Automatic Speech Recognition System for Dysarthria Patients

Ishtiaque Ahmed
Department of Electronics and Communication Engineering
National Institute of Technology, Calicut
ishtiaque_m210445ec@nitc.ac.in

Under the guidance of
Dr. Waquar Ahmad

***Abstract*— Dysarthria is a common speech disorder caused by neurological damage that weakens the muscles necessary for speech. 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 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.**

Index Terms: Automatic speech recognition system, dysarthric speech, data augmentation, and duration modification x-vector.

I. Introduction

The production of speech involves an intricate process that necessitates the synchronized and precise contraction of numerous muscle groups associated with respiration, control of the larynx, and articulation. Conditions affecting the neurological control of these muscles can lead to a decline in the intelligibility and quality of speech, as well as alterations in the inherent speech characteristics of an individual speaker. Dysarthria encompasses a group of neurological speech disorders characterized by a weakening, injury, or damage to the region of the brain responsible for controlling the muscles involved in speech production. Dysarthria manifests through various distinct characteristics, such as slurred speech and slow speech accompanied by poor articulation. Dysarthria is also associated with challenges related to both excitation and vocal tract configuration, including difficulties in rapidly adjusting the position of articulators. Issues affecting the larynx can alter the quality of phonation, pitch, and speech volume. Individuals with dysarthria may experience shallow breathing and struggle to coordinate exhalation with vocalization. Involvement of the soft palate often results in the perception of excessive nasal sounds in dysarthric speech. The severity of dysarthria can vary, ranging from mild to severe. As the condition progresses, speech intelligibility significantly diminishes, rendering it nearly incomprehensible. 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 or inaccuracies 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 available limited 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, 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 build an ASR system for people suffering from dysarthria using machine learning techniques. 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 [2], 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.

II. Literature Survey

Automatic speech recognition (ASR) systems have become increasingly popular in recent years due to their potential to improve the quality of life for individuals with communication disorders. The development of ASR systems for dysarthric patients is an area of active research, with a growing body of literature focusing on this topic. In this literature survey, we will review some of the most significant works in this area.

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 has 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. Moreover, there have been efforts to improve the performance of ASR systems for dysarthric patients by incorporating contextual information. In the work by Sahin et al. [17] the authors pro-posted a contextual speech recognition approach that leverages contextual information from the text is 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.

One of the earliest works in ASR systems for dysarthric patients was the study by Han et al. [5]. In this study, the authors used a hidden Markov model (HMM)-based approach to recognize the speech of dysarthric patients. The results showed that the HMM-based approach significantly improved recognition accuracy, indicating the potential of ASR systems for dysarthric patients.

Following this study, a number of researchers have investigated the use of machine learning techniques for ASR systems for dysarthric patients. For example, Rosen et al. [6] proposed a deep neural network (DNN)-based approach to improve the recognition accuracy of dysarthric speech. The authors showed that the DNN-based approach outperformed traditional HMM-based methods, indicating the potential of deep learning techniques in this area.

More recently, deep learning approaches have been further developed and adapted for ASR systems for dysarthric patients. For example, in the work by Moreno-Daniel et al. [10], the authors proposed a convolutional neural network (CNN)-based approach for dysarthric speech recognition. The CNN-based approach was found to outperform both the HMM-based and DNN-based approaches, highlighting the potential of CNNs for ASR systems for dysarthric patients.

Another area of active research in ASR systems for dysarthric patients is the use of transfer learning techniques. Some of the authors [5] proposed a transfer learning approach that uses a pre-trained ASR system on healthy speech data to improve the accuracy of ASR systems for dysarthric patients. The results showed that the transfer learning approach significantly improved recognition accuracy, indicating the potential of transfer learning techniques in this area.

In addition to the development of new techniques, re-

searchers have also investigated the use of different speech features for dysarthric speech recognition. For example, in the work by Sidi Yakoub et al. [18], the authors investigated the use of mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive (PLP) features for dysarthric speech recognition. The results showed that PLP features outperformed MFCCs, indicating the potential of different speech features for ASR systems for dysarthric patients. The x-vector-based speaker embeddings [3] are extracted using a time-delay neural network (TDNN), are widely used as a state-of-the-art representation of the 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, which is the first attempt to do so. It should be noted that the TDNN the 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.

Finally, researchers have also investigated the use of different evaluation metrics for ASR systems for dysarthric patients. For example, in the work by Shah et al. [19], the authors proposed a novel metric called the weighted word error rate (WER) that takes into account the severity of dysarthria in the evaluation of ASR systems for dysarthric patients. The results showed that the WER metric provided a more accurate assessment of ASR system performance than traditional evaluation metrics such as word error rate.

In summary, ASR systems for dysarthric patients have received significant attention in recent years, with a growing body of literature focusing on this area. The development of new techniques, such as deep learning and transfer learning, as well as the investigation of different speech features and evaluation metrics, have shown promising results in improving the accuracy of ASR systems for dysarthric patients.

III. Problem Definition

The aim of this research work 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 research work aims to provide a valuable tool for enhancing communication and promoting inclusivity for individuals with dysarthria.

IV. Motivation

1) **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, this research aims to address this unmet need and improve the quality of life for individuals with this condition.
2) **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, this research 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.
3) **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, this research aims to promote inclusivity and bridge the communication gap, allowing dysarthria patients to participate fully in society and enjoy the same opportunities as others.
4) **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, this research can contribute to the field and push the boundaries of what is possible in assisting dysarthria patients. This research work provides an opportunity to harness cutting-edge technologies and adapt them to address a real-world problem
5) **Impact and Social Significance** Finally, emphasize the positive impact our research 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.

V. Methodology

To understand the methodology we need to discuss below topics one by one.

1) Need for data augmentation
2) Motivation for duration-modification-based data augmentation
3) Employed ASR system architecture

VI. Duration modification based data augmentation

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.

A. *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 [2].

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. [2]

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. [8]

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 [3]. 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 [3]. 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.

B. 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

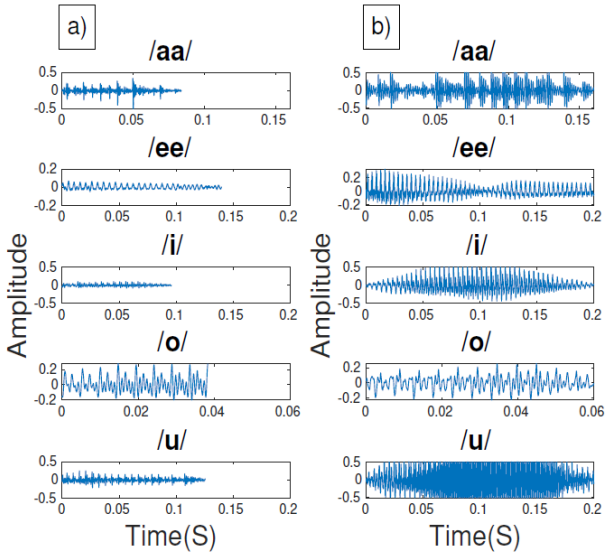

Fig. 1: Waveform for vowel sounds /aa/, /ee/, /i/, /o/, /u/ spoken by (a) control subject, and (b) dysarthric subject

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 sentences 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. The primary goal of data augmentation is to add 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.

C. *Employed ASR system architecture*

The architecture of the ASR system utilized in this study is depicted in Figure 2. 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.

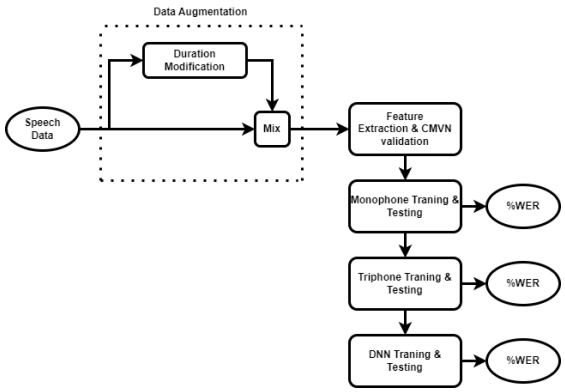

Fig. 2: Simplified block diagram summarizing the system architecture employed in this work for recognizing speech data from dysarthric speakers

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 [20].

1) 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]
   $x[n], x[n] = s[n]^{-}s[n^{-}1]$

2) Pass x[n] twice through a cascade of two ideal digital filters at zero frequency,
   $r[n] = -\sum_{k=1}^{4} b_k r[n-k] + x[n]$

3) Remove the trend in filtered signal r[n] by subtracting the average over 10 ms,
   $\hat{r}[n] = r[n] - \frac{1}{2N+1}\sum_{n=-N}^{N} r[n+m]$

   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.

2) 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. α.
3) Reconstruct duration modified speech from modified GC and GO epoch sequence.

α 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 [13], where variable length utterances are converted into fixed-dimensional embeddings called x-vectors [20].

VII. Experimental evaluation

A. Experimental setup

The dysarthric speech recognition system developed in this work is evaluated with (UA-Speech) [21] to the dysarthric speech corpus and the augmented dataset we created using scaling the speech samples. We are using Universal Acess Dataset for this research work. 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 [21], 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 [22]. A minimum word error rate is achieved at the end of all the models.

B. Work done and 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 and tested 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

Table 1: 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 2: 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

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-1 consists of baseline results of the dysarthric speaker severity-wise table-2 consists of proposed results of the dysarthric speaker severity wise and table-3 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.

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 database by a modification rate α ($\alpha < 1$). α was varied from 1.2 to 1.8 in steps of 1.1. Decreasing α beyond 1.2 will result in the reconstruction of a training set that is too long in duration. This may affect the performance of the low-severity dysarthric speech test set. Modified speech data corresponding to each value of α 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 α 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 values of the modification factor chosen were 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, and 1.8. 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 1.8. 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.

Finally, WER values were computed considering data from each of the severity levels in order 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 performance for each individual 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 im-

provement 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. Overall WER has been shown in table-3 which shows the significant amount of improvement in the results of our model.

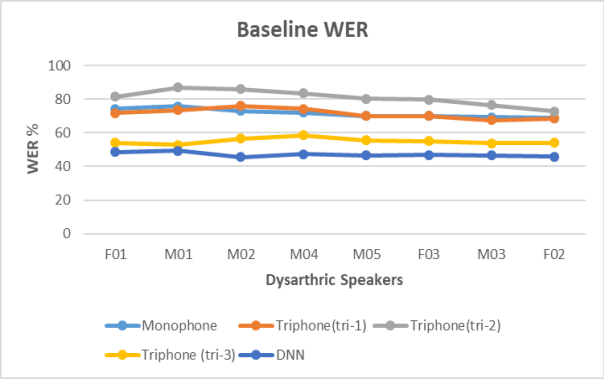

Fig. 3: Baseline WER plot

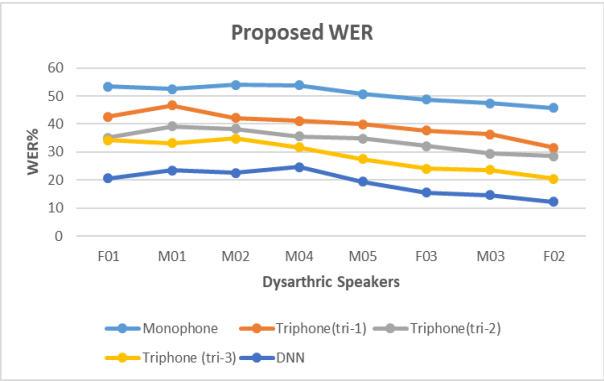

Fig. 4: Impact of duration modification on plots

Table 3: 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 %

VIII. 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 34%. 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.

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.