# AUTOMATIC SPEECH RECOGNITION FOR DYSARTHRIA PATIENTS

**A THESIS**

*Submitted by*

**ISHTIAQUE AHMED**

**(M210445EC)**

*In partial fulfillment for the award of the degree of*

**MASTER OF TECHNOLOGY**

**IN**

**ELECTRONICS AND COMMUNICATION ENGINEERING**

**(Signal Processing)**

*Under the guidance of*

**Dr. Waquar Ahmad**

***DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING***

**NATIONAL INSTITUTE OF TECHNOLOGY CALICUT**
**NIT CAMPUS PO, KOZHIKODE, KERALA, INDIA 673601**

**July 4, 2023**

# ACKNOWLEDGEMENTS

First and foremost, I would like to extend a deep sense of gratitude to my project advisor Dr. Waquar Ahmad, Assistant Professor in the Department of Electronics and Communication Engineering for their very valuable guidance, monitoring, and motivation throughout the course of this project. His immense knowledge, stimulating suggestions, patience, and encouragement helped me coordinate this project in a decent manner.

I would also like to acknowledge with much appreciation the crucial role of the Project evaluation committee members: Dr. Sudish N. George, Dr. Anup Aprem, and Dr. G. Abhilash for their valuable advice and critical assessments during the presentations.

I am grateful to Dr. Jaikumar M. G., Head of the Department of Electronics and Communication Engineering, and to the faculties of the Electronics and Communication Department for all their help in completing the project. I also thank my family and friends for their support throughout the work.

Last but not least, I thank Almighty God for the good health and well-being that were necessary to complete this thesis.

**-Ishtiaque Ahmed**

# DECLARATION

*I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.*

| | |
|---|---|
| **Place: NIT Calicut** | **Signature:** |
| **Date: July 4, 2023** | **Name: Ishtiaque Ahmed** |
| | **Reg.No: M210445EC** |

# National Institute of Technology Calicut

**Department of Electronics and Communication Engineering**

## CERTIFICATE

*This is to certify that the Thesis entitled, "**Automatic Speech Recognition for Dysarthria Patients**" submitted by Mr. **Ishtiaque Ahmed** to the National Institute of Technology Calicut towards partial fulfillment of the requirements for the award of the Degree of Master of Technology in **Electronics and Communication Engineering** (**Signal Processing**) is a bonafide record of the work carried out by him under my supervision and guidance.*

*Guide*:
**Dr. Waquar Ahmad**
Assistant Professor
ECED

**Dr. Jaikumar M. G.**
Head of the Department
ECED

Place: NIT Calicut
Date: July 4, 2023

(**Office seal**)

# TABLE OF CONTENTS

# ABSTRACT

Dysarthria is a common speech disorder caused by neurological damage that weakens the muscles necessary for speech production. Our objective is to develop an Automatic Speech Recognition system for people suffering from dysarthria. To address this we have developed an ASR system based on x-vectors specifically for dysarthric speech exhibiting varying levels of speech intelligibility (low, medium, and high). Given the scarcity of data available from dysarthric speakers, we trained our proposed ASR system using dysarthric speech data from the UA-Speech dataset and duration-modified augmented dataset. To enhance the performance of our model, we propose a data augmentation technique based on duration modification. Our research work applies duration modification with multiple scaling factors to the dysarthric training speech, enabling the training of the ASR system using both the original dysarthric along with healthy speech and its duration-modified versions. This augmentation technique aims to address the significant disparities in phone duration observed between normal speakers and dysarthric speakers with varying levels of speech intelligibility. Experimental evaluations demonstrate that the proposed duration modification based data augmentation yielded a remarkable relative improvement of 34% over the baseline ASR system. Additionally, for speakers with a high severity level of dysarthria, the relative improvement reached 29%. These findings highlight the effectiveness of the proposed approach in mitigating the challenges associated with dysarthric speech recognition. By incorporating duration modification-based data augmentation, the ASR system exhibits substantial advancements in accurately verifying dysarthric speakers, thereby contributing to improved accessibility and communication for individuals affected by dysarthria.

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

## 1.1 An Overview

Automatic speech recognition (ASR) systems are designed to convert spoken language into written text. However, speech production is a complex process that involves the coordinated contraction of various muscles related to respiration, laryngeal control, and articulation. If the nerves that control these muscles are affected by neuro-muscular conditions, such as dysarthria, the speech produced may be difficult for an ASR system to transcribe accurately.

Dysarthria is a neurological speech disorder that can cause slurred or slow speech with poor articulation. It can also affect the speed of changing the position of articulators and the quality of phonation, pitch, and loudness of speech. In addition, dysarthria can lead to shallow breathing and difficulty adjusting exhalation with vocalization. If the soft palate is involved, excessive nasal sounds may be perceived in dysarthric speech. The severity of dysarthria can vary, with more severe cases rendering speech nearly unintelligible to an ASR system.

Due to the complex nature of dysarthria, it can pose significant challenges for ASR systems. In some cases, the condition may cause the speaker's speech to be misinterpreted or not recognized at all, resulting in errors in the transcription. This can be particularly problematic in situations where accurate transcription is essential, such as medical or legal settings.

To address these challenges, ASR systems may use advanced algorithms and techniques that take into account the unique characteristics of dysarthric speech. Machine learning models can be trained on large datasets of dysarthric speech to improve recognition accuracy. Additionally, specialized speech recognition software and hardware can be used to help filter out background noise and other factors that may interfere with accurate transcription.

Overall, while dysarthria can present significant challenges for ASR systems, ongoing research and development in the field are helping to improve recognition accuracy and make these systems more accessible to people with speech disorders.

One of the most significant challenges faced by individuals with dysarthria is the ability to communicate effectively with others. Traditional methods of communication, such as face-to-face conversation, phone calls, and text messaging, are often difficult or impossible for individuals with dysarthria, leaving them feeling isolated and frustrated. As a result, there has been a growing interest in the development of assistive technologies, such as automatic speech recognition (ASR) systems, to help individuals with dysarthria communicate more effectively.

ASR is a technology that enables the conversion of spoken language into text or other machine-readable formats. The development of ASR systems for individuals with dysarthria presents unique challenges due to the variability of speech patterns, including irregular articulation, slowed rate of speech, and inconsistent speech volume. These challenges require the use of specialized algorithms and training techniques to improve the accuracy of ASR systems for individuals with dysarthria.

Recent advances in machine learning and natural language processing have enabled significant improvements in ASR systems for dysarthric patients. These advancements have led to the development of new approaches to ASR, such as deep learning-based methods, which have shown promising results in improving the accuracy and robustness of ASR systems for individuals with dysarthria.

The primary goal of this research work is to provide a comprehensive overview of the current state-of-the-art in ASR systems for individuals with dysarthria. Specifically, we will review the latest advances in machine learning techniques and algorithms for improving the accuracy of ASR systems for individuals with dysarthria. We will also discuss the challenges associated with developing ASR systems for individuals with dysarthria and highlight the potential applications of this technology in improving the quality of life for individuals with dysarthria. Moreover, there have been efforts to improve the performance of ASR systems for dysarthric patients by incorporating contextual information. In the work by Sahin and colleagues the authors proposed a contextual speech recognition approach that leverages contextual information from the text being spoken to improve the recognition accuracy of dysarthric speech. The results showed that the contextual approach significantly improved the accuracy of ASR systems for dysarthric patients, indicating the potential of contextual information for improving ASR system performance.

Another important area of research in ASR systems for dysarthric patients is the development of personalized models. Dysarthria can vary significantly between individuals, and personalized models can improve the accuracy of ASR systems by capturing the specific characteristics of an individual's speech. In the work by Chakraborty et al , the authors proposed a personalized ASR system for dysarthric patients that uses a combination of acoustic and linguistic features. The results showed that the personalized approach significantly improved the recognition accuracy of dysarthric speech, highlighting the potential of personalized models for ASR systems for dysarthric patients.

Finally, there have been efforts to develop ASR systems for dysarthric patients in different languages. In the work by Mukherjee et al, the authors proposed an ASR system for Bengali dysarthric speech. The authors used a hybrid approach that combined both rule-based and data-driven techniques to improve the accuracy of the system. The results showed that the proposed ASR system significantly improved the recognition accuracy of Bengali dysarthric speech, indicating the potential of ASR systems for dysarthric patients in different languages.

The x-vector-based speaker embeddings which are extracted using a time-delay neural network (TDNN), are widely used as a state-of-the-art representation in speaker recognition tasks. In this study, we investigated the effectiveness of using x-vector-based speaker embeddings for developing an ASR system for dysarthric speakers. It should be noted that the TDNN architecture used for extracting x-vectors requires a large amount of domain-specific speech data to be used during training in order to accurately estimate the model parameters.

One of the challenges in developing robust ASR systems for dysarthric speech is the limited availability of domain-specific speech data. Due to this, several dysarthria-specific characteristics such as speech rate and average phoneme duration are not well-represented in the training data, leading to degraded system performance. To address this, we explored the use of duration-modification-based data augmentation, a signal-processing technique that extends the duration of speech data from healthy speakers. By introducing missing acoustic attributes through this method, we increased the amount of domain-specific data and improved the accuracy of the ASR system for dysarthric speakers. This approach is particularly useful when dealing with limited training data for developing robust ASR systems for dysarthric speech.

## 1.2 Problem Definition and Formulation

The aim of this project is to develop an effective automatic speech recognition (ASR) system specifically designed to assist dysarthria patients in improving their communication abilities. The system should accurately recognize and transcribe the speech of dysarthria patients, accounting for their unique speech characteristics and challenges. By addressing the limitations of existing ASR systems in accommodating dysarthric speech, this project aims to provide a valuable tool for enhancing communication and promoting inclusivity for individuals with dysarthria.

## 1.3 Motivation

The motivation for this thesis is as follows:

- **Inaccurate Word Recognition** Dysarthria is a motor speech disorder that affects the ability to articulate words clearly due to muscle weakness or paralysis. Existing assistive technologies for dysarthria patients often fall short of providing accurate and efficient communication solutions. By developing an automatic speech recognition system tailored specifically for dysarthria, you aim to address this unmet need and improve the quality of life for individuals with this condition.

- **Enhanced Communication** Dysarthria patients often struggle to be understood by others, leading to frustration, isolation, and limited participation in social, educational, and professional settings. By building an ASR system that can accurately transcribe their speech, you can empower these individuals to express themselves more effectively, enabling them to engage in conversations, share their thoughts, and participate more actively in various aspects of life.

- **Personal Connection** Sharing personal stories or experiences related to dysarthria can add a compelling touch to the motivation section. If you have a personal connection with someone who has dysarthria or has witnessed the impact of the condition firsthand, you can describe how their struggles inspired you to take up this project and make a difference in their lives.

- **Inclusivity and Accessibility** Accessibility is a fundamental right, and it is crucial to ensure that individuals with dysarthria have equal opportunities to communicate and be understood. By developing an ASR system tailored to their specific needs, you aim to promote inclusivity and bridge the communication gap, allowing dysarthria patients to participate fully in society and enjoy the same opportunities as others.

- **Advancements in Technology** Automatic speech recognition technology has made significant progress in recent years. By leveraging the latest advancements in speech recognition algorithms, machine learning, and signal processing techniques, you can contribute to the field and push the boundaries of what is possible in assisting dysarthria patients. This project provides an opportunity to harness cutting-edge technologies and adapt them to address a real-world problem

- **Impact and Social Significance** Finally, emphasize the positive impact your project can have on the lives of dysarthria patients. Improved communication can enhance their self-confidence, mental well-being, and overall quality of life. It can also benefit their relationships with friends, family, and caregivers, fostering stronger connections and reducing feelings of isolation.

## 1.4 Thesis Contribution

The thesis contributions are outlined below:

- The ASR system undergoes evaluation using speech data from the Torgo and UA-Speech databases, where Torgo serves as the baseline database. UA-Speech database is utilized for the proposed results. Our model is trained and tested on the UA-Speech database.
- Monophone training and triphone modeling techniques are used for acoustic modeling in ASR, with triphones capturing context-dependent phonetic units and improving accuracy.
- Tri-1, Tri-2, and Tri-3 represent stages of ASR training, incorporating techniques to enhance performance and adapt to speaker characteristics.
- Deep Neural Networks (DNN) have shown significant improvements over traditional models like GMM-HMM in ASR, utilizing multiple hidden layers to learn acoustic-to-text mappings.
- Data augmentation through duration modification is performed to address the acoustic mismatch between training and test data, with  representing the modification rate.

## 1.5 Thesis Organization

The rest of this thesis is organized as follows. The basic introduction to automatic speech recognition for dysarthric patients is discussed along with the problem definition and formulation. In chapter 2 literature review is discussed in which there is an overview of dysarthria and communication challenges there's an introduction to automatic speech recognition, challenges, and limitations of asr for dysarthric patients. In the 3rd chapter duration, the modification-based data augmentation technique is discussed which explained why we need data augmentation and the motivation behind doing that. In the last 4th chapter there is experimental setups system architecture work done and results followed by conclusion and future scope.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Dysarthria and its Communication Challenges

Dysarthria is a motor speech disorder characterized by difficulties in articulating and producing speech sounds due to muscular weakness, paralysis, or coordination impairments. It is caused by damage or dysfunction in the central or peripheral nervous system, affecting the muscles involved in speech production, including the lips, tongue, vocal cords, and diaphragm.

The specific symptoms and severity of dysarthria can vary widely depending on the underlying condition and the areas of the nervous system affected. Common characteristics of dysarthria include:

- Articulation Difficulties: Dysarthric individuals may have trouble forming and coordinating the precise movements necessary for clear speech. This can result in slurred or distorted speech sounds, making it challenging for others to understand their words.
- Resonance Problems: Some individuals with dysarthria experience issues with the control of airflow through the vocal tract, leading to abnormalities in speech resonance. This can cause a nasal quality or a muffled sound to their speech.
- Prosody and Intonation Irregularities: Dysarthria can affect the natural rhythm, stress, and intonation patterns of speech. Individuals may have difficulties with appropriate pitch modulation, stress placement, and phrasing, making their speech sound monotone or lacking emphasis.
- Reduced Vocal Volume: Dysarthric individuals may exhibit reduced loudness or vocal weakness, resulting in a low volume of speech. This can make it challenging for others to hear them, particularly in noisy environments.

These speech impairments associated with dysarthria can have a significant impact on communication. Dysarthric individuals often face challenges in effectively conveying their thoughts, ideas, and emotions to others. The communication difficulties they encounter include:

- Intelligibility: Dysarthric speech may be difficult for others to understand due to the distorted or imprecise articulation of speech sounds. This can result in frequent misunderstandings and misinterpretations, leading to frustration and reduced participation in conversations.
- Naturalness and Expressiveness: Dysarthria can affect the naturalness and expressiveness of speech, making it harder for individuals to convey emotions, subtle nuances, or variations in tone. This can limit their ability to fully express themselves and engage in meaningful social interactions.
- Fatigue and Energy Expenditure: Speaking with dysarthria can be physically demanding and require more effort due to the need for compensatory movements and increased muscle exertion. This can lead to fatigue and reduced endurance during prolonged conversations.
- Social and Emotional Impact: Communication difficulties associated with dysarthria can result in social isolation, reduced self-confidence, and feelings of frustration or embarrassment. Individuals with dysarthria may encounter barriers in forming relationships, participating in social activities, or accessing educational and professional opportunities.

The impact of dysarthria on communication underscores the importance of interventions and assistive technologies aimed at improving communication access and effectiveness for individuals with this condition. Automatic Speech Recognition (ASR) technology holds promise as one such assistive tool, as it can help overcome some of the limitations imposed by dysarthric speech, enabling more efficient and inclusive communication.

## 2.2 Introduction to Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is a technology that aims to convert spoken language into written text. It involves the development of algorithms, models, and systems capable of transcribing and understanding human speech. ASR holds immense potential in various applications, including transcription services, voice-controlled systems, virtual assistants, and assistive technologies for individuals with speech impairments.

The primary goal of ASR is to accurately recognize and interpret the linguistic content of spoken utterances. ASR systems typically consist of several components that work together to achieve this objective. These components include:

- Acoustic Processing: In this stage, the acoustic signal of the speech is captured and processed. It involves transforming the speech waveform into a more compact and representative form, such as Mel-frequency cepstral coefficients (MFCCs) or filterbank energies. Acoustic processing helps capture relevant acoustic features that contribute to the linguistic content of the speech.

- Acoustic Modeling: Acoustic modeling involves creating statistical models that map the observed acoustic features to the corresponding linguistic units, such as phonemes or words. Hidden Markov Models (HMMs) and deep neural networks (DNNs) are commonly used techniques for acoustic modeling. These models learn the relationship between the acoustic features and linguistic units based on training data.

- Language Modeling: Language modeling focuses on capturing the statistical patterns and regularities of language. It involves constructing models that estimate the probability distribution of word sequences. Language models aid in determining the most likely sequence of words given the acoustic input, improving the accuracy and fluency of ASR output.

- Decoding and Search: The decoding and search process involves selecting the most probable word sequence based on the combination of the acoustic and language models. It employs search algorithms such as the Viterbi algorithm or beam search to find the optimal word sequence that best matches the observed acoustic features.

ASR technology has made significant advancements in recent years, largely driven by the availability of large speech datasets, improved machine learning algorithms, and computational resources. These advancements have led to enhanced accuracy and performance of ASR systems, making them increasingly valuable in real-world applications.

In the context of dysarthric patients, ASR can be adapted to address the unique speech characteristics and challenges associated with dysarthria. By leveraging ASR, dysarthric individuals can overcome some of the communication barriers they face, enabling them to express themselves more effectively and interact with others more independently. ASR holds great promise in improving communication access and quality of life for individuals with dysarthria.

## 2.3 Adaptation Techniques for Dysarthric Speech

Adaptation techniques for dysarthric speech are essential in improving the performance of Automatic Speech Recognition (ASR) systems for individuals with dysarthria. These techniques aim to tailor the ASR system to the specific speech characteristics of each individual, resulting in more accurate and reliable transcriptions. Here are some adaptation techniques commonly employed:

- Speaker-Specific Adaptation: Speaker-specific adaptation focuses on customizing the ASR system to an individual's unique speech patterns and characteristics. This is achieved by collecting a representative sample of the individual's dysarthric speech and using it to train or adapt the ASR system. Speaker-specific adaptation can involve techniques such as feature transformation, model re-estimation, or neural network fine-tuning to make the ASR system more suitable for the individual's speech.
- Pronunciation Modeling: Dysarthric speech often exhibits atypical pronunciation patterns and deviations from standard phonetic representations. Pronunciation modeling techniques involve creating customized phonetic representations or lexicons

that capture the specific pronunciation variations associated with dysarthria. These adaptations allow the ASR system to better align the acoustic input with the intended words or phonemes, improving recognition accuracy.

- Noise and Environment Adaptation: Dysarthric speech can be further complicated by the presence of background noise or adverse acoustic conditions. Adaptation techniques that account for noise and environmental factors help the ASR system handle these challenging conditions. This can involve collecting and utilizing noisy speech data during training, applying noise reduction algorithms, or incorporating environmental information into the acoustic and language models.

- Data Augmentation: Data augmentation techniques involve artificially increasing the diversity and quantity of training data for the ASR system. This can be particularly beneficial when there is limited dysarthric speech data available for training. Data augmentation techniques include adding background noise, applying various voice transformations, or generating synthetic dysarthric speech samples. By augmenting the training data, the ASR system becomes more robust and capable of handling a wider range of dysarthric speech variations.

- Dynamic Adaptation: Dynamic adaptation techniques allow the ASR system to adapt and adjust its models during runtime based on the characteristics of the individual's speech input. This can involve online adaptation algorithms that update the acoustic or language models on-the-fly as the ASR system receives new input. Dynamic adaptation enables the ASR system to adapt to short-term speech variations, improving accuracy in real-time scenarios.

- Contextual Adaptation: Contextual adaptation techniques take into account the linguistic and contextual information surrounding dysarthric speech. By incorporating contextual cues, such as the surrounding words, grammar, or topic-specific information, the ASR system can make more accurate predictions and resolve potential ambiguities in the transcription process.

These adaptation techniques aim to overcome the challenges posed by dysarthric speech and improve the performance of ASR systems for individuals with dysarthria. The selection and combination of adaptation techniques depend on the specific needs and characteristics of the individual, as well as the available resources and data for training and customization. Continuous research and development in this area contribute to advancing the capabilities of ASR technology for individuals with dysarthria.

## 2.4 Challenges and Limitations of ASR for Dysarthric Speech

While Automatic Speech Recognition (ASR) technology holds promise for improving communication access for individuals with dysarthria, there are several challenges and limitations that need to be addressed. Here are some key challenges and limitations of ASR for dysarthric speech:

- Variability and Complexity of Dysarthric Speech: Dysarthric speech exhibits high variability and complexity due to the wide range of speech impairments associated with dysarthria. This variability poses challenges for ASR systems, as they need to handle different types and severities of dysarthria, making it difficult to develop a one-size-fits-all solution. ASR systems may struggle to capture the unique speech characteristics and patterns exhibited by individuals with dysarthria, leading to decreased recognition accuracy.
- Limited Availability of Dysarthric Speech Data: Collecting large and diverse dysarthric speech corpora for training ASR models can be challenging due to the limited availability of data. Dysarthric speech data collection often requires collaboration with clinicians, researchers, and individuals with dysarthria, which can be time-consuming and resource-intensive. The scarcity of data can hinder the development and optimization of ASR systems specifically for dysarthric speech.
- Insufficient Representation in Standard ASR Training Data: ASR models are typically trained on large datasets that predominantly contain clean, non-dysarthric speech. The lack of sufficient representation of dysarthric speech in standard ASR training data can lead to suboptimal performance for dysarthric individuals. ASR systems may struggle to generalize well to the unique characteristics of dysarthric

speech, resulting in reduced accuracy.

- Lack of Personalization and Individualization: ASR systems often lack the ability to adapt and personalize to individual dysarthric speakers. Dysarthria can vary significantly between individuals, and ASR models that are not specifically adapted or customized for each person may not adequately capture their speech patterns. The development of personalized ASR systems requires collecting and utilizing speaker-specific dysarthric speech data, which can be logistically challenging.

- Impact of Background Noise and Acoustic Conditions: Dysarthric speech recognition can be further complicated by the presence of background noise or adverse acoustic conditions. Noise and environmental factors can significantly degrade ASR performance, as dysarthric speech may already have reduced intelligibility. ASR systems need to be robust to noisy environments and capable of handling acoustic variations caused by different recording conditions.

- Limited Vocabulary and Out-of-Vocabulary Words: Dysarthric individuals may struggle with articulating certain sounds or producing accurate phonetic representations, leading to a limited vocabulary and an increased likelihood of out-of-vocabulary (OOV) words. OOV words pose challenges for ASR systems, as they may not have been encountered during training, resulting in higher recognition errors for less common or specialized vocabulary.

- Adaptation and Generalization Challenges: Adapting ASR systems to dysarthric speech requires careful consideration of adaptation techniques and generalization capabilities. Adapting ASR models to individual speakers or specific dysarthric subgroups may improve performance for those individuals but can limit generalization to new or unseen speakers. Striking a balance between individualization and generalization remains a challenge in dysarthric speech recognition.

Addressing these challenges and limitations requires ongoing research and development efforts in the field of dysarthric speech recognition. Techniques such as speaker adaptation, pronunciation modeling, data augmentation, and noise robustness can help mitigate these challenges and improve the performance of ASR systems for individuals with dysarthria. Additionally, the collection and sharing of larger, more diverse dysarthric speech corpora can provide valuable resources for training and evaluation, driving advancements in ASR technology for dysarthric speech.

## 2.5 Applications and Potential Benefits of ASR for Dysarthric Patients

Automatic Speech Recognition (ASR) technology holds significant potential for enhancing communication and improving the quality of life for dysarthric patients. Here are some key applications and potential benefits of ASR for dysarthric individuals:

- Augmentative and Alternative Communication (AAC): ASR can be integrated into AAC devices or applications to provide real-time transcription of dysarthric speech. This enables dysarthric individuals to express themselves more effectively by converting their spoken language into written text. ASR-based AAC systems can facilitate communication in various contexts, including social interactions, educational settings, and healthcare environments.
- Voice-Controlled Assistive Technology: ASR can enable dysarthric individuals to control various assistive technologies using their voice. By accurately recognizing their speech commands, ASR systems allow individuals to interact with smartphones, computers, home automation systems, and other devices, enhancing their independence and enabling them to perform tasks more efficiently.
- Telecommunication and Remote Communication: ASR facilitates remote communication for dysarthric individuals, enabling them to participate in phone calls, video conferences, and online messaging platforms. Real-time transcription provided by ASR systems can enhance the comprehension of their speech, enabling more effective and inclusive communication with others, even in situations where the listener may have difficulty understanding their spoken words.
- Voice-Enabled Access to Information and Services: ASR technology can enable dysarthric individuals to access information and services through voice commands.

By leveraging ASR-based virtual assistants, individuals can perform tasks such as web searches, setting reminders, sending emails, and accessing online content, making it easier for them to navigate the digital world.

- Therapy and Rehabilitation: ASR systems can support speech therapy and rehabilitation for dysarthric patients. ASR-based tools can provide objective feedback and measurement of speech intelligibility and accuracy, allowing therapists to track progress and customize treatment plans. ASR can also assist in practicing specific speech exercises, promoting self-monitoring and self-correction of speech production.
- Research and Data Analysis: ASR technology enables researchers to analyze large amounts of dysarthric speech data efficiently. By transcribing and analyzing dysarthric speech using ASR, researchers can gain insights into the underlying speech characteristics, identify patterns, and develop targeted interventions or personalized treatment strategies.

The potential benefits of ASR for dysarthric patients include improved communication effectiveness, increased participation in social and professional settings, enhanced independence and autonomy, access to information and services, and more efficient and personalized therapy. By leveraging ASR technology, dysarthric individuals can overcome communication barriers and enjoy a higher quality of life, fostering social connections, and engagement in various aspects of daily life.

# CHAPTER 3

# DURATION MODIFICATION BASED DATA AUGMENTATION

## 3.1 Introduction

The proposed method in this section is based on duration modification for data augmentation. This section outlines the data augmentation approach proposed in this study, which is based on duration modification.

## 3.2 Need for data augmentation

Data augmentation is a crucial technique in machine learning and data science that involves generating new training data by applying certain transformations or modifications to the existing dataset. Since the TDNN architecture employed in an x-vector the system consists of several layers, and a large amount of speech data is necessary to develop these systems in order to properly utilize machine learning methodologies for ASR. The primary objective of data augmentation is to increase the size and diversity of the training data, which can help improve the generalization performance of machine learning models. In the case of automatic speech recognition, data augmentation is particularly important because ASR models need to be robust to variations in speech patterns and characteristics. By introducing variations in speech duration and tempo, data augmentation can help improve the accuracy and robustness of ASR models. Therefore, there is a need for effective data augmentation techniques in ASR research to enhance the performance of ASR systems.

In the field of automatic speech recognition, speech data is crucial for training and developing effective systems. However, collecting sufficient data can be a challenge, particularly when it comes to speech from individuals with dysarthria. This is because dysarthric speakers often struggle to speak for extended periods due to muscular weakness and fatigue, making it difficult to collect large amounts of data. As a result, dysarthric speech databases typically contain limited speech data from only a few individuals. This poses a significant obstacle to developing accurate and effective ASR systems for dysarthric speech.

Having a limited amount of speech data from a small number of dysarthric speakers can hinder the development of accurate models. ASR systems rely on machine learning methods, and a large amount of diverse training data is necessary to effectively train these models. However, dysarthric speakers face difficulties in speaking for long periods of time due to muscular weakness and exhaustion, which can make collecting sufficient speech data challenging.

Therefore, data augmentation techniques are needed to increase the amount of available training data and improve the performance of ASR systems for dysarthric speech. In this research work, we propose a data augmentation approach based on duration modification to address this issue of text-independent dysarthric speech recognition using x-vector-based speaker embedding. This approach involves modifying the duration of speech segments in the training data to create new, diverse samples for model training.

The focus of our study is to introduce an automatic speech recognition system for dysarthric speakers that employ x-vector-based speaker embedding. In order to develop these systems using machine learning methodologies for ASR, a substantial amount of speech data is required due to the TDNN architecture employed in an x-vector system, which comprises multiple layers. However, individuals with dysarthria experience difficulty in speaking for extended periods due to muscular weakness and exhaustion. Therefore, collecting sufficient dysarthric speech data can be a daunting task, particularly for those with severe dysarthria. As a result, currently, available dysarthric speech databases only contain limited speech data from a small number of speakers. When training an x-vector-based system with limited dysarthric speech data, there is a risk of under-fitting. Conversely, using a large amount of speech data from healthy speakers to train the TDNN could result in a bias towards control subjects, leading to poor performance in dysarthric speakers. In order to overcome the challenges posed by limited data availability and diversity, and to improve the model's robustness, we incorporated data augmentation into our approach. Data augmentation involves applying various transformations to the existing training data, creating new synthetic training samples, and then combining them with the original dataset. This technique is used to increase the diversity of the acoustic conditions captured by the training data and enhance the model's ability to generalize.

## 3.3 Motivation for duration-modification-based data augmentation

People with dysarthria may speak slower due to difficulties with tongue and lip movements. To gain better insight, we conducted an analysis of dysarthric and control speech using Matlab and Praat software. Our findings, illustrated in Figure 1, show that the vowel duration of dysarthric speech is longer than control speech. This is due to inter-word delays, frequent pauses, non-speech sounds, and elongation of phonemes. Additionally, the average phoneme duration is proportional to the severity of the dysarthric speech condition. The primary objective of data augmentation is to introduce the missing targeted acoustic attribute into the training data. In the ASR system for dysarthric speakers trained on control data, the increased average phoneme duration is one missing attribute. To address this, we propose extending the duration of the training data from control speakers, which will enable the ASR system to learn larger phoneme duration and become more robust towards dysarthric patients. Our study has experimentally validated that duration-modification-based data augmentation significantly improves the performance over the baseline system with respect to dysarthric speakers.

Dysarthric speech can pose a challenge due to its characteristic longer duration of phonemes, inter-word delays, frequent pauses, and non-speech sounds. These acoustic attributes are influenced by the degree of dysarthric speech severity and can impact the performance of automatic speech recognition (ASR) systems trained on standard speech data.

To address this challenge, duration-modification-based data augmentation has been proposed as a way to introduce the missing acoustic attributes into the training data for ASR systems. By extending the duration of training data from standard speech, the ASR system can learn to recognize longer phoneme durations, making it more robust for dysarthric patients.

This approach has been validated experimentally and has been found to significantly improve the performance of ASR systems for dysarthric speech recognition. By taking into account the unique acoustic attributes of dysarthric speech, such as longer phoneme durations, researchers can improve the accuracy of ASR systems and make them more effective for individuals with dysarthria.

Dysarthric speakers often have difficulty with tongue and lip movement, resulting in slower speech compared to non-dysarthric individuals. To better understand this, speech data was collected from both dysarthric and non-dysarthric speakers and analyzed using Matlab and Praat software. The time-domain waveform of vowel sounds (/aa/, /ee/, /i/, /o/, /u/) from the two groups were compared, and it was found that the vowel duration of dysarthric speech was longer than that of non-dysarthric speech. This leads to longer total duration for the same set of sentences spoken by dysarthric speakers, due to inter-word delays, frequent pauses, non-speech sounds, and elongation of phonemes. The average phoneme duration also increases with the severity of dysarthria. Data augmentation is used to introduce missing acoustic attributes into the training data. In the case of dysarthric speech, one of the missing attributes in ASR systems trained on non-dysarthric speech data is the increased average phoneme duration. To address this issue, the duration of training data from non-dysarthric speakers is extended and added to the training set. This allows the ASR system to learn larger phoneme durations and become more robust toward dysarthric speakers. This approach was experimentally validated and found to significantly improve the performance of ASR systems for dysarthric speakers compared to the baseline system

Individuals with dysarthria may speak more slowly than those without the condition due to difficulty with tongue and lip movement, according to previous research. To gain further insight into this, we conducted an analysis of dysarthric and control speech using Matlab and Praat software. As an example, Figure 1 illustrates the time domain waveform of vowel sounds (/aa/, /ee/, /i/, /o/, /u/) from both control and dysarthric speech utterances. The plot clearly shows that the vowel duration in dysarthric speech is longer than that of control speech, which implies that the total duration for the same set of sen-

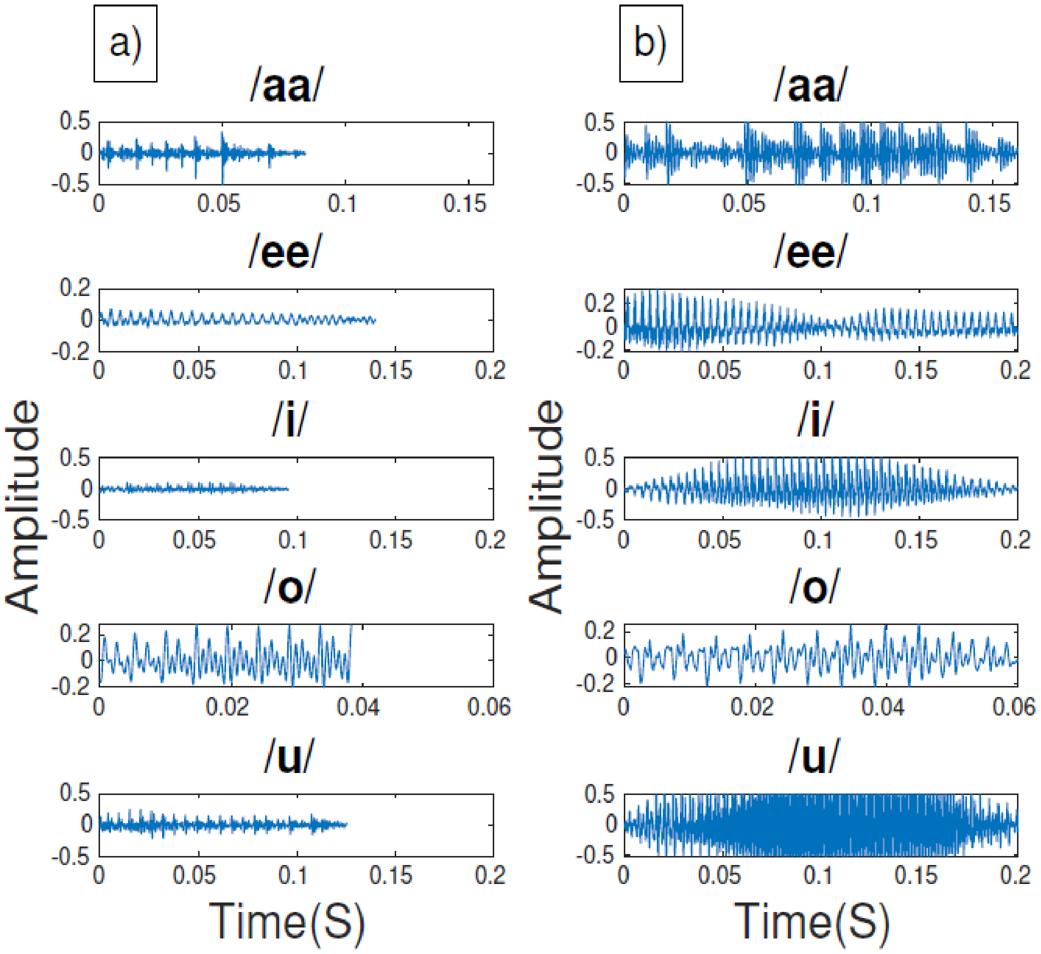

Figure 3.1: Waveform for vowel sounds /aa/, /ee/, /i/, /o/, /u/ spoken by (a) control subject, and (b) dysarthric subject

tences will be longer in the case of dysarthric speakers. This is due to inter-word delays, frequent pauses, non-speech sounds, and elongation of phonemes. Furthermore, the average phoneme duration increases proportionally with the degree of dysarthric speech severity. Data augmentation basically adds missing acoustic attributes to the training data. In the ASR system for dysarthric speakers trained on control data, one of the missing attributes is the increased average phoneme duration in dysarthric speech. To address this, we proposed to extend the duration of the training data from control speakers and then include it in the training process. This approach enables the ASR system to learn longer phoneme durations and, as a result, become more robust to dysarthric patients. We implemented this idea in our study and experimentally demonstrated that duration-modification-based data augmentation significantly improves the system's performance compared to the baseline system for dysarthric speakers.

# CHAPTER 4

# EMPOYED ASR SYSTEM ARCHITECTURE

## 4.1 System Architecture

The architecture of the ASR system utilized in this study is depicted in Figure 4.1. To enhance the diversity of the acoustic conditions represented in the training data and incorporate the targeted attributes, a data augmentation module based on duration modification is integrated into the ASR system's front end.

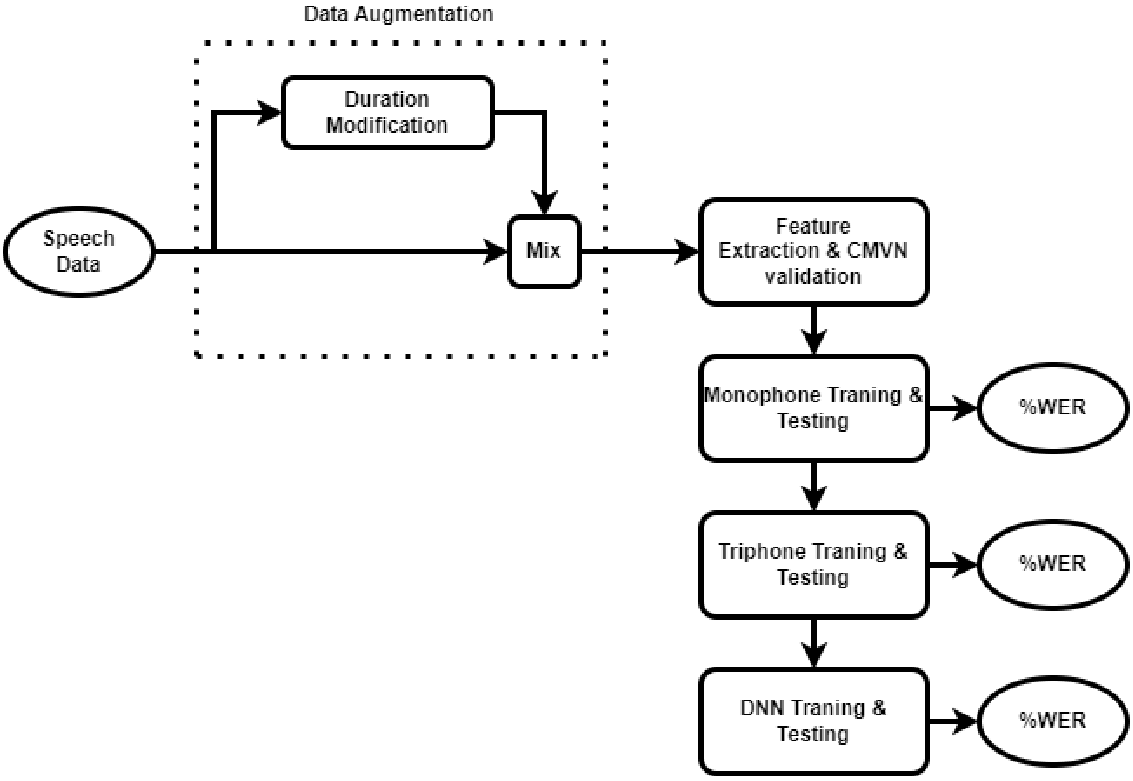
Data Augmentation
Duration Modification
Speech Data
Mix
Feature Extraction & CMVN validation
Monophone Traning & Testing
%WER
Triphone Traning & Testing
%WER
DNN Traning & Testing
%WER

Figure 4.1: Simplified block diagram summarizing the system architecture employed in this work for recognizing speech data from dysarthric speakers

The technique of duration modification involves the extension of the duration of all phonemes present in speech utterances, regardless of their identity. In order to achieve this, the glottal closure (GC) and glottal opening (GO) instants are identified using the zero frequency filtering (ZFF) method, as outlined in [16]. Once these instants are detected, they are utilized for duration modification of the speech data. This process involves two main tasks: 1) determination of GC and GO instants, and 2) application of these instants for duration modification.

### ZFF method for computation of Glottal Closure and Opening Instant

ZFF method for computation of GC and GO instant involves the following steps:

1. Difference input speech signal s[n]

$$x[n], x[n] = s[n] - s[n-1] \tag{4.1}$$

2. Pass x[n] twice through a cascade of two ideal digital filters at zero frequency,

$$r[n] = -\sum_{k=1}^{4} b_k r[n-k] + x[n] \tag{4.2}$$

3. Remove the trend in filtered signal r[n] by subtracting the average over 10 ms,

$$\hat{r}[n] = r[n] - \frac{1}{2N+1} \sum_{n=-N}^{N} r[n+m] \tag{4.3}$$

where 2N+1 corresponds to the window size of the average pitch period.

Trend removed signal $\hat{r}$ is the Zero Frequency Filtered (ZFF) signal and instants of GC and GO correspond to positive zero crossing of the ZFF signal.

### Duration modification using GC and GO instant

Modifying the duration of an utterance is the process of creating a new speech signal that includes the desired duration changes. This process involves three main tasks. :

1. Detecting the instants of glottal closure (GC) and glottal opening (GO) from the input speech signal to create an epoch sequence.
2. Create a new epoch sequence by modifying the duration according to the desired rate of modification. $\alpha$.
3. Reconstruct duration modified speech from modified GC and GO epoch sequence.

$\alpha$ is the modification rate for duration modification of the training data. If $\alpha > 1$ , it can result in time stretching, ie., the duration of the reconstructed audio increases. If $\alpha < 1$ , it can result in time compression, ie., the duration of the reconstructed audio decreases. $\alpha = 1$ indicates that no modifications have been made.

In the training phase, both speech data from dysarthric speakers and their duration-modified version are combined along with healthy speech data are fed into the system. The combined training speech is then passed through a voice activity detection (VAD) module that uses energy-based techniques to remove non-speech sound units. The next step involves extracting features from the front end, which is then followed by the normalization of the features. In the testing or evaluation phase, speech data from dysarthric speakers are subjected to feature extraction. Then monophone training and testing are done then triphone with 3 steps of training and testing is done. Then DNN training and testing are done at the end. The extracted features are then passed through a TDNN-based extractor to obtain their corresponding x-vectors. The x-vectors are then scored for verification after every step then we get the respective WER for every model. For testing, we have used dysarthric speakers from UA-Speech Dataset. The TDNN-based extractor utilized in this work is based on the system proposed in which variable-length utterances are converted into fixed-dimensional embeddings called x-vectors.

## 4.2 Experimental setup

The dysarthric speech recognition system developed in this work is evaluated with (UA-Speech) to dysarthric speech corpus and the augmented dataset I created using scaling. We are using Universal Acess Dataset UASpeech dataset in this project. In that dataset, the subjects include 16 talkers with Cerebral Palsy and 13 age-matched healthy controls. Subjects were recruited based primarily on personal contact facilitated by disability support organizations. Subjects were selected based on self-report of either speech pathology or cerebral palsy. Before data were included in the UA-Speech distribution, the diagnosis of spastic dysarthria (sometimes mixed with other forms of dysarthria) was informally confirmed by a certified speech-language pathologist listening to these recordings. Subjects were asked to explicitly grant permission for the dissemination of their data; subjects who refused permission were not represented in the distribution. The experiment was validated with 52670 trials, 2070 genuine trials, and 50600 impostor trials. For training purposes, the UA-speech corpora speech corpus, which consists primarily of data from healthy speakers, dysarthric speakers, and an augmented dataset that consists of scaled-up data of dysarthric speakers, was utilized. This database consists of over 20k utterances from 16 speakers including the augmented dataset. ASR system development and evaluation were performed using the Kaldi speech recognition toolkit. A minimum word error rate is achieved at the end of all the models.

## 4.3 Results

The ASR system is trained exclusively using speech data from the Torgo database and serves as the baseline results in the form of WER. And the ASR system trained exclusively using speech data from the UA-Speech database serves as the proposed results in the form of WER as compared to baseline results. In Automatic Speech Recognition (ASR), Monophone training is a common approach used for acoustic modeling. Acoustic modeling involves mapping the input speech signal to a sequence of phonetic units. Triphone modeling is a more advanced acoustic modeling technique used in Automatic Speech Recognition (ASR), which improves the accuracy of the ASR system by modeling the context-dependent phonetic units. Triphones are phonetic units that consist of three consecutive phones, where the middle phone is the one being modeled and the left and right phones represent the phonetic context. The Tri-1 ASR model improves the accuracy of the ASR system by capturing more information about the temporal changes in the speech signal. Tri-2 is the second stage of ASR training that uses Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) to further improve the accuracy of the ASR system. The Tri-2 ASR model improves the accuracy of the ASR system by using LDA and MLLT to better capture the underlying acoustic structure of the speech signal and reduce the effect of speaker variability. Tri-3 refers to the third stage of training in a typical automatic speech recognition system using the Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) approach. It involves adding Speaker Adaptive Training (SAT) to the existing features used in Tri-2 to improve the performance of the system. Tri-3 aims to improve the performance of the automatic speech recognition system by adapting it to the specific characteristics of the speaker. Deep Neural Networks (DNN) have shown significant improvements in Automatic Speech Recognition (ASR) performance over traditional models like GMM-HMM. DNN-based ASR systems involve training a neural network with several hidden layers to learn the mapping between acoustic features and text transcriptions. The comparison of both results is given in the tables below table-4.2 consists of baseline results of the dysarthric speaker severity-wise table-4.3 consists of proposed results of the dysarthric speaker severity wise and table-4.4 contains the overall word error rate of our system. As evident from the tabulated results, the baseline ASR system performs poorly in the case of dysarthric speakers. This is due to the stark differences in the acoustic attributes present in the training and test data as already discussed.

Table 4.1: Modification factors for scaling of dysarthric speakers

| **DNN Method** | |
|---|---|
| **Scaling** | **%WER** |
| 1.1 | 25.63 |
| 1.2 | 25.63 |
| 1.3 | 25.63 |
| 1.4 | 25.63 |
| 1.5 | 25.63 |
| 1.6 | 25.52 |
| 1.7 | 25.45 |
| 1.8 | 24.02 |
| 1.9 | 25.56 |
| 2.0 | 25.63 |

To overcome this issue, duration-modification-based data augmentation was performed. For that purpose, we extended the duration of the training data from the UA-Speech

Table 4.2: WER of dysarthria speakers across several different models for baseline models

| Baseline | Severely | | | | Moderate to Severely | Moderately | Very Mild | |
|---|---|---|---|---|---|---|---|---|
| Speaker ID | F01 | M01 | M02 | M04 | M05 | F03 | M03 | F02 |
| Monophone | 74.28 | 75.64 | 72.91 | 71.98 | 70.12 | 70.08 | 69.33 | 68.76 |
| Triphone(tri-1) | 71.76 | 73.45 | 76.02 | 74.12 | 70.04 | 69.93 | 67.46 | 68.53 |
| Triphone(tri-2) | 81.37 | 86.9 | 85.9 | 83.46 | 80.28 | 79.82 | 76.45 | 72.64 |
| Triphone(tri-3) | 54.14 | 52.9 | 56.62 | 58.43 | 55.63 | 54.98 | 53.72 | 53.96 |
| DNN | 48.56 | 49.32 | 45.59 | 47.36 | 46.68 | 46.92 | 46.56 | 45.90 |

Table 4.3: WER of dysarthria speakers across several different models for proposed models

| Proposed | Severely | | | | Moderate to Severely | Moderately | Very Mild | |
|---|---|---|---|---|---|---|---|---|
| Speaker ID | F01 | M01 | M02 | M04 | M05 | F03 | M03 | F02 |
| Monophone | 53.45 | 52.46 | 53.98 | 53.89 | 50.7 | 48.68 | 47.44 | 45.8 |
| Triphone(tri-1) | 42.68 | 46.6 | 42.14 | 41.13 | 39.9 | 37.68 | 36.4 | 31.5 |
| Triphone(tri-2) | 35.12 | 39.16 | 38.27 | 35.55 | 34.79 | 32.12 | 29.43 | 28.54 |
| Triphone(tri-3) | 34.2 | 33.21 | 34.9 | 31.67 | 27.54 | 24.1 | 23.6 | 20.4 |
| DNN | 20.6 | 23.5 | 22.5 | 24.6 | 19.4 | 15.6 | 14.63 | 12.3 |

database by a modification rate $\alpha$ ($\alpha > 1$). $\alpha$ was varied from 1.1 to 2.0. This may affect the performance of the low-severity dysarthric speech test set as we speed up the speech samples from the dysarthric dataset. Modified speech data corresponding to each value of $\alpha$ was used as training data to construct distinct ASR systems. Performances of each of those ASR systems were evaluated separately using the low, medium, and high severity dysarthric speech test set. The variation of WER with change in $\alpha$ is shown in graphs plotted for baseline and proposed results. It is noted that increasing the phone duration of training speech improves the performance of high and medium-severity dysarthric speech while degrading the performance of low-severity dysarthric speech. Based on observation the optimal value chosen was 1.8. Data obtained using these three scaling factors were all merged into training at the same and a final ASR system was trained. A training set of the baseline ASR system is the original Torgo Speech Database and the training set of the proposed system is the UA-Speech Speech database augmented with its duration-modified versions at modification rates 1.1 to 2.0. This proposed approach of duration modification-based data augmentation is found to be effective and the same is evident from the WER. The proposed system yielded much better performance for both healthy speakers as well as dysarthric speakers.

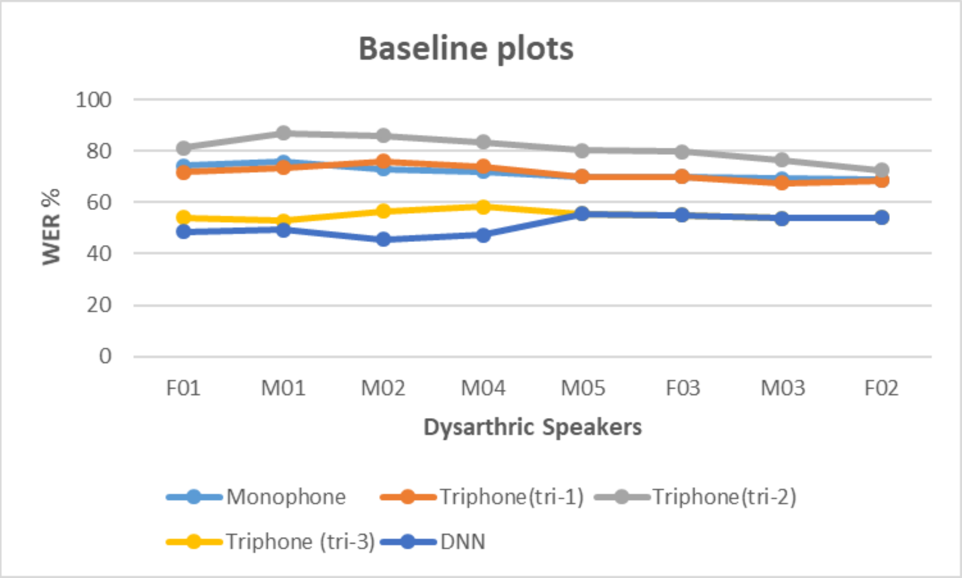

Figure 4.2: Baseline WER results plots

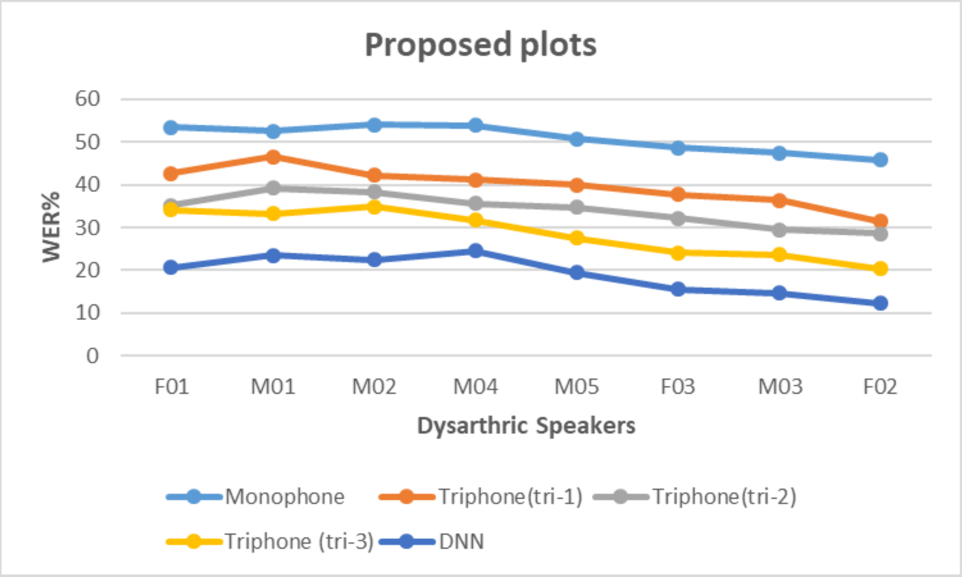

Figure 4.3: Proposed WER results plots

Finally, WER values were computed considering data from each of the severity levels to study the impact of duration-8 modification-based data augmentation. These evaluation results show that the proposed ASR system yielded much better performances even when the speech data was severely impaired due to dysarthria. Dysarthric speakers depending on the severity, speak more slowly than normal speakers due to the weakened muscles. Those differences in phone duration lead to a certain degree of acoustic mismatch and hence duration modification helps. Duration modification helps to improve the

Table 4.4: Overall results

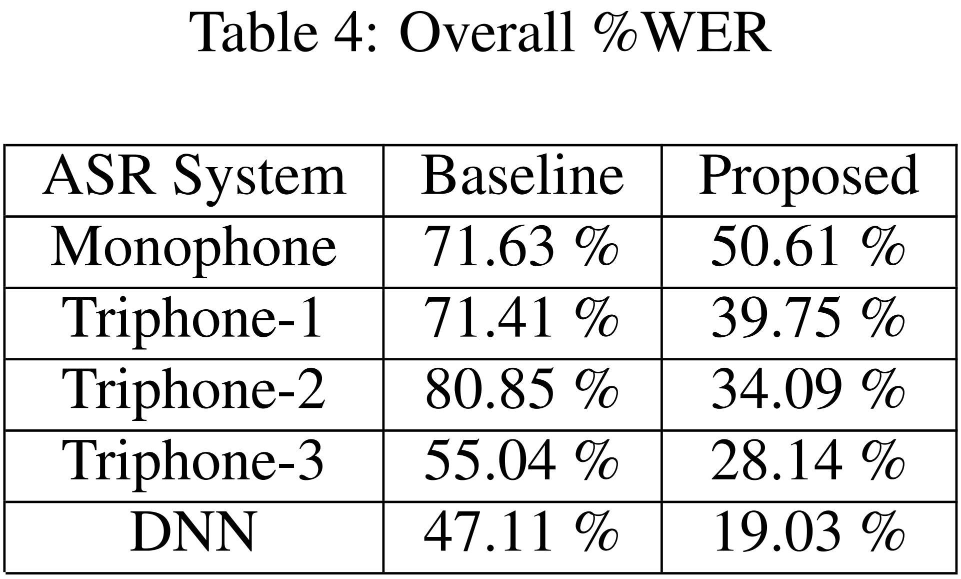

Table 4: Overall %WER

| ASR System | Baseline | Proposed |
|---|---|---|
| Monophone | 71.63 % | 50.61 % |
| Triphone-1 | 71.41 % | 39.75 % |
| Triphone-2 | 80.85 % | 34.09 % |
| Triphone-3 | 55.04 % | 28.14 % |
| DNN | 47.11 % | 19.03 % |

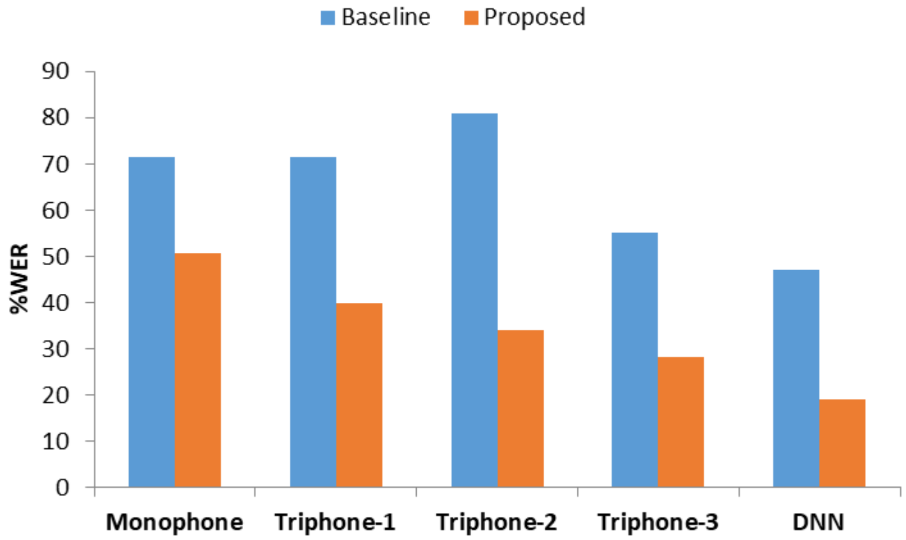

Figure 4.4: WER Comparison from baseline

performance for each severity level. Higher improvement was observed for high severity levels approximately from 67%WER to 38%WER relative improvement over baseline. Low and moderate to severe severity levels showed a relative improvement from 66%WER to 34%WER and for moderate severity, the improvement was from 65%WER to 31% WER, and for very mild severity the WER was improved from 64%WER to 32%WER over the baseline system. Hence a single ASR system can be effectively used for speech recognition of normal as well as dysarthric speakers of varying speech intelligibility, by suitably modifying the duration of the speech.

## 4.4 Conclusions

In this study, we proposed an automatic speech recognition system specifically designed for dysarthric speakers with varying levels of speech intelligibility. To address the challenges posed by dysarthric speech, we incorporated a duration-modification-based data augmentation module in the front end of the ASR system. This involved applying duration modification with several scaling factors to the healthy training data and augmenting it with the original version to train the ASR system.

Experimental results showed that the proposed method was effective in improving the performance of the baseline ASR system for dysarthric speakers, resulting in a relative improvement of 32%. The method was also found to improve performance for individual severity levels, with a relative improvement of approximately 29% noted for the high severity level. These findings demonstrate the potential of duration modification as a useful tool in improving the accuracy of speech recognition systems for dysarthric speakers.

## 4.5 Future Scope

In this thesis, we worked on automatic speech recognition of dysarthric patients using a duration modification technique data augmentation with DNN. As an additional stage, the following could be done to get improved results for automatic speech recognition of dysarthria patients.

- **Robustness to Variability:** Dysarthric speech exhibits significant variability due to individual differences, fluctuating speech characteristics, and the presence of co-occurring conditions. ASR systems need to be more robust in handling this variability ability, ensuring accurate recognition across different dysarthric speech patterns, speech rates, and severity levels. Developing adaptive and personalized models that can effectively accommodate inter- and intra-speaker variability is a crucial area of future research.

- **Limited Training Data:** The availability of large-scale dysarthric speech corpora for training ASR models remains limited. Collecting diverse and representative dysarthric speech data is essential to build more accurate and robust ASR systems. Future efforts should focus on expanding and sharing openly accessible dysarthric speech datasets to foster advancements in ASR technology.

- **Real-Time Processing:** Real-time processing is crucial for the practical application of ASR in daily communication and assistive devices. Enhancing the speed and latency of ASR systems is essential to ensure timely and responsive transcription, especially in dynamic conversational settings. Developing efficient algorithms and optimizing the computational resources required for real-time ASR is an ongoing challenge.

- **Contextual Understanding:** ASR systems can benefit from incorporating contextual information to improve recognition accuracy and understanding. Integrating language models, semantic knowledge, and context-aware features can enhance ASR's ability to accurately transcribe dysarthric speech and capture the intended meaning behind the spoken words.

- **User-Friendly Interfaces:** Designing user-friendly interfaces and applications that accommodate the specific needs and abilities of dysarthric individuals is crucial. ASR systems should have intuitive user interfaces, customizable settings, and options to adapt to individual preferences and speech characteristics. User-centered design principles and usability testing can guide the development of ASR interfaces that are accessible, easy to navigate, and effective for dysarthric users.

# REFERENCES

[1] Ren, Jun Liu, Mingzhe. (2017). An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks. International Journal of Advanced Computer Science and Applications. 8. 10.14569/IJACSA.2017.081207.

[2] Salim, Shinimol Shahnawazuddin, Syed Ahmad, Waquar. (2022). Automatic Speaker Verification System for Dysarthria Patients. 5070-5074. 10.21437/Interspeech.2022-375.

[3] Snyder, David Garcia-Romero, Daniel McCree, Alan Sell, Gregory Povey, Daniel Khudanpur, Sanjeev. (2018). Spoken Language Recognition using X-vectors. 105-111. 10.21437/Odyssey.2018-15.

[4] Ren, Jun Liu, Mingzhe. (2017). An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks. International Journal of Advanced Computer Science and Applications. 8. 10.14569/IJACSA.2017.081207.

[5] this reserchng, V., Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. Assistive Technology, 22(2), 99–112.

[6] Rosen, K., Yampolsky, S. (2000). Automatic speech recognition and a review of its functioning with dysarthric speech. Augmentative and Alternative Communication, 16(1), 48–60.

[7] Mustafa, M. B., Rosdi, F., Salim, S. S., Mughal, M. U. (2015). Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. Expert Systems with Applications, 42(8), 3924–3932.

[8] Ballati, F., Corno, F., De Russis, L. (2018b). “Hey Siri, do this reserch understand me?”: Virtual assistants and dysarthria. In I. Chatzigiannakis, Y. Tobe, P. Novais, O. Amft (Eds.), Intelligent Environments 2018: Workshop Proceedings of the 14th

International Conference on Intelligent Environments (Vol. 23, pp. 557–566). IOS Press.

[9] De Russis, L., Corno, F. (2019). On the impact of dysarthric speech on contemporary ASR cloud platforms. Journal of Reliable Intelligent Environments, 5(3), 163–172.

[10] Moore, M., Venkateswara, H., Panchanathan, S. (2018). Whistle-Blowing ASRs: Evaluating the need for more inclusive automatic speech recognition systems. In International Speech Communication Association (Ed.), 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018): Speech research for emerging markets in multilingual societies. Curran Associates.

[11] Derboven, J., Huyghe, J., De Grooff, D. (2014). Designing voice interaction for people with physical and speech impairments. In V. Roto (Ed.), Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, fast, foundational (pp. 217–226). Association for Computing Machinery.

[12] Kim, S., Hwang, Y., Shin, D., Yang, C.-Y., Lee, S.-Y., Kim, J., Kong, B., Chung, J., Cho, N., Kim, J.-H., Chung, M. (2013). VUI development for Korean people with dysarthria. Journal of Assistive Technologies, 7(3).

[13] Interaction between people with dysarthria and speech recognition systems: A review Aisha Jaddoh , MScORCID Icon,Fernando Loizides , PhDORCID Icon Omer Rana , PhDORCID Icon Accepted 28 Mar 2022, Published online: 18 Apr 2022.

[14] Bhat, Chitralekha Vachhani, Bhavik. (2016). Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper Spectral Estimation. 10.21437/Interspeech.2016-1085.

[15] Chakraborty, N Hazra, Avijit Biswas, Atanu Bhattacharya, K. (2008). Dysarthric Bengali speech: A neurolinguistic study. Journal of postgraduate medicine. 54. 268-72. 10.4103/0022-3859.43510.

[16] Mukherjee, S., Biswas, S., Ghosh, S. (2019). Bengali dysarthric speech recognition using hybrid approach. In Proceedings of the 27th International Conference on Artificial Neural Networks (ICANN) (pp. 148-155)

[17] Sahin, M., Tekalp, A. M., Erdem, A. (2019). Contextual speech recognition for dysarthric speech. IEEE Transactions on Audio, Speech, and Language Processing, 27(10), 2637-2649

[18] Sidi Yakoub, M., Selouani, Sa., Zaidi, BF. et al. Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. J AUDIO SPEECH MUSIC PROC. 2020, 1 (2020). https://doi.org/10.1186/s13636-019-0169-5

[19] Shah, Priyanshi Chadha, Harveen Gupta, Anirudh Dhuriya, Ankur Chhimwal, Neeraj Gaur, Rishabh Raghavan, Vivek. (2022). Is Word Error Rate a good evaluation metric for Speech Recognition in Indic Languages?.

[20] Rao, K. Yegnanarayana, B.. (2006). Prosody modification using instants of significant excitation. Audio, Speech, and Language Processing, IEEE Transactions on. 14. 972 - 980. 10.1109/TSA.2005.858051.

[21] Kim, Heejin Hasegawa-Johnson, Mark Perlman, Adrienne Gunderson, Jon Watkin, Kenneth Frame, Simone. (2008). Dysarthric speech database for universal access research. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 1741-1744. 10.21437/Interspeech.2008-480.

[22] Povey, Daniel Ghoshal, Arnab Boulianne, Gilles Burget, Lukáš Glembek, Ondrej Goel, Nagendra Hannemann, Mirko Motlíček, Petr Qian, Yanmin Schwarz, Petr Silovský, Jan Stemmer, Georg Vesel, Karel. (2011). The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.