Study, Analysis and Recognition of Dysarthric Speech

Thesis submitted in partial fulfilment of the requirements
for the award of the degree of

Doctor of Philosophy

in

Computer Science and Engineering

by

Komal Bharti

Under the supervision of

Prof Pradip K Das

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

January, 2025

Dedicated to

My loving Parents,

and

My dear Husband

Acknowledgment

First and foremost, I would like to express my heartfelt gratitude to my supervisor *Prof. Pradip K. Das* for his consistent support, endless patience, calm nature and positive guidance. His continued encouragement inspired me to do better each day throughout my PhD journey. His guidance has shaped me not only into a better researcher but also a better individual. The life lessons I have learned from him will remain with me forever. I feel truly fortunate to have him as my supervisor and will always be deeply grateful and indebted to him. His encouraging words and guidance have been a source of great motivation throughout my doctoral journey.

I am thankful to my thesis doctoral committee members - *Dr. Pinaki Mitra*, *Prof. S.B Nair*, and *Dr. Ajay Dashora* for their percipient and valuable contributions to my research. Their thoughtful comments, constructive suggestions, and critical evaluations have significantly enhanced the quality, depth, and clarity of my work. Their expertise and feedback have played a crucial role in shaping this thesis. I would like to acknowledge the heads of the Department of CSE during my Ph.D. journey at IITG - Prof. Jatin Deka and Prof. T. Venkatesh for providing me the facilities and resources, including the travel support for the conferences. I appreciate the Technical staff of the Department of Computer Science and Engineering - Mr. Nanu Alan Kachari, Mr. Bhriguraj Borah, Mr. Hemanta Kumar Nath, Mr. Raktajit Pathak, Mr. Pranjit Talukdar, and Mr. Nava Kumar Boro for their welcomely support on any engineering-related issues. I am grateful to - Mrs. Gauri Khuttiya Deori, Mrs. Mithu Lakra, Mr. Gourish Mazumder, and Mr. Prabin Bharali for efficiently handling the administrative work. I am obliged to all the faculty members, the staff, and security personnel for their constant help and support.

I am thankful to all my colleagues and friends during my journey as a Ph.D. scholar. I am indeed thankful to my fellow lab mates at the Speech and Robotics lab - P. Bhagath, Deepankar, Thiyam Susma, Vanshali, Ashita, Himanshu, Sandeep, Darshika, Manpreet, Brajesh, Prarbdh and many more for creating a wonderful experience at my workplace. The engaging discussions, collaborative brainstorming, and collective efforts have greatly contributed to my growth and development as an independent researcher. I have enjoyed gathering with my friends- Sandipan, Imlijungla, Saurav Kumar, Soumya, Ajanta, Lakshita, Swati, Gayathri, Meenu, Minaxi and Suraj in IITG for all the events we have organized and participated in. The serene environment and lush greenery of the campus have always brought me a sense of calm and peace. Whenever I felt stressed, being surrounded by nature helped me relax and rejuvenate.

I am blessed to have good friends - Karnish, Manoj and Saurav with whom I have shared some indelible moments of my life at IITG. The countless cherished moments we shared will always hold a special place in my heart. Their unwavering support and willingness to help me in every situation, no matter how challenging, have been invaluable. Their warmth, care, and affection turned this workplace into a second home for me. This journey would not have been as smooth or memorable without their friendship.

Finally, but most importantly, I would like to thank Almighty God and my family - Mom, Dad, my younger brothers - Asish, and my elder sister Nisha, for their boundless love, support, caring, warmth, and encouragement in all these years. I am truly indebted to them. Last but certainly not least, I would like to express my deepest gratitude to my husband, who has been my pillar of strength throughout my PhD journey. His unwavering support and encouragement kept me focused and dedicated, even during the most challenging times. He stood by me through every hurdle, inspiring me to persevere and dream big with his confidence in my abilities. I will forever be thankful for his love, patience, and belief in me.

January 13, 2025 Komal Bharti

Declaration

I certify that

- The work contained in this thesis is original and has been done by myself and under the general supervision of my supervisor.
- The work reported herein has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (concepts, ideas, text, expressions, data, graphs, diagrams, theoretical analysis, results, etc.) from other sources, I have given due credit by citing them in the text of the thesis and giving their details in the references. Elaborate sentences used verbatim from published work have been clearly identified and quoted.
- I also affirm that no part of this thesis can be considered plagiarism to the best of my knowledge and understanding and take complete responsibility if any complaint arises.
- I am fully aware that my thesis supervisor is not in a position to check for any possible instance of plagiarism within this submitted work.

January 13, 2025 Komal Bharti

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

Prof. Pradip K. Das
Professor
Email : pkdas@iitg.ac.in
Phone : +91-361-2584353

Certificate

This is to certify that this thesis entitled "**Study, Analysis and Recognition of Dysarthric Speech**" submitted by **Komal Bharti**, in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy, to the Indian Institute of Technology Guwahati, Assam, India, is a record of the bonafide research work carried out by him under my guidance and supervision at the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Assam, India. To the best of my knowledge, no part of the work reported in this thesis has been presented for the award of any degree at any other institution.

Date: January 13, 2025
Place: Guwahati

Prof. Pradip K. Das
(Thesis Supervisor)

Study, Analysis and Recognition of Dysarthric Speech

Abstract

Dysarthria is a term derived from two parts: ‘dys’ which signifies having difficulties, and ’arthr’ refer to articulation, is a neurological speech disorder that majorly happens due to cerebral strokes or significant traumatic incidents. It is characterized by a spectrum of speech impairments, including but not limited to unintelligible speech, inconsistent speech pace, atypical speech prosody, slurred speech, poor voice quality and imprecise articulation. As the severity of the condition increases, the coordination between the movements of lips and tongue deteriorates, resulting in highly unintelligible speech. Compared to healthy speech, dysarthric speech is much more challenging to recognize due to inconsistencies in the acoustic signal and limited data availability.

This study introduces a new approach that is based on the combination of recognition, characterization, synthesis and human assessment of dysarthric speech. The goal is to enhance the performance of automatic speech recognition (ASR) systems for this class of people. Additionally, this approach aims to support dysarthric speech assessment process, ensuring that proper treatment is provided with less intervention. We aim to bridge the gap between dysarthric speakers and their interactions with machines, reducing complexity and ultimately improving their quality of life. This study primarily focuses on a speaker-adaptive approach, as the characteristics and behavior of each speaker vary significantly depending on the severity of their condition. For over a decade, researchers have been trying to improve the ASR system and rehabilitation for dysarthric speech, but they still lag behind and still, there is a lot of scope for improvement. Early efforts focused on HMM-based hand-crafted features, which could not handle the variability of dysarthric speech. With the development of new datasets, attention shifted to neural networks and deep learning techniques. This study explores different algorithms and methods that could help them lead better lives.

We investigate the use of the Affinity Propagation (AP) algorithm for dysarthric speech segments to select the most informative feature set, which captures key information about the speaker. As we switch to larger datasets, we explore the LSTM-RNN architecture with a fusion of multiple audio descriptors. This additional information is provided to the classifier to enhance its ability to accurately identify each word and we achieved 83.11% overall accuracy. However, data scarcity is a persistent challenge in the field of dysarthria. Given the limitations of available datasets, it is difficult for researchers to develop a robust ASR system. To address this, we utilized the existing UASpeech dataset and expanded it by generating new words that retain the same characteristics and behaviors of the speakers. This contribution enables the expansion of the dataset in both breadth and depth. This thesis investigates the detection and severity classification of dysarthric speech using the Audio Spectrogram Transformer, which is an essential component for the entire treatment and recovery. Our model achieved an accuracy of 99.64%, surpassing the state-of-the-art results. We also explore dysarthric speech at the phoneme level to pinpoint the specific areas where speakers face difficulties. We identified the phoneme sets in which they most frequently misarticulate using the Goodness of Pronunciation (GOP) algorithm. These insights will assist both speakers and speech pathologists in facilitating early recovery. To ensure the usability of the UASpeech database, we annotated every audio file and obtained the ground truth information for the dataset. This annotation process allowed us to accurately capture the essential details of the data, making it suitable for further research and analysis. Based on various experimental results, it is clear that the proposed method and findings significantly enhance the quality of life for dysarthric individuals.

Contents

Listing of figures

List of Tables

List of Abbreviations

Terms	**Abbreviations**
ASR	Automatic Speech Recognition
ABI	Acquired Brain Injury
TBI	Traumatic Brain Injury
SLP	Speech Language Pathologist
WHO	World Health Organization
LSTM	Long-Short Term Memory
RNN	Recurrent Neural Network
HMM	Hidden Markov Model
NN	Neural Network
WRA	Word Recognition Accuracy
AST	Audio Spectrogram Transformer
SD	Speaker Dependent
SI	Speaker Independent
OOV	Out of Vocabulary
MOS	Mean Opinion Score
FDA	Frenchay Dysarthria Assessment
DS-MOS	Dysarthria Similarity-MOS
WSOLA	Waveform Similarity-based Synchronized OverLap-Add
GoP	Goodness of Pronunciation
GMM	Gaussian Mixture Model
IPA	International Phonetic Alphabet
MFA	Montreal Forced Aligner
DTW	Dynamic Time Warping
RI	Random Index
AP	Affinity Propagation
LPCC	Linear Prediction Cepstral Coefficient
SP	Speech Pathologist
MFCC	Mel Frequency Cepstral Coefficients
CNN	Convolution Neural Network
DNN	Deep Neural Network

WER	Word Error Rate
API	Application Programming Interface
CPU	Central Processing Unit
GPU	Graphics Processing Unit
CUDA	Compute Unified Device Architecture
MAF	Multiple Acoustic Frames
AAM	Active Appearance Model
WFST	Weighted Finite State Transducers
KPOW	Korean Phonetically Optimized Words
VC	Voice Conversion
TTS	Text-to-Speech
LPCNet	Linear Predictive Coding Network
VAE-VC	Variational Autoencoder-based Voice Conversion
DBN	Deep Belief Networks
GNRE	Glottal-to-Noise Excitation Ratio
HNR	Harmonics-to-Noise Ratio
CP	Cerebral Palsy
ALS	Amyotrophic Lateral Sclerosi
LCD	Liquid-Crystal Display
RIFF	Resource Interchange File Format
PLPCC	Perceptual Linear Predictive Cepstral Coefficients
STE	Short Term Energy
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
PPQ	Period Perturbation Quotient
APQ	Amplitude Perturbation Quotient
SHIRC	Speech and Hearing Institute & Research Centre
AIISH	All India Institute of Speech and Hearing
NIMHANS	National Institute of Mental Health and NeuroSciences
IISc	Indian Institute of Science
UQ	Uncertainty Quantification

The secret of getting ahead is getting started.

Mark Twain, American writer

1 Introduction

Speech and language are fundamental to human life and differentiate us from other animals. From childhood to late adulthood, with the help of speech, we learn a lot of things and adapt ourselves to a particular environment. Speech is the use of the human voice as a medium for language. Through language, individuals express their identities, beliefs, emotions and personality. This is the most suitable and straightforward technique for conveying our messages and ideas to someone. Speech forms the basis of most human interactions. This trait develops instinctively along with other signs of average growth and development. Speech comprises sounds, letters, language and verbal or non-verbal gestures that aid humans in communication. Speech and language affect individuals, significantly shaping their daily lives.

1.1 Production of speech sounds

Generally, we don't stop to think about the various movements involved when we speak a word—we hear it, observe it, interpret the rules and then produce it. However, producing speech involves many quick, precise movements that blend together to form words. The speech process starts in the brain in the form of a message followed by a lexico-grammatical structure. To convert this message to an utterance, we need to represent this message in the form of sound with the help of a speech production mechanism. Speech production is a complex process that involves lungs, vocal folds, trachea and nasal cavities. An air stream from the lungs is expelled towards the trachea. The Larynx (voice box) is present between lungs and mouth and is responsible for the phonation process. It has two folds of tissues known as vocal folds or vocal cords. There are some gaps between these folds known as glottis. The larynx vibrates to create a voiced sound. When glottis is open, vibration of the vocal folds is reduced and creates voiceless sounds (breathing). This vibration travels through the vocal tract (oral and nasal cavity) and takes its shape and gets resonated. Then, the tongue, lips, teeth and jaw work together to make the sound into recognizable speech sounds and the processed sound waves come out as spoken words. These coordinated movements produce the individual sounds that come together to form a meaningful word.

The basic elements of speech include articulation, fluency and voice. Articulation refers to how these speech sounds are produced. We should be able to produce an "S"

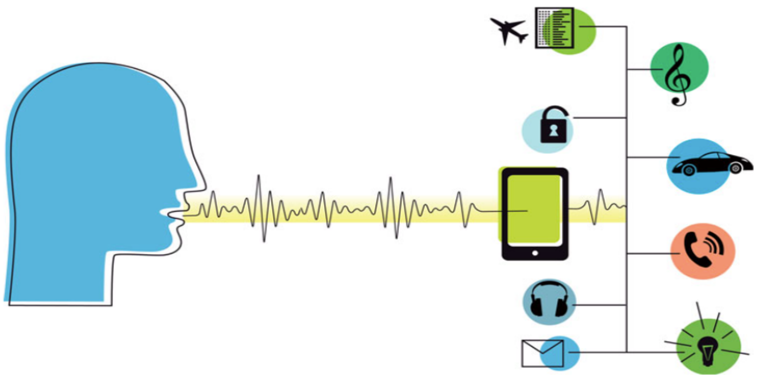


Figure 1.1: Application of speech recognition technology

sound to say “sun”. The pattern or rhyme of our speech is known as fluency. Voice is produced through the use of vocal cords and breathing patterns and differs from person to person that how they are using it to generate voice. Voice can be high or low-pitched, loud or soft, depending on the usage of vocal cords.

1.2 Importance of speech technology

Nowadays, speech technology is applied across a wide range of fields, including home automation, security systems, public places, navigation, healthcare, drones, satellites and voice assistance like Cortana, Alexa and Siri as illustrated in Figure 1.1. These applications help to improve the quality of life by making everyday tasks easier and more convenient. Speech technology is changing the way we communicate with machines and with each other, creating more intuitive and efficient experiences. It also makes technology more accessible, allowing more people to benefit from its advancements in various fields. While most people have the privilege of using speech technology to improve their lives, there are many all around the world who cannot fully benefit from it due to speech disorders. These individuals face challenges that prevent them from taking advantage of the same advancements.

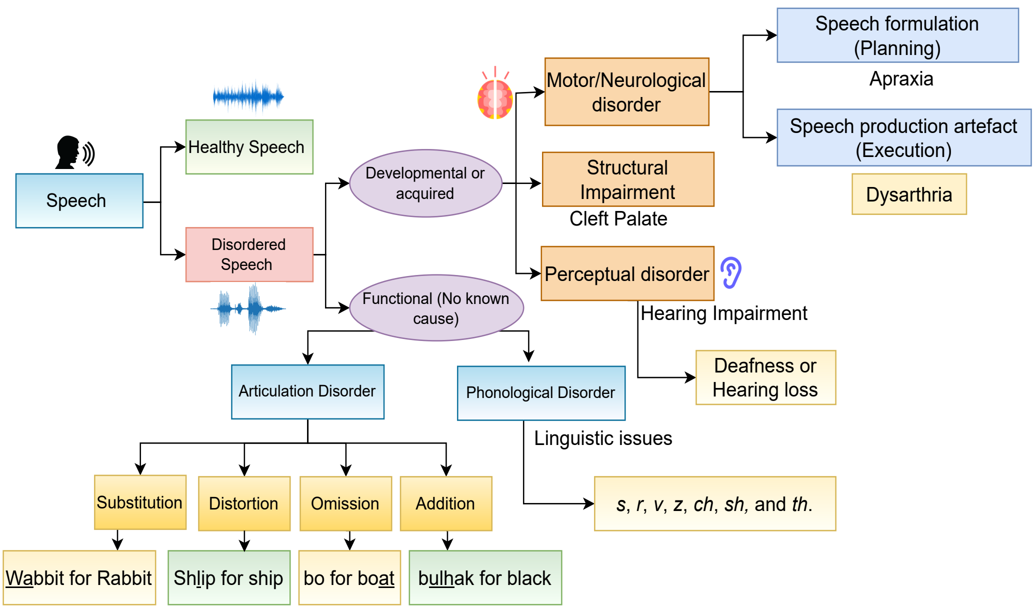


Figure 1.2: Structure of speech disorders classification

1.3 Speech disorders

A particular condition that creates problems in the formation of speech sounds while communicating with others is termed as speech disorder. This is a broad term that includes any problem related to speech sounds. At an early age, when children start learning languages by listening to others, it is obvious that they often mispronounce words, misspell them or struggle to say certain words correctly. While this garble can seem amusing at that age, if the same issues continue into pre-teen years, it could indicate a serious speech problem. This impairment create problems for others to understand their speech and impacts the individual's ability to speak fluently. They struggle to communicate on a daily basis leading to reduced self-confidence and hesitation to engage in social interactions. Based on the causes behind speech disorders, it is categorized into **Organic speech sound disorders** and **Functional speech sound disorders**. Organic speech disor-

Type of Dysfluency	Description	Example
Repetition of word	Person Who Stutter (PWS) repeats the whole word.	How-How is my mother?
Repetition of part-word	PWS repeats the sound or syllable.	H-H-How is my mother?
Prolongation	PWS continues a sound for a long time.	How is my mmmmm-mother?
Interjection	PWS add some meaningless words to a sentence.	How is ummm my mother?
Repetition of phrases	PWS repeats the whole phrase.	How is-How is my mother?
Broken words	PWS gives pauses within words.	How is my mo-(pause)-ther?
Blocks	PWS gives a long silence between words.	How is my (block) mother?

Table 1.1: Several types of dysfluencies

ders results from underlying neurological (Dysarthria and Apraxia), structural (physical abnormality in the face or mouth, such as a cleft lip or palate), or perceptual (hearing impairment) cause and can develop with the age or can be acquired at any point in life due to injury or illness. Functional speech sound disorders on the other hand, are idiopathic, meaning they have no identifiable physical cause and majorly classified into articulation and phonological disorders [34][36][32]. A detailed diagram is provided in Figure 1.2 and a few common speech disorders are explained in the next subsection.

1.3.1 Stuttering

When there is a constant abnormal halt in the regular flow of speech or several repetitions in the syllable, this behavior is called stuttering. Stuttering is also known as stammering. This is a problem in which fluency plays a vital role. If a person is suffering from this problem, they may repeat the first part of the word like “pr-pr-pr-oblem” for “problem” or may incessantly hold on to a particular sound like “compu-pu-pu-pu-ter” for word “computer”. Anyone can tackle the dysfluency while speaking for a while, but that is not stuttering. There are several types of dysfluencies shown in Table 1.1.

1.3.2 Lisp

This is a condition in which persons finds difficulty in learning to produce a particular sound or a few specific sounds in their speech. It is common for functional speech disorders to have difficulty in the pronunciation of ‘s’, ‘z’, ‘r’, ‘l’ and ‘th’. Lisping refers to the particular replacement of words involving the letters “s” and “z” with “th”. A person who lisps usually has their tongue between their teeth or touches them. Speech-Language Pathology Graduate Programs outlines that there are four professional categories of lisps:

1. Frontal lisp: The frontal lisp is the most common form of lisping, which occurs when a person puts their tongue too far forward.
2. Lateral lisp: When air moves over the sides of the tongue while speaking, it produces a slurred sound and is termed as lateral lisp.
3. Palatal lisp: With palatal lisps, certain sounds are uttered with the tongue touching the roof of the mouth.
4. Dental lisps: It occurs when the tongue pushes against the teeth and this is easily confused with a frontal lisp.

A person’s intelligibility is not overly affected by a few cases of lisping, as this is an isolated speech characteristic. A person with a lisp is usually easily understood by most people.

1.3.3 Apraxia

Apraxia is a neurological disorder that affects motor speech production. People with this disorder face difficulty moving their auditory muscles and they can not construct the structures needed to form speech sounds into words. Their brain does not coordinate with the movement to speak that particular word because the brain gets a weak signal or no signal at all [36][54]. A major issue occurs when the brain struggles to determine which part of the speech production system to move, when to move it and for how long. Imagine someone trying to say a long word with many syllables, but the brain is not helping to sequence and emphasize each syllable properly. This confusion over how to produce and control the correct sounds causes the brain to randomly change the sounds the person is attempting to produce. Every time they try to sequence those sounds, they have to start over, which becomes exhausting and frustrating. Apraxia of speech usually includes errors at phoneme level, including dysfluency, atypical prosody and consonant and vowel substitutions. There are many more speech disorders with different characteristics and causes. We discuss about **Dysarthria**, the main focus of this thesis, in the next section.

1.4 Dysarthria

Dysarthria is a term derived from two parts: 'dys' which signifies having difficulties and 'arthr' which refers to articulation, a neurological motor speech disorder characterized by inadequate synchronization of speech production subsystems. Neurodegenerative diseases like cerebral palsy and Parkinson's disease typically cause dysarthria or can be acquired through neurological injuries such as stroke, severe accidents, paralysis or brain tumors. Serious injury to the left hemisphere of the brain reduces the coordination and movement of speech-related muscles, consequently deteriorating speech quality. Speech and language problems can also occur with right hemisphere brain damage, but these are rare [5][157][35][16].

It is characterized by a spectrum of speech impairments, including but not limited to unintelligible speech, inconsistent speech pace, atypical speech prosody, poor voice quality, imprecise articulation, low audibility, slurred or mumbled voice, inter and intra-speaker variability and irregular speech rate [*][164]. As severe cases of dysarthria progress, the speech pace can drop down to 15 words per minute or even lower and the rate of speech can be 10-17 times slower than normal speech. Individuals with dysarthria fully comprehend syntactically flawless sentences and know what they want to communicate and how to express it. However, weakened muscles make it challenging for them to produce phonetically correct sentences. They experience changes in pitch or voice quality as well. A slow speech is not only more laborious for the speaker and listener, but it also creates several acoustical problems. There are several other types of dysfluencies commonly associated with dysarthria (especially when coupled with Apraxia) such as hesitation (e.g. false starts) and repetition. It is nearly impossible to grasp an utterance when there is severe dysfluency in the phrasing.

In this condition, it becomes more difficult for individuals to communicate with others and eventually, they feel neglected and unworthy over time. Notably, as a result of severe accidents, they have several other physical impairments also making them dependent on others. Dysarthria is not a life-threatening disorder, but it affects the livelihood of patients in many aspects, including social, physical and emotional challenges. As severity increases, they are more likely to rely on others for their daily activities and household chores. Speech-based assistance and keyboard or joystick-based applications also do not help much due to lack of muscle coordination and trembling hands. According to R. Duffy [156] Dysarthria is formally defined as:

"A collective name for a group of neurologic speech disorders resulting from abnormalities in the strength, speed, range, steadiness, tone, or accuracy of movements required

[*] https://pubs.asha.org/doi/10.1044/jshr.3502.296

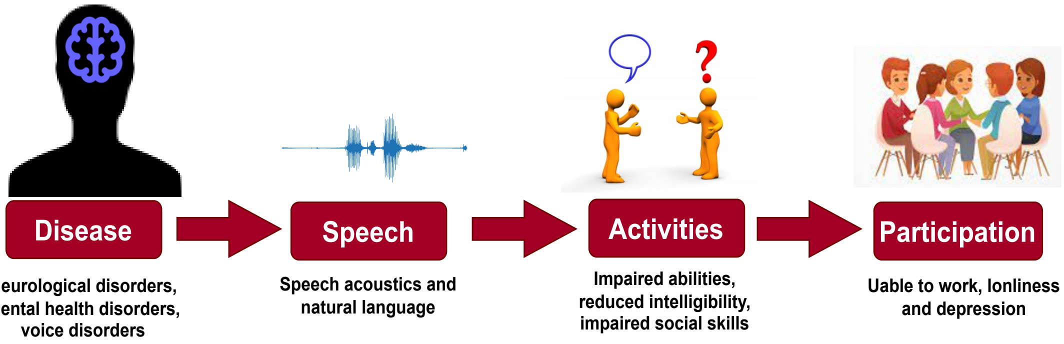

Disease
Neurological disorders, mental health disorders, voice disorders
Speech
Speech acoustics and natural language
Activities
Impaired abilities, reduced intelligibility, impaired social skills
Participation
Uable to work, lonliness and depression
?

Figure 1.3: Speech based biomarkers

for control of the respiratory, phonatory, resonatory, articulatory and prosodic aspects of speech production."

1.5 Dysarthria and its causes

For human speech to be produced, several components of the vocal apparatus must work together in a highly coordinated manner. The initiation of speech production begins in various regions of the brain, where ideas are organized, planned and sequenced through a coordinated process. Neural signaling regulates speech production, which sends a coordinated and timed message to the musculoskeletal system. There is, however, a possibility that the musculoskeletal structure will not work correctly if the central or peripheral nervous system is damaged. When the speech muscles are damaged, they become weak and uncoordinated, resulting in a group of motor speech disorders. As a consequence of such neurological impairments, motor speech disorders often result in dysarthria [137]. Rather than only affecting the musculoskeletal structure due to neural damage, dysarthria affects various other body parts as well. Swigert et al.[142] identify five systems that can be affected by dysarthria: respiration, phonation, resonance, articulation and prosody. Based on the severity and extent of dysarthria, it can also impact multiple aspects of the

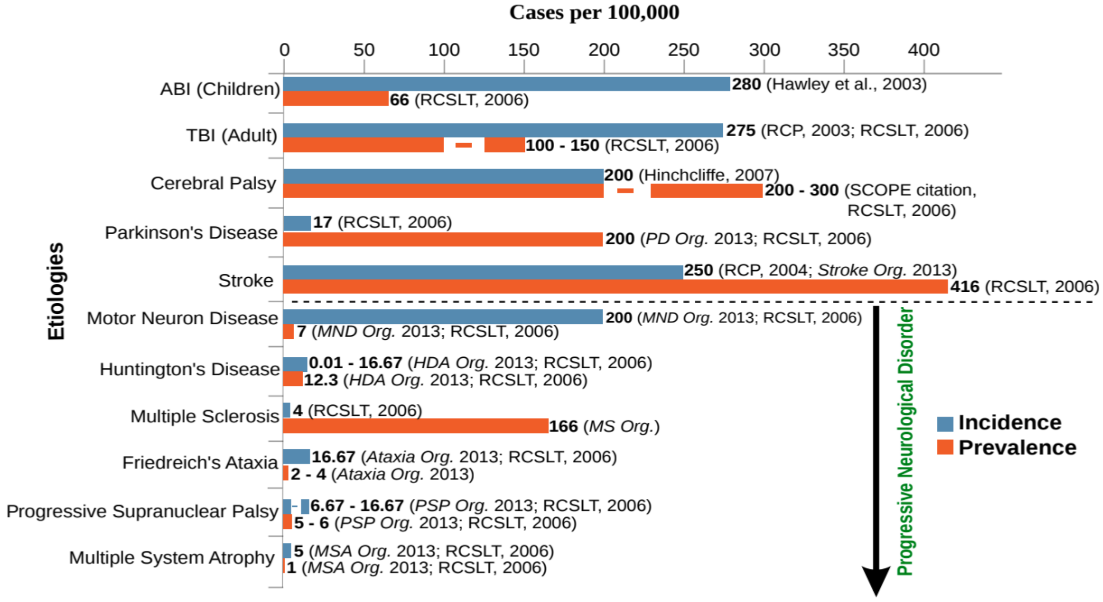
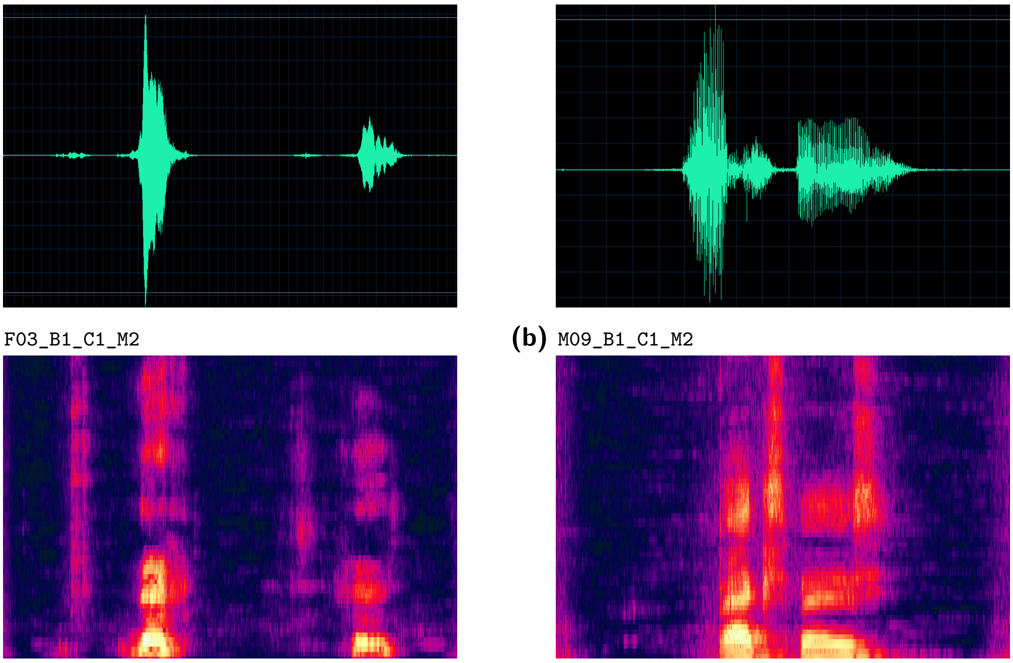


Figure 1.4: Incidence and prevalence of dysarthria [125]

supraglottal, laryngeal and subglottal systems. The nature of dysarthria can be either developmental or acquired. Developmental dysarthria occurs as a result of brain damage during fetal development or at birth. Conditions such as cerebral palsy can lead to dysarthria in children. Acquired dysarthria occurs due to brain damage sustained later in life.

1.6 Statistics of dysarthria etiologies

The occurrence and prevalence of the various forms and etiologies of dysarthria in the UK are not estimated by government statistics. According to rough estimates, 1% of the UK population is diagnosed with neurological illness each year, comprising both progressive and non-progressive disorders and not all neurological disorders result in dysarthria. The incidence is the number of newly diagnosed cases of an etiology over time and prevalence is the actual number of cases resulting in dysarthria over time or on the specific day the



(a) F03_B1_C1_M2

(b) M09_B1_C1_M2

(c) F03_B1_C1_M2

(d) M09_B1_C1_M2

Figure 1.5: Waveforms (a) (b) and spectrograms (c) (d) of the word "Backspace" from two different speakers F03 and M09 with different intelligibility rates

data was collected. With 416 cases of dysarthria per 100,000 people, stroke is the most significant disorder frequency in the UK. Cerebral palsy follows, with a prevalence of 200–300 cases per 100,000 people. Progressive neurological disorders are the umbrella term for the etiologies range from motor neuron disease to multiple system atrophy. Referring to Figure 1.4, it has been reported that the major reason behind dysarthria is stroke, Acquired Brain Injury (ABI) and Traumatic Brain Injury (TBI) having 250, 280 and 275 cases per 100,000 people, respectively. According to Indian Census 2011, the total speech-disabled population is 19,98,692[*].

[*] https://censusindia.gov.in/nada/index.php/catalog/43469

1.7 Challenges associated with dysarthric speech

Dysarthria is a degenerative speech disorder, meaning that once the symptoms are accurately identified, we can promptly take steps toward recovery and medical treatment. Dysarthria is associated with various symptoms. The following subsections discuss dysarthric speech symptoms.

- **Inconsistent speech intelligibility**: Depending on the severity of dysarthric speech, intelligibility refers to the extent to which a listener can accurately comprehend the speaker's intended message. Based on the intelligibility of speech, dysarthria is categorized into very low, low, mid and high intelligibility ranges. However, this categorization varies in literature and there is no fixed threshold defined.
- **Speaker variability**: In dysarthric speech, both inter and intra-speaker variability can be found. Fluctuations in speech performance within the same speaker are influenced by fatigue, emotional state, task complexity and age. Differences in speech patterns of the same word are quite large among all speakers due to the different severity levels and environmental factors. This difference is clearly visible between the waveform and spectrogram in Figure 1.5 for the same word "Backspace" taken from two different patients with different intelligibility rates. Speaker F03 has severe dysarthria with a 6% intelligibility rate, while speaker M09 has mild dysarthria with an 86% intelligibility rate.
- **Imprecise articulation**: They can't control their tongue, lip and jaw movement, which leads to unclear and distorted sound, making it difficult to understand for another listener. They face issues in the pronunciation of vowel sounds.
- **Abnormal voice quality**: Dysarthric speakers experience an abnormal speech rate,

often marked by multiple pauses and breakdowns. They may struggle with controlling pitch and loudness, have inadequate breath support and exhibit abnormal prosody, all of which disrupt the natural flow of speech. These issues affect the speaker's ability to maintain a steady rhythm, making their speech sound disjointed and harder to understand.

- **Limited dataset availability**: There are three English datasets available for dysarthric speech, yet none of the datasets provide a diverse range of speakers and their conversational speech is also very less, making them incapable of handling ASR (Automatic Speech Recognition) challenges. Moreover, the limited number of speakers can not capture the acoustic nature and inter-speaker variability. There are a lot of factors that make the data collection process lengthy and tedious.

  Speakers are very emotional; and when they find someone willing to listen, they often want to share all of their challenges. Factors such as their native language and literacy rate significantly influence the data collection process. Additionally, due to muscle fatigue, they are unable to talk for extended periods. Even when all conditions for data collection are met, the speaker or their family may refuse to allow speech data to be collected.

1.8 Types of dysarthria

Depending on where the brain damage occurs, the type of dysarthria differs. Dysarthria is generally irreversible but there are some possibilities to improve speech quality and communication skills through specific speech therapies. Physical activity like the movement of the arm, leg or face can also become restricted depending on the part of the injured brain. In each hemisphere of the brain, motor control and sensation are handled

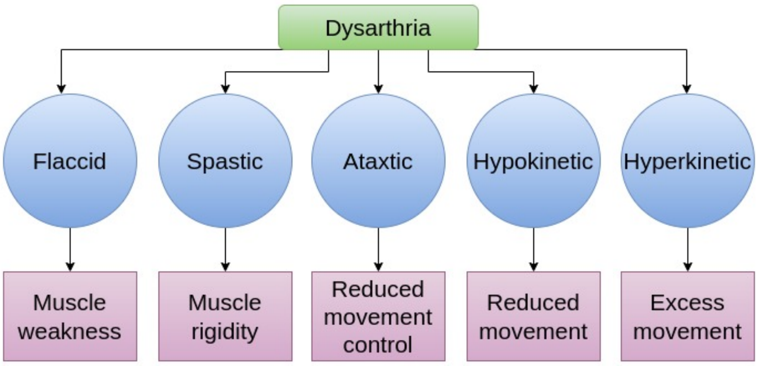


Figure 1.6: Different types of dysarthria and their symptoms

by the opposite hemisphere. So, the person with dysarthria may have physical deficits on their right side of the body also. Figure 1.6 describes different kinds of dysarthria with their pathophysiological reason [17][16].

1.8.1 Flaccid dysarthria

Dysarthria caused by damaged lower motor neurons is called flaccid dysarthria involving spinal and cranial nerves. Flaccid dysarthria majorly happens due to damage to the peripheral nervous system. A spinal nerve connects the spinal cord to other parts of the body and cranial nerves connect brain and neck cells. Individuals suffering from Flaccid dysarthria can be characterized by a breathy voice, reduced speed, tongue fasciculation and muscle weakness.

1.8.2 Spastic dysarthria

It is possible for people who suffer from spastic dysarthria to have speech problems along with muscle weakness and abnormal reflexes. This involves areas that control movements. Those with spastic dysarthria have damaged motor neurons in the central nervous system. An upper motor neuron is damaged on one or both sides of the brain.

Individuals suffering from Spastic dysarthria can be characterized by muscle rigidity, strained vocal quality and “grunting” at the end of phrases.

1.8.3 Ataxic dysarthria

A person with ataxic dysarthria produces slurred speech and is deficient in speech coordination. A person can have this type of dysarthria if their cerebellum is damaged. In the brain, the cerebellum receives sensory information and regulates movement. Irregular articulatory breakdowns, slurred speech with excessive and equal stress and harsh voice are the characteristics of Ataxtic dysarthria.

1.8.4 Hypokinetic dysarthria

This is caused by a malfunction in the extrapyramidal system of the brain. This system includes brain areas where subconscious muscle movements are coordinated. There is a strong correlation between this type and Parkinson’s disease. People with Hypokinetic dysarthria are characterized by rapid rate, reduced loudness, mono pitch, flat pitch, distorted vowels and mono loudness.

1.8.5 Hyperkinetic dysarthria

In hyperkinetic dysarthria, brain regions referred to as basal ganglia are damaged, causing the disorder. In the center of the brain, the Basel ganglia form essential connections between neurons. Different areas of the brain communicate with each other through these connections. Speech production is mostly unpredictable in this type of dysarthria.

1.9 Assessment and treatment

Accurate diagnosis and appropriate treatment are vital for enhancing communication abilities and overall quality of life. Speech therapy is a widely practiced and effective treatment for dysarthria, offering personalized interventions and support based on each individual's unique needs and goals. The therapy involves a range of techniques and exercises to address the underlying causes of dysarthria. These include vocal exercises, articulation drills, breathing exercises and language development activities.

The availability of Speech-Language Pathologists (SLPs) is limited and the cost of treatment is often high. As a result, motor speech therapy tends to be a lengthy process, involving multiple sessions with SLPs. To achieve an effective and accurate outcome, individuals must attend numerous appointments. Unfortunately, many people with speech disorders struggle to meet this requirement due to factors like financial limitations, transportation expenses, scheduling conflicts and other health problems. Major challenges can be summarised as:

- Unavailability or rarity of Speech-Language Pathologists (SLPs)
- Treatment costs are high and time taking
- Financial constraints and transportation
- Judgment Inconsistency
- Outcomes of therapies are sometimes not noticeable
- Inter-judge and intra-judge variations can often give variable results.

Despite these challenges, there has been a continuous effort over the past five decades to establish a solid foundation for effectively managing and treating various motor speech disorders. However, a complete discussion of the treatment and management approach is outside the scope of this thesis, but a brief summary is important.

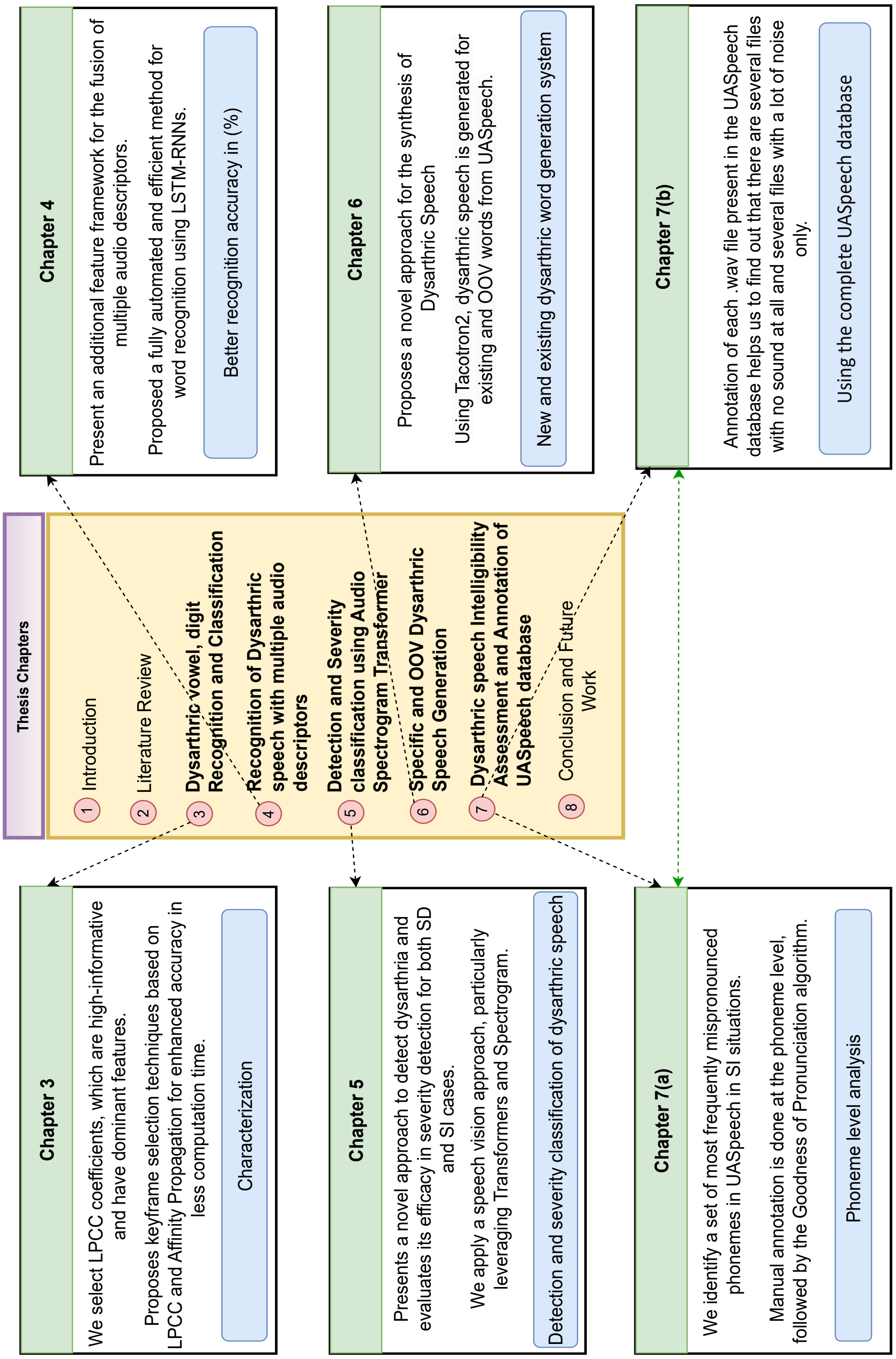

Thesis Chapters
1 Introduction
2 Literature Review
3 Dysarthric vowel, digit Recognition and Classification
4 Recognition of Dysarthric speech with multiple audio descriptors
5 Detection and Severity classification using Audio Spectrogram Transformer
6 Specific and OOV Dysarthric Speech Generation
7 Dysarthric speech Intelligibility Assessment and Annotation of UASpeech database
8 Conclusion and Future Work
Chapter 3
We select LPCC coefficients, which are high-informative and have dominant features.
Proposes keyframe selection techniques based on LPCC and Affinity Propagation for enhanced accuracy in less computation time.
Characterization
Chapter 4
Present an additional feature framework for the fusion of multiple audio descriptors.
Proposed a fully automated and efficient method for word recognition using LSTM-RNNs.
Better recognition accuracy in (%)
Chapter 5
Presents a novel approach to detect dysarthria and evaluates its efficacy in severity detection for both SD and SI cases.
We apply a speech vision approach, particularly leveraging Transformers and Spectrogram.
Detection and severity classification of dysarthric speech
Chapter 6
Proposes a novel approach for the synthesis of Dysarthric Speech
Using Tacotron2, dysarthric speech is generated for existing and OOV words from UASpeech.
New and existing dysarthric word generation system
Chapter 7(a)
We identify a set of most frequently mispronounced phonemes in UASpeech in SI situations.
Manual annotation is done at the phoneme level, followed by the Goodness of Pronunciation algorithm.
Phoneme level analysis
Chapter 7(b)
Annotation of each .wav file present in the UASpeech database helps us to find out that there are several files with no sound at all and several files with a lot of noise only.
Using the complete UASpeech database

Figure 1.7: Layout of the thesis

1.10 Motivation

According to the fact sheets of the World Health Organization (WHO) updated on 7th March 2023, 1.3 billion people are estimated to have significant disabilities, which equates to 16% of the world's population. According to the Census 2011 by the Government of India, 2.21% of the total Indian population, which is around 21.9 million people, have a speaking disability. This is a significant number of individuals that need to be supported and taken care of.

Apart from this, my personal experience with dysarthria had a profound impact on me. My grandmother suffered from dysarthria following a stroke at the age of 74. She often tried to communicate, but due to the severity of her condition, none of us in the family could understand what she was trying to convey. The family witnessed the pain that she suffered very closely every day. Later we came to know that there are a lot of people like her suffering from dysarthria. Depending on the severity, characteristics and behavior of each patient differ. Apart from the emotional aspect, a major contribution can be done to improve the quality of life in a positive way. For normal speech, we achieved 98-99% recognition accuracy but the accuracy of dysarthric speech is still far behind. The goal is to develop a system that can understand them and act accordingly. That will definitely do wonders for them. As many of them suffer from other physical injuries, they are required to use a wheelchair and have to be dependent on others for daily chores. The end goal is to create a personalized environment for them in which they can feel safe, comfortable and be understood. This would undoubtedly boost their confidence and make daily life easier. In this way, they can socialize themselves with others and get rid of loneliness and depression.

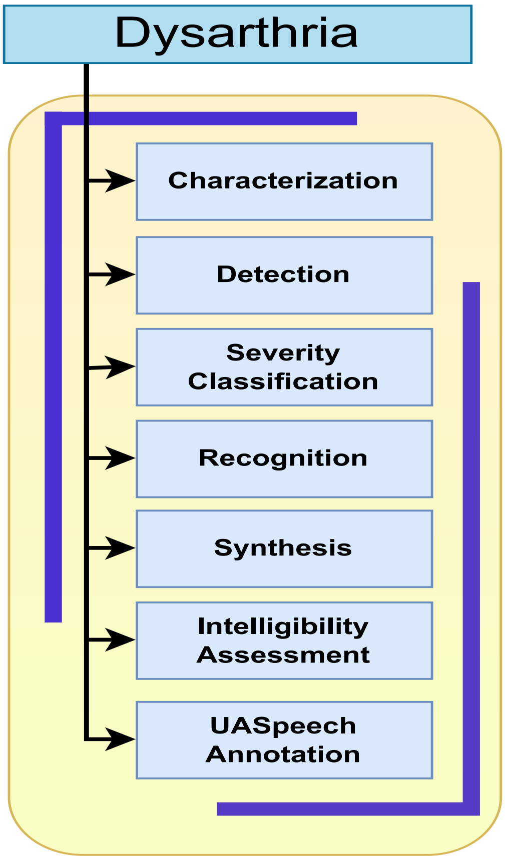


Figure 1.8: Scope of the thesis

1.11 Goals and contributions

Motivated by above observations, the objective of this thesis is to study the behavior and characteristics of dysarthric speech and reduce its complexity to improve the quality of life for affected individuals. There is a lack of study and dataset availability compared to normal speech. This study aims to help bridge that gap. This thesis, titled "Study, Analysis and Recognition of Dysarthric Speech," focuses entirely on dysarthric speech and associated treatment plans. While exploring dysarthria, we found that there are several issues that are addressed in each contribution:

- **Dysarthric speech characterization and classification based on affinity propagation:** Dysarthric speech is hard to recognize and understand because of its enormous variability and low intelligibility rate [158]. We propose a model to narrow down this gap and improve the recognition rate by using deep feature analysis and informative feature extraction techniques. Linear Prediction Cepstral Coefficient (LPCC) is used as a feature extraction technique and a pre-clustering-based algorithm, Affinity Propagation (AP), is used to select the relevant features. We found that this performs effectively with LPCCs for feature selection, significantly reducing modeling computation time without compromising on recognition accuracy. The model which takes 3ms to train using all the features in the vowel dataset, has been reduced to 0.9ms after applying the selected features. Similarly, for the digit dataset, the time has been reduced from 18.9ms to 1.9ms. This approach helped to identify the most impactful feature set, that contains majority of the relevant information.
- **Fusion of multiple audio descriptors for the recognition of dysarthric speech:** In this contribution, we implement a dysarthric speech recognition system in a speaker-adaptive manner with LSTM-RNN architecture, leveraging additional acoustic features to help the model effectively learn and adapt to each speaker's unique speech patterns. Feature selection is based on MFCC's combined with different variants of Jitter and Shimmer. Jitter and Shimmer are acoustic characteristics of voice signals and they are caused by irregular vocal fold vibrations. Their acoustic characteristics of voice signals offer valuable insights into individual voice characteristics to capture essential features and significantly enhance the capability of dysarthric speech recognition systems. In addition to ASR, we also analyze dysarthric speakers with very low intelligibility rates and found some prominent insights into their

Intelligibility	Very Low	Low	Mid	High	Overall
WRA(%)	68.94	74.81	90.76	93.93	83.11

Table 1.2: WRA% for each intelligibility group

speech characteristics. Table 1.2 presents the Word Recognition Accuracy (WRA)% achieved for each intelligibility group.

- **Dysarthric speech detection and severity classification using Audio Spectrogram Transformer:** Detection and evaluation of the dysarthric speech severity level offers valuable insights into a patient's progress, helps pathologists in therapy planning, medication and supports the functionality of automated dysarthric speech recognition systems [33]. In this study, we conducted experiments on dysarthric speech detection followed by severity classification for speaker-dependent and speaker-independent scenarios. Our findings highlight the effectiveness of Speech-Vision approaches, particularly those leveraging transformers and spectrograms. Audio Spectrogram Transformer has been taken as a base model in this contribution marking the development of a convolution-free, exclusively attention-driven model for audio classification. While various deep learning techniques have been explored in this domain, this work distinguishes itself by introducing detection and classification using an audio spectrogram. For dysarthric speech detection, we achieved 99.64% accuracy for dysarthric speech detection. We also achieved 95.60% and 78.97% for severity classification in speaker-dependent and speaker-independent contexts respectively.

- **Dysarthric speech synthesis system using Tacotron2 for specific and Out-Of-Vocabulary words:** To develop a robust ASR model, an adequate amount of training data is essential. Since the available dysarthric speech data is limited, we aim to generate new data using the existing dataset. In this contribution, we propose a speaker-adaptive dysarthric speech synthesis technique using the Tacotron2 model. We used this model to generate dysarthric speech utterances that already exist in the UASpeech database and for OOV words as well to expand the vocabulary size. By generating dysarthric speech from textual input, we achieved the highest Mean Opinion Score (MOS) of 4.575±0.039 for known words and 4.045±0.048 for OOV words for the selected speakers. This approach successfully adapted to the inherent variability and intelligibility differences among speakers. To ensure the fidelity of our generated speech, we incorporated Dynamic Time Warping (DTW) to demonstrate the similarity between the original and generated waveforms. Additionally, we applied Waveform similarity-based Synchronized OverLap-Add (WSOLA) on the OOV words, ensuring a flawless evaluation process. This approach allows us to address the challenge of limited data availability by augmenting the database and vocabulary size.

- **Intelligibility assessment of dysarthric speech based on Goodness of Pronunciation:** Individuals with dysarthria struggle a lot to control the movements of the speech-related muscles. With appropriate treatment, medication and consistent practice, the severity of speech difficulties can be alleviated. However, intelligibility assessment is an initial and crucial step that helps clinicians to understand the severity of the disorder [10][36][27][81]. We propose a speaker-independent approach at the phoneme level to identify the most frequently mispronounced phonemes among speakers from the UASpeech database. This method utilizes the Goodness of Pronunciation (GoP) algorithm, followed by GMM-GoP and NN-GoP score calculations. We found that they have difficulty in the articulation of phonemes /ŋ/, /æ/, /əʊ/, /ɪ/ and /aʊ/. Further, we calculated the correlation between the obtained

GoP score and dysarthric speech intelligibility using Kendell's rank correlation coefficients and found a positive correlation for both GMM-GoP and NN-GoP, which establish the reliability of the observed relationships. Additionally, this chapter provides an in-depth analysis of the UASpeech database, focusing on its structure, content and quality. We thoroughly explore the data characteristics, including speaker voice, noise levels and the acoustic features present in each recording. The UASpeech database lacks pre-existing annotations, requiring us to dedicate a significant amount of time to manually label the data.

1.12 Organization of the thesis

The thesis consists of eight chapters organized as follows:

- **Chapter 1:** Introduces the thesis content layout and emphasizes its key contributions and objectives.
- **Chapter 2:** This chapter investigates the origin of dysarthria and reviews various approaches that have been applied in previous research to address the challenges of dysarthric speech, highlighting key research gaps in the process.
- **Chapter 3:** Dysarthric speech characterization and classification based on affinity propagation at the feature level is presented in this chapter and result obtained are analyzed.
- **Chapter 4:** In this chapter, experiments with the fusion of multiple audio descriptors (MFCC's, Jitter and Shimmer) is performed for the recognition of dysarthric speech. Using this architecture, a speaker-dependent automatic speech recognition system is implemented.
- **Chapter 5:** This chapter explores for the detection and severity classification task

of dysarthric speech in both speaker-dependent and speaker-independent scenarios using transformer and deep learning methods.

- **Chapter 6:** This chapter generates existing and out-of-vocabulary words for six dysarthric speakers using Tacotron2 model with the aim of increasing the existing database size and retaining the same characteristics as the original speakers.
- **Chapter 7:** This chapter delves into the actual problem associated with dysarthric speakers at the phoneme level with the help of the Goodness of Pronunciation (GoP) algorithm. Additionally, this chapter presents a comprehensive annotation of the UASpeech database, emphasizing its organization, content and quality.
- **Chapter 8:** This chapter concludes this thesis with possible future research directions.

The beautiful thing about learning is that no one can take it away from you.

B.B. King, American blues guitarist

2 Literature survey

Dysarthric speech has a lot of challenges and in each contribution, we have endeavored to reduce its complexity with the aim of improving dysarthric ASR systems. Several challenges have been carried out in various perspectives, such as the improvement of ASR performance, sufficient amount of data collection, accurate detection, severity classification, better intelligibility assessment and detailed analysis and characterization. Early research on dysarthric speech recognition primarily focused on phonetic and acoustic variations. The key studies during the 1980s-1990s revolved around understanding the

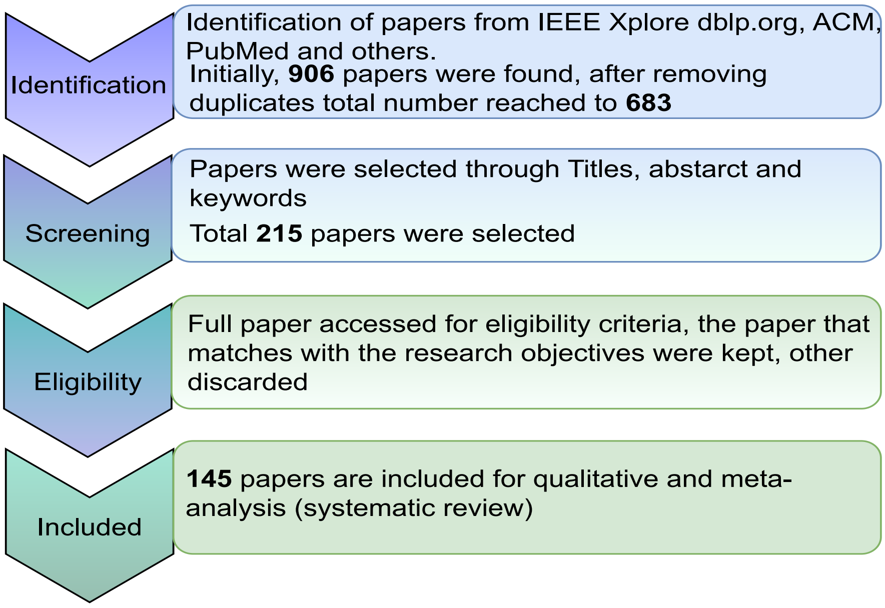

Identification
Identification of papers from IEEE Xplore dblp.org, ACM, PubMed and others.
Initially, **906** papers were found, after removing duplicates total number reached to **683**
Screening
Papers were selected through Titles, abstarct and keywords
Total **215** papers were selected
Eligibility
Full paper accessed for eligibility criteria, the paper that matches with the research objectives were kept, other discarded
Included
145 papers are included for qualitative and meta-analysis (systematic review)

Figure 2.1: Search study process

challenges posed by reduced speech intelligibility and inconsistent articulation patterns [76][86][77]. Due to data unavailability and limited computational power, researchers were bound to do the analysis only for a few dysarthric audio samples, which was inadequate for developing robust ASR systems.

2.1 Search strategy

A thorough search was performed across multiple electronic databases, including ACM, dblp, SpringerLink, PubMed, Scopus, IEEE, MDPI, Elsevier and other prominent conferences in the field. The search utilized keywords such as ”Dysarthric” AND ”Dysarthria” AND ”Recognition” AND ”severity levels,” as well as ”Assessment” AND ”Detection” AND ”Intelligibility.” Figure 2.3 details the search strategy process. A total of 906 papers were initially retrieved from all the databases. After removing duplicates, 683 unique papers were retained. Following a detailed screening of titles, abstracts and keywords, additional papers were excluded. In this process, review and survey papers, as well as studies on other types of disorders, were not considered. During the eligibility phase, only papers directly relevant to the objectives of this thesis were retained. Numerous papers related to tools and software for therapy were also excluded. Ultimately, 145 papers were identified as being within the scope of this thesis.

2.2 Automatic speech recognition for dysarthria

The literature shows many attempts have been undertaken to develop ASR technologies using dysarthric speech datasets. Initially, a statistical causal model was introduced for the assessment of dysarthric speech, along with the utility of a computer-based speech recognition system [143]. A new figure of merit was proposed to quantify both the intelligibility of dysarthric speech and the performance of the speech recognition system. However, their primary focus was on linguistic and perceptual analysis. They observed that dysarthric speech exhibited distinct articulatory error patterns when compared to typical speech. Additionally, they identified that the primary sources of these errors, in terms of manner of articulation, predominantly originate from either stops or fricatives. Early research has primarily concentrated on acoustic modeling and hand-crafted features, aiming to grasp the unique sound features of dysarthric speech [41]. When ASR was first applied to individuals with dysarthria, Hidden Markov Models (HMMs) were almost exclusively used to capture the variability and parameters were generally trained to the general population. The recognition task involved a minimal dataset (few speakers with dysarthria and a good number of regular speakers) [18] [14] [100] [124] [26].

They found that the speech recognition system's performance can change depending on the type of speech it is dealing with, i.e. isolated words, connected words or continu-

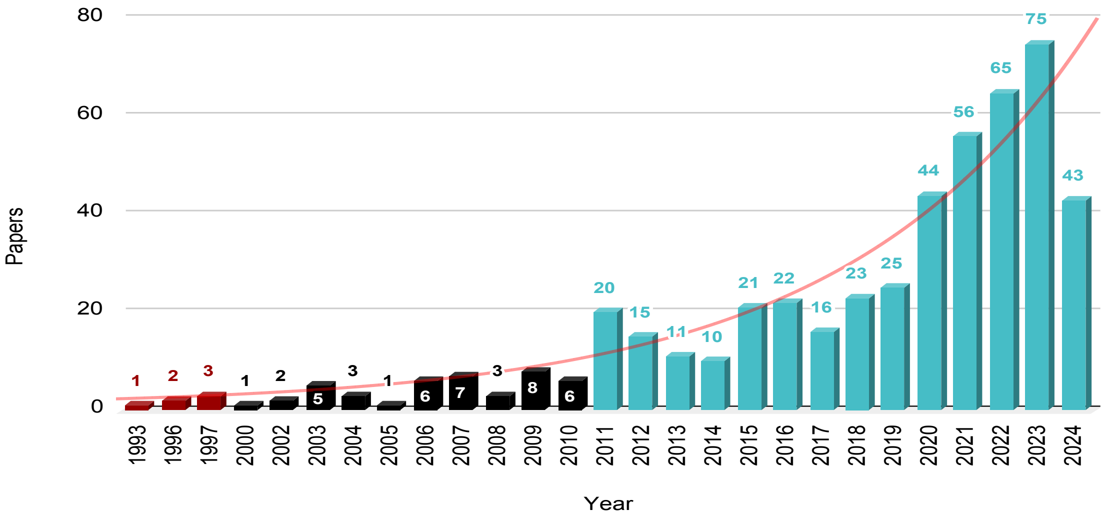


Figure 2.2: Number of papers for the improvement of dysarthric speech recognition over the years

ous speech [87][118][68]. Nevertheless, HMMs perform poorly due to overlapping classes and insufficient training data. To address these challenges, Rajeswari and Chandrakala [99] studied fixed dimensional vector representations that provided better discrimination for dysarthric speech than the conventional HMMs. Also, HMMs suffer from other limitations that impede their performance. Formant analysis was widely explored, but it didn't help much because of the limited data and the complexity of dysarthric speech. These limitations, coupled with the rise of neural networks, have motivated DSR researchers to shift their focus toward exploring deep learning-based approaches [146]. In the early twenties, the development of specialized datasets UASpeech [78] enabled researchers to work with larger amounts of data. These datasets provided a foundation for improving recognition algorithms. As shown in Figure 2.2, there has been an exponential increase in the number of publications related to dysarthria after dataset availability. Different types of acoustic models like Gaussian Mixture Model (GMM)-Hidden Markov Models (HMMs), Support Vector Machine (SVM) and Artificial Neural Networks (ANNs) were explored [57][153][7].

2.3 Dysarthric speech recognition using ANN

Few researchers [68][131][115] reviewed that in the severe case of dysarthria, adapting only the HMM acoustic model trained for the general population is not sufficient. Neural networks have emerged as a powerful tool in speech recognition, offering significant advantages over traditional methods [62]. Unlike HMMs, neural networks are highly flexible and capable of learning complex patterns through progressive updates during training. This dynamic learning process allows neural networks to build on prior knowledge and adjust to variations in the speech signal over time. Among the earliest studies in this context, Jayaram and K. Abdelhamied [68] utilized Artificial Neural Networks (ANNs) to recognize dysarthric speech. They developed, trained and tested two multilayer neural networks based on isolated words spoken by a dysarthric speaker. Using a feedforward neural network combined with Fourier spectral coefficients reduced the error rate by 40% on isolated word recognition for dysarthric speech. While the overall accuracy was not disclosed in this study, experiments were conducted on a limited dataset containing only 22 utterances of 20 words with one subject having severe dysarthria. ASR accuracy has been improved by focusing on dysarthric speech and the types of errors they generally make. Polur et al. [112] investigated an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric speech signals where backward state transition were also allowed. This study aimed to investigate the applicability and utility of using a hybrid Mel Cepstrum-based HMM/ANN model for recognizing dysarthric speech and was evaluated on speech samples obtained from three moderately intelligible cerebral palsy speakers on a 25-word vocabulary.

2.4 Deep learning based dysarthric ASR

Deep Neural Networks (DNNs) have gained widespread popularity for their effectiveness in recognizing dysarthric speech, providing a more robust approach for handling its unique complexities [45][106]. These models could automatically learn features from raw speech data and capture temporal dependencies. Convolutional Neural Networks (CNNs) have been used effectively in this domain because they are capable of picking out small details by looking at nearby parts and then summarizing them [4]. Long Short-Term Memory (LSTM) models were also explored as they are good at understanding longer patterns in speech and are capable of addressing a problem that older recurrent networks suffered. However, many researchers didn't delve deeply into the feature level [13][82]. In several speech recognition tasks, LSTM is combined with other neural network architectures, such as CNNs, attention mechanisms and transformer models to enhance its performance and capabilities for speech recognition. These methods improved performance by learning from raw data rather than relying solely on hand-crafted features. Deep learning models act as a black box and often discard minute details while feature extraction techniques [160][144].

In recent times, people have become more interested in a type of speech recognition that uses various kinds of information, known as multi-modal ASR [41]. In dysarthric speech, audio and visual features are proposed by the UASpeech database where we can use visual and articulatory features at the same time [165]. However, there hasn't been much research on using different types of data together for dysarthric speech recognition because there is limited dataset available that has both types of data. Chikoto Miyamoto et al.[92] propose an audio-visual method to recognize disordered speech. They use Multiple Acoustic Frames (MAF) for acoustic feature extraction and an Active Appearance

Model (AAM) for visual speech recognition. AAMs are used to track the facial feature points. The proposed method improved 7.3% in the recognition rate compared to the audio-only methods.

Given the scarcity of dysarthric speech, transfer learning has gained popularity in training deeper models. Seyed Reza et al.[128] proposed a speech vision system to address three challenges of the dysarthric ASR system. In the first challenge, they addressed variation and inaccuracy of phonemes by transforming word utterances into visual-feature representations and tried to recognize the shape of words instead of phonemes by using shape detectors such as CNNs. The second challenge focused on the unavailability of dysarthric speech by using data augmentation and transfer learning. In the third challenge, they identified the difficulties with labeling dysarthric phonemes and solved this issue by using speech vision. They achieved a Word Recognition Accuracy of 61.11% and 64.71% for the two-step experiments and the system surpassed recognition of mild and severe dysarthric speech, achieving state-of-the-art results. The progress of dysarthric speech ASR system over the years has been summarized in Table 2.1.

Table 2.1: Work done on recognition of dysarthric speech

Author and Year	Algorithm	Number of subjects	Database	Result
Deller Jr et al. 1991 [26]	HMM	10 digits and 196 common words with 3 speakers having CP	None	92% for digits 88.3% for words
Jayaram and Abdelhamied 1995 [68]	ANN	One male subject with low intelligibility	None	Recognition rate of 78.25% for a 10 word vocabulary, only one subject was taken.
Espana-Bonet and Fonollosa 2015 [39]	DNN-HMM	15 subjects (8 dysarthric speakers with varying severity and 7 control speakers without any disorder)	TORGO	Improved WER by 3% for control speakers and 13% for dysarthric speaker

Table 2.1: Work done on recognition of dysarthric speech (continued)

Author and Year	Algorithm	Number of subjects	Database	Result
Rajeswari and Chandrakala 2016 [116]	HMM-SVM	11 males and 4 females with all kinds of intelligibility	UASpeech	Overall 87.91%for LL-SVM and 73% for TP-SVM
Sanders et al. 2002 [124]	HMM	Two male speakers and speech intelligibility is low	None	Average WER of 28.2%
Green et al. 2003 [55]	HMM	Three subjects Intelligibility information is not provided	None	Average word recognition accuracy is 94.33% for 10 digit vocabulary
Hasegawa-Johnson et al. 2006 [60]	HMM	Three subjects (Two males and one female), Speech Intelligibility is low	None	Average word recognition accuracy of 91% for 10 digit vocabulary.
Polur and Miller 2006 [112]	ANN/ HMM	Three male subjects and speech intelligibility is only moderate	None	Average recognition rate of 97% over a 25 word vocabulary
Hawley et al. 2007 [61]	HMM	Seven subjects(5 males, 2 females), Speech Intelligibility: low	None	Average word recognition accuracy is 95.4% for 10 digit word vocabulary
Morales et al. 2007 [93]	HMM/ WFST	Ten male subjects and with moderate intelligibility	Nemours	Maximum Accuracy of 56%
Selouani et al. 2009 [126]	HMM	Four subjects(only males), Speech Intelligibility information is not provided	Nemours	Average word recognition accuracy is 69.1%
Seong et al. 2012 [127]	HMM/ WFST	Ten male subjects and with moderate intelligibility	Nemours	Lowest Word Error rate of 28.2%
Shahamiri and Salim 2014 [130]	M-N ANN	Sixteen subjects(12 males and 4 females) with all types of speech intelligibility	UASpeech	Average Recognition rate of 75.41%
Shahamiri and Salim 2014 [131]	ANN	Seven Subjects (2 males, 5 females), with all kinds of intelligibility	UASpeech	Average recognition rate of 59.72% and average error rate of 25.2%
Kim et al. 2015 [79]	WFST/ PSSLM	144 were dysarthric and 30 were non-dysarthric control speakers	(KPOW) database [84]	Root Mean Square Error (RMSQ) 26.96 for baseline model

Table 2.1: Work done on recognition of dysarthric speech (continued)

Author and Year	Algorithm	Number of subjects	Database	Result
Chen et al. 2021 [15]	Conformer	14 speakers with 8 types of etiologies, 7 of which are ALS	LibriSpeech, Google voice search	Average WER of 9.9%
Shahamiri 2021 [128]	Speech Vision	15 subjects (4 females and 11 males) with all kinds of intelligibility rates	UASpeech	Improved recognition accuracies by 67% of UASpeech speakers
Sahane et al. 2021 [122]	KNN/ SVM	1400 samples of normal and dysarthric speech	UASpeech	Recognition rate of 87.75% by using multi-taper MFCC
Shahamiri et al. 2023 [129]	Attention-based Transformer	15 subjects (4 females and 11 males) with all kind of speech intelligibility	UASpeech	Average WRA of 67%

2.5 Strategies to overcome data scarcity

ASR has made remarkable progress recently due to advances in deep learning, which uses a large amount of training data[21][123]. Due to the athetoid symptoms, collection of data from an individual with an articulation disorder is challenging. Alternate ways to expand the training data size, techniques such as speech synthesis, voice conversion (VC) and data augmentation were explored [53]. Selection and concatenation of small units of speech from a large speech database was the state-of-the-art method for many years, but the generated speech was robotic and unnatural compared to human speech. Aihara et al. [2] proposed an approach of partial least square-based Voice conversion (VC), which is useful for addressing the data scarcity problem. They use the phoneme-discriminative feature to convert dysarthric speech into normal speech.

In the past few years, deep learning-based advanced acoustic models have made far-reaching changes in speech synthesis and speech transformation systems [66][166][113][1] [114][152][80]. DeepSpeech [51] is also a text-to-speech engine based on deep neural networks. The system is faster than real-time speed but the naturalness of synthesized voice is not specified. Char2Wav[138] is an end-to-end model for speech synthesis based on an encoder-decoder model with attention and a bidirectional recurrent neural network. The model was trained in English and Mexican Spanish. Josh et al. establish a synthesis approach for dysarthric speech by combining Transformer-TTS, LPCNet-Based TTS and CycleVAE-VC aiming to achieve enhanced speech intelligibility [89]. A recent publication [136] explored dysarthric speech synthesis using DNN-HMM by inserting additional pauses in the TORGO dataset.

2.6 Detection of dysarthria and its severity levels

A person's speech intelligibility measures the amount of their speech that can be comprehended by a normal listener and is typically used to gauge their dysarthria severity[12]. Detecting dysarthric speech and classifying its severity has been a primary research focus for many years, as it serves to be the foundation for rehabilitation and medical intervention for patients. Numerous studies have been conducted to investigate dysarthric speech severity by capturing essential acoustic data related to prosody, vocal tract dynamics and excitation source information [140][163][71][139]. Deep Belief Networks (DBNs) [45] were compared with MFCC giving a marginal improvement in dysarthric severity classification using a multi-layer perceptron neural network. The combination of Glottal-to-Noise Excitation Ratio (GNER) and Harmonics-to-Noise Ratio (HNR) with MFCC was pursued in [106], as both of these metrics can determine the degree of noise caused by the disorder.

Dahmani et al. [22] introduced a novel method to differentiate dysarthric speech from healthy controlled speech using rhythm metrics based on vocalic and intervocalic intervals' durations on the Nemours dataset [91]. They applied a Gaussian Bayes classifier for this classification task. However, the rhythm metrics they extracted did not yield promising results in expressing the severity level of dysarthria. Garima et al. used a genetic algorithm to select prosodic features and applied SVM to classify dysarthric speech severity [153].

Machine learning algorithm models, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have been also utilized with the aim of achieving high classification accuracy rates and attained accuracy has generally been favorable, but the computation time was quite large for the training process. Chitralekha et al. explored Bi-directional LSTM (BLSTM) for a binary classification of dysarthric and non-dysarthric speech using the transfer learning method and achieved an improvement of 6% compared to the traditional machine learning method [7].

In the long run, CNN was exhausted and researchers started to add an attention layer with CNN. To address the challenge of capturing long-range global context, some researchers have introduced a hybrid model by incorporating a self-attention mechanism with CNNs. Clinical approaches for dysarthria detection, such as evaluations by SLPs, professional assessments, analysis of articulation error patterns and manual evaluation methods, also exist. However, these fall outside the scope of this thesis.

2.7 Intelligibility assessment of dysarthric speech

In the early methods of dysarthric speech intelligibility assessment, the focus was on measuring the percentage of correctly understood words by a Speech-Language Pathologist (SLP). These evaluations were subjective, time-consuming and required the expertise of trained professionals. To overcome these challenges automated assessments were introduced, offering a more reliable and cost-effective alternative. These automated systems not only address the limitations of traditional manual methods but also enable personalized treatment plans and support remote diagnosis, enhancing accessibility and efficiency in patient care.

Generally, automated intelligibility assessment approaches are divided into: reference-free and reference blind. Reference-free approaches to intelligibility assessment seek to build classification models without relying on prior knowledge of healthy speech. Instead, they focus on extracting acoustic features that are thought to be strongly linked to speech intelligibility [44][106][73][153]. Reference-based approaches rely on using healthy speech signals as a benchmark to assess intelligibility. These methods utilize data from healthy speakers to help the models learn the features of intelligible speech, which then serve as a foundation for predicting the intelligibility of dysarthric speech [148][88][25]. Existing work related to intelligibility assessment are summarized in Table 2.2.

Table 2.2: Work done on automated intelligibility assessment of dysarthric speech

Author and Year	Model	Number of subjects	Dataset	Classification Accuracy
Bhat et al. 2017 [8]	ANN	Trained and tested on same 10 subjects	UASpeech	96.4% accuracy, 4 categories
Tong et al. 2020 [147]	CNN	Trained on 10 subjects and tested on 5 different subjects	UASpeech	99.6% accuracy, 4 categories

Table 2.2: Work done on automated intelligibility assessment of dysarthric speech (continued)

Author and Year	Model	Number of subjects	Database	Classification Accuracy
Gupta et al. 2021 [58]	ResNet	Tested and trained on same 8 subjects	UASpeech	98.9% accuracy, 4 categories
Joshy and Rajan 2021 [70]	SVM,DNN CNN,LSTM	Tested and trained on same 15 subjects	UASpeech	82.9%, 93.6%, 93.2% 75.1% accuracy, 4 categories
Joshy and Rajan 2006 [70]	SVM,DNN CNN,LSTM	Tested and trained on same 8 subjects	TORGO	82.7%, 95.1%, 96.2% 85.9% accuracy, 3 categories
Bhat and Strik 2020 [7]	BLSTM	Leave one out with 15 subjects	TORGO	85.2% accuracy, 2 categories
Kadi et al. 2016 [73]	GMM/ SVM	Trained on 11 Nemours and 8 TORGO, test on same subjects	Nemours TORGO	78.8% accuracy, 3 categories

2.8 Database

There are a limited number of dysarthric speech databases available that are being used to study articulation errors and develop ASR systems. However, these datasets do not explicitly define word accents, leading to variations in pronunciation even for the same words. Currently, only three publicly available English-language corpora exist for dysarthric speech.

2.8.1 UASpeech

The University of Illinois generated UASpeech* database, which consists of speech samples from 19 individuals affected by cerebral palsy, exhibiting a wide range of intelligibility levels from 2% to 95% [78]. Speakers consist of both males and females and their ages

*https://experts.illinois.edu/en/publications/dysarthric-speech-database-for-universal-access-research

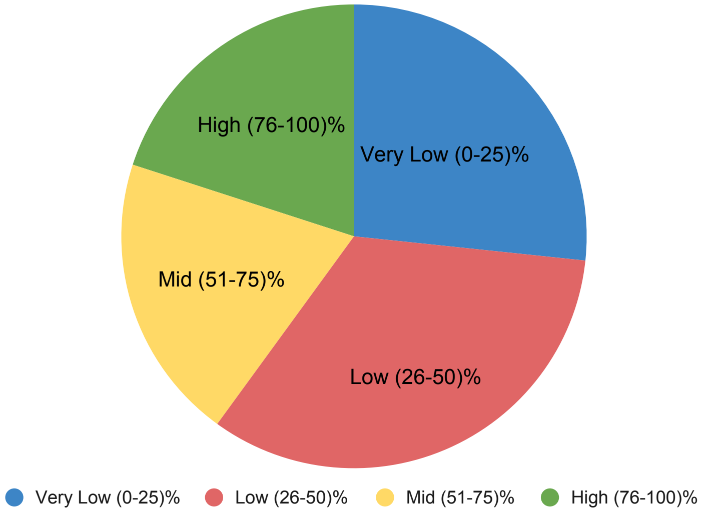


Figure 2.3: Distribution of speakers in UASpeech database

range from 18 to 58 years. The data was recorded using an eight-microphone array, with 1.5 inches of spacing between each microphone and complemented by video recordings to capture visual features. Throughout the recording sessions, participants were seated comfortably in front of a laptop computer and they were instructed to read isolated words displayed on PowerPoint slides. The entire recording process was divided into three blocks, each containing 255 words. Among these, 155 words were repeated across the blocks and 100 words were in uncommon categories. Common words comprised 10 digits, 26 radio alphabets, 19 computer commands and 100 words sourced from the Brown Corpus. Subsequently, each of the 19 speakers recorded a total of 765 isolated words.

To assess the intelligibility of dysarthric patients, five impartial listeners, proficient in American English aged between 18 and 40, were selected for each speaker. Listeners were instructed to listen to real words spoken by an individual with a speech disorder. The speech files were presented on a web page and listeners used headphones in a quiet room. Each listener transcribed the words and the percentage of correct responses were calculated. Speaker intelligibility was determined by averaging the correct percentages across five listeners. The classification of speakers intelligibility into four categories (very low, low, mid and high) was based on the average percent accuracy for each speaker. The recorded speech files were saved in the .wav format and the entire dataset is publicly available for further research and analysis.

2.8.2 TORGO

This database is an outcome from the collaboration between the Departments of Computer Science and Speech-Language Pathology at the University of Toronto in 2008. TORGO* was developed primarily to evolve ASR models suited to people with atypical speech production. This database contains detailed physiological information to learn hidden articulatory parameters.

This database currently contains information on seven dysarthric subjects (four males and three females) having Cerebral Palsy (CP) or Amyotrophic Lateral Sclerosis (ALS) with a wide intelligibility range. The age of the subjects was between 16 to 50. A 19 inch LCD screen was placed in front of all subjects where English text was visualized and the distance between subject and screen was 60 cm. During the data collection process, each subject took a short quiz covering general demographic information and health-related questions. Speech-language pathologists assessed the motor functions of each experimental subject using the standardized Frenchay Dysarthria Assessment (FDA)[50][121]. The database includes three types of stimuli non-words, short words and restricted sentences.

*https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html

2.8.3 Nemours

The Nemours database [91] consists of short nonsense sentences and few meaningful sentences spoken by 11 male speakers with different levels of severity. There are a total of 814 short nonsense sentences. Short nonsense sentences mean sentences with void meanings like ”The X is Ying the Z”, where X and Y are different. X is selected from a set of 74 monosyllabic nouns and Y is selected from a set of 37 disyllabic verbs without replacement. This process generates a total 37 sentences. Reversing X and Y also creates another set of 37 sentences. So, in total 74 sentences with each noun and verb were spoken by each subject. All nonsense sentence was saved in standard RIFF format. In addition, each speaker recorded two connected speech: the ”Grandfather” passage and the ”Rainbow” passage paragraphs to the database. They broke out the longer speech passage into sentences to make things more convenient and each sentence was saved as a separate waveform file. All nonsense sentences were recorded first, followed by speech passages. They placed a large sheet with nonsensical sentences in front of talker and experimenter read each sentence before being repeated by the subject. Including breaks, the recording session took an average of two to three hours. A speech pathologist conducted a few assessment sessions for all the recording sessions [75]. Recording of all sessions took place in a small dampened room.

After reviewing all the existing work, we came to know that there are still many challenges that need to be solved, to improve the quality of life for individuals with dysarthria. They need solutions that are practical, feasible, usable and affordable. In each chapter, we focus on different problems faced by dysarthric speakers and work on reducing them so that they can live a life more comfortably like healthy speakers.

All things are difficult before they are easy

Dr. Thomas Fuller

3

Dysarthric speech characterization and classification based on AP

3.1 Introduction

Speech analysis is the detailed study of voice signals, which are characterized by various parameters such as amplitude, duration, pitch, tone and speaker-specific variability. In this chapter, the analysis emphasizes at feature level to explore the unique acoustic and phonetic characteristics of dysarthric speech, which typically include irregularities in articulation, variations in pitch and loudness and distorted phoneme production. This process is fundamental to advancements in speech recognition, synthesis and various applications in healthcare and technology.

This field seeks to understand and interpret the nuanced qualities of human speech by breaking down signals into measurable components. It is done by extracting features from the speech signal. There are various methods for feature extraction, such as Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Predictive Cepstral Coefficients (PLPCCs), Linear Predictive Cepstral Coefficients (LPCCs), etc. Feature extraction is an essential process in speech analysis and classification, but feature selection is an important key to get accurate and time-efficient models, which affects the overall run-time of the classification process [49][150] [3] [23] [44]. In this work, we utilize LPCCs alongside the pre-clustering algorithm "Affinity Propagation" to efficiently classify dysarthric speech with minimal computation time. We chose LPCCs over other feature extraction techniques because LPCC captures fine details related to vocal tract features, that are relevant for disordered speech where articulation may vary from person to person. Additionally, LPCC has lower computational requirements, making it a practical choice for clustering algorithms.

Pre-clustering is utilized to select features based on clustering scores. These scores help to identify a subset of features that are most effective for classification. The Affinity Propagation (AP) clustering algorithm has gained considerable attention in recent years due to its efficiency, robustness to initialization and ability to produce clusters with lower error rates compared to other exemplar-based methods [83][168][169]. Although Affinity Propagation was introduced by Frey and Dueck in 2007, has gained recognition but remains unexplored for dysarthric speech recognition [155]. Given the unique challenges of dysarthric speech, we aim to leverage Affinity Propagation for accurate recognition, using it to identify key speech characteristics. This approach helps in selecting dominant features, ultimately enhancing classification accuracy.

3.2 Affinity propagation

Affinity Propagation is a clustering algorithm that uses message passing between data points to identify clusters [103][49]. Exemplars in affinity propagation refer to the most representative data points in each cluster. Each data point is connected to one and only one exemplar. For message passing between data points, three types of matrices are used: Similarity matrix, Responsibility matrix and Availability matrix.

Similarity Matrix: AP starts with a similarity matrix that indicates how closely each data point relates to every other point. This matrix is used to calculate affinities between data points, where each entry shows how well a point could represent another point.

Responsibility and Availability Messages: In each iteration, AP updates two key values: Responsibility: Represents how suited a point might be as an exemplar for another point. Availability: Indicates how appropriate it would be for a point to consider another as its exemplar, based on feedback from other points. These values are updated iteratively to refine the relationships between data points and candidates for exemplars.

Message Passing: During each iteration, messages are passed between points, updating the responsibility and availability values until they converge. This allows the algorithm to identify a consistent set of exemplars that best represent the data.

Once the messages stabilize, the algorithm designates a subset of data points as exemplars and each point is assigned to the exemplar with the highest availability and responsibility, forming clusters around these representatives. The algorithm outputs clusters where each one has an exemplar (representative data point) at its center. AP is notable for its efficiency, scalability and lack of need for a pre-defined number of clusters, making it well-suited for complex data. For this experiment, we use digit and vowel samples from all speakers from the UASpeech database 2.8.1.

3.3 Methodology

In the initial phase of our experiment, we had unprocessed audio data. The raw audio data is divided into small chunks that are called frames, each representing a short time segment with 320 samples per second. To prevent information loss, the frames overlap by 25%, meaning each new frame starts with the last 80 samples of the previous frame.

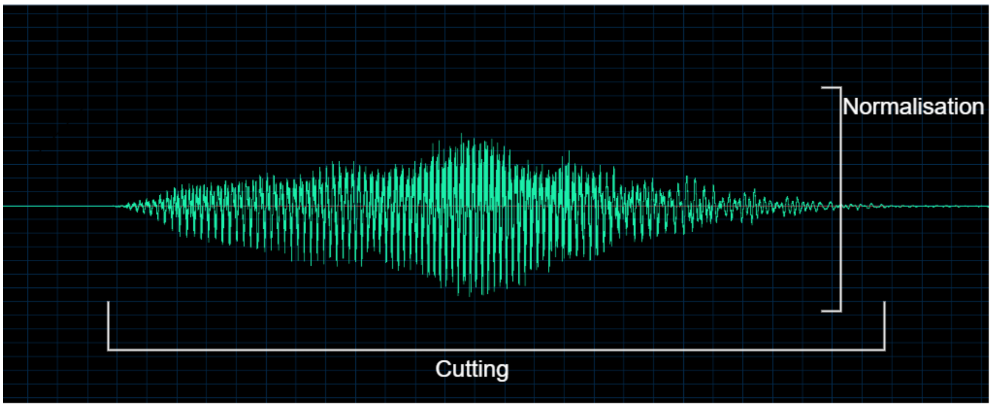


Figure 3.1: Pre-processing of audio files

Dysarthric speech typically contains significant pauses and silences at the start and end of recordings, which can increase data size unnecessarily and impact processing efficiency. To address this, we have trimmed the silent, unnecessary frames from each audio sample, focusing only on the relevant, active speech sections. This has ensured that the analysis targets meaningful speech data rather than non-informative silence, ultimately leading to more efficient and accurate processing. This trimming has been achieved by calculating the Short Term Energy (STE) for all frames of 320 samples each. We have identified the start of stable frames where the STE has increased by 400%, marking this as the beginning of speech. Similarly, the endpoint has been determined using the same threshold. Additionally, we normalized each sample by scaling it to a range between +5000 and -5000, ensuring consistency across maximum and minimum amplitude levels as shown in Figure 3.1.

3.3.1 Feature Extraction

After pre-processing the audio signal is converted into distinctive parametric values, LPCCs that capture speech characteristics. We calculated R_i for each frame using the auto-correlation method. This is the degree of similarity between a given time series and a lagged version of itself over successive intervals of time.

$$R_i = \sum_{m=0}^{N-i-1} S(m) * S(m+i) \tag{3.1}$$

Equation 3.1 calculates the autocorrelation for each lag i and varies from 0 to 12. N represents the total number of samples in the audio frame. S(m) represents the speech signal sample at position m in the frame and S(m+i) represents the speech signal sample at position m+i, which is offset by i. So, for each frame, we get 13 values namely R_0, R_1, R_2 ...R_{12}. Then, we proceed with the Levinson-Durbin algorithm to compute the linear prediction coefficients.

$$E^{(0)} = R(0) \tag{3.2}$$

$$K_i = [R(i) - \sum_{j=1}^{i-1} \{\alpha_j^{(i-1)} R(i-j)\}]/E^{(i-1)} \tag{3.3}$$

$$\alpha_i^i = K_i \tag{3.4}$$

$$\alpha_j^i = \alpha_j^{i-1} - K_{i\alpha_{i-j}}^{(i-1)} \tag{3.5}$$

$$E^i = (1 - k_i^2) * E^{(i-1)} \tag{3.6}$$

Using the above equations we get 13 A_i's from each frame. The A_i's are not used directly as they have high variability, so they are converted into more efficient form known as Cepstral Coefficients.

$$C_0 = log[R_0]^2 \tag{3.7}$$

$$C_i = A_i + \sum_{k=1}^{i-1} C_k A[i-k] \tag{3.8}$$

Equation 3.8 converts LPCs into LPCCs where A_i is the i^{th} LPC, which reflects the vocal tract parameters and C_i is the i^{th} Linear Predictive Cepstral Coefficient. Using above equation we get 12 C_i values C_1, C_2, C_3...C_{12} for each frame. We gather all C_1 values from each frame into a single feature set, then repeat this process for each C_i coefficient, resulting in a total of 12 distinct feature sets. The feature set of vowel /a/ is shown in Figure 3.2. Each feature is the set C_i for each frame from a voice signal.

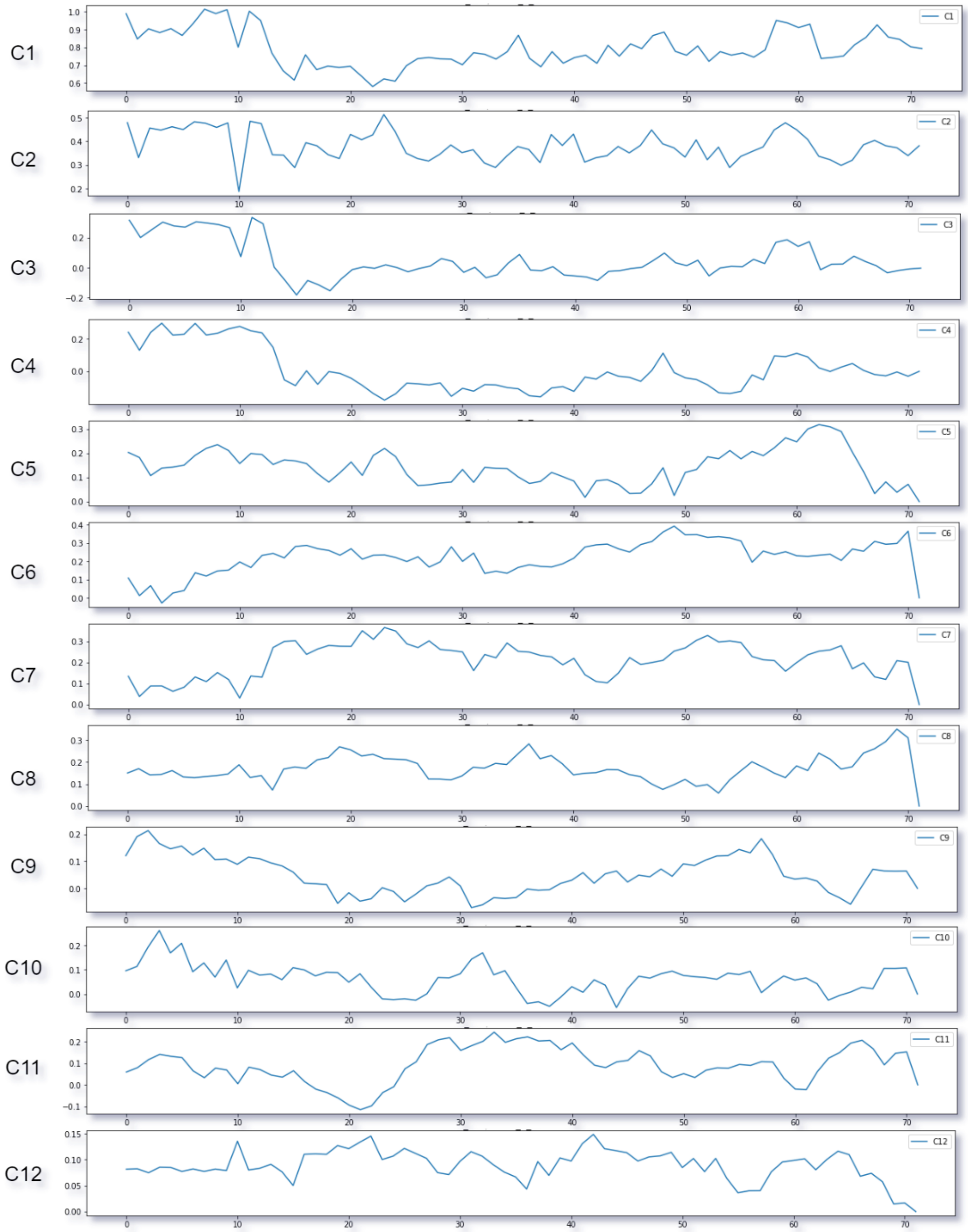


Figure 3.2: Features of vowel /a/

3.3.2 Feature selection

After feature extraction, we focus on identifying the most informative feature set that best distinguishes each class. We use an affinity propagation algorithm to select features and calculate the clustering score for each feature set. We then choose the feature sets with the highest scores, as a higher clustering score indicates that the features better differentiate between classes.

The Affinity propagation algorithm finds the number of clusters and the mapping of data points to those clusters. It does not require the initial centroids or number of clusters in advance. It represents each data point as a node in a network, where messages with real values are exchanged between points to refine clusters iteratively. This process continues until high-quality clusters and exemplars (representative data points) emerge. First, the similarity matrix is generated with the data points by negating the sum of the squares of the differences between data points. The similarity matrix contains information of preferences and similarity and is stored in the diagonal and the non-diagonal elements which represent the similarity values between data points. Preferences are the data points that are suitable to be an exemplar. Let x_1 through x_n be a set of data points and S be a function that quantifies the similarity between any two points n and k. Equation 3.9 S(n,k) indicates how close n is to k.

$$S(n,k) = -|x_n - x_k|^2 \tag{3.9}$$

Initially, we construct an availability matrix with zero elements. The responsibility matrix is generated using Equation 3.10.

$$r(i,k) \leftarrow s(i,k) - max_{k',k' \neq k}\{a(i,k') + s(i,k')\} \tag{3.10}$$

where ‘i’ are rows and ‘k’ are columns of the responsibility matrix. “Responsibilities” r(i,k) is sent from data points to candidate exemplars to indicate how strongly each data point favors the candidate data exemplar. Then, we calculate the availability matrix using two separate formulas for diagonal and non-diagonal elements respectively.

$$a(k,k) \leftarrow \sum_{i',i' \neq k} max\{0, r(i',k\} \tag{3.11}$$

$$a(i,k) \leftarrow min\{0, (r(k,k) + \sum_{i',i' \notin (i,k)} max\{0, r(i',k)\})\} \tag{3.12}$$

where ‘i’ is row and ‘k’ is column of the availability matrix. In summary, the equation sums all positive values in a column, excluding the row matching the column index. The “availability” a(i,k) represents the degree to which each candidate exemplar is available to act as a cluster center for a given data point. The proposed approach is shown in Figure 3.3. In this approach, cepstral coefficients are treated as data points. Clustering scores are computed for each feature set and the best-performing sets are selected. Specifications for Affinity Propagation are given in Table 3.1.

No.	**Specification**	**Description**
1.	Damping Factor	0.5
2.	Affinity Function	Euclidean
3.	Number of iteration to converge	15
4.	Maximum number of iteration	200

Table 3.1: Specifications for affinity propagation algorithm

3.3.3 Classification

After selecting the features for the labeled data points, we modeled and classified dysarthric vowel and digit sounds using selected features. The dataset has been divided into

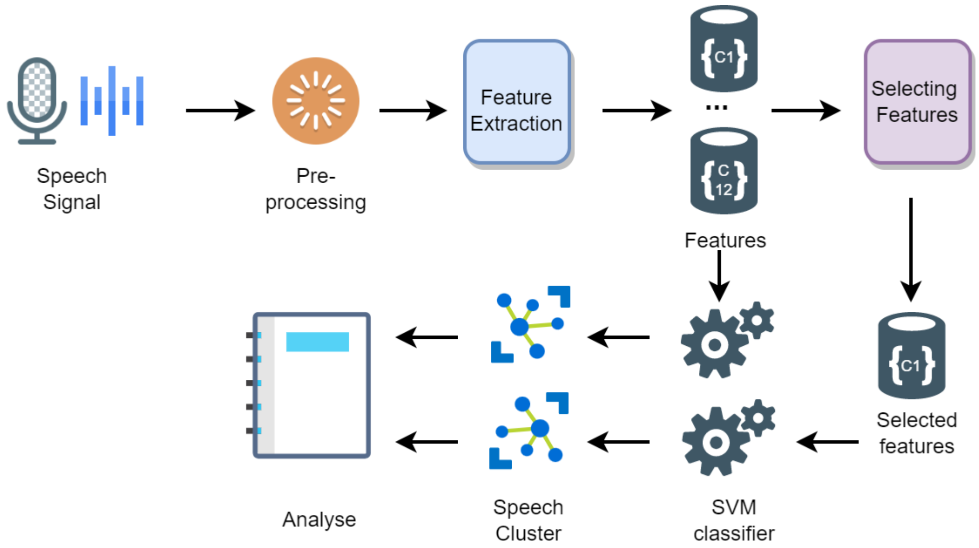


Figure 3.3: Proposed approach

training and testing samples, with the training set being used to build the classification model and the testing set for evaluating its performance. To validate the feature selection process, we have used Support Vector Machines (SVM), a well-known and effective classification algorithm. Specifically, we have carried out classification in two ways:

- All feature sets: Initially, we have used all available feature sets without any selection process to establish a baseline classification accuracy for both vowels and digits.
- Selected features: Subsequently, we have applied SVM on the features selected through our clustering-based approach.

Interestingly, we have achieved a similar classification accuracy with the selected features as with all the feature sets. This result demonstrates that the selected features have effectively captured the critical distinctions between classes (vowels and digits) while reducing redundancy and computational overhead. By identifying and using the most informative features, we have streamlined the classification process and ensured accurate recognition of both vowels and digits. This approach highlights the robustness of our feature selection method and the efficiency of SVM in achieving accurate classification. Parameters for SVM are shown in Table 3.2.

No.	**Specification**	**Description**
1.	Regularization	1.0
2.	Kernel	RBF
3.	gamma	scale
4.	Cache size	200 MB

Table 3.2: Parameters of SVM

3.4 Experiments and results

Using AP, we have calculated the clustering scores, represented as the random index, as shown in Table 3.3. Based on the results in the table, we have identified that features c_1 and c_3 possess the highest clustering scores for vowel data and have selected them for further analysis. These selected features have been used for classification using SVM, achieving an accuracy of 75%. To verify the effectiveness of the feature selection, we have also applied SVM using all features and have achieved the same accuracy of 75%. For the digit dataset, we have similarly selected the best features, as shown in Table 3.4, based on the pre-clustering scores. Features c_4 and c_{10} have been identified as the most significant. Applying SVM to these selected features, we have achieved 100% accuracy. This result has been further validated by applying SVM using all features, which has also yielded an accuracy of 100%.

Fig. 3.4 has shown the comparison between the original clusters and the clustering results obtained using Affinity Propagation. The model's training time for the vowel dataset, using all features has been recorded as 3 ms, which has been significantly reduced to 0.9 ms after selecting the optimal features. Similarly, for the digit dataset, the

Feature	Score
C_1	**0.81**
C_2	0.78
C_3	**0.83**
C_4	0.72
C_5	0.70
C_6	0.73
C_7	0.64
C_8	0.70
C_9	0.75
C_{10}	0.68
C_{11}	0.66
C_{12}	0.69

Table 3.3: Clustering Score(Random Index) of each feature set for dysarthric vowel dataset

Feature	Score
C_1	0.87
C_2	0.78
C_3	0.87
C_4	**0.93**
C_5	0.08
C_6	0.83
C_7	0.85
C_8	0.89
C_9	0.84
C_{10}	**0.90**
C_{11}	0.86
C_{12}	0.83

Table 3.4: Clustering Score(Random Index) of each feature set for dysarthric digit dataset

modeling time has decreased from 18.9 ms to 1.9 ms following feature selection.

3.5 Conclusion

The experiments conducted in this system have focused on applying a pre-clustering algorithm technique to datasets of English vowels and digits spoken by a dysarthric patient. LPCCs have been used as features and classification has been performed using SVM. It has been observed that pre-clustering through Affinity Propagation works effectively with LPCCs for feature selection, significantly reducing modeling time without compromising accuracy. In the future, this method needs to be validated on other datasets and this approach could be extended by integrating it with other feature extraction techniques. When implementing the same model for the entire UASpeech database, the program has terminated abruptly and the system crashed. This issue has likely occurred because the Affinity Propagation (AP) algorithm has a lot of matrix calculations that may be unable to handle the large number of rows and columns associated with the data. In the

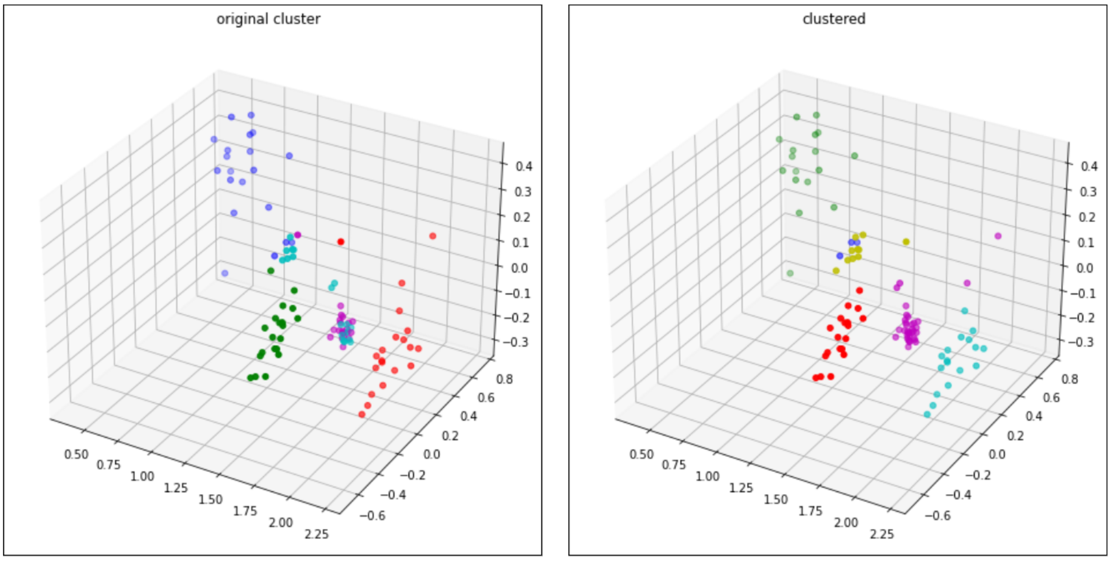


Figure 3.4: Original clustering vs Clustering by affinity propagation

subsequent chapter, we have addressed this limitation by adopting alternative methods to effectively manage a large number of classes, thereby overcoming this problem.

Every new beginning comes from some other beginning's end

Lucius Annaeus Seneca, Roman philosopher

4

Fusion of multiple audio descriptors for the recognition of dysarthric speech

4.1 Introduction

Effective communication and speech are an integral part of our day-to-day life. Speaking involves the coordination of multiple muscles and when these muscular systems are lacking in strength and coordination, the act of speaking becomes more challenging[110].

Developing a robust ASR system for dysarthric patients holds the potential to bring out remarkable transformations for individuals facing speech challenges [11]. A major challenge behind dysarthric speech recognition is data scarcity. The databases that are available have a limited number of speakers and vocabulary and in addition, the variability among the same word utterances is very high. In severe cases of dysarthria, a person starts to avoid communication with others and feels neglected, as their intended messages are often not comprehensible to others. Furthermore, it has been observed that many dysarthric patients face additional physical challenges beyond their speech difficulties. These challenges often include impaired motor coordination, such as clumsiness in their hands and stumbling movements in their legs. These physical limitations significantly hinder their ability to operate digital devices, further restricting their access to assistive technologies and digital platforms that could otherwise help them communicate more effectively. Such compounded difficulties highlight the need for accessible solutions tailored to the unique needs of dysarthric individuals.

Taking everything into account, speech recognition technology holds promise for giving them new opportunities to communicate and significantly enhance their quality of life. Individuals with dysarthria can utilize speech recognition technology as a comprehensive therapy program as well to help them speak more clearly [37]. They can receive feedback on their speaking clarity, pronunciation, intonation and other areas that might require enhancements. This practice can assist them in improving their speech through interactive exercises.

In this contribution, we aim to create a personalized automatic speech recognition system for dysarthric speech by combining different acoustic features followed by LSTM-RNN architecture. The feature selection process utilizes MFCC in combination with various Jitter and Shimmer variants. Jitter and Shimmer are acoustic characteristics of voice signals and they are caused by irregular vocal fold vibrations. Their acoustic characteristics of voice signals offer valuable insights into individual voice characteristics to capture essential features and significantly enhance the capability of dysarthric speech recognition systems [46][145][95].

4.2 Literature background

Researchers have been trying to improve dysarthric speech recognition for a long time [87][118][68]. Initial research concerning the recognition of dysarthric speech has primarily concentrated on acoustic modeling and hand-crafted features, aiming to grasp the unique sound features of dysarthric speech [41]. Formant analysis was widely explored, but it didn't help much because of the limited data and the complexity of dysarthric speech. In early twenties, with the development of specialized datasets such as UASpeech, enabled researchers to work with larger amounts of data. These datasets provided a foundation for improving recognition algorithms.

Additionally, Deep Neural Networks (DNNs) and HMM-based acoustic models became popular for recognizing dysarthric speech due to their ability to learn features from raw data and capture temporal dependencies [45] [106]. Convolutional Neural Networks are effective in identifying local details and summarizing them [4], while LSTM networks addressed limitations of older recurrent models and captured longer speech patterns, though they lacked deep feature-level exploration [13][82]. These methods improved performance by learning from raw data rather than relying solely on hand-crafted features. Deep learning models act as a black box and often discard minute details while feature extraction techniques. So, in the case of dysarthric speech, where each patient has unique characteristics, it's important to focus on the feature extraction part to give the model more detailed and useful information.

Several research studies have been done with a focus on developing speaker-independent ASR systems. Since each speaker possesses distinct characteristics, behavior and intelligibility rates, we decided to proceed with a Speaker-Dependent (SD) system. For dysarthric speakers, choosing a solely speaker-dependent approach appears to be more effective mainly in the case of severe dysarthria.

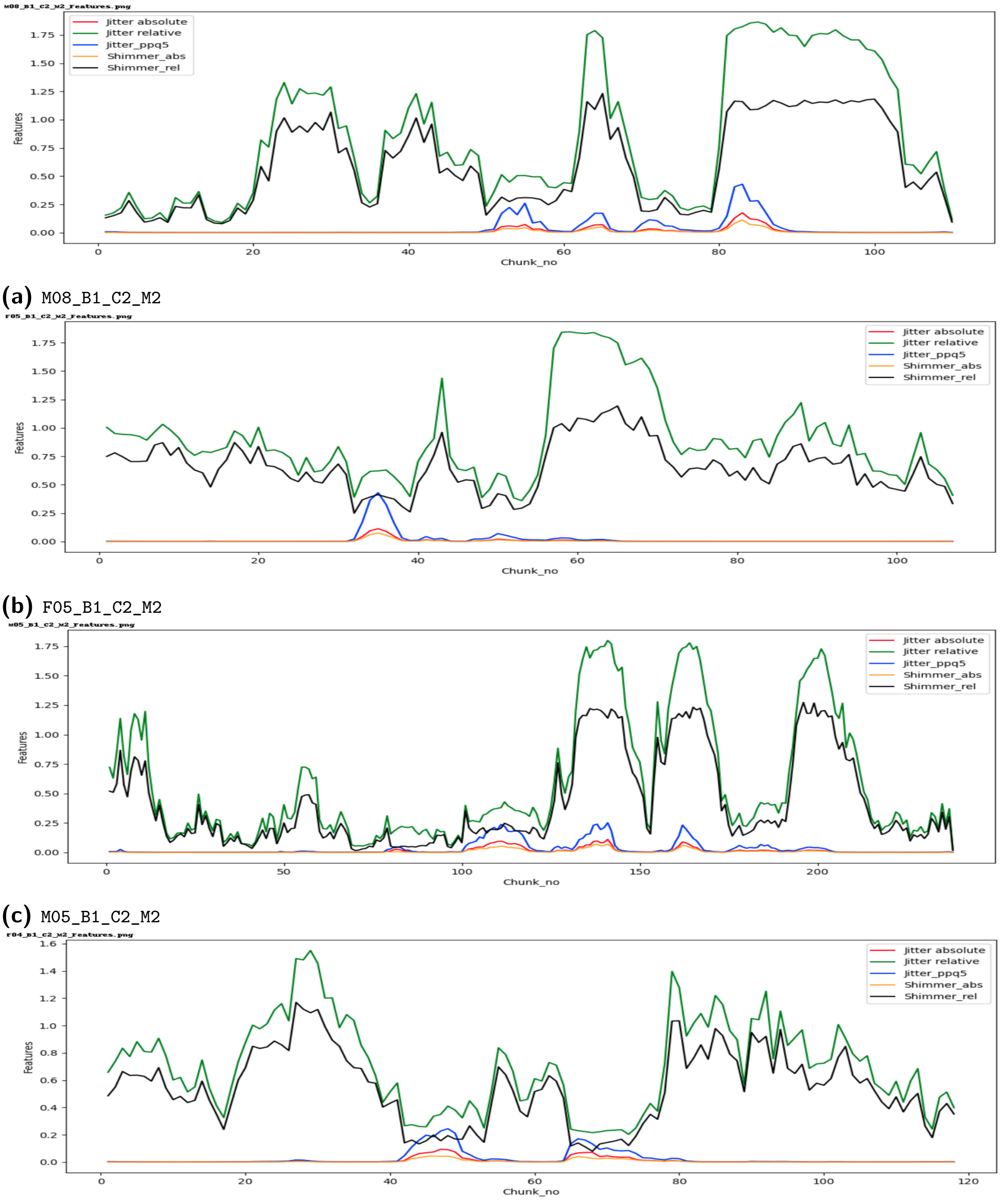
4.3 Methodology

Our objective is to develop a resilient ASR system in a speaker-dependent context. On the basis of availability, we selected 15 speakers from UASpeech database and the vocabulary size is set to 155 words. These 155 words include all digits, alphabets, computer commands and common words, which are sufficient to cover all kinds of phonemes. Table 4.1 has the description of the selected speakers with their intelligibility rates. Using just one acoustic feature was not enough to capture all the details.

4.3.1 Feature Extraction

The original audio files in the database contain periods of silence mainly in the initial part. We truncate the silence part from the beginning and end part, followed by normalization, windowing function and DC Shift correction. We use Mel-Frequency Cepstral Coefficients (MFCCs), variants of Jitter (Jitter absolute, Jitter relative and Jitter ppq5) and Shimmer (Shimmer absolute, Shimmer relative and Shimmer apq5) to represent the characteristics of a signal.

Jitter Absolute: Jitter refers to the cycle-to-cycle variation in the fundamental frequency (F0) of a speech signal. It is often used as an acoustic feature to assess voice



(a) `M08_B1_C2_M2`

(b) `F05_B1_C2_M2`

(c) `M05_B1_C2_M2`

(d) `F04_B1_C2_M2`

Figure 4.1: Feature set of the word “Backspace” from four different speakers

quality, particularly in pathological speech. Fundamental frequency is defined as the frequency at which the vocal folds vibrate during voiced speech sounds. Equation 4.1 measures the average absolute difference between consecutive pitch periods, where N is the total number of pitch periods in the audio file and T_1 Duration of the i^{th} pitch period.

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - T_{i+1}| \tag{4.1}$$

Jitter Relative: Jitter relative measures the average difference between consecutive pitch periods relative to the mean pitch period as shown in Equation 4.2.

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

Jitter(PPQ5)(Five-point Period Perturbation Quotient): Jitter ppq5 is the average absolute difference between a period and the average of that period and its four neighbors. This is the average perturbation across five consecutive pitch periods.

$$Jitter(PPQ5) = \frac{1}{N-4}\sum_{i=3}^{N-2} \frac{|(T_{i-2} + T_{i-1} + T_i + T_{i+1} + T_{i+2}/5 - T_i|}{T_i} \tag{4.3}$$

Shimmer refers to the cycle-to-cycle variation in the amplitude of the speech signal. It is another acoustic measure often used to analyze voice quality in distorted speech. It reflects how much the loudness of the speech signal fluctuates from one cycle to another, which is often irregular in individuals with speech disorders like dysarthria. The three types of shimmer used here are:

Shimmer Absolute: This is the basic shimmer measurement, representing the average difference in amplitude between consecutive vocal cycles as shown in Equation 4.4 where N is total number of glottal cycles in the audio file and A_i is peak amplitude of the i^{th}

glottal cycle.

$$Shimmer(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i+1}| \tag{4.4}$$

Shimmer Relative: Represents the average relative difference in amplitude between consecutive cycles compared to the overall amplitude.

$$Shimmer(Relative) = \frac{\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\sum_{i=1}^{N} A_i} \tag{4.5}$$

Shimmer APQ5 (Amplitude Perturbation Quotient over 5 cycles): This variant calculates the average absolute difference in amplitude between a cycle and the average amplitude of its five neighboring cycles. This is more robust against local amplitude variations compared to absolute shimmer.

$$Jitter(PPQ5) = \frac{1}{N-4}\sum_{i=3}^{N-2} \frac{|(A_{i-2} + A_{i-1} + A_i + A_{i+1} + A_{i+2}/5 - A_i|}{A_i} \tag{4.6}$$

Incorporating this additional feature enriches our understanding of each speaker's distinctive traits and behaviors. Figure 4.1 represents different variants of jitter and shimmer used in this experiment from four speakers with high and mid intelligibility rates, all uttering the word “backspace”. The jitter relative and shimmer relative curves are higher than the other features, indicating the stress patterns of each individual.

For MFCC computation, Hamming window is applied to each frame to reduce spectral leakage and the Discrete Fourier Transform (DFT) is applied to each windowed frame to obtain the magnitude spectrum. The magnitude spectrum is converted to a Mel spectrum by applying a set of triangular or cosine filters that span the frequency range. The number and width of the filters can vary depending on the application, but typically there are

20-40 filters per frame. The logarithm of the mel spectrum is taken to compress the dynamic range and mimic the human hearing response. The Discrete Cosine Transform (DCT) is then applied to the log Mel spectrum to obtain the MFCCs. The DCT reduces the correlation between adjacent coefficients and produces a compact representation of the spectral shape. Usually, only the first 13 coefficients are retained, as they contain most of the information. Utilizing 13 MFCCs alongside jitter and shimmer from each frame across all data provides a comprehensive 2-dimensional feature vector.

Let the feature matrix be $M \in \mathbb{R}^{\mathbb{N}\times 13}$ where $\mathbb{N}$ is the number of frames. Let the jitter value for each frame be J_a, J_r and J_{ppq5} and shimmer be S_a, S_r and S_{apq5} respectively. Let F_i be the feature vector of i^{th} frame.

$$F_i = [M_{i,1}, M_{i,2}, ..M_{i,F}, J_{ai}, J_{ri}, J_{ppq5i}, S_{ai}, S_{ri}, S_{apq5i}]$$

$M_{i,j}$ represents the j^{th} MFCC coefficient for the i^{th} frame. This vector has a dimension of $F+6$, for each frame.

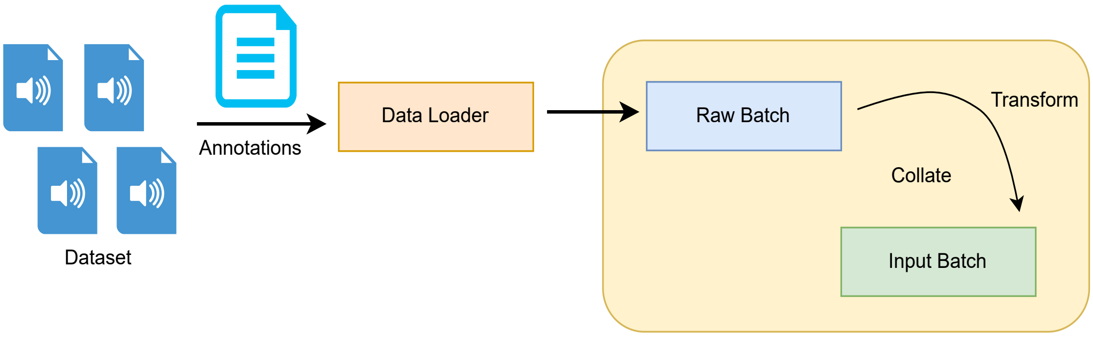


Figure 4.2: The figure depicts the process followed by data loader to provide an input batch tensor in every iteration

We partitioned the data into smaller subsets known as mini-batches, which were loaded sequentially. Each input batch has multiple audio files, each containing a waveform of a different length. The LSTM needs all the inputs given in a batch to be of the same shape. We defined a collate function to pad all the transformed inputs in a batch to make all the matrices equal in length, i.e., the maximum of all the matrices as shown in Figure 4.2. DataLoader fetches UASpeech data from the annotations and batches the inputs into a single input tensor.

In each batch, we have multiple audio files and their waveforms can vary in length. However, LSTMs require all inputs in a batch to have the same shape. After preprocessing and feature extraction, each audio file is converted into a 2D matrix where the number of columns represents the features (e.g., 13 MFCCs). The lengths of these matrices can differ from file to file. To handle this, we use a collate function that pads all the matrices in a batch to the length of the longest matrix. This ensures that all inputs in the batch have the same shape, allowing us to train the model with a batch size greater than 1.

4.4 Model architecture

LSTM is a type of recurrent neural network that learns long-term dependencies in sequential data. It consists of three main components: an input gate, a forget gate and an output gate. The input gate decides what information to store in the memory cell, the forget gate decides what information to erase from the memory cell and the output gate decides what information to output from the memory cell. The memory cell is able to retain information over long periods of time without suffering from the vanishing gradient problem [65].

For each word, we had seven audio files, we used 5 of them for training and two for testing. In total $15 \times 155 \times 5 = 11,635$ audio files were used for training. We train our models separately for each speaker. The model we trained uses 3 layers of LSTM cells

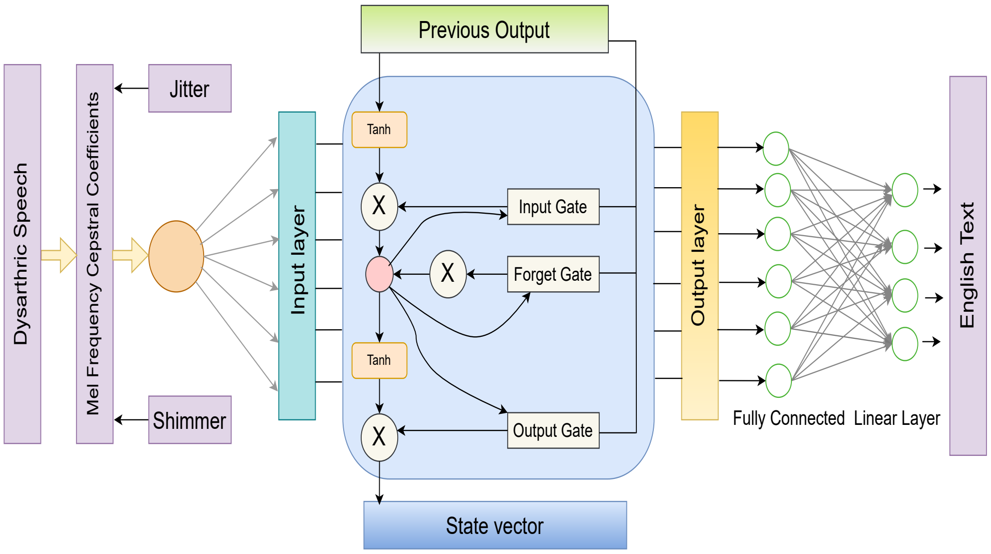
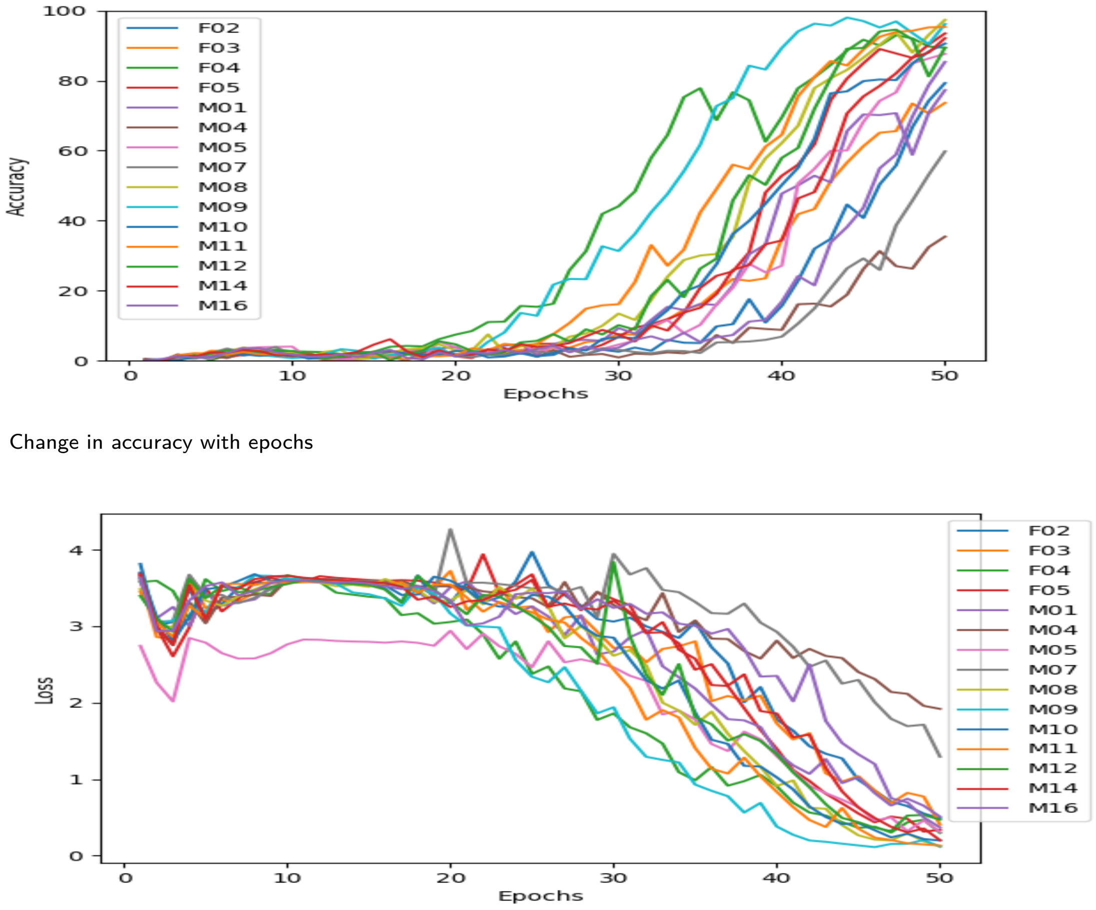


Figure 4.3: The figure shows an LSTM followed by a fully connected layer. The data flow starts from the dysarthria-impaired speech audio files followed by feature extraction and LSTM to produce English text as the output

stacked upon each other. In this case, the hidden state (output) of the previous cell acts as the input to the subsequent cell. The hidden state of the final layer is considered as the output of the LSTM. The output of LSTM is then fed to a fully connected layer with output size as the number of classes in the classification, that is 155. The fully connected layer can be seen as a linear transformation from the output of LSTM to the classes. We applied a softmax layer to the output of the fully connected layer and used the Cross-Entropy Loss in our experiment as the criterion for calculating the loss. Applying softmax layer on the logits converts the vector with a random number to a probability distribution. Figure 4.3 shows our proposed methodology flowchart. We used LSTM-RNN architecture connected by a fully connected layer with the goal to classify dysarthric words effectively, followed by a personalized ASR system. Table 4.1 presents the results for each speaker. For the robustness of the model, we tested with $15 \times 155 \times 2 = 4,650$ audio files. We also repeated the testing process several



(a) Change in accuracy with epochs

(b) Change in loss with epochs

Figure 4.4: Accuracy and loss curve for each speaker

times by swapping out the testing files with other utterances of the same words and then calculated the average results. During both training and testing phases, exclusively dysarthric speech utterances were used, despite the availability of control data within the UASpeech database.

4.4.1 Simulation setup

To perform the above-stated operation on large volume data, the CPU takes a huge amount of time. The usage of GPU ensures faster computation speed by parallelizing the basic operations. Because of these reasons, the simulations are run on a remote host with access to GPU. The GPU has Python and CUDA installed initially. A container orchestration API is used to run our code on a remote server. We used Kubernetes to run the docker image on the remote host. The total parameter size is $5,85,652$ and the learning rate is 0.001. For each training model, we fixed the epoch size to 50 because no significant improvement was seen after that. We have used Adam optimizer in our simulations as it is faster and more efficient.

4.5 Results and discussion

The effectiveness of ASR is often assessed using two criteria: Word Error Rate (WER) and Word Recognition Accuracy (WRA). We calculate WRA for each individual speaker and the average WRA for each group of intelligibility rate. Since the experiment covers speakers with different levels of intelligibility, the accuracy for speakers with mid and high intelligibility is remarkable. Even for speakers with low and very low intelligibility, the accuracy stands out. Overall, we got a WRA of 83.11% for 155 words. The accuracy graph can be seen in Figure 4.4a and their loss curve in Figure 4.4b respectively. For speaker M08, having a 93% intelligibility rate, we achieved a 97.34% WRA. Generally, speakers with high intelligibility rates have good WRA because their speech deterioration is in the initial stages. However, speakers with lower intelligibility have more distorted speech with many pauses and stops in between, making it very challenging to recognize their spoken words. For those speakers, audio trimming and Hamming windows

Table 4.1: Intelligibility rate of speakers along with the WRA(%) of the LSTM-RNN model and baseline method (VL: Very Low, L: Low, M: Mid, H: High)

Speaker	Intell.(%)	**WRA(%)**	[40]	[128]
M01	VL(15%)	77.29	61.53	28.39
M04	VL(2%)	35.44	59.33	10.94
F03	VL(6%)	73.63	70.66	32.47
M12	VL(7.4%)	89.42	69.33	40.65
Very Low Intelligibility Average WRA%		**68.94**	-	28.11
M07	L(28%)	59.80	81.33	69.46
F02	L(29%)	79.30	81.33	67.02
M16	L(43%)	85.35	70.00	55.91
Low Intelligibility Average WRA%		74.81	-	69.44
M05	M(58%)	87.62	90.00	64.95
F04	M(62%)	89.36	71.33	53.98
M11	M(62%)	95.30	92.00	49.68
Mid Intelligibility Average WRA%		**90.76**	-	56.49
M08	H(93%)	97.34	95.33	86.67
M09	H(86%)	96.15	86.66	85.16
M10	H(93%)	90.57	92.00	89.68
F05	H(95%)	93.47	94.00	94.41
M14	H(90%)	92.12	91.33	86.45
High Intelligibility Average WRA%		**93.93**	-	88.47
Average WRA%		**83.11**	-	-

significantly improved the concentration of energy within the selected frames, leading to better recognition performance. We didn't exactly compare our work with two similar approaches [40][128] because in the methods, number of classes, configuration and word selection are different, even though the goals were similar. However comparative analysis is represented in Table 4.1. Like us, they also developed an automatic speech recognition system for dysarthric speech using the UASpeech database in a speaker-dependent context. They applied a multi-net artificial neural networks approach for recognizing dysarthric speech with a vocabulary of 25 words.

However, when we tested the model with Out of Vocabulary (OOV) words, accuracy was lower. This situation can be attributed to several factors. The LSTM network contains an extensive number of parameters to learn. Additionally, only 7 instances of each word were available, as the UASpeech dataset contained only this limited amount of data per word contributing to initial underfitting.

In addition to the WRA, we have conducted the articulatory analysis for the speaker with very low intelligibility rate, since a large portion of this category were misclassified. For speaker M04, vowel duration was noticeably longer than the other speakers, leading to extended stretches in speech. Pronunciation issues with stops and fricatives were also observed in speakers M04 and M01. In many words, the final consonant was deleted, like *beat → [b iy]*. For speakers F03 and M04, there were many cases of stop consonant substitution, such as *farm → [p aa r m]*. We found several flaws associated with each speaker. This analysis motivates us to adjust the model accordingly and do the analysis for the remaining speakers as well. Currently, we are looking for the data so that we can improve the model's performance for dysarthric speech. By combining multiple audio descriptors, we achieved good WRA for most of the speakers, which was not yet reported for the dysarthric speech recognition system.

4.6 Conclusion

This chapter has investigated how well a personalized speech recognition training and adjustment techniques can handle dysarthric speech. Our approach revolves around a speaker-dependent strategy for recognizing dysarthric speech. Overall, the study underscores the absence of a universal solution and emphasizes that this is particularly true for individuals with dysarthria. We explored LSTM-RNN with MFCC and different variants of jitter and shimmer features and it works well to handle the variability among each speaker. Our experiment reveals that our proposed method successfully enhances the accuracy and is capable of recognizing words spoken by them. This will give them confidence and surely impact dysarthric speakers' lives in a positive way.

There are things known and there are things unknown and in between are the doors of perception.

Aldous Huxley, English writer

5

Dysarthric speech detection and severity classification using AST

5.1 Introduction

Humans are social creatures by nature and mutual dependence is required for growth in this environment; hence, communication is an essential aspect of life. People with disabilities are twice as likely as the general population to suffer disorders such as anxiety, asthma, heart disease, stroke or poor oral health [120][105][101][102]. Dysarthria is not a life-threatening disorder, but it affects the livelihood of patients in many aspects, including social, physical and emotional challenges [94][42]. As severity increases, they are more likely to rely on others for their daily activities and household chores. They can use a keyboard or joystick-based application, but due to a lack of muscle coordination and trembling hands, that also does not help much. This chapter studies the detection and severity classification methods for dysarthric speech.

To manage dysarthria effectively, it is critical to detect the condition at an initial stage. Early detection of dysarthria allows patients to begin therapy sessions timely, which will definitely improve their communication and reduce the disorder's impact on their lives. Assessing the severity of dysarthria is an essential diagnostic step that can provide valuable insights into the patient's condition, disease progression and potential treatment options. It can also assist clinicians in determining the appropriate course of medication and therapy sessions. However, classifying the intelligibility and severity of dysarthric speech poses several challenges due to variable speech features and subjective judgments.

In clinical practice and research, we frequently utilize severity ratings to explore speech difficulties. However, existing methods for determining the intensity of speech have not been properly evaluated and there is no commonly accepted definition for classification system [141]. Speech difficulties caused by dysarthria are commonly described by both clinicians and researchers. Speech-Language Pathologists (SLPs) often use two common informal methods to measure speech intelligibility: (a) Estimating the percentage of a patient's speech that others can understand and (b) using descriptive labels like 'normal', 'mild', 'moderate', 'severe' or 'profound'. We require further research to establish the reliability and accuracy of these assessments and to understand the factors

Table 5.1: The distribution of speech severity based on the intelligibility range (%) in the literature.

Article	Profound	Severe	Moderate	Mild	Normal
[9]	0-60	60-70	70-80	80-90	90-100
[19]	-	0-45	45-75	75-100	-
[29]	0-25	25-50	50-75	75-100	-
[31]	-	0-40	40-70	70-100	-
[20]	0-25	25-50	50-75	75-100	-
[43]	-	0-50	50-75	75-100	-
[141]	0-50	50-80	80-90	90-95	95-100

that influence how we perceive the severity of someone's speech issue.

There is often a correlation between speech intelligibility and speech severity in dysarthria literature, yet these two measures are unconnected. Kaila et al. highlighted the relationship between speech intelligibility and severity very elegantly [141]. Dysarthria severity pertains to the extent of motor impairment affecting speech production, while intelligibility refers to how well listeners comprehend the speaker. Speech intelligibility relies on speech efficiency, voice quality and speaking rate of patients. In common practice, severity is determined based on the intelligibility rate. Table 5.1 shows that researchers have used different cutoff points to define severity levels based on speech intelligibility scores. Nevertheless, inconsistencies exist not just in the assigned ranges of intelligibility for each category but also in the approaches used to measure intelligibility across various studies.

5.2 Motivation and related work

In conventional Speech-Language Pathology (SLP) practice, the severity of speech disorders was typically evaluated using the standardized rating scales given by the Frenchay Dysarthria Assessment (FDA) [38]. This evaluation process incorporates a combination of acoustic, physiological and perceptual measures. However, it is worth noting that while treating patients with dysarthria, the use of physiological measurements can be demanding and require specialized equipment and expertise and on the other hand, perceptual measures can vary considerably depending on the clinician's level of experience and listening skills. Additionally, this would be costly and time-consuming, limiting its use in remote rehabilitation. In order to maintain homogeneous interpretation across SLPs, it is necessary to conduct dynamic assessments to determine speech severity rates. It is, therefore, necessary to develop a system that automatically classifies dysarthria severity levels. Automated severity assessment methods are cost-effective, traceable, reliable and allow remote monitoring of rehabilitation progress for patients. Researchers have been exploring various approaches to achieve accurate results for dysarthric speech severity classification [140][163][71].

Akshatha et al. [72] employed CNN to automatically detect early-stage ALS from highly intelligible speech. They utilize both time-domain and frequency-domain CNNs to categorize speech from a group of 13 patients with early-stage ALS. The frequency-based CNN showed better performance at predicting ALS at the individual level, compared to the time-based CNN. In [13], researchers utilized joint spectro-temporal features extracted from a Mel-scale spectrogram for dysarthria severity estimation. Their findings demonstrated that a time-frequency CNN that captures both spectral and temporal information outperforms CNN that captures only temporal or spectral information separately. This highlights the importance of jointly considering both aspects to achieve superior dysarthria severity estimation results. In the long run, CNN was saturated and researchers started to add an attention layer with CNN for end-to-end audio classification.

Over the last decade, deep learning techniques have been extensively explored for end-to-end audio classification, emphasizing direct mapping from spectrogram to corresponding labels. To address the challenge of capturing long-range global context, some researchers have introduced a hybrid model by incorporating a self-attention mechanism with CNN. The proposed method builds upon the transformer architecture, which has been previously explored for language and audio processing, but in this case, it is combined with CNN. Researchers have experimented with various combinations of transformer and CNN, such as implementing a transformer on top of CNN and incorporating a transformer within each block of CNN. Hybrid models that combine CNNs with attention mechanisms have demonstrated remarkable effectiveness in delivering precise outcomes across various tasks like audio event classification, emotion recognition and command recognition. In the vision domain, purely attention-based models have demonstrated remarkable success, leading to the question of whether CNNs are still necessary for audio classification. The approach put forward in this study is entirely attention-based and does not rely on convolution, making it a unique and innovative method for the task at hand. The main objective of this work is to address the tasks of dysarthric speech detection in a speaker-dependent manner and severity classification in both speaker-dependent and speaker-independent contexts, specifically:

- Speaker-dependent dysarthric speech detection: This involves detecting whether a speaker's speech is affected by dysarthria, a motor speech disorder.
- Speaker-dependent dysarthric speech severity classification: After detecting dysarthric speech, system aims to classify the severity level of dysarthria for specific speaker.
- Speaker-independent dysarthric speech severity classification: In this task, the objective is similar to the second task, but the system could classify dysarthria severity for speakers who were not part of the training set, making it more generalizable.

The proposed architecture for these tasks which has been designed outperform all previous methods in terms of accuracy for dysarthric speech detection and severity classification. The experiments utilize the publicly accessible UASpeech dataset for all objectives. [67] found that as dysarthria severity increased, listeners' confidence in transcribing dysarthric speech decreased. So, we can say that as severity increases, speech intelligibility decreases.

The entire UASpeech database is utilized for the detection and severity classification experiments in which data is split in a random fashion for training, testing and validation. Based on the data available in UASpeech, our dataset consisted of audio files from 16 individual speakers, each contributing 765 isolated words, with 7 audio files available for each word. For the detection task, we included both controlled and dysarthric speech samples from these 16 speakers, resulting in a total of 171,360 audio files (2×16×765×7 = 171,360). However, for the severity classification task, we exclusively used dysarthric speech data, totaling 85,680 audio files (16×765×7 = 85,680). The following section elaborates on the research methodology and model architecture, which is then followed by a comprehensive report detailing the experimental procedures and the outcomes achieved for each specific objective.

5.3 Methodology

Lately, the Transformer architecture has gained significant popularity in the realm of image processing. To adapt it for audio processing, a modification has been made wherein the Audio Spectrogram Transformer (AST) is designed in such a way that instead of taking an image as an input AST utilizes logarithmic Mel spectrograms derived from speech signals. The baseline of AST draws inspiration from the architecture of the Visual

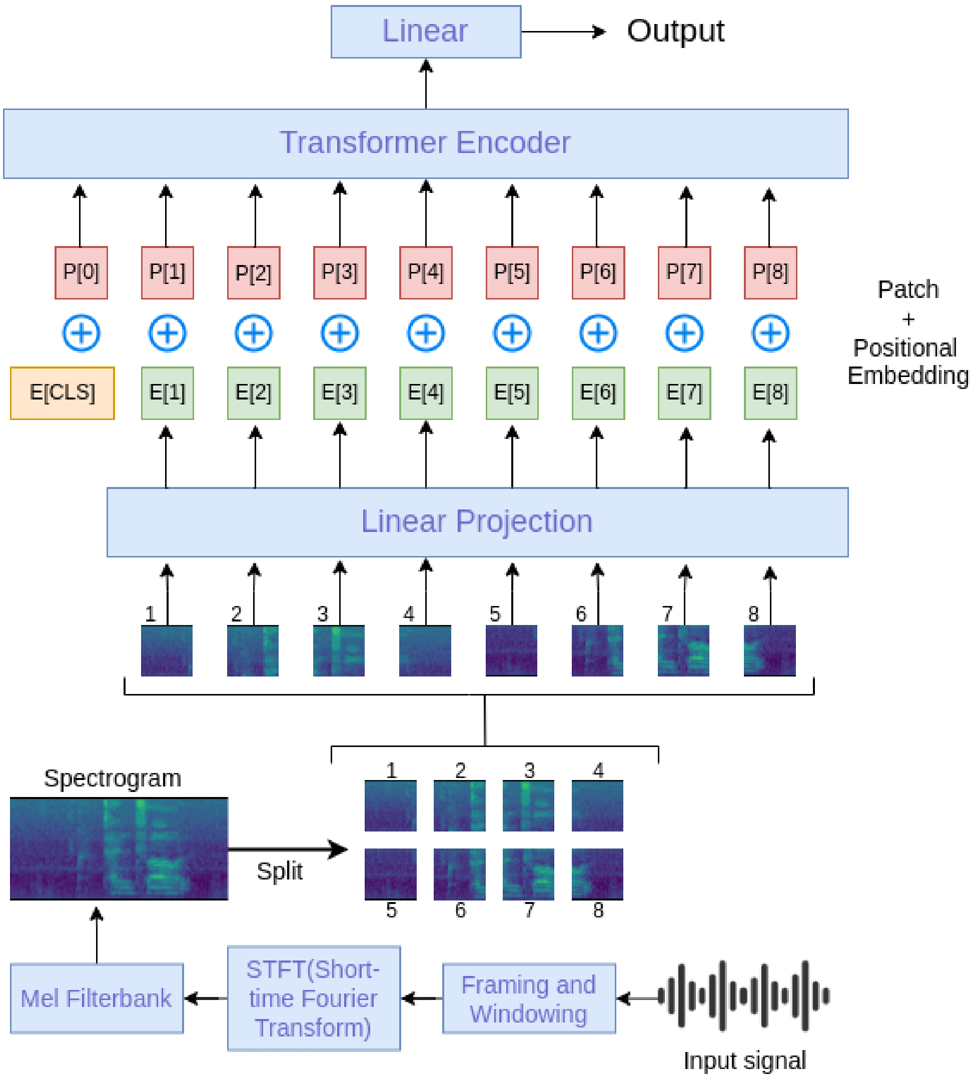


Figure 5.1: Audio Spectrogram Transformer

Transformer (ViT) [52][30].

5.3.1 Audio Spectrogram Transformer

To achieve this transformation, the input audio waveform is initially converted into a sequence of 128-dimensional vectors using librosa, called log Mel filterbank (fbank) features [90]. The process of Mel filterbank feature extraction involves handling the input audio waveform with a duration of ‘t’ seconds. This process includes dividing the audio input into manageable chunks every 10 ms and performing a Short-Time Fourier Transform (STFT) with a 25 ms Hamming window on each chunk generating a sequence of 128-dimensional vectors that illustrate the evolution of the input audio waveform [111]. The STFT determines the power spectrum of each chunk, which is then passed through a set of filters. These filters are designed to mimic human hearing and are spaced non-linearly in frequency to better capture the properties of speech and other sounds. The non-linear spacing of Mel filterbanks helps to focus on the frequency regions that are more relevant for discriminating between different phonemes and sounds in speech, contributing to improved robustness to noise as well. The distribution of the filters in the Mel scale tends to concentrate more on the lower frequencies, which often contain critical information for speech understanding. The filters in the Mel filterbank are designed to capture the distribution of energy across different frequency bands. Human perception of sound is not linear with respect to frequency and the Mel scale is a perceptual scale that approximates the human ear’s response to different frequencies. The set of filters helps to map the raw frequency content of the audio signal into a representation that aligns better with human perception. The resulting sequence of vectors forms a spectrogram, which serves as the input to the AST model.

Next, the spectrogram is split into a sequence of smaller N patches, each having a size of 16×16. These patches are extracted with a 6-step overlap in both time and frequency dimensions. The overlap refers to the degree of overlap between consecutive patches derived from the spectrogram. A 6-step overlap in both time and frequency dimensions means that when extracting patches from the spectrogram, each patch is offset by 6 steps in both the time and frequency directions compared to the previous patch. This overlap is used to ensure that information from adjacent patches is shared, which is beneficial for capturing temporal and frequency-related patterns in the data. To determine the total number of patches, denoted as N, the following formula is utilized:

$$N = 12 * (100 * t - 16)/10 \tag{5.1}$$

Here, 't' represents the duration of the audio waveform in seconds. The value of N corresponds to the effective input sequence length for the transformer at that particular stage. Each patch is then treated as an individual input token and the transformer processes them separately.

Thereafter, using a linear projection layer, each 16x16 patch that was extracted from the spectrogram is flattened into a 1D patch embedding of size 768. This dimensionality reduction technique helps retain essential features while reducing the complexity of the input data. To preserve the spatial structure of the original 2D audio spectrogram, a trainable positional embedding of size 768 is added to each patch embedding. It means that for each patch, there is a learnable vector of 768 elements that represents its position in the 2D space. These embeddings are trainable, signifies that the model can adjust them during training to best capture the spatial relationships in the data. By incorporating positional embeddings, the model retains information about the spatial location of each patch, which is otherwise lost during the flattening process. The positional embeddings are required since the model does not naturally capture the order information of the input data, making them necessary to capture the spatial structure of the input. By encoding the spatial structure, the model becomes capable of distinguishing between patches that are close together and those that are far apart. Along with the other model parameters, the positional embeddings are learned during training and updated during backpropagation.

The sequence begins by appending a unique token named [CLS]. In classification tasks, the [CLS] token is a special symbol that is used to represent the entire sequence and allows the model to make predictions based on the full input sequence. The sequence is then fed to the transformer. In this case, only encoder layers are being used as we are doing detection and classification instead of recognition tasks. The output of the Transformer encoder, specifically the hidden state of the [CLS] token, serves as the representation of the spectrogram. To generate the final prediction, a linear layer with a sigmoid activation function is applied to map the audio spectrogram to the labels. The combination of the linear layer and sigmoid activation allows the model to learn a mapping from the features extracted from the audio spectrogram to a prediction of positive or negative class. For the detection task, label 0 corresponds to controlled voice, while label 1 corresponds to dysarthric voice.

5.3.2 Adaptation of ViT in AST

In Visual Transformer (ViT), the input image is divided into non-overlapping patches and each patch is treated as an individual token. Similarly, in audio processing, the spectrogram is divided into overlapping segments, treating each segment as a token. This allows the transformer to capture local patterns in both vision and audio. ViT uses positional embeddings to provide the model with information about the spatial arrangement of patches. Similarly, in audio processing, positional embeddings are used to convey information about the temporal order of spectrogram tokens. In this way, the architecture of the ViT serves as inspiration for the baseline of the Audio Spectrogram Transformer (AST).

The AST model is designed in such a way that it is able to transfer the 2D spatial knowledge from a pre-trained ViT to the AST even when the input shapes are different. The AST takes advantage of transfer learning by utilizing pre-trained weights from the ViT architecture, which enables it to use expertise gained from a sizable dataset of images to improve its performance on the audio classification job. Additionally, since the network has already picked up useful features from the images it was trained on, using pre-trained weights reduces the amount of training data required for the AST to perform well. Given the limited availability of dysarthric speech in UASpeech, transfer learning enables the model to acquire valuable representations from a larger visual dataset and subsequently adapt these representations to the audio domain.

The positional embedding of ViT architecture is fixed in size since it employs a fixed-size input image, but while dealing with audio data it can be of variable length. As audio signals vary in length, maintaining the sequential nature of the data becomes essential for capturing temporal relationships within the audio signal. To accommodate variable-length audio sequences, the model employs padding, where shorter audio sequences are padded with zeros to match the length of the longest sequence in the dataset. The AST, analyses 16x16-pixel patches in variable-length audio spectrograms.

The adaptation of positional embedding from the ViT to AST architecture involves the utilization of cut and bi-linear methods. These techniques enhance the model's capacity to adeptly handle audio data characterized by diverse sequence lengths, facilitating the capture of temporal dependencies within input sequences. By effectively capturing

Table 5.2: Experimental setup and model parameters used for experiments

Parameter	Value
Input Normalisation	Dataset mean -4.268 and std 4.569
Number of Classes	2 and 4
Frequency Stride	10
Time Stride	10
Loss Function	BCE and CE
Learning Rate Scheduler	MultiStepLR with 0.5 decay
Training Device	cuda
Total Parameter Number	87.728 million
Optimizer	Adam and SGD
Input Method	JSON file

temporal relationships, mitigating the impact of padding and enhancing generalization across varying sequence lengths, these methods contribute significantly to the model's performance in processing audio data.

In a more technical sense, the patch embedding layer is likened to a single convolution layer with an extensive kernel and stride size and the projection layer within each transformer block is equivalent to a 1x1 convolution. It is important to note, though, that the design diverges from conventional CNNs, which typically employ multiple layers having smaller kernel and stride sizes. Transformer models are often labeled as "convolution-free" to distinguish them from traditional CNN architecture.

5.4 Experiments and results

The experimental setup involved using spectrograms, which are visual representations of the frequency content of a signal over time. The Audio Spectrogram model was employed for all tasks performed. To execute the experiments, we utilized a Kubernetes cluster, specifically utilizing an Nvidia A100 GPU with 40GB RAM for training the model for

Table 5.3: Classification result with glottal features and CNN+LSTM model by Narendra et al. [98] for dysarthric speech detection

Input	Accuracy	Sensitivity	Specificity
Raw Speech	74.19	69.26	81.48
Glottal flow	**77.57**	73.13	82.48

all scenarios. We summarize our experimental setup in Table 5.2. The subsequent subsection provides an in-depth description of each objective.

5.4.1 Speaker-dependent dysarthric speech detection

Initially, we conducted dysarthric speech detection using a limited subset of the UASpeech dataset, specifically encompassing 16% of its content. The training set consisted of 11,486 audio files, while the testing and validation sets contained 5,744 and 5,743 audio files, respectively. After training the model for 3 epochs, we achieved an accuracy of 94%. To further improve the results, we extended training to 10 epochs and the accuracy significantly improved to 96.86%. These outcomes demonstrated better performance compared to some early experiments conducted by Narendra et al. [98]. They applied CNN+LSTM model on raw speech and glottal flow as shown in Table 5.3.

Subsequently, we conducted the same experiment using the entire UASpeech dataset. Remarkably, the accuracy improved to an impressive **99.64%**, surpassing the performance presented by [135]. The detailed results are documented in Table 5.4, while the loss curve is visualized in Figure 5.2. Additionally, Figure 5.3 showcases a comparison graph illustrating the accuracies of the previous state-of-the-art CNN-GRU model alongside our proposed AST model. The graph clearly demonstrates the superior accuracy achieved by the AST model over the previous approach.

Table 5.5 presents a comprehensive comparison of various approaches for dysarthric

Table 5.4: Training and validation metrics for dysarthric speech detection using AST

Epoch	Validation				Training loss	Validation loss	learning rate
	Accuracy(%)	AUC	Avg Precision	Avg Recall			
1	93.0039	0.988149	0.890938	0.966523	0.314101	0.55059	0.001
2	97.6181	0.998203	0.937057	0.997080	0.100899	0.517632	0.0005
3	99.2443	0.999622	0.943930	0.999259	0.029803	0.508477	0.00025
4	99.4219	0.999764	0.948494	0.999717	0.015913	0.532678	0.000125
5	99.5299	0.999825	0.967261	0.997639	0.009256	0.506424	6.25e-05
6	99.4742	0.999758	0.976602	0.993125	0.006116	0.506301	3.125e-05
7	99.5960	0.999848	0.986126	0.999197	0.004430	0.505693	1.5625e-05
8	99.6169	0.999820	0.992635	0.997011	0.003522	0.505428	7.8125e-06
9	99.6204	0.999800	0.995619	0.996684	0.002949	0.505475	3.90625e-06
10	99.6378	0.999750	0.995623	0.993877	0.002666	0.505362	1.953125e-06

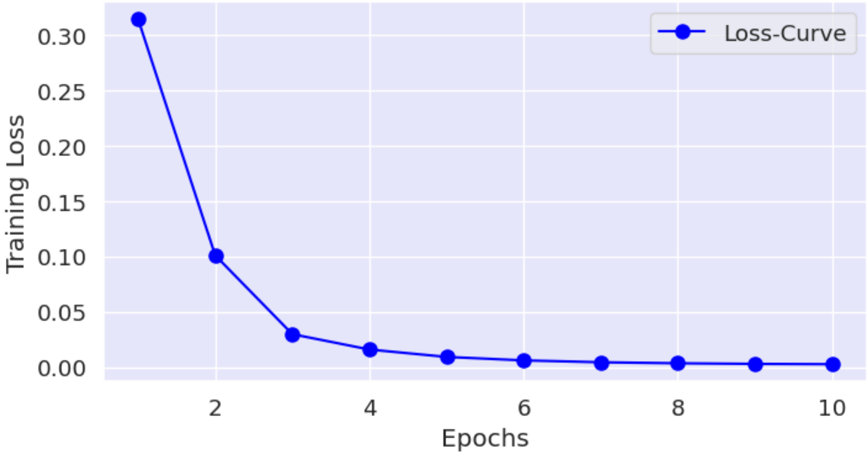


Figure 5.2: Loss-curve for dysarthric speech detection by AST.

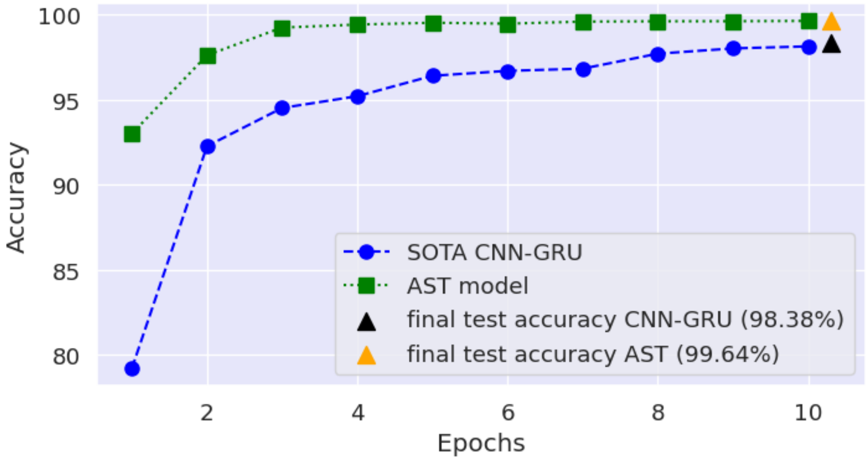


Figure 5.3: Comparison between the SOTA CNN-GRU results [135] and current method results for dysarthric speech detection.

speech detection on the UASpeech corpus. It includes details about the authors, their classification methods and the respective maximum accuracy achieved by each method.

Various authors have employed different approaches for dysarthric speech detection on the UASpeech corpus. We utilized the Audio Spectrogram Transformer (AST) approach, achieving the highest accuracy of 99.64% for dysarthric speech detection. The AST model is a neural network architecture that is specifically designed to process audio

Table 5.5: Performance comparison for dysarthric speech detection for UASpeech corpus

Author	Classification Method	Accuracy
Hernandez et al. (2019) [63]	SVM	72%
Narendra et al. (2019) [96]	SVM	96.38%
Narendra et al. (2020) [98]	CNN-LSTM	77.57%
Rajeswari et al. (2022) [117]	CNN	95.95%
Dong-Her et al. (2022) [135]	CNN-GRU	98.38%
Present work (2023)	**AST**	**99.64**%

spectrograms for speech recognition tasks. Overall, the results in Table 5.5 demonstrate the effectiveness of deep learning approaches, particularly those based on CNNs and their variations, are effective for dysarthric speech detection on the UASpeech corpus. Among these approaches, the AST method achieved the highest accuracy. Our model surpassed all previous dysarthric speech detection accuracies, highlighting its superiority in this task. These approaches have the potential to improve the accuracy and efficiency of dysarthria screening in clinical settings.

5.4.2 Speaker-dependent severity classification

After dysarthric speech detection, we proceeded with severity classification experiments. The severity levels for dysarthria in the UASpeech dataset are grouped into four categories, which are based on the assessment of speech-language pathologists. Table 5.6 includes the severity categories of UASpeech along with corresponding speaker IDs. These severity levels form the foundation for assessing the effectiveness of dysarthric speech severity classification models in accurately predicting the level of dysarthria exhibited by the speaker.

Initially, we performed severity classification using Binary Cross Entropy (BCE) as the loss function, Automated Dynamic Analysis of Mechanical Systems(ADAMS) as the

Table 5.6: UASpeech dataset distribution according to severity level

Severity	UASpeech
Very Low	F05, M08, M09, M10, M14
Low	F04, M05, M11
Medium	F02,M07, M16
High	F03, M04, M12, M01

optimizer and a batch size of 16 for 30 epochs, resulting in an accuracy of 84%. In pursuit of better performance, we switched to Stochastic Gradient Descent (SGD) as the optimizer, which significantly improved the accuracy to 93.6%. Continuing our efforts to enhance the model, we implemented a dynamic learning rate strategy, reducing it after every 3rd epoch. Additionally, we opted for the Cross-Entropy (CE) loss function. The choice of the Cross-Entropy (CE) loss function is motivated by its suitability for classification tasks, including severity classification models. This is derived from the principle of maximum likelihood estimation. It encourages the predicted probability distribution to be close to the true distribution of the labels. These combined adjustments resulted in the highest accuracy achieved so far, reaching an impressive **95.6%**. However, further experiments were conducted to enhance the model, such as experimenting with the learning rate by decreasing it after every 4 epochs. Unfortunately, this alteration did not yield the desired results and the accuracy dropped to 89%. Figure 5.4 shows the accuracy versus loss graph plotted for 30 epochs during the classification experiments. Additionally, we evaluated the model's performance on both the validation and test sets and the corresponding confusion matrices are provided in Figure 5.5 and 5.6.

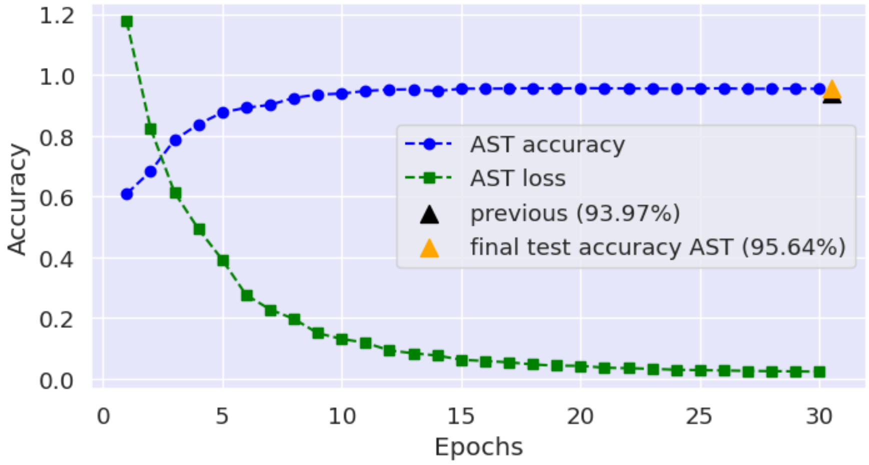
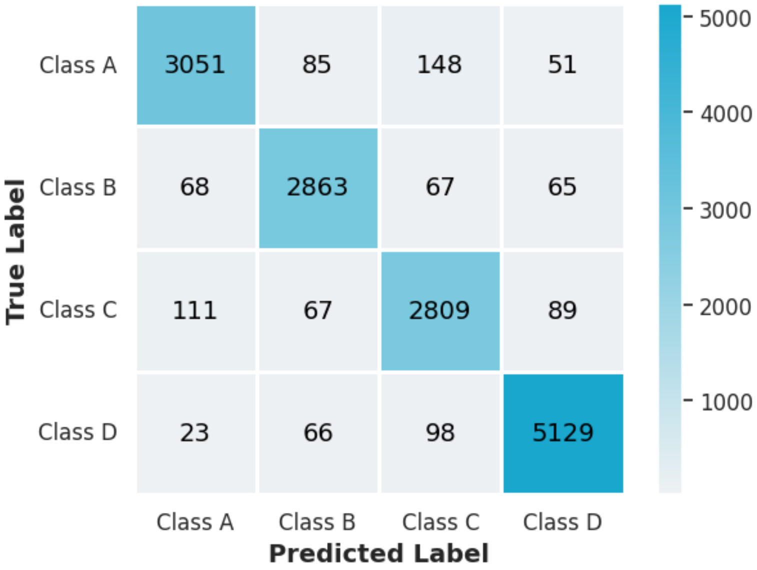


Figure 5.4: Accuracy and loss curve for severity level classification by AST.



True Label \ Predicted Label	Class A	Class B	Class C	Class D
Class A	3051	85	148	51
Class B	68	2863	67	65
Class C	111	67	2809	89
Class D	23	66	98	5129

Figure 5.5: Validation confusion matrix

5.4.3 Speaker-independent binary severity classification

Initially, the severity classification was carried out using a speaker-dependent model, where the training and testing data included only one specific speaker. However, to enhance practicality and applicability, it was desirable to develop a speaker-independent

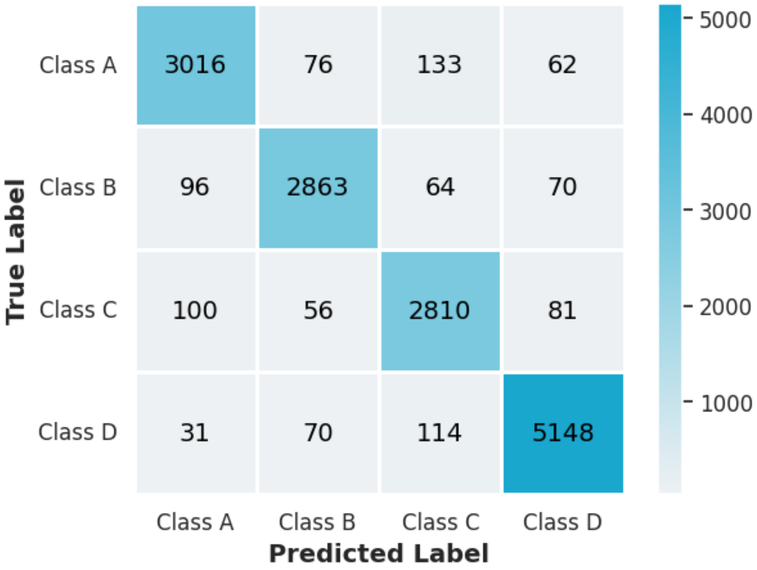


Figure 5.6: Test confusion matrix

model. This means creating a model capable of accurately classifying the severity of any speaker's voice without requiring prior knowledge of that specific speaker. Therefore, we aimed to create a speaker-independent model.

Table 5.7: Speaker distribution for train and test sets based on severity level

Severity	**Speakers**	
	Train	**Test**
High	F02,M16,F03,M12,M01,M07	M04
Low	F05,M09,M10,M14,F04,M11,M05	M08

Table 5.6 highlights that the intermediate classes exhibit a smaller speaker count compared to the border classes, indicating an imbalance in the UA-Speech database. In pursuit of creating a speaker-independent model capable of accurately classifying severity, we merged the "low" and "very low" severity classes, as well as the "high" and "medium" severity classes into one. This decision was motivated by the limited

availability of data for the “low” and “medium” severity classes, which could potentially lead to reduced model accuracy if these classes were treated separately. By combining them, we effectively increased the amount of data available for training, enabling us to build a more robust model capable of classifying severity independently of the speaker’s voice. The speaker’s data taken for training and testing can be seen in Table 5.7. This approach allowed us to improve the model’s performance across different speakers.

For speaker-independent classification, we trained the model for 10 epochs, using Binary Cross Entropy (BCE) as the loss function and utilizing the ADAM optimizer. The resulting accuracy was 62.5%. To enhance the model’s performance, we switched to the Stochastic Gradient Descent (SGD) optimizer and reduced the learning rate every two epochs during training. These modifications had a significant impact on the model’s accuracy, which increased the accuracy to **78.97%**. The dynamic learning rate strategy plays a crucial role in enhancing the accuracy of the speaker-independent severity classification model. Adjusting the learning rate allows the optimization process to converge more efficiently. Initially, a higher learning rate helps the model make large updates to its parameters, potentially escaping from local minima. As the optimization progresses, reducing the learning rate helps the model converge more precisely to the optimal solution. By reducing the learning rate, the model becomes more sensitive to smaller gradients and makes finer adjustments to its parameters. This is particularly useful in later stages of training when the model is close to convergence. Figure 5.7 shows the validation curves of accuracy and loss along with the marker of final testing accuracy as well as a comparison with the previous highest accuracy mentioned in [149] and [71].

Table 5.8 shows the results for speaker-independent severity level classification for dysarthric speech and Table 5.9 presents a comparison between the current results and

Table 5.8: Training and validation metrics for speaker-independent severity-level classification of dysarthria using AST

Epoch	Validation				Training loss	Validation loss	learning rate
	Accuracy(%)	AUC	Avg Precision	Avg Recall			
1	72.8976	0.8015	0.6404	0.8753	0.51792	0.6454	0.001
2	71.2854	0.8212	0.7163	0.8088	0.20765	0.6436	0.001
3	77.1023	0.8467	0.6818	0.8963	0.11326	0.6191	0.0005
4	71.1764	0.8172	0.7134	0.8127	0.08471	0.6438	0.0005
5	71.7211	0.8253	0.7157	0.8085	0.05499	0.6398	0.00025
6	78.3006	0.8623	0.7199	0.8676	0.04382	0.6110	0.00025
7	77.6252	0.8559	0.7204	0.8601	0.02992	0.6133	0.000125
8	76.5577	0.8461	0.7144	0.8506	0.02458	0.6181	0.000125
9	77.2331	0.8456	0.7253	0.8429	0.01694	0.6159	0.0000625
10	76.4705	0.8445	0.7236	0.8351	0.01497	0.6178	0.0000625

Table 5.9: Performance comparison for dysarthric speech speaker-dependent and speaker-independent severity classification

Work	Approach	Results
A.Tripathi, S.Bhosale, and S.K.Kopparapu[149]	Deep Speech posteriors with SVM	97.40%(SD) 65.20%(binary)
Amlu Anna Joshy and Rajeev Rajan[71]	i_MFCC with DNN	93.97%(SD) 70.52%(binary)
Current work	Spectrogram and Transformers, (AST), Speech-vision	95.6%(SD) **78.97%**(binary)

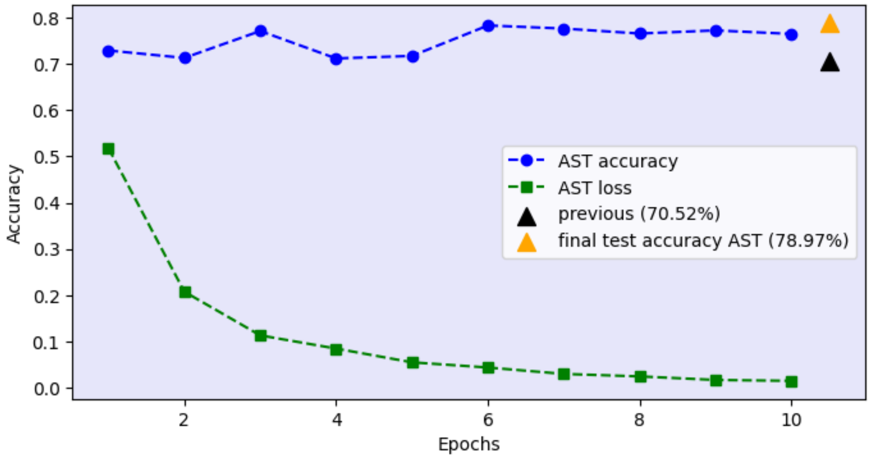


Figure 5.7: Accuracy and loss-curve for speaker-independent dysarthric speech detection by AST

the prior findings, highlighting that our outcomes demonstrate superior performance compared to the earlier results. Fig. 5.8 shows the validation and test confusion matrices which provide a visual representation of the model's performance in classifying severity levels.

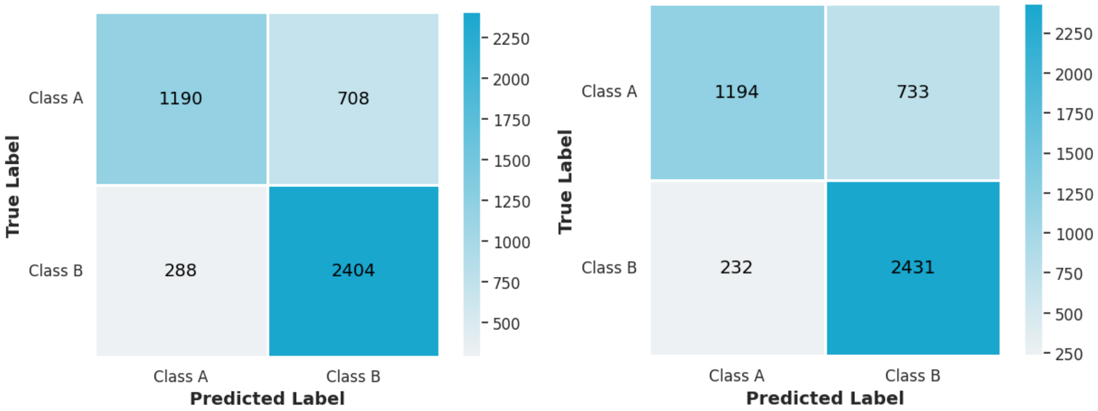


Figure 5.8: Validation and Test confusion matrix for speaker-independent severity-level classification

5.5 Conclusion

This chapter represents a comprehensive exploration of various deep-learning models employing the detection and classification of dysarthria severity levels. We use the Speech Vision approach with Audio Spectrogram Transformer (AST) for the above tasks and achieve 99.64% accuracy for the detection surpassing the performance of previous state-of-the-art models. For severity classification, our model achieved an accuracy of 95.6%. Additionally, we developed a speaker-independent model, which demonstrated a notable accuracy of 78.97%. These findings emphasize the potential of these advanced techniques to improve the accuracy and performance of dysarthria-related tasks significantly.

The model's capabilities can be extended by incorporating additional features such as phonetic and prosodic information. This inclusion would enable capturing more intricate details regarding the speech patterns of individuals with dysarthria, enhancing the model's overall performance and accuracy. To address the challenge of limited data, data augmentation and speech synthesis techniques offer valuable solutions.

This research can serve as a valuable resource for both patients and clinicians in accurately identifying the exact level of speech severity, thereby enabling them to track and

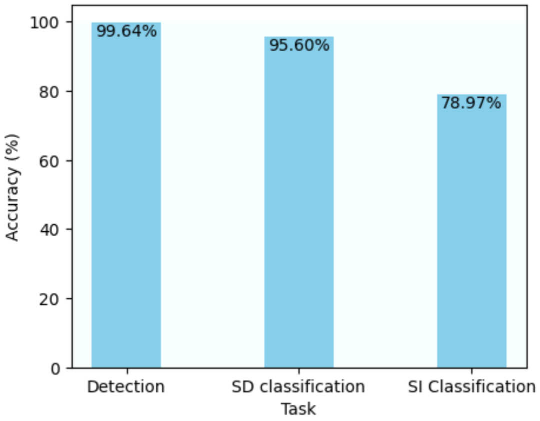


Figure 5.9: Accuracy of different objectives performed.

assess progress toward improvements. In severe cases, where individuals often encounter significant difficulty in articulating words, such advancements can be transformative and have a life-changing impact.

The scientific method is nothing more than a systematic way of thinking

Thomas Henry Huxley, English biologist

6

Dysarthric speech synthesis system using Tacotron2 for specific and OOV words

Automatic Speech recognition system has been explored extensively for normal speech [105][101] and has achieved high accuracy but dysarthric speech recognition remains relatively underdeveloped and needs to be explored. A primary factor contributing to this gap is the scarcity of suitable data [124]. Currently, only three publicly available dysarthric speech datasets exist. However, none of these datasets provide a diverse range of speakers and their conversational speech is also very less, making them incapable of handling ASR challenges. Moreover, the limited number of speakers can not capture the acoustic nature and inter-speaker variability. To establish a robust ASR system, a substantial volume of training data is needed and in case of impaired speech, data collection is quite difficult. There are a lot of factors behind these difficulties, including the medical condition of the patients, their willingness to participate and the emotional aspects.

Despite advancements at the feature level and modeling techniques, we identified that the primary obstacle to achieve better accuracy lies in the limited availability of suitable data. To address this issue, we have decided to prioritize and proceed with an extensive data collection process. This step aims to bridge the gap in data availability, enabling the development of more robust and accurate models capable of effectively recognizing dysarthric speech across different speakers and severity levels.

6.1 Outreach for data collection

I visited the Speech and Hearing Institute & Research Centre (SHIRC)* Kolkata, India on 4th November 2022. The research center is situated in Tollygunge, which is quite developed and known. SHIRC is a full-fledged organization devoted to the care of children with speech and hearing impairments in West Bengal and Eastern India. They have four other rural centers funded by several agencies. SHIRC provides an integrated service to speech and hearing-impaired children through parent/professional collaboration, home training, preparation for integrated and special education, vocational training and day-care programs. Each and every disabled child is encouraged to become self-sufficient and contributing members of the society. I talked to a few of them and most of them

*https://www.shirc.org/

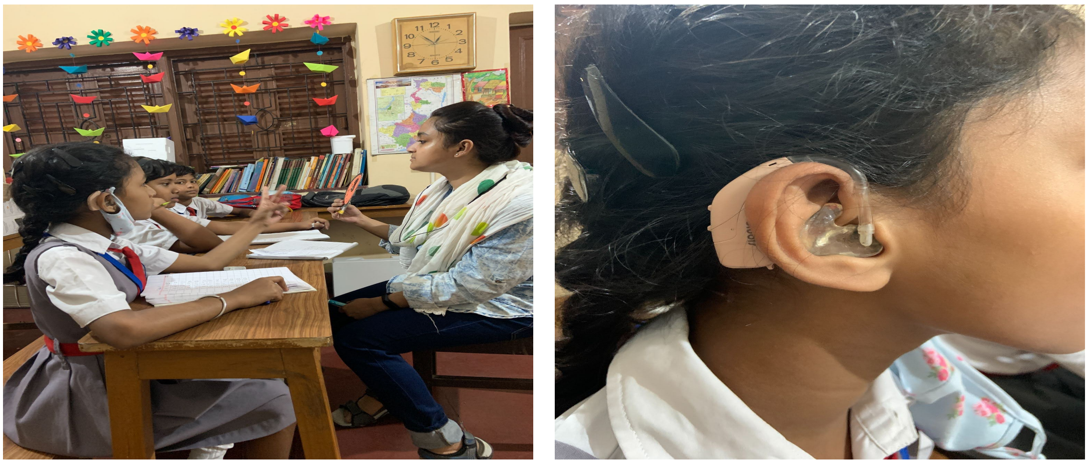


Figure 6.1: Learning environment at SHIRC

are intellectually fit and use sign language to make themselves understand to others. They are just like average children who grasp new things very easily. Due to society's negligence, they distance themselves from new people. There are different categories of children with several speech disfluencies and hearing aids. They categorize them based on a similar problem and do assessments at certain intervals. A few categories of disabled children are:

- Hearing Impairment
- Cerebral Palsy
- Emotionally Stressed Child
- Mentally Retarded
- Stammering and Articulation Defects
- Slow Learners and Language Learning Difficulties

Approximately 400 hearing-impaired children attend its special education centers. An average of 25 children are examined in the outpatient department per day. When necessary, the center offers early intervention, diagnostic, therapeutic treatment and fitting services. Several rooms for different activities are decorated by the Ananta project staff members, which are fully funded by the Government as shown in Figure 6.1. I tried to record digits with my mobile, but the environment was too noisy. They also tried to record in a quiet environment by taking the children into a closed room, but that also didn't help. Few of them knew English and most of them were from Bengali medium.

In conclusion, the data was of use, but the limited data and environmental conditions did not support using them in the experiments. We made significant efforts to collaborate with AIISH Mysore, NIMHANS Bangalore and IISc Bangalore for data collection, including several meetings with them. However, despite our efforts, these collaborations did not yield substantial results, which led us to look for alternate solutions to address the limitation of data scarcity.

6.2 Justification for synthetic data generation

We realized that relying solely on physical data collection would not be very effective, as it is a lengthy and time-consuming process. Therefore, we shifted our focus toward synthesizing data from the existing database. Speech synthesis, mostly abbreviated as Text-to-Speech (TTS), helps to increase the number of utterances that are already present in the dataset's vocabulary and they are capable of generating Out-Of-Vocabulary (OOV) words exhibiting the same characteristics, traits and naturalness as the speaker possesses. We propose a speaker-adaptive dysarthric speech synthesis technique using the Tacotron2 model in a speaker-dependent scenario. Utilizing the available UASpeech database, we train our model to generate speech from text inputs while preserving the distinctive characteristics and natural fluency of the original speaker. The objective of this contribution is to expand the database, resulting in a substantial increase in the number of utterances and new words. This expansion will greatly benefit model training, ultimately leading to the development of robust automatic speech recognition systems for dysarthric speech. We used this model to generate dysarthric speech utterances that already exist in the UASpeech database and for OOV words as well to expand the vocabulary size.

6.3 Motivation

In the past few years, deep learning-based advanced acoustic models have made far-reaching changes in speech synthesis and speech transformation systems [66]. Researchers have explored that Deep Neural Networks (DNNs) [166][113][1] based speech synthesis can provide better and more realistic speech than HMM-based speech synthesis [114][152][80]. Within certain extensive TTS systems, the WaveNet [104] which is a generative model for raw audio was designed to generate waveforms directly in the time domain, but linguistic features, predicted log fundamental frequency and phoneme durations require laborious feature engineering and extensive domain expertise to handle the input parameters. Char2Wav[138] is an end-to-end model for speech synthesis based on an encoder-decoder model with attention and a bidirectional recurrent neural network.

TTS engines has been implemented for normal speech majorly. If individuals with impaired speech could synthesize their own voice solely through text input, it could offer remarkable benefits for them. Tacotron2 incorporates an attention mechanism that allows the model to focus on relevant parts of the input text while generating speech. This attention mechanism helps the model to adapt to the variable and diverse patterns present in dysarthric speech even with limited training data. With this inspiration, we use the Tacotron2 model for the synthesis of dysarthric speech which has not yet been reported. End-to-end speech synthesis holds promise in converting text into spoken words with naturalness, but we were not sure how well it would work for dysarthric speech. Tacotron2 introduced a new way of generating speech directly from text, bypassing the traditional pipeline methods that require multiple steps, such as text analysis, sound processing and timing adjustments.

6.4 Database description

For our investigation, we selected six speakers, including both male and female voices from the UASpeech database. During the speaker selection process, we intentionally tried to encompass a diverse range of intelligibility rates, ensuring comprehensive coverage across distinct categories. Details of chosen speakers and their corresponding intelligibility rates are provided in Table 6.1. UASpeech database consists of 3*(10 digits, 26 radio alphabet letters, 19 computer commands and 100 common words) and 300 uncommon words. For each word, 7 utterances are provided. They claim that each speaker is recorded utilizing a total of 455 different words. However, upon closer analysis, it was noticed that some uncommon words were repeated among the selected speakers. Specifically, words like ‘Moustache’, ‘Choking’, ‘Watch’, ‘Vouchsafe’ and ‘Powwow’ were repeated twice each. To avoid redundancy, only one instance of these words were considered. For each word, we took 7 utterances, resulting in a total of 7*450=3150 utterances (recordings) per speaker. Throughout the experiment and training phases, exclusively dysarthric utterances were utilized, even though controlled speech data was available for each speaker.

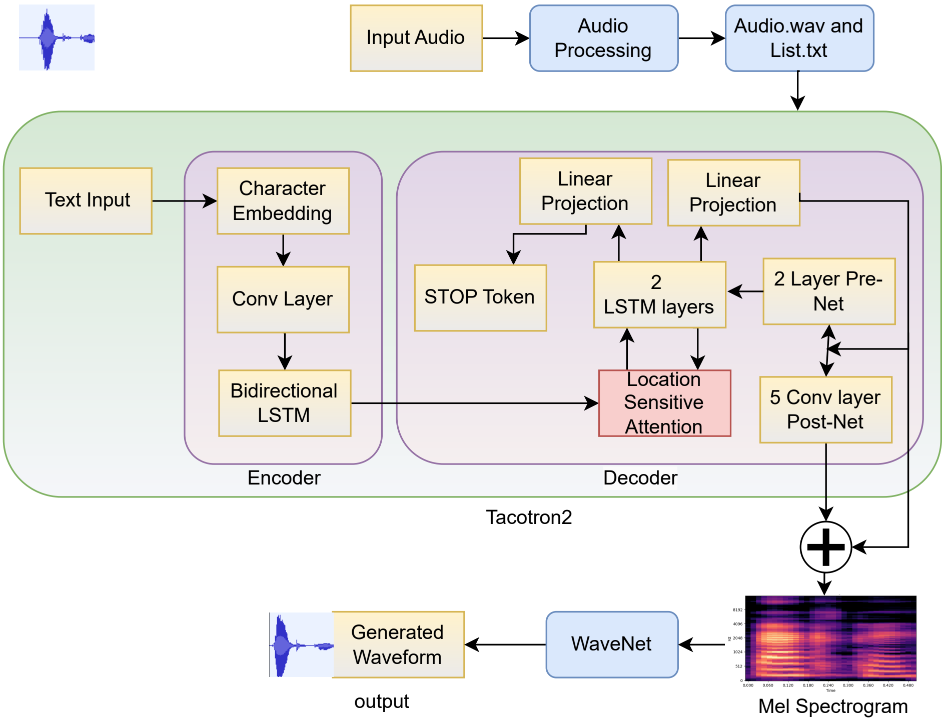


Figure 6.2: Proposed architecture

6.5 Proposed method

Due to the distinctive speaking style of individuals with articulation disorders, it is crucial to develop TTS systems customized to the specific requirements of dysarthric individuals. In the proposed method, we train the model separately for each individual, which allows them to learn their specific speech patterns and characteristics. Moreover, we also generate speech for OOV words exhibiting the same sound as the speaker characterization exhibits. We made a few changes to the Tacotron2 model to accommodate the individual characteristics of each speaker.

6.5.1 Pre-processing

The Tacotron2 model is initially trained on the LJSpeech dataset [154]. Before using dysarthric speech data with the model, we followed several steps to prepare it. First, we renamed all the audio files in sequence from 1 to the number of audio files present in each folder and we simultaneously created a list.txt file that matched each audio file to its transcribed text. After normalizing the audio files to a sample rate of 22050, 16-bit depth and mono channel, we updated the metadata for each audio file. 22050 Hz sample rate is a practical compromise that provides sufficient quality for speech while keeping file sizes and processing requirements manageable and it aligns with established standards in audio processing.

6.5.2 Tacotron2

For dysarthric speech synthesis, our baseline method is Tacotron2 [133] which is a modified version of [154]. Tacotron2 is a sequence-to-sequence neural network-based model that maps character embedding to Mel spectrograms. In sequence to sequence model, input sequence $(x_1, x_2 ... x_r)$ is converted to an output sequence $(y_1, y_2 ... y_r)$ where each y_t is determined by $y_1, y_2 ... y_{t-1}$. These two sequences may be of same or different length. The network basically consists of three parts: encoder, decoder with attention and WaveNet vocoder. The encoder changes a sequence of letters into a 512-dimensional representation by taking the character embedding as input and process them through a series of 3 convolution layers. After this convolutional process, the hidden representations are passed through a bidirectional Long Short-Term Memory (LSTM) layer to generate encoded features. The decoder functions as an autoregressive recurrent neural network equipped with a location-sensitive attention network. Its primary task is to predict the Mel spectrogram based from the encoded input sequence. Location-sensitive attention helps the model to move smoothly through the input data, avoiding situations where certain parts might be repeated or overlooked by the decoder. In severe cases of dysarthria there might be a lot of pauses, stops and repetitions which is handled by this module. The attention probabilities are calculated and passed through 2 LSTM layers. Along with this probability, the prediction made in the previous time step by a 2-layer pre-net are concatenated and output is projected through a linear transform to make predictions for the target spectrogram frame.

Intelligibility	High	Mid	low
Male	M08 93%	M05 58%	M07 28%
Female	F05 95%	F04 62%	F02 29%

Table 6.1: Speakers and their intelligibility rates from UASpeech database

Finally, the Mel spectrogram that was predicted goes through a special 5-layer convolutional post-net. This post-net predicts a kind of extra information that is added to the prediction. This helps in making the final result better. Then, we use a modified version of WaveNet to turn the Mel spectrogram into the actual sound waveform that we can hear. The model architecture is presented in Figure 6.2 and parameters are listed in Table 6.3.

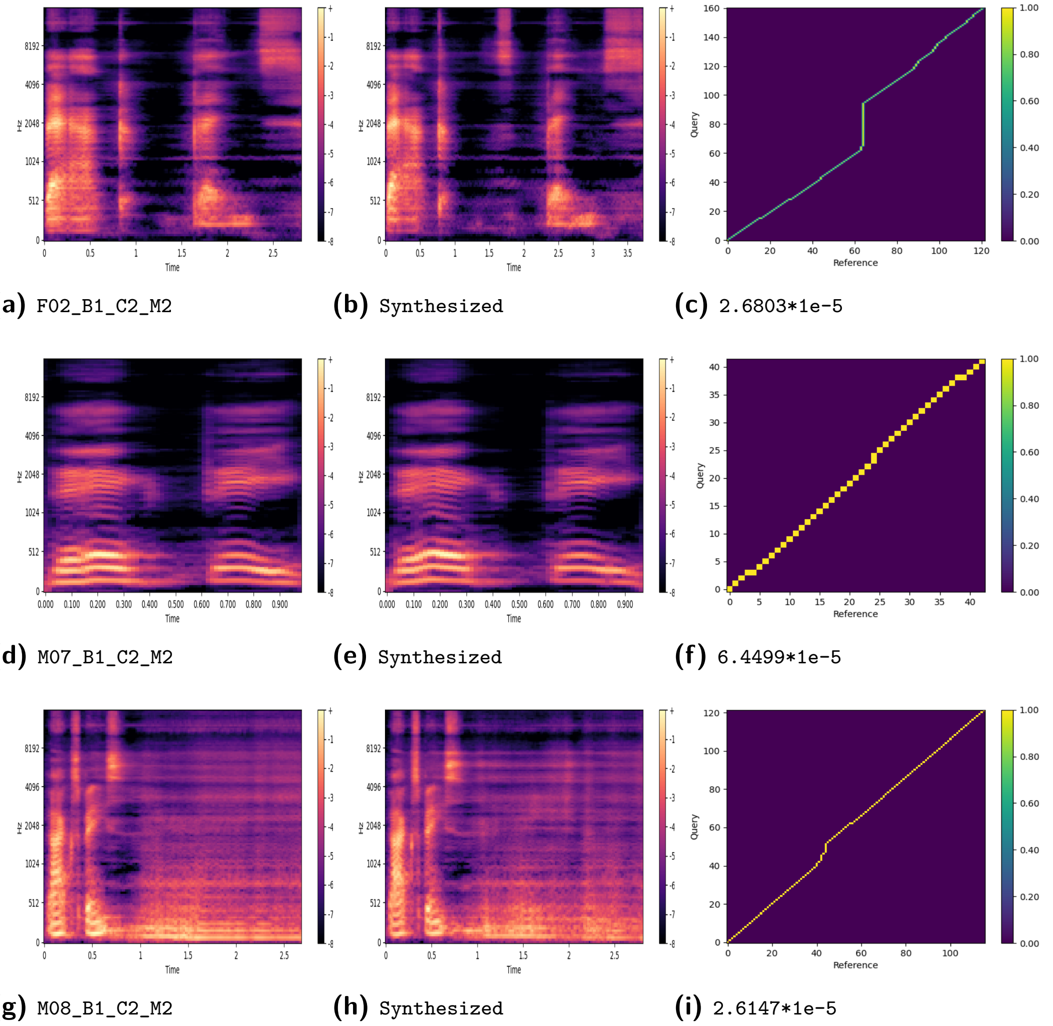
6.6 Evaluations and discussion

To incorporate synthetic data into subsequent experiments, it is necessary to verify their correctness and accuracy. First, we employed perceptual evaluation methods, particularly the Mean Opinion Score (MOS), which has been a superior technique for evaluating synthesized speech quality for many years. Second, we substantiated the correctness of our generated speech through a mathematical approach: Dynamic Time Warping (DTW). This comprehensive approach, including both perceptual and mathematical evaluations, offers a robust validation mechanism for our synthesized speech, lending credibility to its accuracy and suitability for subsequent experimentation.

We evaluated the model using words present in the UASpeech corpus and OOV words, which are words that have never been spoken by any dysarthric patient. The outcomes of these tests were outstanding. Testing the model on the words that have already been spoken by dysarthric patients allowed us to generate similarity graphs between the original and synthesized speech but in the case of OOV words we did not have an original audio file to compare against, we conducted MOS tests instead.

6.6.1 Dynamic Time Warping

Dynamic Time Warping (DTW) is a traditional method used to find the similarity between the source and target spectrograms [167]. The sequences are warped in a nonlinear way in the time dimension to figure out how similar they are. We apply DTW on the spectrogram of an original audio file and synthesized audio file. For most of the words, the matrix aligned perfectly, but we also encountered cases where the matrix alignment wasn't ideal, especially for a speaker with low intelligibility. Figure 6.3 shows the DTW alignment matrix for three speakers. In the DTW matrix the reference serves as the foundational template for comparison and query is the one being contrasted against the reference. The horizontal and vertical axis corresponds to the frames of the reference spectrogram and query spectrogram respectively.



(a) F02_B1_C2_M2 **(b)** Synthesized **(c)** 2.6803*1e-5

(d) M07_B1_C2_M2 **(e)** Synthesized **(f)** 6.4499*1e-5

(g) M08_B1_C2_M2 **(h)** Synthesized **(i)** 2.6147*1e-5

Figure 6.3: Original and generated spectrogram for word "Backspace" with their DTW similarity score (In row) for each speaker

6.6.2 Mean Opinion Score

The effectiveness of text-to-speech systems relies heavily on reliable and valid listening tests. Mahesh et al. briefly explored the MOS scale and made adjustments based on the speech's naturalness, intelligibility and listening effort [151]. The MOS test conducted for this experiment is an attempt to find the degree of perceptible speech

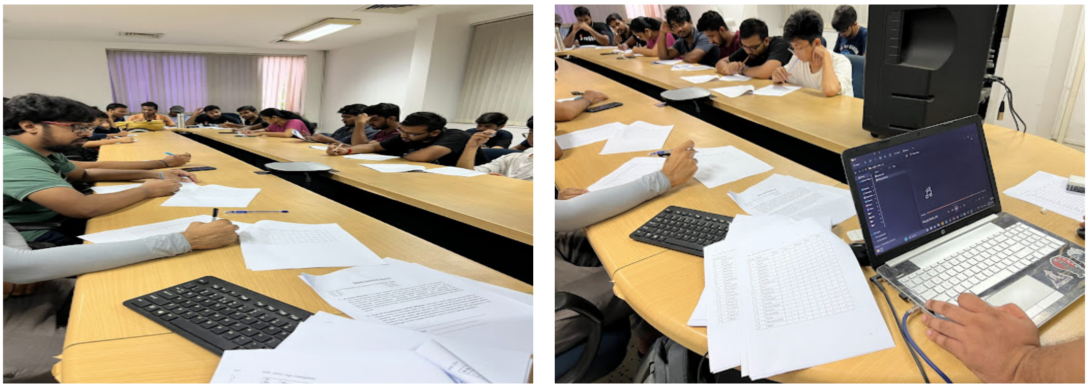


Figure 6.4: A MOS test has been conducted in the CSE department with 20 participants, each of them has provided their individual feedback in a designated booklet

deviation of synthesized dysarthric words compared to original words. MOS scores reflect the perceptual similarity of the synthesized dysarthric words after listening to the corresponding uttered words. Traditional MOS measures the quality of speech and in this case, actually MOS is used to compare the speech quality to the original recording's quality. We termed this test as Dysarthric Similarity MOS (DS-MOS) test. We conduct DS-MOS test for each speaker among 20 listeners on a 5-point scale ranging from 1 being the lowest similarity to 5 being the highest.

6.6.3 Known words

In the context of known words, achieving similarity in pronunciation and intelligibility took precedence over the naturalness factors because the speaker we selected for the experiment covered a range of intelligibility levels, from low to high. Our objective is to increase the utterance of existing words spoken by dysarthric patients while preserving their unique properties intact. We chose 20 words from the UASpeech database, each paired with its original .wav file and a synthesized .wav file. The listeners then listened to both files sequentially and provided their feedback on how similar they found them to be. For a different speaker, we repeated the process with another set of 20 words, using the same group of listeners. This procedure was consistently followed for each subsequent speaker. During our DS-MOS tests, we have observed that listeners often find it challenging to distinguish between machine-generated speech and speech produced by the patients themselves. The overall performance evaluation was determined by calculating the mean values. The synthesized speech achieved a similar quality and aligned well with the speaker's characteristics. For M08 and F04 we receive the best MOS ratings of 4.575 ± 0.039 and 4.347 ± 0.052 even though F04 belongs to mid intelligibility rate demonstrating a comparable performance to professionally recorded speech, which achieves a MOS of 4.58 [133]. Comparing our results with a similar study [89] we conclude that our approach demonstrates significant improvements across all intelligibility rates, irrespective of the fact that they reported speech synthesis for a single dysarthric speaker without specifying their intelligibility rates.

6.6.4 Out-of-Vocabulary words

In the OOV test, we chose a consistent set of 11 OOV words (commonly used in lab) that remain unchanged across all speakers. The audience initially listened to 20 words from the database to familiarize themselves with the speaker's attributes. Subsequently, they evaluate the synthesized voice based on the speaker's speaking rate, behavior and characteristics. This evaluation was done on a scale from 1 to 5, where a score of 1 indicated that the spoken word didn't seem like it came from that speaker at all and a score of 5 meant it matched the speaker's characteristics very closely. The results were strikingly impressive. For a few words, it posed a considerable challenge for us to distinguish that the generated voice was artificial. The synthesized speech remarkably resembled the way the person themselves would utter those OOV words. In Table 6.2,

Speaker	MOS		
	UASpeech	OOV	WSOLA
M07	4.062 ± 0.053	3.745±0.081	3.772±0.040
F02	3.750±0.059	3.656±0.069	3.609±0.092
M05	4.225±0.090	3.690±0.006	3.731±0.069
F04	**4.347±0.052**	3.690±0.053	3.727±0.007
M08	**4.575±0.039**	**4.045±0.048**	**4.077±0.017**
F05	3.540±0.030	3.013±0.029	3.104±0.083

Table 6.2: Result of MOS test with 20 listening subjects

third column presents OOV results for each speaker. Notably, for speaker M08, the highest score obtained was 4.045 ± 0.048.

6.6.5 WSOLA

In addition, the MOS test was conducted on the generated OOV words passed through the waveform similarity overlap-and-add technique (WSOLA) for alpha value 1. WSOLA is a digital audio effect that alters the duration of an audio signal while preserving its original qualities [56]. The aim behind this conversion was to stretch the speaking rate of the generated dysarthric speech to reduce the difficulty of evaluation since each speaker exhibits different speaking rates and we didn't have the same word to do a comparison with. MOS test result on this stretched audio file is given in Table 6.2. Throughout the MOS evaluation, 20 Indian speakers with English language speaking skills participated in listening to 360 sources and 252 generated audio segments, which ensured its reliability and rigorousness.

Category	**Parameter**	**Value**
Audio Processing	hop_length	200
	win_length	800
	n_fft	800
Encoder	encoder_kernel_size	5
	encoder_embedding_dim	512
	encoder_n_convolutions	3
Decoder	n_frames_per_step	1
	prenet_dim	256
	decoder_rnn_dim	1024
	max_decoder_steps	1300
	p_decoder_dropout	0.1
	p_attention_dropout	0.1
	gate_threshold	0.5
Attention	attention_dim	128
	attention_rnn_dim	1024
Location Layer	attention_location_kernel_size	31
	attention_location_n_filters	32
Mel-post Processing	postnet_embedding_dim	512
	postnet_n_convolutions	5
	postnet_kernel_size	5
Optimization	weight_decay	1e-6
	learning_rate	1e-5
	batch_size	1
	grad_clip_thresh	1.0
	mask_padding	True
	use_saved_learning_rate	False
	epochs	450

Table 6.3: Parameters list

6.7 Conclusion

In this contribution, we utilized the Tacotron2 model for synthesizing dysarthric speech to expand both the number of utterances and vocabulary size and succeeded in this endeavor. Tacotron 2 effectively utilizes UASpeech data to produce high-quality speech synthesis. This is definitely beneficial for dysarthric speech synthesis, where obtaining large amounts of data is in demand. We evaluate its effectiveness using DTW and MOS tests achieving satisfactory MOS ratings for all speakers. The speech synthesis system is tailored specifically to address the unique requirements of individuals with dysarthria. Overall results demonstrate that it can be used in further experiments to enhance the database size, which is a good idea followed by a dysarthric speech recognition system.

The only way to do great work is to love what you do.

Steve Jobs, Apple Inc

7

Intelligibility assessment of dysarthric speech based on GoP

7.1 Introduction

Intelligibility assessment of dysarthric speech is an essential step in understanding the patient's severity as it forms the basis for all subsequent medication and speech therapy. Early checkups with appropriate treatment and regular exercises significantly reduce the severity of dysarthria. There are two primary methods for automatic dysarthric speech assessment. The first method looks into hand-crafted speech features, such as voice quality, articulation characteristics, prosody and their combinations to capture the nuances of dysarthric speech [159][48]. The second method employs neural networks, that focus on raw speech data to predict outcomes. However, the black-box nature of neural networks can overlook features that may be valuable to speech pathologists. Hence it is essential to study the behavior of patients at the phoneme level to know where the exact problem lies instead of generalization of all features. In the traditional subjective method, speech intelligibility is assessed perceptually using articulation tests and manual acoustic analysis, which are costly, laborious and subject to many listener biases [59]. It also highly depends on the previous experience and expectations of the speech pathologist and their perceptual skills, which can lead to disparity among different speech pathologists. Due to the shortcomings of traditional evaluation methods, there is a need for automatic assessment techniques to evaluate dysarthric speech intelligibility. This study focuses on assessing the intelligibility of dysarthric speech by conducting an in-depth analysis at the phoneme level.

The motivation for this study comes from the fact that it has been seen after a subsequent number of perception tests and listening to the audio files that certain phonemes and sounds were consistently mispronounced by specific dysarthric speakers. This repeated mispronunciation pattern leads to the point that if we can isolate and figure out those mispronounced phonemes, we can be able to find out the actual pronunciation problem which is again specific to a specific speaker. To address this, the focus shifted to an approach that leverages the Goodness of Pronunciation (GoP) algorithm, offering a systematic way to evaluate the pronunciation quality of each phoneme. In this study, we explore the Goodness of Pronunciation (GoP) algorithm for dysarthric speech, which analyzes which phonemes are distorted and up to what extent each phoneme is atypical. Dysarthric speech often exhibits distorted pronunciation at the phoneme level, such as substitutions, omissions or distortions of specific sounds. By assessing each phoneme, the GoP algorithm identifies which sounds are most affected by the speakers. In Fig 7.1, spectrograms of “Coil” and “Foil” appear similar, but in actuality, there are minor differences that completely change their meaning. One distorted phoneme in a word is sufficient to decrease the intelligibility rate for dysarthric speakers. The GoP scores highlight specific patterns of mispronunciation, such as errors in vowels or consonants. This information supports depth analysis and helps to address the severity of dysarthria, aiding in rehabilitation. Studying at the phoneme level is essential for dysarthric speech.

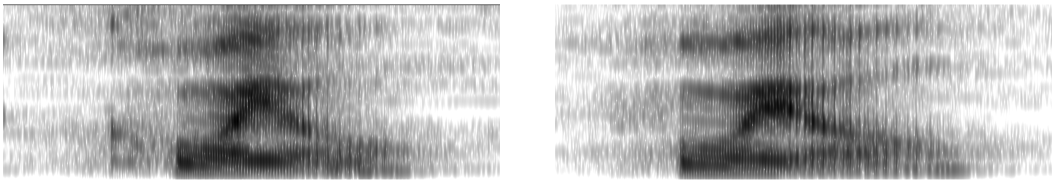


Figure 7.1: Spectrogram of word ‘Coil’ and ‘Foil’ respectively from same speaker M07 having 28% speech intelligibility

This approach does not require a huge amount of parallel datasets for training, having a positive aspect for the assessment of dysarthric speech, as data scarcity and speaker variability cause a significant challenge in this field [6][148][24]. Although GoP is frequently used to evaluate non-native speech pronunciation [28][69][74][97], some research has confirmed that it may be useful in evaluating speech disorders too [107][108][132]. We utilized the GoP algorithm and calculated the final GoP score using both GMM-GoP and NN-GoP rather than the baseline GoP algorithm. We then correlated the obtained scores with dysarthria severity using Kendall’s coefficient (τ) and found a weak positive correlation. Each phoneme’s score was analyzed, revealing significant insights at the phoneme level. The phoneme scores clearly indicate which phonemes are most distorted, highlighting those that should be specifically targeted while treating dysarthria.

7.2 Background

Earlier studies attempt to combine dysarthric speech assessment and the utility of computer-based speech recognition in a single model [119][143]. These studies emphasize feature representation and prediction using a phonologically structured sparse linear model [79][64][75]. In several studies, researchers extract different types of feature vectors from the raw signal, like prosodic features, acoustic and lexical features, vowel articulation features, audio descriptors and multi-tapered spectral estimation and classify them using genetic algorithms or Artificial Neural Networks (ANN) [153][8][109]. A language-independent approach was also investigated for automatic corner vowel detection, utilizing a language-universal phoneme recognizer, followed by statistical analysis of the formant data. Speaker-dependent and speaker-independent intelligibility assessment system based on DeepSpeech posteriors was also explored with the SVM classifier [85][149][47]. A recent study enhanced the GoP score using Uncertainty Quantification (UQ), achieving the best performance with the prior-normalized maxlogit GoP. The experiment was conducted across English, Korean and Tamil languages [162]. This paper inspired us to conduct a detailed analysis of dysarthric speech at the phoneme level.

7.3 Goodness of Pronunciation

The Goodness of Pronunciation algorithm uses a probabilistic approximation to calculate likelihood ratio between canonical and spoken phonemes. The purpose of the GoP measure is to generate a score for each phoneme within an utterance [134]. Since we are using this algorithm for dysarthric speech where each speaker possesses different

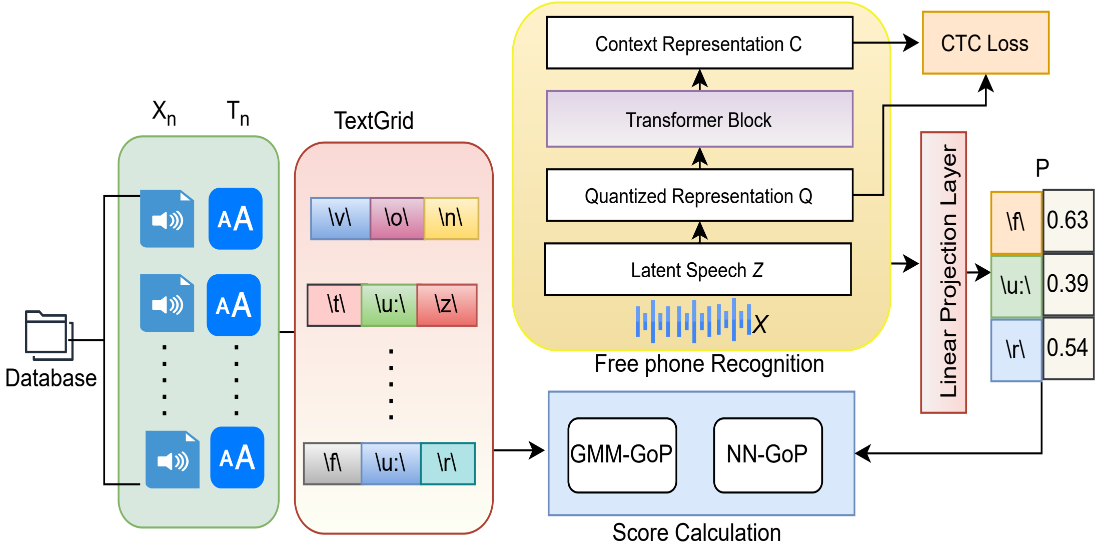


Figure 7.2: Architecture of the proposed framework

characteristics we define GoP as the degree of similarity between produced and correct pronunciation of phonemes in terms of smoothness and the ease with which the pronunciation can be understood. For each phone q with its corresponding acoustic segment $O^{(q)}$, there is a set of HMMs available to determine the likelihood that is $P(O^{(q)}|q)$. The quality of pronunciation of any phoneme p is defined as the log of the posterior probability given the corresponding acoustic segment $O^{(p)}$. That is

$$GoP(p) = \left|log(P(p|O^{(p)}))\right|/N_F(p) \tag{7.1}$$

$$GoP(p) = \left|log\left[\frac{(P(O^{(p)}|p)P(p)}{\sum_{q\in Q} P(O^{(p)}|q)P(q))}\right]\right|/N_F(p) \tag{7.2}$$

where $P(p)$ is the prior probability of phoneme p, Q is set of all phonemes and $N_F(p)$ is the number of frames in the acoustic segment $O^{(p)}$. The GoP algorithm has three main components: 1) Forced phone alignment phase, 2) Free phone recognition phase and

3) score is calculated as the difference between the log-likelihoods of the two preceding phases for each forced aligned phone. A higher absolute score signifies a larger difference between actual pronunciation and its expected model. The architecture of the proposed work is shown in Figure 7.2. In the subsequent subsection, each component of GoP is explained.

7.3.1 Forced phone alignment phase

During this phase, the system is directed to match the spoken words with their expected sounds, ensuring that the timing and pronunciation align correctly. It uses a pronunciation dictionary to break words down into phones and aligns them with specific moments in the recording. Basically, it matches orthographic transcriptions to spoken audio at the phoneme level. Initially, we had raw speech data with a list of words that were being spoken in audio files. We used the first 100 words from the UASpeech database, which is enough to cover all the phonemes. The selection of words is same for all speakers. For each audio file, its corresponding text grid file is generated using Montreal Forced Aligner (MFA) with International Phonetic Alphabet (IPA) notation[*]. MFA is designed to work with the Kaldi speech recognition toolkit, which is widely used in the research community for various speech-related tasks. A TextGrid file is generated for each word that contains start and end times for each phoneme with "Min" and "Max" values. These values corresponding to each phoneme represent the actual start and end time of occurrence of that event within the acoustic signal. During the forced alignment phase, the phoneme boundaries from the TextGrid are used to calculate the likelihood of each phoneme's acoustic features which is critical for accurate GoP computation.

We have done the additional task of manually verifying the phoneme boundaries to en-

[*] https://en.wikipedia.org/wiki/International_Phonetic_Alphabet_chart

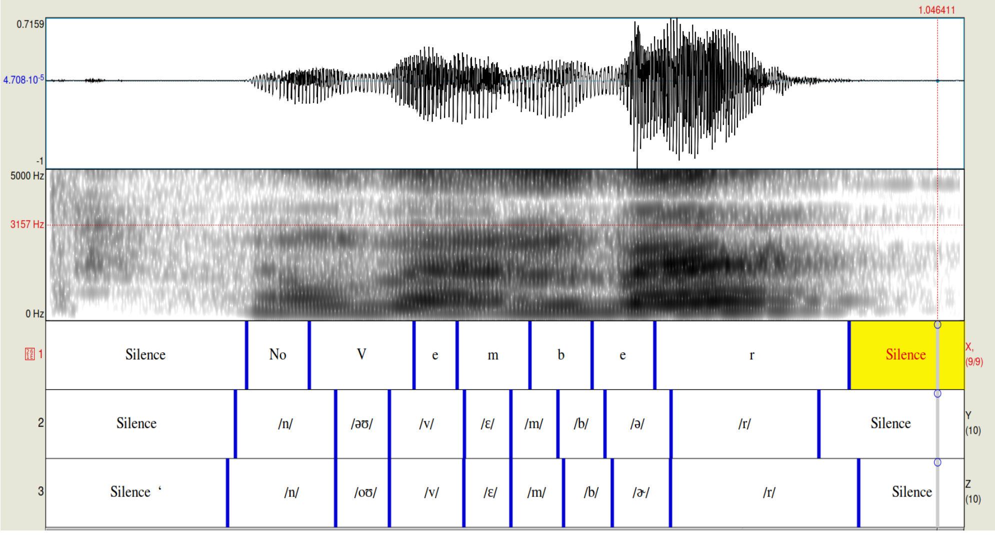

1.046411
0.7159
4.708·10⁻⁵
-1
5000 Hz
3157 Hz
0 Hz
1 Silence No V e m b e r Silence
2 Silence /n/ /əʊ/ /v/ /ɛ/ /m/ /b/ /ə/ /r/ Silence
3 Silence ‘ /n/ /oʊ/ /v/ /ɛ/ /m/ /b/ /ɚ/ /r/ Silence
X, (9/9)
Y (10)
Z (10)

Figure 7.3: Different tiers of annotation for the word "November"

sure that they are accurate reducing the likelihood of errors, as this phase lays the groundwork for subsequent analyses and evaluations of pronunciation quality. TextGrid files are organized into tiers, each representing a distinct layer of annotation. The TextGrid file consists of three tiers and the data from the third tier—annotated by an expert—is used to extract the "Min" and "Max" values for further experimental settings. The reason behind the three tiers is to represent the overlapping features of phonemes in dysarthric speech. As shown in Figure 7.3, word "November" is annotated in three stages. Tier 1 is by naive listener, tier 2 by MFA and tier 3 by a speech expert. In this way, we create a Textgrid file T for each audio file X.

7.3.2 Free phone recognition phase

This phase recognizes the phoneme sequence of the input speech signal without using any reference transcription and calculates log probability. It tries to determine which phoneme sequence was actually spoken, based purely on the acoustic features. For data preparation, the TextGrid file is first converted into a structured CSV format, containing

Table 7.1: Database description

Categories	**Speaker**	**Intelligibility %**
Very Low	M01, M04, F03	(0-25)
Low	M06, M07, F02, M16	(25-50)
Mid	M05, M11, F04	(50-75)
High	M08, M09, F05	(75-100)

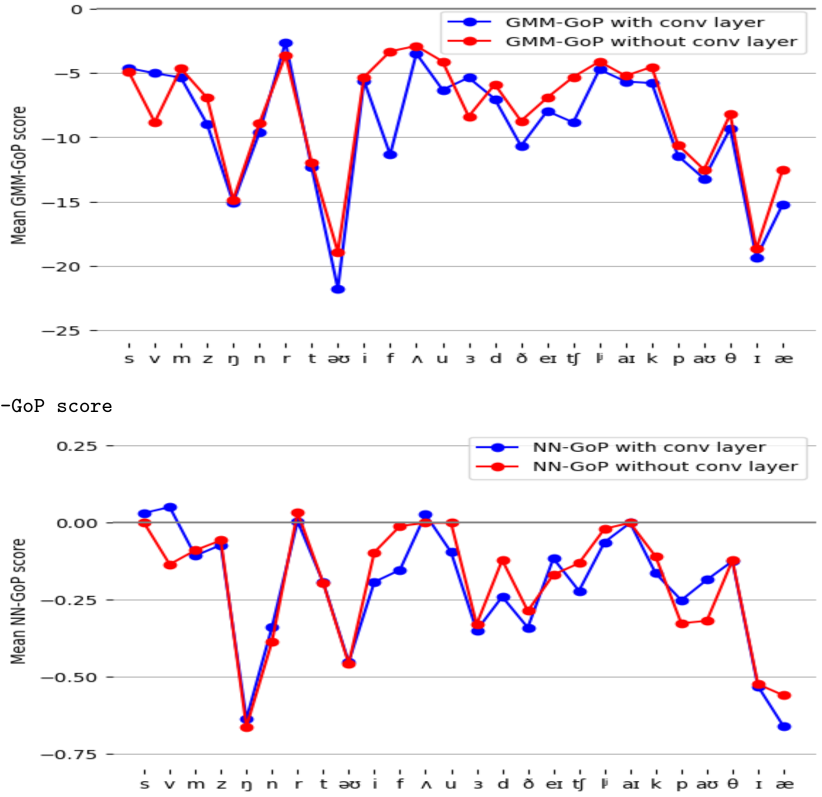
phone labels and the corresponding phoneme boundaries for each data file. For phoneme representation learning, Wav2Vec2 model architecture is utilized, which is trained to operate directly on raw waveform data instead of using traditional models like MFCC or spectrograms [161]. In this experiment, we select the first 100 words (out of 455) from 13 speakers, ensuring comprehensive coverage of all phonemes. Each word has 7 utterances available, so we used a total of $(7 * 100 * 13) * 2 = 18,200$ audio files for training. The model is trained on both healthy and dysarthric speech data so that the model learns to extract useful features from the audio by predicting masked portions of the input. Selected speaker with their intelligibility rate is shown in Table 7.1.

A stack of convolutional layers takes raw audio X and converts it to latent speech representation $Z_1, Z_2, ..., Z_T$ for T time steps. This process transforms the raw audio into a sequence of latent feature vectors where each vector corresponds to a small segment of the original audio. They are then fed to the transformer to build representation $C_1, C_2, ..., C_T$ to capture contextual information and dependencies over the entire sequence. The output of the multi-layer convolutional feature encoder is discretized to q_t with a quantization module Q to represent the targets. In the free phone recognition phase quantization module converts continuous acoustic representations into discrete states. This is particularly useful for aligning the observed acoustic features of speech with predefined phoneme models.

In addition, Linear projection layer is used for dimensionality reduction before passing it to the downstream classifiers. The linear projection layer maps high-dimensional latent feature embeddings into a lower-dimensional space. In GoP this layer produces accurate phoneme probabilities for GoP score computation. We fine-tuned the model according to dysarthric speech and its labeled data, followed by adding a classifier at the top of the Wav2Vec2.0 model. This layer maps the contextual representations to phoneme labels. We use the standard train and test split and follow the standard protocol of collapsing phone labels to 26 classes. The model predicts logits for the input audio features and the loss is calculated using the CTC loss function. The model is 82.3% accurate to predict phonemes accurately. The model is trained on both healthy and dysarthric speech data so that the model learns to extract useful features from the audio by predicting masked portions of the input. In addition, we implement a variant of the above-explained architecture where feature embedding obtained from Wav2Vec2.0 is passed through the linear head for classification instead of the linear projection layer and logits are computed using the head layer, achieving 81.5% phoneme accuracy.

7.3.3 Score computation

We measure intelligibility scores for each phoneme using GMM-GoP (Gaussian Mixture Model-GoP) and NN-GoP (Neural Network-GoP). The mean GoP score for each phoneme across all speakers is shown in Figure 7.4. Due to the significant variation in speaker characteristics in the UASpeech database, we divided it into two groups based on intelligibility. We combined low and very low-intelligible speakers into one category, while the remaining speakers with mid and high intelligibility were placed in another one. GMM-GoP is effective for handling low-intelligibility speakers as it captures broad statistical patterns, whereas NN-GoP offers better flexibility and adapts to variability in speech, especially for mid-to-high intelligibility speakers, while capturing finer phonetic



(a) `GMM-GoP score`

(b) `NN-GoP score`

Figure 7.4: GoP score for each phoneme, where larger difference indicates potential mispronunciation

details. Combining GMM-GoP and NN-GoP leverages the strengths of both approaches and helps to address the full spectrum of dysarthric speech to better evaluate GoP.

The GMM-GoP score is calculated by comparing the log-likelihoods of phoneme sequences obtained from the forced alignment phase and the free phone recognition phase. This comparison reflects the model's ability to detect phonetic variations present in dysarthric speech, indicating how well the pronunciation matches the expected phoneme sequence. This method relies on the statistical model that represents phonetic units. We evaluated speakers with low and very low intelligibility with this method and found that mostly all scores were negative due to major severity. In contrast, NN-GoP computes the score based on neural network outputs, using softmax probabilities. NN-GoP is

Table 7.2: Most distorted phonemes according to the obtained score

Method	Speaker	Phoneme	Score
GMM-GoP	M01, M04, F03, M06, M07, F02, M16	əʊ	-18.9067
		ɪ	-18.6762
		ŋ	-14.8451
		aʊ	-12.5230
		æ	-12.5273
		t	-11.9644
		p	-10.5817
NN-GoP	M05, M11, F04, M08, M09, F05	ŋ	-0.6610
		æ	-0.6601
		ɪ	-0.5302
		əʊ	-0.4566
		n	-0.3850
		p	-0.3275
		əʊ	-0.3174

more flexible in handling variability in speech, making it better suited for evaluating dysarthric speakers with higher intelligibility. We evaluate mid and high-intelligibility speakers with NN-GoP.

In Figures 7.4a and 7.4b, the scores highlight distinct trends based on speaker intelligibility. Figure 7.4a, which represents results for speakers with low to mid intelligibility, shows more extreme scores due to higher misclassification rates and poorer recognition. In contrast, Figure 7.4b corresponds to speakers with mid to high intelligibility and exhibits a narrower range of scores, with a mix of positive and negative values. Most phoneme scores in Figure 7.4b are close to zero, indicating relatively better pronunciation. Overall, speakers with greater variability in their speech tend to score lower across both figures. In both cases, we identify a set of phonemes that dysarthric speakers find most challenging. For GMM-GoP we find /əʊ/, /ɪ/ and /ŋ/ with higher score and for NN-GoP /ŋ/, /æ/ and /ɪ/ as most distorted one. Table 7.2 provides details on the distorted phonemes along with their GoP scores.

7.4 Results and discussion

To make sure that the obtained phoneme score is associated well with dysarthric speech intelligibility we compute the correlation between the obtained GoP score and dysarthric speech severity rate using Kendall's rank correlation coefficient (τ). Both GMM-GoP and NN-GoP systems show positive correlations between GoP scores and dysarthria severity.

- τ(GMM-GoP) = 0.4558
- τ(NN-GoP) = 0.4142

This correlation reflects that the speakers with more severe dysarthria tend to have lower GoP scores (indicating greater pronunciation difficulties). Interestingly, both GMM-GoP and NN-GoP, despite their different underlying features, produced very similar (τ) values. This suggests that both models demonstrate a comparable level of correlation between dysarthria severity and GoP scores. This consistency strengthens the reliability of the observed relationship. We found that certain phonemes have more impact on the severity level of the speaker. A higher absolute value reflects more deviation from the actual pronunciation. If we know a few sets of phonemes for which the chances of deviation from the actual scenario are high, it would be valuable to identify the problem at ground level and clinicians can provide correct medical procedures without any delay. This will surely help patients as well to recover early. In future, we will implement an interface for the patient on which they can score themselves for each utterance; this will keep them engaged in a positive way and they will practice regularly with a target without several visits to the clinic.

7.5 Annotation of UASpeech database

The second part of this chapter focuses on the annotations of the UASpeech database. Annotating audio data is an essential step as it provides comprehensive and detailed information about each audio file. These annotations describe how the audio sounds and help us understand its quality. This information is critical for analyzing and improving speech recognition systems. While performing the annotation process, we noticed that the database contains certain files that are problematic. For instance, some audio files are entirely silent, contain no human sound, or are heavily impacted by noise. These kinds of files can significantly reduce the accuracy of any speech recognition system. Therefore, it becomes essential to identify such files, filter them out and exclude them from further analysis and experiments.

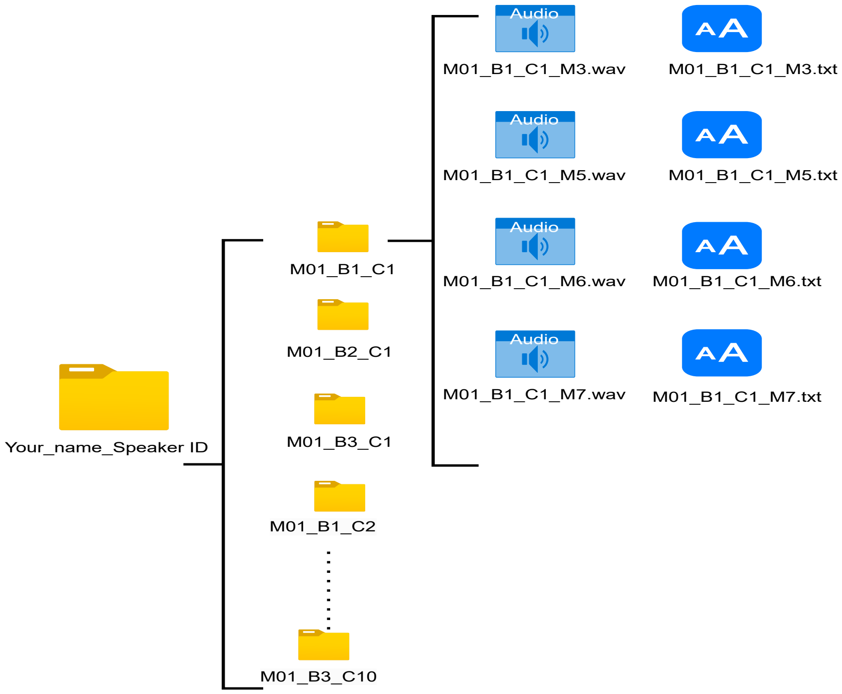


Figure 7.5: Folder structure for the annotation of UASpeech database

As of now, we have completed the annotation process for audio files from 10 speakers in the UASpeech database. The annotation for the remaining speakers is still in progress. Once we finish annotating the entire database, we plan to make the annotated dataset publicly available. This way, other researchers will also be able to download it and use it for their work, potentially contributing to advancements in the field. To ensure clarity and ease of use, we have organized the annotated database using a well-defined folder structure as shown in Figure 7.5. The organization is as follows:

- **Parent folders for each speaker**: Each speaker has a dedicated parent folder. The name of this folder corresponds to the speaker's ID as specified in the UASpeech database.

- **Subfolders for blocks and words**: Within each speaker's parent folder, there are multiple subfolders. These subfolders are categorized by blocks and word IDs, which further help in structuring the data.

- **Audio and annotation files**: Each of these subfolders contains all the utterances for a particular word in .wav format. Additionally, for every .wav file, there is an associated .txt file. This text file contains the annotations specific to that audio file, providing detailed insights about its content and quality.

This systematic organization makes it easier to navigate through the annotated database, ensuring that researchers can access the required files efficiently. The annotations, combined with this folder structure, aim to facilitate further research and experimentation in improving speech recognition systems.

While annotating the audio recordings of dysarthric speakers in the UASpeech database, we observed a significant pattern among speakers with low intelligibility rates. These speakers often struggle with producing clear and accurate pronunciations of the target

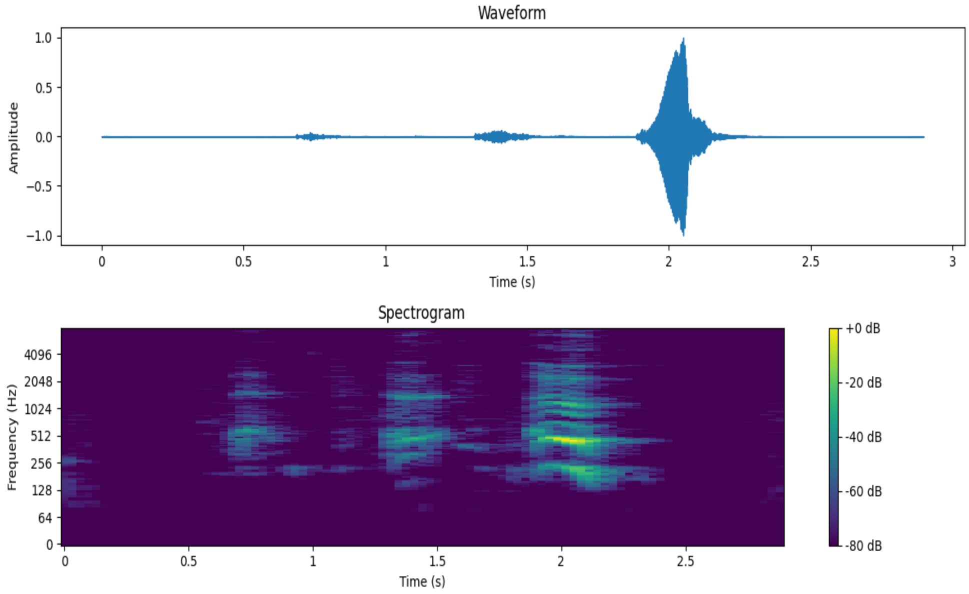


Figure 7.6: Waveform and spectrogram for the word 'Echo' reveal that the audio contains almost negligible sound

words. In many cases, a majority of the audio clips contain noticeable mispronunciations, where the intended word is either partially articulated, distorted, or replaced by unintelligible sounds. This observation highlights the challenges associated with dysarthric speech, where the physical and neurological impairments affect the speaker's ability to produce consistent and intelligible speech. As a result, the annotated data for these speakers often reflects frequent deviations from the expected pronunciation. For example, certain phonemes may be omitted, substituted, or slurred, making the speech harder to comprehend both for human listeners and for automated systems. In Figure 7.6, the waveform and spectrogram for the word echo are shown. This audio clip contains no sound at the beginning and towards the end, only the /oʊ/ sound is audible.

By thoroughly documenting the mispronunciations and other speech anomalies in the annotation process, we aim to create a dataset that accurately represents the challenges faced by dysarthric speakers, enabling more effective research and innovation in this field.

7.6 Conclusion

This study proposes a Goodness of Pronunciation algorithm for the assessment of dysarthric speech and identifies the most distorted phonemes in the UASPeech database. Despite several challenges associated with dysarthric speakers, we successfully annotated all audio files with accurate phoneme boundaries and developed a speaker-independent phoneme recognition system, followed by GoP score calculation. These findings will benefit both speech pathologists and patients as correct assessment is essential for recovery and if it fails at the initial stage, the entire process after that will be ineffective. With a GoP score, patients can practice particular sounds/words at home as well with accurate pronunciation references, ultimately improving their speech articulation.

8

Conclusion

This thesis explores the unique characteristics of dysarthric speech and investigates methods to address its complexities. Dysarthric speech presents several challenges, including variability in pronunciation, reduced intelligibility and limited data availability, making its analysis and processing a challenging task. These issues significantly hinder the development of effective automated systems personalized to individuals with dysarthria. Recognizing the societal importance of such systems, this thesis emphasizes the need for advanced solutions that can assist individuals suffering from dysarthria in communication and daily interactions. To address these challenges, this thesis employs and evaluates a variety of algorithms and feature representation techniques. The findings and approaches presented in this work contribute to the growing efforts to bridge the communication gap for individuals with dysarthria, reflecting the broader societal need for inclusive technological solutions. In this chapter, the summary of the contributions made in this thesis and future scope is presented.

8.1 Summary of thesis

Dysarthria is a neurological disorder that presents individuals with significant communication challenges in their daily lives. To alleviate these difficulties, this research focuses on developing an automated speech recognition system tailored to their needs, which holds great potential to improve their quality of life.

Initially, the research focused on analyzing, characterizing and recognizing dysarthric speech at the feature level. During this process, it became evident that detection and intelligibility assessment are equally crucial at this stage, as they form the foundation for recovery and further medical intervention for individuals with dysarthria. After exploring various methods, we found that AST is effective for handling the detection and assessment of dysarthric speech.

In a later chapter, we address the issue of data scarcity, a common challenge in dysarthric speech research. To overcome this, we generated synthetic dysarthric speech data that replicates the characteristics and behaviors of actual speakers, thereby increasing the volume of data available for training. Finally, we conducted an intelligibility assessment at the phoneme level using the GoP algorithm and provided detailed annotations for the UASpeech database. These contributions aim to improve the understanding, processing and recognition of dysarthric speech, advancing the development of automated systems for individuals with dysarthria.

This thesis is dedicated entirely to the study of dysarthric speech, driven by the singular objective of improving the lives of individuals with dysarthria. The goal is to empower them to lead dignified and independent lives while enabling them to contribute meaningfully to society. Every person's emotions, feelings and sense of autonomy are profoundly important and even small positive changes in the lives of dysarthric speakers can have a transformative impact. By addressing their communication challenges, this work aspires to make a lasting difference, fostering inclusion and respect for their abilities and potential.

8.2 Future scope of research

The work presented in the chapters of this thesis contributes broad scope and proclaims several directions for future research endeavors. This section discusses some of the potential directions for future extension of the thesis work.

Database collection and annotation: The accuracy of dysarthric speech recognition lags significantly behind that of normal speech recognition. A major contributing factor is the scarcity of data across all domains of pathological speech. With the limited data available, researchers must optimize and fine-tune models accordingly. Collecting pathological speech data poses challenges on multiple fronts, from accessibility to variability in speech patterns. Despite these obstacles, data collection remains the most effective way to expand the size and diversity of the database, which is crucial for improving the performance of speech recognition systems for individuals with dysarthria. Database annotation is equally important to establish an efficient dysarthric ASR system. Accurate and detailed annotations provide valuable insights into the unique characteristics of dysarthric speech, enabling models to better understand and process it. Together, data collection and proper annotation are fundamental to advancing the performance and reliability of speech recognition systems for individuals with dysarthria.

Automatic speech recognition with unseen data: The performance of dysarthric speech recognition on unseen data has been notably poor. In the future, our goal is to develop a model capable of operating in real-time, allowing new speakers to have their words recognized without the need for prior training on their specific speech patterns. Such a system would be highly beneficial in real-world applications. By enhancing the accuracy of recognition for unseen data, we can work towards bridging the gap between the performance of healthy and dysarthric speech recognition systems, making them more inclusive and effective.

Identifying the speaker among group conversation: Throughout this thesis, from the initial to the final contribution, the study, analysis and recognition of dysarthric speech have been conducted primarily on word-based datasets. The exploration of dysarthric speech at the conversational level was not pursued due to limitations in the available datasets. In the future, we aim to address this gap by collecting conversational speech involving two or more dysarthric speakers. This will enable us to identify speaker-specific contributions and delve deeper into each speaker's unique patterns and features. Such an approach will pave the way for a more advanced and comprehensive analysis of dysarthric speech in real-world conversational settings.

References

[1] (2017). Deep voice: Real-time neural text-to-speech, author=Arık, Sercan Ö and Chrzanowski, Mike and Coates, Adam and Diamos, Gregory and Gibiansky, Andrew and Kang, Yongguo and Li, Xian and Miller, John and Ng, Andrew and Raiman, Jonathan and others. In *International conference on machine learning*, pages 195–204. PMLR.

[2] Aihara, R., Takiguchi, T., and Ariki, Y. (2017). Phoneme-Discriminative Features for Dysarthric Speech Conversion. In *Interspeech*, pages 3374–3378.

[3] Al-Qatab, B. A. and Mustafa, M. B. (2021). Classification of Dysarthric Speech According to the Severity of Impairment: an Analysis of Acoustic Features. *IEEE Access*, 9:18183–18194.

[4] An, K., Kim, M., Teplansky, K., Green, J., Campbell, T., Yunusova, Y., Heitzman, D., and Wang, J. (2018). Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks.

[5] Baylor, C., Yorkston, K., Strand, E., Eadie, T., Duffy, J., Beukelman, D., Golper, L., Miller, R., and Spencer, K. (2005). Committee for Development of Practice Guidelines in Dysarthria.

[6] Bharti, K., Haque, S., and Das, P. K. (2024). A Novel Dysarthric Speech Synthesis system using Tacotron2 for specific and OOV words. In *2024 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5.

[7] Bhat, C. and Strik, H. (2020). Automatic Assessment of Sentence-Level Dysarthria Intelligibility Using BLSTM. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):322–330.

[8] Bhat, C., Vachhani, B., and 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)*, pages 5070–5074. IEEE.

[9] Blaney, B. and Hewlett, N. (2007). Dysarthria and Friedreich's ataxia: What can intelligibility assessment tell us? *International journal of language & communication disorders / Royal College of Speech & Language Therapists*, 42:19–37.

[10] Bunton, K., Kent, R., Duffy, J., Rosenbek, J., and Kent, J. (2008). Listener Agreement for Auditory-Perceptual Ratings of Dysarthria. *Journal of speech, language, and hearing research : JSLHR*, 50:1481–95.

[11] Celin, T. A. M., Nagarajan, T., and Vijayalakshmi, P. (2020). 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.

[12] Chandrashekar, H., Karjigi, V., and Sreedevi, N. (2020a). Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech. *Ieee transactions on neural systems and rehabilitation engineering*, 28(12):2880–2889.

[13] Chandrashekar, H. M., Karjigi, V., and Sreedevi, N. (2020b). Spectro-Temporal Representation of Speech for Intelligibility Assessment of Dysarthria. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):390–399.

[14] Chen, F. and Kostov, A. (1997). Optimization of dysarthric speech recognition. In *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 'Magnificent Milestones and Emerging Opportunities in Medical Engineering' (Cat. No.97CH36136)*, volume 4, pages 1436–1439 vol.4.

[15] Chen, Z., Ramabhadran, B., Biadsy, F., Zhang, X., Chen, Y., Jiang, L., Chu, F., Doshi, R., and Moreno, P. J. (2021). Conformer Parrotron: A Faster and Stronger End-to-End Speech Conversion and Recognition Model for Atypical Speech. In *Interspeech*, pages 4828–4832.

[16] Clark, H., Duffy, J., Strand, E., Hanley, H., and Solomon, N. (2022). Orofacial Muscle Strength across the Dysarthrias. *Brain Sciences*, 12:365.

[17] Clark, H., Duffy, J., Whitwell, J., Ahlskog, J., Sorenson, E., and Josephs, K. (2013). Clinical and imaging characterization of progressive spastic dysarthria. *European journal of neurology : the official journal of the European Federation of Neurological Societies*, 21.

[18] Coleman, C. and Meyers, L. (1991). Computer recognition of the speech of adults with cerebral palsy and dysarthria. *Augmentative and Alternative Communication*, 7(1):34–42.

[19] Connaghan, K. and Patel, R. (2017). The Impact of Contrastive Stress on Vowel Acoustics and Intelligibility in Dysarthria. *Journal of Speech, Language, and Hearing Research*, 60:1–13.

[20] Cote-Reschny, K. and Hodge, M. (2010). Listener Effort and Response Time When Transcribing Words Spoken by Children with Dysarthria. *Journal of Medical Speech-Language Pathology*, 18:24–34.

[21] Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2011). Large vocabulary continuous speech recognition with context-dependent DBN-HMMs. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4688–4691. IEEE.

[22] Dahmani, H., Selouani, S. A., O'shaughnessy, D., Chetouani, M., and Doghmane, N. (2013). Assessment of dysarthric speech through rhythm metrics. *Journal of King Saud University - Computer and Information Sciences*, 25:43–49.

[23] Darley, F. L., Aronson, A. E., and Brown, J. R. (1969). Clusters of deviant speech dimensions in the dysarthrias. *Journal of speech and hearing research*, 12(3):462–496.

[24] Darling-White, M. and Huber, J. (2020). The Impact of Parkinson's Disease on Breath Pauses and Their Relationship to Speech Impairment: A Longitudinal Study. *American Journal of Speech-Language Pathology*, 29:1–13.

[25] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

[26] Deller Jr, J., Hsu, D., and Ferrier, L. J. (1991). On the use of Hidden Markov Modelling for recognition of dysarthric speech. *Computer Methods and Programs in Biomedicine*, 35(2):125–139.

[27] Dietsch, A., Solomon, N., Sharkey, L., Duffy, J., Strand, E., and Clark, H. (2014). Perceptual and instrumental assessments of orofacial muscle tone in dysarthric and normal speakers. *The Journal of Rehabilitation Research and Development*, in press.

[28] Doremalen, J., Cucchiarini, C., and Strik, H. (2013). Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch. *The Journal of the Acoustical Society of America*, 134:1336–47.

[29] dos Santos Barreto, S. and Zazo Ortiz, K. (2016). Protocol for the Evaluation of Speech Intelligibility in Dysarthrias: Evidence of Reliability and Validity. *Folia Phoniatrica et Logopaedica*, 67(4):212–218.

[30] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[31] Doyle, P., Leeper, H., Kotler, A., Thomas-Stonell, N., O'Neill, C., Dylke, M., and Rolls, K. (1997). Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of rehabilitation research and development*, 34:309–16.

[32] Duffy, J. (2008). Motor Speech Disorders and the Diagnosis of Neurologic Disease: Still a Well-kept Secret? *ASHA Leader*, 13:10–11+12.

[33] Duffy, J. (2010). *Distinguishing among motor speech disorders is important: the role of speech pathology in neurologic diagnosis*, pages 271–282.

[34] Duffy, J. (2016). Motor Speech Disorders: Where Will We Be in 10 Years? *Seminars in speech and language*, 37:219–224.

[35] Duffy, J. and Kent, R. (2001). Darley"s contributions to the understanding, differential diagnosis, and scientific study of the dysarthrias. *Aphasiology*, 15:275–289.

[36] Duffy, J., Strand, E., and Josephs, K. (2014). Motor Speech Disorders Associated with Primary Progressive Aphasia. *Aphasiology*, 28:1004–1017.

[37] Enderby, P. (2014). Use of the extended therapy outcome measure for children with dysarthria. *International Journal of Speech-Language Pathology*, 16(4):436–444.

[38] Enderby, P. M. (1983). *Frenchay Dysarthria Assessment.* Pro-Ed.

[39] Espana-Bonet, C. and Fonollosa, J. A. (2016). Automatic speech recognition with deep neural networks for impaired speech. In *Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH 2016, Lisbon, Portugal, November 23-25, 2016, Proceedings 3*, pages 97–107. Springer.

[40] et al., S. (2014). A Multi-Views Multi-Learners Approach Towards Dysarthric Speech Recognition Using Multi-Nets Artificial Neural Networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(5):1053–1063.

[41] et al., Y. (2022). Multi-Modal Acoustic-Articulatory Feature Fusion For Dysarthric Speech Recognition. In *ICASSP 2022*, pages 7372–7376.

[42] FAAP, R. and Ruben, R. (2000). Redefining the Survival of the Fittest: Communication Disorders in the 21st Century. *The Laryngoscope*, 110:241 – 241.

[43] Fager, S. and Burnfield, J. (2015). Speech Recognition for Environmental Control: Effect of Microphone Type, Dysarthria and Severity on Recognition Results. *Assistive Technology*, 27:150626083720004.

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

[45] Farhadipour, A., Veisi, H., Asgari, M., and Keyvanrad, M. (2018). Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks. *ETRI Journal*, 40.

[46] Farrús, M., Hernando, J., and Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. pages 778–781.

[47] Fontan, L., Pellegrini, T., Olcoz, J., and Abad, A. (2015). Predicting disordered speech comprehensibility from Goodness of Pronunciation scores.

[48] Franciscatto, M., Fabro, M., Lima, J. C., Trois, C., Moro, A., Maran, V., and Keske-Soares, M. (2020). Towards a Speech Therapy Support System Based on Phonological Processes Early Detection. *Computer Speech Language*, 65:101130.

[49] Frey, B. J. and Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976.

[50] Ghio, A., Giusti, L., Blanc, E., and Pinto, S. (2020). French adaptation of the “Frenchay Dysarthria Assessment 2” speech intelligibility test. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, 137(2):111–116.

[51] Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J., and Zhou, Y. (2017). Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems*, 30.

[52] Gong, Y., Chung, Y.-A., and Glass, J. (2021). AST: Audio Spectrogram Transformer. pages 571–575.

[53] Gopalakrishnan, T., Imam, S. A., and Aggarwal, A. (2022). Fine Tuning and Comparing Tacotron 2, Deep Voice 3, and FastSpeech 2 TTS Models in a Low Resource Environment. In *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, pages 1–6.

[54] Graff-Radford, J., Jones, D., Strand, E., Duffy, J., and Josephs, K. (2013). Pure Apraxia of Speech from Stroke. pages S29–S29.

[55] Green, P. D., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M. S., and Parker, M. (2003). Automatic speech recognition with sparse training data for dysarthric speakers. In *Interspeech*, pages 1189–1192.

[56] Grofit, S. and Lavner, Y. (2008). Time-Scale Modification of Audio Signals Using Enhanced WSOLA With Management of Transients. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):106–115.

[57] Guerra, E. and Lovely, D. (2003). Suboptimal Classifier for Dysarthria Assessment. pages 314–321.

[58] Gupta, S., Patil, A. T., Purohit, M., Parmar, M., Patel, M., Patil, H. A., and Guido, R. C. (2021). Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, 139:105–117.

[59] Gurugubelli, K. and Vuppala, A. K. (2019). Perceptually Enhanced Single Frequency Filtering for Dysarthric Speech Detection and Intelligibility Assessment. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6410–6414.

[60] Hasegawa-Johnson, M., Gunderson, J., Perlman, A., and Huang, T. (2006). HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 3, pages III–III. IEEE.

[61] Hawley, M. S., Enderby, P., Green, P., Cunningham, S., Brownsell, S., Carmichael, J., Parker, M., Hatzis, A., O'Neill, P., and Palmer, R. (2007). A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics*, 29(5):586–593.

[62] Haykin, S. (1998). *Neural networks: a comprehensive foundation.* Prentice Hall PTR.

[63] Hernandez, A. and Chung, M. (2019). Dysarthria Classification Using Acoustic Properties of Fricatives.

[64] Hernandez, A., Lee, H.-Y., and Chung, M. (2019). Acoustic analysis of fricatives in dysarthric speakers with cerebral palsy. *Phonetics and Speech Sciences*, 11:23–29.

[65] Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6:107–116.

[66] Hosom, J.-P., Kain, A., Mishra, T., van Santen, J., Fried-Oken, M., and Staehely, J. (2003). Intelligibility of modifications to dysarthric speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

[67] Hustad, K. (2007). Effects of Speech Stimuli and Dysarthria Severity on Intelligibility Scores and Listener Confidence Ratings for Speakers with Cerebral Palsy. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, 59:306–17.

[68] Jayaram, G. and Abdelhamied, K. (1995). Experiments in dysarthric speech recognition using artificial neural networks. *Journal of rehabilitation research and development*, 32:162–162.

[69] Joshi, S. and Rao, P. (2013). Acoustic models for pronunciation assessment of vowels of Indian English. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–6.

[70] Joshy, A. A. and Rajan, R. (2021). Automated dysarthria severity classification using deep learning frameworks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 116–120. IEEE.

[71] Joshy, A. A. and Rajan, R. (2022). Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1147–1157.

[72] K, A., T N, L., Bhat, S. U., R, S., and H M, C. (2023). Automatic Early Detection of Dysarthria using Deep Neural Network. In *2023 International*

Conference on Smart Systems for applications in Electrical Sciences (ICSSES), pages 1–4.

[73] Kadi, K. L., Selouani, S. A., Boudraa, B., and Boudraa, M. (2016). Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. *Biocybernetics and Biomedical Engineering*, 36(1):233–247.

[74] Kanters, S., Cucchiarini, C., and Strik, H. (2009). The Goodness of Pronunciation algorithm: a detailed performance study. *Speech Communication.*

[75] Kent, R., Weismer, G., Kent, J., and Rosenbek, J. (1989). Toward Phonetic Intelligibility Testing in Dysarthria. *The Journal of speech and hearing disorders*, 54:482–99.

[76] Kent, R. D. and Rosenbek, J. C. (1983). Acoustic patterns of apraxia of speech. *Journal of Speech, Language, and Hearing Research*, 26(2):231–249.

[77] Kent, R. D., Weismer, G., Kent, J. F., Vorperian, H. K., and Duffy, J. R. (1999). Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of communication disorders*, 32(3):141–186.

[78] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J. R., Huang, T. S., Watkin, K. L., and Frame, S. (2008). Dysarthric speech database for universal access research. In *Interspeech*, volume 2008, pages 1741–1744.

[79] Kim, M. J., Kim, Y., and Kim, H. (2015). 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.

[80] Kim, S.-J., Kim, J.-J., and Hahn, M. (2006). Implementation and evaluation of an HMM-based Korean speech synthesis system. *IEICE transactions on information and systems*, 89(3):1116–1119.

[81] Kim, Y., Kent, R., Kent, J., and Duffy, J. (2010). Perceptual and Acoustic Features of Dysarthria in Multiple System Atrophy. *Journal of Medical Speech-Language Pathology*, 18:66–70.

[82] Kumbhar, H. S. and Bhandari, S. U. (2019). Speech Emotion Recognition using MFCC features and LSTM network. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–3.

[83] Le Cam, L. M. and Neyman, J. (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 4. Univ of California Press.

[84] Lim, Y. and Lee, Y. (1995). Implementation of the POW (phonetically optimized words) algorithm for speech database. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 89–92. IEEE.

[85] Liu, Y., Penttilä, N., Ihalainen, T., Lintula, J., Convey, R., and Räsänen, O. (2021). Language-Independent Approach for Automatic Computation of Vowel Articulation Features in Dysarthric Speech Assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2228–2243.

[86] Lu, F.-L. and Matteson, S. (2014). Speech tasks and interrater reliability in perceptual voice evaluation. *Journal of Voice*, 28(6):725–732.

[87] Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall.

[88] Martínez, D., Green, P., and Christensen, H. (2013). Dysarthria intelligibility assessment in a factor analysis total variability space. In *Proc Interspeech 2013*. ISCA.

[89] Matsubara, K., Okamoto, T., Takashima, R., and et al., T. (2021). High-Intelligibility Speech Synthesis for Dysarthric Speakers with LPCNet-Based TTS and CycleVAE-Based VC. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7058–7062.

[90] McFee, B., Raffel, C., Liang, D., Ellis, D., Mcvicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. pages 18–24.

[91] Menendez-Pidal, X., Polikoff, J., Peters, S., Leonzio, J., and Bunnell, H. (1996). The Nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1962–1965 vol.3.

[92] Miyamoto, C., Komai, Y., Takiguchi, T., Ariki, Y., and Li, I. (2010). Multimodal speech recognition of a person with articulation disorders using AAM and MAF. In *2010 IEEE International Workshop on Multimedia Signal Processing*, pages 517–520. IEEE.

[93] Morales, S. O. C., Cox, S. J., et al. (2007). Modelling confusion matrices to improve speech recognition accuracy, with an application to dysarthric speech. In *INTERSPEECH*, pages 1565–1568.

[94] Mueller, J., Wenning, G., Verny, M., Mckee, A., Ray Chaudhuri, K., Jellinger, K., Poewe, W., and Litvan, I. (2001). Progression of Dysarthria and Dysphagia in Postmortem-Confirmed Parkinsonian Disorders. *Archives of neurology*, 58:259–64.

[95] Mulfari, D., Meoni, G., Marini, M., and Fanucci, L. (2021). Machine learning assistive application for users with speech disorders. *Applied Soft Computing*, 103:107147.

[96] N P, N. and Alku, P. (2019). Dysarthric speech classification from coded telephone speech using glottal features. *Speech Communication*, 110.

[97] N P, N. and Alku, P. (2020). Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features. *Computer Speech Language*, 65:101117.

[98] Narendra, N. P. and Alku, P. (2020). Glottal Source Information for Pathological Voice Detection. *IEEE Access*, 8:67745–67755.

[99] Natarajan, R. and Chandrakala, S. (2016). Generative Model-Driven Feature Learning for dysarthric speech recognition. *Biocybernetics and Biomedical Engineering*, 36.

[100] Noyes, J. and Frankish, C. (1992). Speech recognition technology for individuals with disabilities. *Augmentative and Alternative Communication*, 8(4):297–303.

[101] of Health, U. and Services, H. (2016). Department of Health and Human Services, Statistics on Voice, Speech, and Language.

[102] Office of the Registrar, G. and Census Commissioner, I. (2011). Disabled population by type of disability, age and sex.

[103] Ooi, C. A., M, H., Yaacob, S., and Chee, L. (2012). Classification of speech dysfluencies with MFCC and LPCC features. *Expert Syst. Appl.*, 39:2157–2165.

[104] oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio.

[105] Organization, W. H. (2023). Disability and health.

[106] Paja, M. and Falk, T. (2012). Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 1:62–65.

[107] Pellegrini, T., Fontan, L., Mauclair, J., Farinas, J., Alazard-Guiu, C., Robert, M., and Gatignol, P. (2015). Automatic Assessment of Speech Capability Loss in Disordered Speech. *ACM Transactions on Accessible Computing*, 6:1–14.

[108] Pellegrini, T., Fontan, L., Mauclair, J., Farinas, J., and Robert, M. (2014). The Goodness of Pronunciation algorithm applied to disordered speech.

[109] Perez, M., Jin, W., Le, D., Carlozzi, N., Dayalu, P., Roberts, A., and Mower Provost, E. (2020). Classification of Huntington Disease using Acoustic and Lexical Features.

[110] Perrotta, G. (2020). Dysarthria: Definition, clinical contexts, neurobiological profiles and clinical treatments, journal = Archives of Community Medicine and Public Health. pages 138–141.

[111] Podder, P., Khan, T., Khan, M., and Rahman, M. (2014). Comparative Performance Analysis of Hamming, Hanning and Blackman Window. *International Journal of Computer Applications*, 96:1–7.

[112] Polur, P. D. and Miller, G. E. (2006). Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals. *Medical engineering & physics*, 28(8):741–748.

[113] Qian, Y., Fan, Y., Hu, W., and Soong, F. K. (2014). On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3829–3833. IEEE.

[114] Qian, Y., Soong, F., Chen, Y., and Chu, M. (2006). An HMM-based Mandarin Chinese text-to-speech system. In *Chinese Spoken Language Processing: 5th International Symposium, ISCSLP 2006, Singapore, December 13-16, 2006. Proceedings*, pages 223–232. Springer.

[115] Raghavendra, P., Rosengren, E., and Hunnicutt, S. (2001). An investigation of different degrees of dysarthric speech as input to speaker-adaptive

and speaker-dependent recognition systems. *Augmentative and Alternative Communication - AUGMENT ALTERN COMMUN*, 17:265–275.

[116] Rajeswari, N. and Chandrakala, S. (2016). Generative model-driven feature learning for dysarthric speech recognition. *Biocybernetics and Biomedical Engineering*, 36(4):553–561.

[117] Rajeswari, R., Thirupathi, D., and Selvaraj, S. (2022). Dysarthric Speech Recognition Using Variational Mode Decomposition and Convolutional Neural Networks. *Wireless Personal Communications*, 122.

[118] Rosen, K. and Yampolsky, S. (2009). Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication*, 16:48–60.

[119] Rudzicz, F. (2009). Phonological features in discriminative classification of dysarthric speech. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4605–4608.

[120] Rudzicz, F. (2011). Articulatory Knowledge in the Recognition of Dysarthric Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):947–960.

[121] Rudzicz, F., Namasivayam, A. K., and Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation*, 46:523–541.

[122] Sahane, P., Pangaonkar, S., and Khandekar, S. (2021). Dysarthric speech recognition using multi-taper mel frequency cepstrum coefficients. In *2021 International Conference on Computing, Communication and Green Engineering (CCGE)*, pages 1–4. IEEE.

[123] Sainath, T. N., Kingsbury, B., Mohamed, A.-r., Dahl, G. E., Saon, G., Soltau, H., Beran, T., Aravkin, A. Y., and Ramabhadran, B. (2013). Improvements to deep convolutional neural networks for LVCSR. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 315–320. IEEE.

[124] Sanders, E., Ruiter, M., Beijer, L., and Strik, H. (2002). Automatic recognition of Dutch dysarthric speech, a pilot study.

[125] Sehgal, S. (2018). Dysarthric speech analysis and automatic recognition using phase based representations.

[126] Selouani, S.-A., Sidi Yakoub, M., and O'Shaughnessy, D. (2009). Alternative speech communication system for persons with severe speech disorders. *EURASIP Journal on Advances in Signal Processing*, 2009:1–12.

[127] Seong, W. K., Park, J. H., and Kim, H. K. (2012). Dysarthric speech recognition error correction using weighted finite state transducers based on context-dependent pronunciation variation. In *Computers Helping People with Special Needs: 13th International Conference, ICCHP 2012, Linz, Austria, July 11-13, 2012, Proceedings, Part II 13*, pages 475–482. Springer.

[128] Shahamiri, S. R. (2021). 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.

[129] Shahamiri, S. R., Lal, V., and Shah, D. (2023). Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering.*

[130] Shahamiri, S. R. and Salim, S. S. B. (2014a). A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artificial neural networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(5):1053–1063.

[131] Shahamiri, S. R. and Salim, S. S. B. (2014b). Artificial neural networks as speech recognizers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. *Advanced Engineering Informatics*, 28(1):102–110.

[132] Shahin, M. and Ahmed, B. (2019). Anomaly detection based pronunciation verification approach using speech attribute features. *Speech Communication*, 111:Pages 29–43.

[133] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

[134] Sheoran, K., Bajgoti, A., Gupta, R., Jatana, N., Geetikadhand, Gupta, C., Dadheech, P., Yahya, U., and Aneja, N. (2023). Pronunciation Scoring With Goodness of Pronunciation and Dynamic Time Warping. *IEEE Access*, PP.

[135] Shih, D.-H., Liao, C.-H., Wu, T.-W., Xu, X.-Y., and Shih, M.-H. (2022). Dysarthria Speech Detection Using Convolutional Neural Networks with Gated Recurrent Unit. *Healthcare*, 10.

[136] Soleymanpour, M., Johnson, M. T., Soleymanpour, R., and Berry, J. (2022). Synthesizing Dysarthric Speech Using Multi-Speaker Tts For Dysarthric Speech Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7382–7386.

[137] Solomon, N., Lorell, D., Robin, D., Rodnitzky, R., and Luschei, E. (1995). Tongue strength and endurance in mild to moderate Parkinson's disease. *Journal of Medical Speech-Language Pathology*, 3:15–26.

[138] Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., and Bengio, Y. (2017). Char2wav: End-to-end speech synthesis.

[139] Sriranjani R, Umesh S, R. M. (2015). Automatic Severity Assessment of Dysarthria using State-Specific Vectors. *Biomed Sci Instrum*, 51:99–106.

[140] Stipancic, K., Palmer, K., Rowe, H., Yunusova, Y., Berry, J., and Green, J. (2021). "You Say Severe, I Say Mild": Toward an Empirical Classification of Dysarthria Severity. *Journal of Speech, Language, and Hearing Research*, 64:1–18.

[141] Stipancic, K. L., Yunusova, Y., Berry, J. D., and Green, J. R. (2018). Minimally Detectable Change and Minimal Clinically Important Difference of a Decline in Sentence Intelligibility and Speaking Rate for Individuals With Amyotrophic Lateral Sclerosis. *Journal of Speech, Language, and Hearing Research*, 61(11):2757–2771.

[142] Swigert, N. (2010). *The Source for Dysarthria.* LinguiSystems.

[143] Sy, B. and Horowitz, D. (1993). A statistical causal model for the assessment of dysarthric speech and the utility of computer-based speech recognition. *IEEE Transactions on Biomedical Engineering*, 40(12):1282–1298.

[144] Takashima, Y., Takiguchi, T., and Ariki, Y. (2019). End-to-end dysarthric speech recognition using multiple databases. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6395–6399. IEEE.

[145] Teixeira, J. and Fernandes, P. (2014). Jitter, Shimmer and HNR Classification within Gender, Tones and Vowels in Healthy Voices. *Procedia Technology*, 16.

[146] Tirumala, S. S. and Shahamiri, S. R. (2017). A Deep Autoencoder approach for Speaker Identification. In *Proceedings of the 9th International Conference on Signal Processing Systems*, ICSPS 2017, page 175–179.

[147] Tong, H., Sharifzadeh, H., and McLoughlin, I. (2020). Automatic assessment of dysarthric severity level using audio-video cross-modal approach in deep learning.

[148] Tripathi, A., Bhosale, S., and Kopparapu, S. K. (2020a). A novel approach for intelligibility assessment in dysarthric subjects. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6779–6783. IEEE.

[149] Tripathi, A., Bhosale, S., and Kopparapu, S. K. (2020b). Improved Speaker Independent Dysarthria Intelligibility Classification Using Deepspeech Posteriors. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6114–6118.

[150] Tyagi, V., Dev, A., and Bansal, P. (2023). Analysis and Classification of Dysarthric Speech. In *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

[151] Viswanathan, M. and Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: development and assessment of a modified Mean Opinion Score (MOS) scale. *Computer speech & language*, 19(1):55–83.

[152] Vu, T. T., Luong, M. C., and Nakamura, S. (2009). An HMM-based Vietnamese speech synthesis system. In *2009 Oriental COCOSDA International Conference on Speech Database and Assessments*, pages 116–121. IEEE.

[153] Vyas, G., Dutta, M. K., Prinosil, J., and Harár, P. (2016). An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, pages 515–518.

[154] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

[155] Wang, Y.-T., Kent, R., Duffy, J., and Thomas, J. (2005a). Dysarthria associated with traumatic brain injury: Speaking rate and emphatic stress. *Journal of communication disorders*, 38:231–60.

[156] Wang, Y.-T., Kent, R., Duffy, J., and Thomas, J. (2005b). Dysarthria in Traumatic Brain Injury: A Breath Group and Intonational Analysis. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, 57:59–89.

[157] Wang, Y.-T., Kent, R., Duffy, J., Thomas, J., and Weismer, G. (2004). Alternating motion rate as an index of speech motor disorder in traumatic brain injury. *Clinical linguistics phonetics*, 18:57–84.

[158] Wang, Y.-T., Kent, R., Kent, J., Duffy, J., and Thomas, J. (2009). Acoustic analysis of voice in dysarthria following stroke. *Clinical linguistics phonetics*, 23:335–47.

[159] Witt, S. and Young, S. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2):95–108.

[160] Xiong, F., Barker, J., and Christensen, H. (2018). Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. In *speech communication; 13th ITG-symposium*, pages 1–5. VDE.

[161] Xu, X., Kang, Y., Cao, S., Lin, B., and Ma, L. (2021). Explore wav2vec 2.0 for Mispronunciation Detection. pages 4428–4432.

[162] Yeo, E., Choi, K., Kim, S., and Chung, M. (2023a). Speech Intelligibility Assessment of Dysarthric Speech by using Goodness of Pronunciation with Uncertainty Quantification.

[163] Yeo, E. J., Choi, K., Kim, S., and Chung, M. (2023b). Automatic Severity Classification of Dysarthric Speech by Using Self-Supervised Model with Multi-Task Learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

[164] Yorkston, K., Spencer, K., and Duffy, J. (2003). Behavioral management of respiratory/phonatory dysfunction from dysarthria: A systematic review of the evidence. *Journal of Medical Speech-Language Pathology*, 11:xiii–xxxviii.

[165] Yu, C. t. (2023). Multi-Stage Audio-Visual Fusion for Dysarthric Speech Recognition With Pre-Trained Models. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1912–1921.

[166] Zen, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7962–7966. IEEE.

[167] Zhang, J., Ren, J., Li, L., Gu, J., and Zhang, D. (2022). Defect identification of layered adhesive structures based on dynamic time warping and simulation analysis. *Infrared Physics & Technology*, 120:103943.

[168] Zhang, K. and Gu, X. (2014). An Affinity Propagation Clustering Algorithm for Mixed Numeric and Categorical Datasets. *Mathematical Problems in Engineering*, 2014.

[169] Zhang, X., Gao, J., Lu, P., and Yan, Y. (2008). A novel speaker clustering algorithm via supervised affinity propagation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4369–4372.

Publications

Journals

1. **Komal Bharti**, S Agri, P. K. Das. " Dysarthric Speech Detection and Severity Classification using Audio Spectrogram Transformer" in International Association of Engineers, Engineering Letters, January 2024. [Accepted]
2. **Komal Bharti** and P. K. Das. *Characteristics, Analysis and Annotation of UASpeech dataset.* [Preparing for submission]

Conferences

1. **Komal Bharti** and P. K. Das, "A Survey on ASR Systems for Dysarthric Speech," 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), Delhi, India, 2022, pp. 1-6, doi: 10.1109/AIST55798.2022.10065162.
2. **Komal Bharti**, Agri, S., P. K. Das. "Dysarthric Speech Characterization and Classification Based on Affinity Propagation". In: Swain, B.P., Dixit, U.S. (eds) Recent Advances in Electrical and Electronic Engineering. ICSTE 2023. https://doi.org/10.1007/978-981-99-4713-3_54
3. **Komal Bharti**, S. Haque and P. K. Das. "A Novel Dysarthric Speech Synthesis system using Tacotron2 for specific and OOV words," 2024 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 2024, pp. 1-5, doi: 10.1109/SPCOM60851.2024.10631612.
4. **Komal Bharti** and P. K. Das. "Fusion of Multiple Audio Descriptor for the Recognition of Dysarthric Speech," 27th International Conference of the International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA) 2024, Taiwan. [Presented]

5. **Komal Bharti** and P. K. Das. "Intelligibility Assessment of Dysarthric Speech based on Goodness of Pronunciation" in 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing (Interspeech 2025). [Submitted]

6. P. Bhagath, **K Bharti**, A. Kotiya and P. K. Das, "Feature Selection using Pre-clustering via Affinity Propagation for Speech Classification in Low-resource Languages," 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, 2021, pp. 1-6, doi: 10.1109/IICAIET51634.2021.9573696.

Other publication (Not related to thesis work)

1. **Komal Bharti**, P. Shrivastava and P. K. Das, "Comparison of Multiple Classifiers for Audio-Visual Speaker Recognition System," 2023 International Conference on Computational Intelligence, Networks and Security (ICCINS), Mylavaram, India, 2023, pp. 1-6, doi: 10.1109/ICCINS58907.2023 .10450017.