Masters of Applied Science. Queen's University Kingston; Ontario, Canada:*, 2011. URL https://api.semanticscholar.org/CorpusID:16053020. 25, 43

[69] Richard Hummel, Wai-Yip Chan, and Tiago H. Falk. Spectral features for automatic blind intelligibility estimation of spastic dysarthric speech. In *Proc. Interspeech 2011*, pages 3017–3020, 2011. URL https://doi.org/10.21437/Interspeech.2011-755. 50, 51, 57, 59

[70] T B Ijitona, J J Soraghan, A Lowit, G Di-Caterina, and H Yue. Effects of acoustic features modifications on the perception of dysarthric speech — preliminary study (pitch, intensity and duration modifications). In *IET 3rd International Conference on Intelligent Signal Processing (ISP 2017)*, pages 1–6, 2017. URL https://doi.org/10.1049/cp.2017.0363. 22, 42

[71] Takaaki Ishii, Hiroki Komiyama, Takahiro Shinozaki, Yasuo Horiuchi, and Shingo Kuroiwa. Reverberant speech recognition based on denoising autoencoder. In *Interspeech*, 2013. URL https://api.semanticscholar.org/CorpusID:28819623. 93

[72] Navdeep Jaitly and Geoffrey E. Hinton. Vocal Tract Length Perturbation (VTLP) improves speech recognition. In *Proc. ICML workshop on deep learning for audio, speech and language*, volume 117, page 21, 2013. URL https://api.semanticscholar.org/CorpusID:14140670. 133

[73] Parvaneh Janbakhshi, Ina Kodrasi, and Hervé Bourlard. Subspace-based learning for automatic dysarthric speech detection. *IEEE Signal Processing Letters*, 28:96–100, 2021. URL https://doi.org/10.1109/LSP.2020.3044503. 27, 44

[74] Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss. Simulating Dysarthric Speech for Training Data Augmentation in Clinical Speech Applications. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6009–6013, 2018. URL https://doi.org/10.1109/ICASSP.2018.8462290. 122

[75] Zengrui Jin, Mengzhe Geng, Xurong Xie, Jianwei Yu, Shansong Liu, Xunying Liu, and Helen M. Meng. Adversarial Data Augmentation for Disordered Speech Recognition. In *Interspeech*, 2021. URL https://api.semanticscholar.org/CorpusID:236772851. 123, 138

[76] Zengrui Jin, Xurong Xie, Mengzhe Geng, Tianzi Wang, Shujie Hu, Jiajun Deng, Guinan Li, and Xunying Liu. Adversarial Data Augmentation Using VAE-GAN for Disordered Speech Recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. URL https://doi.org/10.1109/ICASSP49357.2023.10095547. 35, 48, 124

[77] Amlu Anna Joshy and Rajeev Rajan. Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1147–1157, 2022. URL https://doi.org/10.1109/TNSRE.2022.3169814. 29, 46

[78] Jayaraman D K and Das J M. Dysarthria. In *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing*, 2023. URL https://www.ncbi.nlm.nih.gov/books/NBK592453/. 8

[79] Kamil Lahcene Kadi, Sid Ahmed Selouani, Bachir Boudraa, and Malika Boudraa. Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. *Biocybernetics and Biomedical Engineering*, 36(1):233–247, 2016. ISSN 0208-5216. doi: https://doi.org/10.1016/j.bbe.2015.11.004. URL https://www.sciencedirect.com/science/article/pii/S020852161500087X. 51, 55, 57

[80] KL Kadi, SA Selouani, B Boudraa, and M Boudraa. Discriminative Prosodic Features to Assess the Dysarthria Severity Levels. In *Proceedings of the World*

Congress on Engineering, volume 3, 2013. URL https://www.iaeng.org/publication/WCE2013/WCE2013_pp2201-2205.pdf. 26, 43, 59, 75, 77, 84

[81] Alexander Kain, Xiaochuan Niu, John-Paul Hosom, Qi Miao, and Jan P. H. van Santen. Formant re-synthesis of dysarthric speech. In *Proc. 5th ISCA Workshop on Speech Synthesis (SSW 5)*, pages 25–30, 2004. URL https://www.isca-archive.org/ssw_2004/kain04_ssw.pdf. 75

[82] Alexander B. Kain, John-Paul Hosom, Xiaochuan Niu, Jan P.H. van Santen, Melanie Fried-Oken, and Janice Staehely. Improving the intelligibility of dysarthric speech. *Speech Communication*, 49(9):743–759, 2007. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2007.05.001. URL https://www.sciencedirect.com/science/article/pii/S0167639307000854. 74, 83, 93, 102

[83] Naoyuki Kanda, Ryu Takeda, and Yasunari Obuchi. Elastic spectral distortion for low resource speech recognition with deep neural networks. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 309–314, 2013. URL https://doi.org/10.1109/ASRU.2013.6707748. 126

[84] Panagiota Karanasou, Yongqiang Wang, Mark John Francis Gales, and Philip C. Woodland. Adaptation of deep neural network acoustic models using factorised i-vectors. In *Interspeech*, 2014. URL https://api.semanticscholar.org/CorpusID:2642709. 64

[85] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. A Comparative Study on Transformer vs RNN in Speech Applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, December 2019. URL http://dx.doi.org/10.1109/ASRU46091.2019.9003750. 123

[86] Shigeki Karita, Nelson Yalta, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration. In *Interspeech*, 2019. URL https://api.semanticscholar.org/CorpusID:202736731. 132

[87] Hemant Kumar Kathania, Mittul Singh, Tam'as Gr'osz, and Mikko Kurimo. Data augmentation using prosody and false starts to recognize non-native children's speech. In *Interspeech*, 2020. URL https://api.semanticscholar.org/CorpusID:221377331. 133

[88] Hemant Kumar Kathania, Viredner Kadyan, Sudarsana Reddy Kadiri, and Mikko Kurimo. Data Augmentation Using Spectral Warping for Low Resource

Children ASR. *J. Signal Process. Syst.*, 94(12):1507–1513, dec 2022. ISSN 1939-8018. doi: 10.1007/s11265-022-01820-0. URL https://doi.org/10.1007/s11265-022-01820-0. 127, 133

[89] Elaine Kearney and Frank H. Guenther. Articulating: the neural mechanisms of speech production. *Language, Cognition and Neuroscience*, 34(9):1214–1229, 2019. doi: 10.1080/23273798.2019.1589541. URL https://doi.org/10.1080/23273798.2019.1589541. PMID: 31777753. 6

[90] R. D. Kent, J. F. Kent, G. Weismer, R. E. Martin, R. L. Sufit, B. R. Brooks, and J. C. Rosenbek. Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. *Clinical Linguistics & Phonetics*, 3(4):347–358, 1989. doi: 10.3109/02699208908985295. URL https://doi.org/10.3109/02699208908985295. 24

[91] Ray D. Kent. Hearing and believing. *American Journal of Speech-Language Pathology*, 5(3):7–23, 1996. doi: 10.1044/1058-0360.0503.07. URL https://pubs.asha.org/doi/abs/10.1044/1058-0360.0503.07. 21

[92] Ray D Kent. Research on speech motor control and its disorders: A review and prospective. *Journal of Communication Disorders*, 33(5):391–428, 2000. ISSN 0021-9924. doi: https://doi.org/10.1016/S0021-9924(00)00023-X. URL https://www.sciencedirect.com/science/article/pii/S002199240000023X. 58

[93] Ray D Kent, Gary Weismer, Jane F Kent, Houri K Vorperian, and Joseph R Duffy. Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of Communication Disorders*, 32(3):141–186, 1999. ISSN 0021-9924. doi: https://doi.org/10.1016/S0021-9924(99)00004-0. URL https://www.sciencedirect.com/science/article/pii/S0021992499000040. 23, 42

[94] Raymond D. Kent. *The MIT Encyclopedia of Communication Disorders*. The MIT Press, 09 2003. ISBN 9780262277020. doi: 10.7551/mitpress/4658.001.0001. URL https://doi.org/10.7551/mitpress/4658.001.0001. 50, 58

[95] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon R. Gunderson, Thomas S. Huang, Kenneth L. Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Interspeech*, 2008. URL https://api.semanticscholar.org/CorpusID:8961222. 17, 39, 53, 77, 84, 86, 98, 103, 106, 113, 116, 130

[96] Jangwon Kim, Naveen Kumar, Andreas Tsiartas, Ming Li, and Shrikanth Narayanan. Intelligibility classification of pathological speech using fusion of multiple high level descriptors. In *Proc. Interspeech 2012*, pages 534–537, 2012. doi: 10.21437/Interspeech.2012-103. URL https://doi.org/10.21437/Interspeech.2012-103. 43, 51, 59

[97] Jangwon Kim, Naveen Kumar, Andreas Tsiartas, Ming Li, and Shrikanth S. Narayanan. Automatic intelligibility classification of sentence-level pathological speech. *Computer Speech & Language*, 29(1):132–144, 2015. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2014.02.001. URL https://www.sciencedirect.com/science/article/pii/S088523081400014X. 26, 44, 59, 68, 69, 71

[98] M. J. Kim, Y. Kim, and H. Kim. Automatic Intelligibility Assessment of Dysarthric Speech Using Phonologically-Structured Sparse Linear Model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4): 694–704, April 2015. ISSN 2329-9290. URL https://doi.org/10.1109/TASLP.2015.2403619. 59

[99] Min Jung Kim, Soyeong Pae, and Chang Il Park. Assessment of phonology and articulation for children (apac), 2007. 19

[100] Myung Jong Kim and Hoirin Kim. Automatic Assessment of Dysarthric Speech Intelligibility Based on Selected Phonetic Quality Features. In Klaus Miesenberger, Arthur Karshmer, Petr Penaz, and Wolfgang Zagler, editors, *Computers Helping People with Special Needs*, pages 447–450, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-31534-3. URL https://doi.org/10.1007/978-3-642-31534-3_66. 50, 59

[101] Myung Jong Kim and Hoirin Kim. Combination of multiple speech dimensions for automatic assessment of dysarthric speech intelligibility. In *Proc. Interspeech 2012*, pages 1323–1326, 2012. URL https://doi.org/10.21437/Interspeech.2012-317. 26, 59

[102] Myung Jong Kim, Joohong Yoo, and Hoirin Kim. Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models. In *Proc. Interspeech 2013*, pages 3622–3626, 2013. URL https://doi.org/10.21437/Interspeech.2013-320. 24, 32, 47, 50, 59, 75, 83, 93, 103

[103] Myungjong Kim, Younggwan Kim, Joohong Yoo, Jun Wang, and Hoirin Kim. Regularized Speaker Adaptation of KL-HMM for Dysarthric Speech Recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(9):1581–1591, 2017. URL https://doi.org/10.1109/TNSRE.2017.2681691. 32, 47

[104] Myungjong Kim, Beiming Cao, Kwanghoon An, and Jun Wang. Dysarthric Speech Recognition Using Convolutional LSTM Neural Network. In *Proc. Interspeech 2018*, pages 2948–2952, 2018. URL https://doi.org/10.21437/Interspeech.2018-2250. 33, 36, 48

[105] Myungjong Kim, Beiming Cao, and Jun Wang. Multi-view Representation Learning via Canonical Correlation Analysis for Dysarthric Speech Recognition. In Kevin Deng, Zhengtao Yu, Srikanta Patnaik, and John Wang,

editors, *Recent Developments in Mechatronics and Intelligent Robotics*, pages 1085–1095, Cham, 2019. Springer International Publishing. ISBN 978-3-030-00214-5. URL https:/doi.org/10.1007/978-3-030-00214-5_133. 47

[106] Yunjung Kim, Gary Weismer, Raymond D. Kent, and Joseph R. Duffy. Statistical Models of F2 Slope in Relation to Severity of Dysarthria. *Folia Phoniatrica et Logopaedica*, 61(6):329–335, 10 2009. ISSN 1021-7762. doi: 10.1159/000252849. URL https://doi.org/10.1159/000252849. 22, 42

[107] Yunjung Kim, Raymond D. Kent, and Gary Weismer. An Acoustic Study of the Relationships Among Neurologic Disease, Dysarthria Type, and Severity of Dysarthria. *Journal of Speech, Language, and Hearing Research*, 54(2): 417–429, 2011. doi: 10.1044/1092-4388(2010/10-0020). URL https://pubs.asha.org/doi/abs/10.1044/1092-4388%282010/10-0020%29. 22, 42

[108] Tomi H. Kinnunen, Rahim Saeidi, Johan Sandberg, and Maria Hansson. What Else is New Than the Hamming Window? Robust MFCCs for Speaker Recognition via Multitapering. In *Interspeech*, 2010. URL https://api.semanticscholar.org/CorpusID:1876329. 56

[109] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Proc. Interspeech 2015*, pages 3586–3589, 2015. URL https://doi.org/10.21437/Interspeech.2015-711. 113, 122, 126

[110] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224, 2017. URL https://doi.org/10.1109/ICASSP.2017.7953152. 122

[111] Sonja A. Kotz and Michael Schwartze. Chapter 57 - Motor-Timing and Sequencing in Speech Production: A General-Purpose Framework. In Gregory Hickok and Steven L. Small, editors, *Neurobiology of Language*, pages 717–724. Academic Press, San Diego, 2016. ISBN 978-0-12-407794-2. doi: https://doi.org/10.1016/B978-0-12-407794-2.00057-2. URL https://www.sciencedirect.com/science/article/pii/B9780124077942000572. 6

[112] Yukie Kozaki-Yamaguchi, Noriko Suzuki, Yukihiro Fujita, Hidemi Yoshimasu, Masato Akagi, and Teruo Amagasa. Perception of Hypernasality and its Physical Correlates. *Oral Science International*, 2(1):21–35, 2005. ISSN 1348-8643. doi: https://doi.org/10.1016/S1348-8643(05)80004-7. URL https://www.sciencedirect.com/science/article/pii/S1348864305800047. 128

[113] Jody Kreiman, Bruce R. Gerratt, and Brian Gabelman. Jitter, shimmer, and noise in pathological voice quality perception. *The Journal of the Acoustical*

Society of America, 112(5_Supplement):2446–2446, 10 2002. ISSN 0001-4966. doi: 10.1121/1.4780067. URL https://doi.org/10.1121/1.4780067. 81

[114] Alice S-Y. Lee, Valter Ciocca, and Tara L. Whitehill. Acoustic correlates of hypernasality. *Clinical Linguistics & Phonetics*, 17(4-5):259–264, 2003. doi: 10.1080/0269920031000080091. URL https://doi.org/10.1080/0269920031000080091. PMID: 12945600. 128

[115] Xinwei Li, Yuanyuan Zhang, Xiaodan Zhuang, and Daben Liu. Frame-Level Specaugment for Deep Convolutional Neural Networks in Hybrid ASR Systems. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, pages 209–214. IEEE, 2021. doi: 10.1109/SLT48900.2021.9383626. URL https://doi.org/10.1109/SLT48900.2021.9383626. 122

[116] Julie M. Liss, Laurence White, Sven L. Mattys, Kaitlin Lansford, Andrew J. Lotto, Stephanie M. Spitzer, and John N. Caviness. Quantifying Speech Rhythm Abnormalities in the Dysarthrias. *Journal of Speech, Language, and Hearing Research*, 52(5):1334–1352, 2009. doi: 10.1044/1092-4388(2009/08-0208). URL https://pubs.asha.org/doi/abs/10.1044/1092-4388%282009/08-0208%29. 22, 42

[117] Julie M. Liss, Sue LeGendre, and Andrew J. Lotto. Discriminating Dysarthria Type From Envelope Modulation Spectra. *Journal of Speech, Language, and Hearing Research*, 53(5):1246–1255, 2010. doi: 10.1044/1092-4388(2010/09-0121). URL https://pubs.asha.org/doi/abs/10.1044/1092-4388%282010/09-0121%29. 23, 42

[118] Huei-Mei Liu and Chin-Hsing Tseng. The perceptual and acoustical analysis of speech intelligibility in Mandarin-speaking adults with cerebral palsy. *The Journal of the Acoustical Society of America*, 100(4_Supplement):2827–2827, 10 1996. ISSN 0001-4966. doi: 10.1121/1.416657. URL https://doi.org/10.1121/1.416657. 21, 42

[119] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *Proc. Interspeech 2013*, pages 436–440, 2013. URL https://doi.org/10.21437/Interspeech.2013-130. 93, 103

[120] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth. PEAKS – A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5):425–437, 2009. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2009.01.004. URL https://www.sciencedirect.com/science/article/pii/S016763930900003X. 21, 24, 50, 58, 59, 75, 93

[121] Victor M Makuto. Articulation Of Consonants In The Speech Of Cerebral Palsy Learners Of English As A Second Language. *International Journal of Innovative Research and Advanced Studies (IJIRAS)*, Volume 3 Issue 12, November 2016, 2016. ISSN ISSN: 2394-4404. URL https://api.semanticscholar.org/CorpusID:29623919. 127

[122] Marco Marini, Mauro Viganò, Massimo Corbo, Marina Zettin, Gloria Simoncini, Bruno Fattori, Clelia D'Anna, Massimiliano Donati, and Luca Fanucci. IDEA: An Italian Dysarthric Speech Database. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1086–1093, 2021. URL https://doi.org/10.1109/SLT48900.2021.9383467. 17, 41

[123] TA Mariya Celin, P Vijayalakshmi, and T Nagarajan. Data augmentation Techniques for Transfer Learning-Based Continuous Dysarthric Speech Recognition. *Circuits, Systems, and Signal Processing*, 42(1):601–622, 2023. URL https://doi.org/10.1007/s00034-022-02156-7. 36, 48, 123, 124

[124] David Martínez, Phil D. Green, and Heidi Christensen. Dysarthria intelligibility assessment in a factor analysis total variability space. In *Proc. Interspeech 2013*, pages 2133–2137, 2013. URL https://doi.org/10.21437/Interspeech.2013-505. 26, 43, 59

[125] Viviana Mendoza Ramos, Hector A. Kairuz Hernandez-Diaz, Maria E. Hernandez-Diaz Huici, Heidi Martens, Gwen Van Nuffelen, and Marc De Bodt. Acoustic features to characterize sentence accent production in dysarthric speech. *Biomedical Signal Processing and Control*, 57: 101750, 2020. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2019.101750. URL http://www.sciencedirect.com/science/article/pii/S1746809419303313. 23, 42

[126] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell. The Nemours database of dysarthric speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1962–1965 vol.3, Oct 1996. URL https://doi.org/10.1109/ICSLP.1996.608020. 17, 39, 84, 113

[127] K. T. Mengistu and F. Rudzicz. Adapting acoustic and lexical models to dysarthric speech. *In Proc. ICASSP 2011*, pages 4924–4927, May 2011. ISSN 1520-6149. URL https://doi.org/10.1109/ICASSP.2011.5947460. 75

[128] Kinfe Tadesse Mengistu and Frank Rudzicz. Comparing Humans and Automatic Speech Recognition Systems in Recognizing Dysarthric Speech. In Cory Butz and Pawan Lingras, editors, *Advances in Artificial Intelligence*, pages 291–300, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-21043-3. URL https://doi.org/10.1007/978-3-642-21043-3_36. 31, 47

[129] Catherine Middag, Jean-Pierre Martens, Gwen Van Nuffelen, and Marc De Bodt. DIA: a tool for objective intelligibility assessment of pathological speech. In *6th International workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, pages 165–167. Firenze University Press, 2009. URL https://biblio.ugent.be/publication/828696. 50

[130] Catherine Middag, Tobias Bocklet, Jean-Pierre Martens, and Elmar Nöth. Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment. In *Proc. Interspeech 2011*, pages 3005–3008, 2011. URL https://doi.org/10.21437/Interspeech.2011-752. 26, 43, 50, 59

[131] Santiago Omar Caballero Morales and Stephen J Cox. Modelling Errors in Automatic Speech Recognition for Dysarthric Speakers. *EURASIP Journal on Advances in Signal Processing*, 2009(1):1–14, 2009. URL https://doi.org/10.1155/2009/308340. 75

[132] Mumtaz Begum Mustafa, Siti Salwah Salim, Noraini Mohamed, Bassam Al-Qatab, and Chng Eng Siong. Severity-Based Adaptation with Limited Data for ASR to Aid Dysarthric Speakers. *PLOS ONE*, 9(1):1–11, 01 2014. URL https://doi.org/10.1371/journal.pone.0086285. 32, 50, 59, 93, 103

[133] Mumtaz Begum Mustafa, Fadhilah Rosdi, Siti Salwah Salim, and Muhammad Umair Mughal. Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. *Expert Systems with Applications*, 42(8):3924–3932, 2015. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2015.01.033. URL https://www.sciencedirect.com/science/article/pii/S0957417415000482. 24, 47

[134] Narendra N P and Paavo Alku. Dysarthric Speech Classification Using Glottal Features Computed from Non-words, Words and Sentences. In *Proc. Interspeech 2018*, pages 3403–3407, 2018. URL https://doi.org/10.21437/Interspeech.2018-1059. 26, 44

[135] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia. Dysarthric speech recognition using a convolutive bottleneck network. *12th International Conference on Signal Processing (ICSP)*, pages 505–509, Oct 2014. ISSN 2164-5221. URL https://doi.org/10.1109/ICOSP.2014.7015056. 32, 47, 75, 102, 112

[136] Michael Neumann, Oliver Roesler, Jackson Liscombe, Hardik Kothare, David Suendermann-Oeft, David Pautler, Indu Navar, Aria Anvar, Jochen Kumm, Raquel Norel, Ernest Fraenkel, Alexander V. Sherman, James D. Berry, Gary L. Pattee, Jun Wang, Jordan R. Green, and Vikram Ramanarayanan.

Investigating the Utility of Multimodal Conversational Technology and Audiovisual Analytic Measures for the Assessment and Monitoring of Amyotrophic Lateral Sclerosis at Scale. In *Proc. Interspeech 2021*, pages 4783–4787, 2021. URL https://doi.org/10.21437/Interspeech.2021-1801. 27, 45

[137] Thai-Son Nguyen, Sebastian Stüker, Jan Niehues, and Alex Waibel. Improving Sequence-To-Sequence Speech Recognition Training with On-The-Fly Data Augmentation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7689–7693, 2020. URL https://doi.org/10.1109/ICASSP40776.2020.9054130. 122

[138] Mauro Nicolao, Heidi Christensen, Stuart Cunningham, Phil Green, and Thomas Hain. A Framework for Collecting Realistic Recordings of Dysarthric Speech - the homeService Corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1993–1997, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1315. 18, 37, 113, 147

[139] Zeinab Noorian, Chloé Pou-Prom, and Frank Rudzicz. On the importance of normative data in speech-based assessment. *ArXiv*, abs/1712.00069, 2017. URL https://api.semanticscholar.org/CorpusID:21172694. 113

[140] Mamyrbayev Orken, Oralbekova Dina, Alimhan Keylan, Turdalykyzy Tolganay, and Othman Mohamed. A study of transformer-based end-to-end speech recognition system for Kazakh language. *Scientific Reports*, 12, 2022. URL https://api.semanticscholar.org/CorpusID:248890655. 129, 130

[141] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia González-Rátiva, and Elmar Nöth. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 342–347, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/7_Paper.pdf. 17, 41, 113

[142] Juan Rafael Orozco-Arroyave, Juan Camilo Vásquez-Correa, Jesús Francisco Vargas-Bonilla, R. Arora, N. Dehak, P.S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei, A. Vann, N. Vogler, T. Bocklet, M. Cernak, J. Hannink, and Elmar Nöth. NeuroSpeech. *SoftwareX*, 8:69 – 70, 2018. ISSN 2352-7110. doi: https://doi.org/10.

1016/j.softx.2017.08.004. URL http://www.sciencedirect.com/science/article/pii/S2352711017300341. Digital Signal Processing & SoftwareX - Joint Special Issue on Reproducible Research in Signal Processing. 37, 147

[143] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. URL https://doi.org/10.1109/TKDE.2009.191. 35

[144] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6440–6444, 2016. URL https://doi.org/10.1109/ICASSP.2016.7472917. 113

[145] Elisabet Rosengren Parimala Raghavendra and Sheri Hunnicutt. An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *Augmentative and Alternative Communication*, 17(4):265–275, 2001. doi: 10.1080/aac.17.4.265.275. URL https://doi.org/10.1080/aac.17.4.265.275. 75

[146] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech*, 2019. URL https://api.semanticscholar.org/CorpusID:121321299. 122, 123, 128, 134

[147] Sree Hari Krishnan Parthasarathi, Björn Hoffmeister, Spyridon Matsoukas, Arindam Mandal, Nikko Strom, and Srinivas Garimella. fMLLR based feature-space speaker adaptation of DNN acoustic models. In *Interspeech*, 2015. URL https://api.semanticscholar.org/CorpusID:10087552. 117

[148] Rupal Patel. Prosodic control in severe dysarthria. *Journal of Speech, Language, and Hearing Research*, 45(5):858–870, 2002. doi: 10.1044/1092-4388(2002/069). URL https://pubs.asha.org/doi/abs/10.1044/1092-4388%282002/069%29. 21, 31, 122

[149] Rupal Patel and Pamela Campellone. Acoustic and Perceptual Cues to Contrastive Stress in Dysarthria. *Journal of Speech, Language, and Hearing Research*, 52(1):206–222, 2009. doi: 10.1044/1092-4388(2008/07-0078). URL https://pubs.asha.org/doi/abs/10.1044/1092-4388%282008/07-0078%29. 42, 77, 81

[150] Vijayaditya Peddinti, Guoguo Chen, Vimal Manohar, Tom Ko, Daniel Povey, and Sanjeev Khudanpur. JHU ASpIRE system: Robust LVCSR with TDNNS, iVector adaptation and RNN-LMS. In *2015 IEEE Workshop on Automatic*

Speech Recognition and Understanding (ASRU), pages 539–546, 2015. URL https://doi.org/10.1109/ASRU.2015.7404842. 67

[151] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proc. Interspeech 2015*, pages 3214–3218, 2015. URL https://doi.org/10.21437/Interspeech.2015-647. 104, 106

[152] Geoffroy Peeters, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5): 2902–2916, 11 2011. ISSN 0001-4966. doi: 10.1121/1.3642604. URL https://doi.org/10.1121/1.3642604. 51

[153] K. J. Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Sept 2015. URL https://doi.org/10.1109/MLSP.2015.7324337. 113

[154] Matthew L. Poole and Adam P. Vogel. Chapter 42 - Linking motor speech function and dementia. In Colin R. Martin and Victor R. Preedy, editors, *Genetics, Neurology, Behavior, and Diet in Dementia*, pages 665–676. Academic Press, 2020. ISBN 978-0-12-815868-5. doi: https://doi.org/10.1016/B978-0-12-815868-5.00042-6. URL https://www.sciencedirect.com/science/article/pii/B9780128158685000426. 13

[155] M. Portnoff. Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(3):243–248, Jun 1976. ISSN 0096-3518. URL https://doi.org/10.1109/TASSP.1976.1162810. 84, 97, 105

[156] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The Kaldi speech recognition toolkit. *IEEE 2011 workshop on automatic speech recognition and understanding*, (EPFL-CONF-192584), 2011. URL https://www.danielpovey.com/files/2011_asru_kaldi.pdf. 67, 79, 87, 98, 105, 117

[157] Luke Prananta, Bence Halpern, Siyuan Feng, and Odette Scharenborg. The Effectiveness of Time Stretching for Enhancing Dysarthric Speech for Improved Dysarthric Speech Recognition. In *Proc. Interspeech 2022*, pages 36–40, 2022. URL https://doi.org/10.21437/Interspeech.2022-190. 34, 48

[158] G. A. Prieto, R. L. Parker, D. J. Thomson, F. L. Vernon, and R. L. Graham. Reducing the bias of multitaper spectrum estimates. *Geophysical Journal*

International, 171(3):1269–1281, 12 2007. ISSN 0956-540X. doi: 10.1111/j.1365-246X.2007.03592.x. URL https://doi.org/10.1111/j.1365-246X.2007.03592.x. 52, 76

[159] Y. Qian, R. Ubale, V. Ramanaryanan, P. Lange, D. Suendermann-Oeft, K. Evanini, and E. Tsuprun. Exploring ASR-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 569–576, Dec 2017. URL https://doi.org/10.1109/ASRU.2017.8268987. 67

[160] Constance Dean Qualls. Chapter 8 - Neurogenic disorders of speech, language, cognition-communication, and swallowing. In Dolores E. Battle, editor, *Communication Disorders in Multicultural and International Populations (Fourth Edition)*, pages 148 – 163. Mosby, Saint Louis, 2012. ISBN 978-0-323-06699-0. doi: https://doi.org/10.1016/B978-0-323-06699-0.00017-0. URL http://www.sciencedirect.com/science/article/pii/B9780323066990000170. 8, 58

[161] Liborio Rampello, Luigi Rampello, Francesco Patti, and Mario Zappia. When the word doesn't come out: A synthetic overview of dysarthria. *Journal of the Neurological Sciences*, 369:354–360, 2016. ISSN 0022-510X. doi: https://doi.org/10.1016/j.jns.2016.08.048. URL https://www.sciencedirect.com/science/article/pii/S0022510X16305391. 128

[162] Siddharth Rathod, Monil Charola, and Hemant A. Patil. Transfer Learning Using Whisper for Dysarthric Automatic Speech Recognition. In Alexey Karpov, K. Samudravijaya, K. T. Deepak, Rajesh M. Hegde, Shyam S. Agrawal, and S. R. Mahadeva Prasanna, editors, *Speech and Computer*, pages 579–589, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-48309-7. URL https://doi.or/10.1007/978-3-031-48309-7_46. 36, 48

[163] Emily E Redd. The Effect of an Artificially Flattened Fundamental Frequency Contour on Intelligibility in Speakers with Dysarthria. *All Theses and Dissertations. Paper 3229.*, 2012. URL http://scholarsarchive.byu.edu/etd/3229. 83

[164] Frank Rudzicz. 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, page 255–256, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595935731. doi: 10.1145/1296843.1296899. URL https://doi.org/10.1145/1296843.1296899. 83

[165] Frank Rudzicz. Learning mixed acoustic/articulatory models for disabled speech. In *Proc. NIPS 2010, – Workshop on Machine Learning for Assistive*

Technologies at the 24th annual conference on Neural Information Processing Systems, pages 70–78, 2010. URL https://www.cs.toronto.edu/~frank/Download/Papers/rudzicz_nips10.pdf. 14, 30, 74, 75, 83, 92, 102, 112, 122

[166] Frank Rudzicz. Articulatory Knowledge in the Recognition of Dysarthric Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):947–960, 2011. URL https://doi.org/10.1109/TASL.2010.2072499. 75, 92

[167] Frank Rudzicz. Adjusting dysarthric speech signals to be more intelligible. *Computer Speech & Language*, 27(6):1163–1177, 2013. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2012.11.001. URL https://www.sciencedirect.com/science/article/pii/S0885230812001003. Special Issue on Speech and Language Processing for Assistive Technology. 33, 47, 75, 83, 84, 93, 97, 102, 105, 112

[168] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The TORGO Database of Acoustic and Articulatory Speech from Speakers with Dysarthria. *Lang. Resour. Eval.*, 46(4):523–541, dec 2012. ISSN 1574-020X. doi: 10.1007/s10579-011-9145-0. URL https://doi.org/10.1007/s10579-011-9145-0. 17, 39, 53, 60, 84, 113, 114

[169] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *The Journal of the Acoustical Society of America*, 129(1):350–367, 02 2011. ISSN 0001-4966. doi: 10.1121/1.3514381. URL https://doi.org/10.1121/1.3514381. 17, 41, 113

[170] Hasim Sak, Andrew W. Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Interspeech*, 2014. URL https://api.semanticscholar.org/CorpusID:6263878. 62

[171] Justin Salamon and Juan Pablo Bello. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017. URL https://doi.org/10.1109/LSP.2017.2657381. 113

[172] Michael Saxon, Ayush Tripathi, Yishan Jiao, Julie M. Liss, and Visar Berisha. Robust Estimation of Hypernasality in Dysarthria With Acoustic Model Likelihood Features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2511–2522, 2020. URL https://doi.org/10.1109/TASLP.2020.3015035. 128

[173] Guilherme Schu, Parvaneh Janbakhshi, and Ina Kodrasi. On Using the UA-Speech and TORGO Databases to Validate Automatic Dysarthric Speech

Classification Approaches. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. URL https://doi.org/10.1109/ICASSP49357.2023.10095981. 30

[174] Siddharth Sehgal and Stuart Cunningham. Model adaptation and adaptive training for the recognition of dysarthric speech. In Jan Alexandersson, Ercan Altinsoy, Heidi Christensen, Peter Ljunglöf, François Portet, and Frank Rudzicz, editors, *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 65–71, Dresden, Germany, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-5112. URL https://aclanthology.org/W15-5112. 31, 50, 59, 75, 83, 92, 93, 103

[175] S. Selva Nidhyananthan, R. Shantha Selva kumari, and V. Shenbagalakshmi. Assessment of dysarthric speech using Elman back propagation network (recurrent network) for speech recognition. *International Journal of Speech Technology*, 19(3):577–583, Sep 2016. ISSN 1572-8110. URL https://doi.org/10.1007/s10772-016-9349-1. 103

[176] Seyed Reza Shahamiri. Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:852–861, 2021. URL https://doi.org/10.1109/TNSRE.2021.3076778. 35, 48, 123, 124

[177] Seyed Reza Shahamiri and Siti Salwah Binti Salim. Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of mfcc parameters and studying a speaker-independent approach. *Advanced Engineering Informatics*, 28(1):102–110, 2014. ISSN 1474-0346. doi: https://doi.org/10.1016/j.aei.2014.01.001. URL https://www.sciencedirect.com/science/article/pii/S1474034614000020. 32, 47, 75, 92

[178] S Shahnawazuddin, Nagaraj Adiga, Hemant Kumar Kathania, and B Tarun Sai. Creating speaker independent asr system through prosody modification based data augmentation. *Pattern Recognition Letters*, 131:213–218, 2020. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2019.12.019. URL https://www.sciencedirect.com/science/article/pii/S0167865519303952. 126

[179] Arun Kumar Shanmugam and Ramalatha Marimuthu. Chapter 15 - A Critical Analysis and Review of Assistive Technology: Advancements, Laws, and Impact on Improving the Rehabilitation of Dysarthric Patients. In Hemanth D. Jude, editor, *Handbook of Decision Support Systems for Neurological Disorders*, pages 263–281. Academic Press, 2021. ISBN 978-0-12-822271-3. doi: https://doi.org/10.1016/B978-0-12-822271-3.00001-3. URL https://www.sciencedirect.com/science/article/pii/B9780128222713000013. 151

[180] Harsh Vardhan Sharma and Mark Hasegawa-Johnson. State-Transition Interpolation and MAP Adaptation for HMM-based Dysarthric Speech Recognition. In Melanie Fried-Oken, Kathleen F. McCoy, and Brian Roark, editors, *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pages 72–79, Los Angeles, California, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-1310. 31

[181] Harsh Vardhan Sharma and Mark Hasegawa-Johnson. Acoustic model adaptation using in-domain background models for dysarthric speech recognition. *Computer Speech & Language*, 27(6):1147–1162, 2013. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2012.10.002. URL https://www.sciencedirect.com/science/article/pii/S0885230812000848. Special Issue on Speech and Language Processing for Assistive Technology. 31, 47, 75, 83, 92, 103

[182] Shi Yan. https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714. Viewed July 2024. 64

[183] Prashanth Gurunath Shivakumar and Panayiotis Georgiou. Perception Optimized Deep Denoising AutoEncoders for Speech Enhancement. In *Proc. Interspeech 2016*, pages 3743–3747, 2016. URL https://doi.org/10.21437/Interspeech.2016-1284. 93, 103

[184] Alexander B. Silva, Jessie R. Liu, Lingyun Zhao, Deborah F. Levy, Terri L. Scott, and Edward F. Chang. A Neurosurgical Functional Dissection of the Middle Precentral Gyrus during Speech Production. *Journal of Neuroscience*, 42(45):8416–8426, 2022. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1614-22.2022. URL https://www.jneurosci.org/content/42/45/8416. 7

[185] Sabine Skodda, Wenke Grönheit, and Uwe Schlegel. Intonation and Speech Rate in Parkinson's Disease: General and Dynamic Aspects and Responsiveness to Levodopa Admission. *Journal of Voice*, 25(4):e199–e205, 2011. ISSN 0892-1997. doi: https://doi.org/10.1016/j.jvoice.2010.04.007. URL https://www.sciencedirect.com/science/article/pii/S0892199710000767. 17, 41, 113

[186] Glenn T. Stebbins and Christopher G. Goetz. Factor structure of the unified Parkinson's disease rating scale: Motor examination section. *Movement Disorders*, 13(4):633–636, 1998. doi: https://doi.org/10.1002/mds.870130404. URL https://movementdisorders.onlinelibrary.wiley.com/doi/abs/10.1002/mds.870130404. 19

[187] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Two-step acoustic model adaptation for dysarthric speech recognition. In *ICASSP 2020*

- *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6104–6108, 2020. URL https://doi.org/10.1109/ICASSP40776.2020.9053725. 32, 47

[188] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology*, 9: 1112–1122, 2013. ISSN 2212-0173. doi: https://doi.org/10.1016/j.protcy.2013.12.124. URL https://www.sciencedirect.com/science/article/pii/S2212017313002788. CENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies. 75, 77

[189] D.J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982. URL https://doi.org/10.1109/PROC.1982.12433. 51, 55, 78

[190] Kris Tjaden and Greg Wilding. Rate and loudness manipulations in dysarthria. *Journal of Speech, Language, and Hearing Research*, 47(4):766–783, 2004. URL https://doi.org/10.1044/1092-4388(2004/058). 22, 42, 126

[191] Kris Tjaden and Greg Wilding. The Impact of Rate Reduction and Increased Loudness on Fundamental Frequency Characteristics in Dysarthria. *Folia Phoniatrica et Logopaedica*, 63(4):178–186, 10 2010. ISSN 1021-7762. doi: 10.1159/000316315. URL https://doi.org/10.1159/000316315. 77

[192] Kelvin Tran, Lingfeng Xu, Gabriela Stegmann, Julie Liss, Visar Berisha, and Rene Utianski. Investigating the Impact of Speech Compression on the Acoustics of Dysarthric Speech. In *Proc. Interspeech 2022*, pages 2263–2267, 2022. URL https://doi.org/10.21437/Interspeech.2022-10817. 23, 42

[193] Ming Tu, Visar Berisha, and Julie Liss. Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks. In *Proc. Interspeech 2017*, pages 1849–1853, 2017. URL https://doi.org/10.21437/Interspeech.2017-1222. 28, 45, 103

[194] Rosanna Turrisi, Arianna Braccia, Marco Emanuele, Simone Giulietti, Maura Pugliatti, Mariachiara Sensi, Luciano Fadiga, and Leonardo Badino. EasyCall Corpus: A Dysarthric Speech Dataset. In *Proc. Interspeech 2021*, pages 41–45, 2021. URL https://doi.org/10.21437/Interspeech.2021-549. 17, 41

[195] Bhavik Vachhani, Chitralekha Bhat, Biswajit Das, and Sunil Kumar Kopparapu. Deep Autoencoder Based Speech Features for Improved Dysarthric Speech Recognition. In *Proc. Interspeech 2017*, pages 1854–1858, 2017. URL

https://doi.org/10.21437/Interspeech.2017-1318. 11, 33, 47, 103, 108, 112, 120

[196] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu. Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. In *Proc. Interspeech 2018*, pages 471–475, 2018. URL https://doi.org/10.21437/Interspeech.2018-1751. 11, 34, 48, 122, 124, 138

[197] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964. 131

[198] VoxForge. http://www.voxforge.org/home/downloads. Viewed July 2024. 114

[199] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, 1989. URL https://doi.org/10.1109/29.21701. 104

[200] Chengyi Wang, Yu Wu, Yujiao Du, Jinyu Li, Shujie Liu, Liang Lu, Shuo Ren, Guoli Ye, Sheng Zhao, and Ming Zhou. Semantic Mask for Transformer Based End-to-End Speech Recognition. In *Proc. Interspeech 2020*, pages 971–975, 2020. URL https://doi.org/10.21437/Interspeech.2020-1778. 122

[201] Jun Wang, Prasanna V. Kothalkar, Beiming Cao, and Daragh Heitzman. Towards Automatic Detection of Amyotrophic Lateral Sclerosis from Speech Acoustic and Articulatory Samples. In *Proc. Interspeech 2016*, pages 1195–1199, 2016. URL https://doi.org/10.21437/Interspeech.2016-1542. 28, 45

[202] Jun Wang, Prasanna V. Kothalkar, Myungjong Kim, Andrea Bandini, Beiming Cao, Yana Yunusova, Thomas F. Campbell, Daragh Heitzman, and Jordan R. Green. Automatic prediction of intelligible speaking rate for individuals with als from speech acoustic and articulatory samples. *International Journal of Speech-Language Pathology*, 20(6):669–679, 2018. doi: 10.1080/17549507.2018.1508499. URL https://doi.org/10.1080/17549507.2018.1508499. PMID: 30409057. 26, 44

[203] Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, Christian Fuegen, Geoffrey Zweig, and Michael L. Seltzer. Transformer-Based Acoustic Modeling for Hybrid Speech Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP). IEEE, May 2020. URL http://dx.doi.org/10.1109/ICASSP40776.2020.9054345. 123

[204] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. ESPnet: End-to-End Speech Processing Toolkit. In *Proc. Interspeech 2018*, pages 2207–2211, 2018. URL https://doi.org/10.21437/Interspeech.2018-1456. 123, 132, 133

[205] Ka Ho Wong, Yu Ting Yeung, Edwin H. Y. Chan, Patrick C. M. Wong, Gina-Anne Levow, and Helen Meng. Development of a Cantonese dysarthric speech corpus. In *Proc. Interspeech 2015*, pages 329–333, 2015. URL https://doi.org/10.21437/Interspeech.2015-149. 17, 40, 113

[206] Chung-Hsien Wu, Hung-Yu Su, and Han-Ping Shen. Articulation-Disordered Speech Recognition Using Speaker-Adaptive Acoustic Models and Personalized Articulation Patterns. *ACM Transactions on Asian Language Information Processing*, 10(2), jun 2011. ISSN 1530-0226. doi: 10.1145/1967293.1967294. URL https://doi.org/10.1145/1967293.1967294. 75

[207] Feifei Xiong, Jon Barker, and Heidi Christensen. Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalised Dysarthric Speech Recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5836–5840, 2019. URL https://doi.org/10.1109/ICASSP.2019.8683091. 34, 48, 122, 126, 138

[208] Feifei Xiong, Jon Barker, Zhengjun Yue, and Heidi Christensen. Source Domain Data Selection for Improved Transfer Learning Targeting Dysarthric Speech Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7424–7428, 2020. URL https://doi.org/10.1109/ICASSP40776.2020.9054694. 35, 48, 124

[209] Lingfeng Xu, Julie Liss, and Visar Berisha. Dysarthria detection based on a deep learning model with a clinically-interpretable layer. *JASA Express Letters*, 3(1):015201, 01 2023. ISSN 2691-1191. doi: 10.1121/10.0016833. URL https://doi.org/10.1121/10.0016833. 29, 46

[210] S. Xue and Z. Yan. Improving latency-controlled BLSTM acoustic models for online speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5340–5344, March 2017. URL https://doi.org/10.1109/ICASSP.2017.7953176. 63

[211] Wei Xue. *Measuring the intelligibility of pathological speech through subjective and objective procedures*. [Doctoral dissertation, Radboud University] https://hdl.handle.net/2066/289696, 2023. 24, 46

[212] Wei Xue, Catia Cucchiarini, Roeland van Hout, and Helmer Strik. Measuring the intelligibility of dysarthric speech through automatic speech recognition in a pluricentric language. *Speech Communication*, 148:23–30, 2023. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2023.02.004. URL https://www.sciencedirect.com/science/article/pii/S0167639323000274. 30

[213] Emre Yilmaz, Mario Ganzeboom, Lilian Beijer, Catia Cucchiarini, and Helmer Strik. A Dutch Dysarthric Speech Database for Individualized Speech Therapy Research. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 792–795, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1127. 17, 40, 113

[214] Sarah E. Yoho, Tyson S. Barrett, and Stephanie A. Borrie. The Influence of Sensorineural Hearing Loss on the Relationship Between the Perception of Speech in Noise and Dysarthric Speech. *Journal of Speech, Language, and Hearing Research*, 66(10):4025–4036, 2023. doi: 10.1044/2023_JSLHR-23-00115. URL https://pubs.asha.org/doi/abs/10.1044/2023_JSLHR-23-00115. 103

[215] Kathryn M. Yorkston and David R. Beukelman. Ataxic dysarthria. *Journal of Speech and Hearing Disorders*, 46(4):398–404, 1981. doi: 10.1044/jshd.4604.398. URL https://pubs.asha.org/doi/abs/10.1044/jshd.4604.398. 22

[216] Steve J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book Version 3.4*. Cambridge University Press, 2006. 86

[217] Xu Yuanchao, Cai Zhiming, and Kong Xiaopeng. Improved pitch shifting data augmentation for ship-radiated noise classification. *Applied Acoustics*, 211:109468, 2023. ISSN 0003-682X. doi: https://doi.org/10.1016/j.apacoust.2023.109468. URL https://www.sciencedirect.com/science/article/pii/S0003682X23002669. 126

[218] Zhengjun Yue, Erfan Loweimi, Heidi Christensen, Jon Barker, and Zoran Cvetkovic. Dysarthric Speech Recognition From Raw Waveform with Parametric CNNs. In *Proc. Interspeech 2022*, pages 31–35, 2022. URL https://doi.org/10.21437/Interspeech.2022-163. 48, 124

[219] Emre Yılmaz, Mario Ganzeboom, Catia Cucchiarini, and Helmer Strik. Multi-Stage DNN Training for Automatic Recognition of Dysarthric Speech. In *Proc. Interspeech 2017*, pages 2685–2689, 2017. URL https://doi.org/10.21437/Interspeech.2017-303. 103, 112

[220] Emre Yılmaz, Mario Ganzeboom, Catia Cucchiarini, and Helmer Strik. Multi-Stage DNN Training for Automatic Recognition of Dysarthric Speech. In *Proc. Interspeech 2017*, pages 2685–2689, 2017. URL https://doi.org/10.21437/Interspeech.2017-303. 32, 47

[221] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter, and H. Ney. A comprehensive study of deep bidirectional LSTM RNNS for acoustic modeling in speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2462–2466, March 2017. URL https://doi.org/10.1109/ICASSP.2017.7952599. 63

[222] Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi H. Kinnunen, Zhenhua Ling, and Tomoki Toda. Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *ArXiv*, abs/2008.12527, 2020. URL https://api.semanticscholar.org/CorpusID:221370890. 127

Radboud Universiteit