NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY POLITEHNICA BUCHAREST

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

Doctoral School of Automatic Control and Computers

Ph.D.
THESIS

Sarkhell Sirwan Nawroly

ÎMBUNĂTĂȚIREA SISTEMELOR DE RECUNOAŞTERE A VORBIRII PENTRU VORBITORII CU DISARTRIE PRIN AUGMENTAREA DATELOR BAZATĂ PE ZGOMOTS

ENHANCING SPEECH RECOGNITION SYSTEMS FOR DYSARTHRIC SPEAKERS THROUGH NOISE-BASED DATA AUGMENTATION

THESIS COMMITTEE

Prof. Habil. Dr. Eng. Decebal Popescu
Politehnica Univ. of Bucharest — PhD Supervisor

BUCHAREST 2024

Abstract

Dysarthria is a speech disorder that affects articulation, speed, and prosody, making verbal communication challenging. Automatic speech recognition (ASR) can enhance accessibility for individuals with dysarthria, improving independence and economic opportunities. However, ASR development faces a significant hurdle—data scarcity, which limits the system's effectiveness.

This thesis proposes a noise-based data augmentation approach to address this challenge. Unlike traditional ASR models that treat noise as a disruptive factor, this research leverages low-frequency noise (e.g., from vehicles) to generate additional training data while preserving speech intelligibility. The study was conducted in three phases: first, analyzing noise characteristics suitable for augmentation; second, applying noise augmentation at specific signal-to-noise ratio (SNR) levels to train ASR models using deep neural network-hidden Markov model (DNN-HMM) architectures; and third, extending the method to sentence-level recognition. A category-based training approach grouped speakers by severity, and a two-stage transfer learning strategy further refined the models by mitigating acoustic variability and enhancing speaker-specific adaptation.

The proposed approach significantly improved ASR performance, especially for severe dysarthric speakers, where conventional models struggle due to acoustic disparities. By strategically augmenting dysarthric speech data, this research introduces a novel and effective method for ASR enhancement, paving the way for more inclusive speech recognition technology.

Table of contents

List of tables

List of figures

Chapter 1

INTRODUCTION

1.1 Introduction

Speech is the primary mode of verbal communication for human beings, relying on both a phonetic plan (the brain's formulation of the intended message) and a motor plan (the physical execution of that message using the speech organs). The motor plan involves several stages: initiation, where air is expelled from the lungs; phonation, in which the vocal folds vibrate to produce sound; resonance, as the sound travels through the throat and oral or nasal cavities; and articulation, where various sound units are formed based on the movement and positioning of the speech organs. Disruptions in any part of this motor plan lead to speech disorders like dysarthria, a neuro-motor condition that impairs clarity and intelligibility. Dysarthric speakers, therefore, face not only communication challenges but also limitations in education, employment, and social interactions due to their speech impairment.

1.2 Dysarthria

Dysarthria is a motor-speech articulation disorder that significantly impacts an individual's ability to produce clear and intelligible speech. It can be either developmental or acquired, often resulting from neurological conditions such as cerebral palsy (CP), stroke, traumatic brain injury (TBI), multiple sclerosis (MS), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), or brain tumors. These conditions damage the nervous system and impair the control of speech muscles, affecting the strength, coordination, and precision of articulation. The disorder weakens the motor control system responsible

for speech, leading to slow, imprecise, or uncoordinated articulatory movements, which in turn affects the clarity and natural rhythm of speech production.

One of the primary challenges associated with dysarthria is reduced intelligibility, making it difficult for listeners to accurately perceive spoken words. Due to impaired neuromuscular control, individuals with dysarthria may involuntarily substitute, delete, distort, or insert phonemes during speech, altering the intended message. For instance, as noted in previous studies [11], a person with dysarthria who struggles with lip and tongue coordination might attempt to say "tip," but it could be perceived as "sip" or "hip." Similarly, more complex words like "decide" could be misheard as "sigh" or "say" due to phonemic distortions and coarticulatory errors. Such inconsistencies in articulation create communication barriers, requiring listeners to rely heavily on context to understand the intended speech content.

The severity of dysarthria varies, ranging from mild slurring of speech to a complete inability to produce intelligible words. Based on the degree of impairment, dysarthria is generally classified into four categories: mild, moderate, moderate-to-severe, and severe. In mild cases, speech may be slightly distorted but still understandable, whereas in severe cases, speech can become almost unintelligible, requiring alternative or augmentative communication (AAC) aids. Some individuals experience speech rate reductions, monotonic pitch variations, or prolonged vowel durations, making their speech sound unnatural or robotic. These characteristics contribute to increased listener effort and often necessitate frequent repetition or clarification from the speaker.

To address their communication challenges, various speech therapy techniques and assistive communication technologies have been developed to enhance the communication abilities of individuals with dysarthria. Traditional approaches involve speech-language therapy, which focuses on improving articulation, breath control, and speech rhythm through repetitive exercises and training. However, for individuals with severe dysarthria, Augmentative and Alternative Communication (AAC) aids such as text-to-speech devices, voice amplifiers, and speech-generating applications have proven beneficial in facilitating effective communication. Recent advancements in automatic speech recognition (ASR) technology have also shown promise in adapting to dysarthric speech patterns, potentially offering more accessible solutions in the future. Understanding

the speech challenges faced by dysarthric individuals is crucial for developing more effective intervention strategies and fostering greater inclusive in communication.

This research aims to address this gap by developing augmented dysarthric speech data to improve ASR accuracy. By incorporating noise-based data augmentation techniques, this study seeks to enhance the robustness of ASR models in recognizing and processing dysarthric speech more effectively. Improved ASR systems can significantly enhance the usability of AAC aids, enabling individuals with dysarthria to communicate more clearly and efficiently. In turn, this can lead to improved social interactions, greater independence, and a higher overall quality of life for individuals affected by this condition. The following section will provide a detailed discussion on ASR systems.

1.3 Automatic speech recognition (ASR) system – an overview

Automatic speech recognition (ASR) system performs automatic conversion of speech-input to text-output. The performance of the ASR system is evaluated using Word Error Rate (WER).

$$Word\ error\ rate\ (WER) = \frac{D+S+I}{N} * 100\% \tag{1.1}$$

where, N is the total number of words, and D, S, and I are the number of deleted, substituted, and inserted words, respectively.

Modern ASR systems, especially those based on deep learning models such as Deep Neural Networks (DNNs), have achieved remarkable success in large vocabulary, continuous, speaker-independent speech recognition. These advancements are driven by the availability of massive datasets, robust machine learning algorithms, and computational power. Commercial ASR platforms like Google Assistant, Apple Siri, Amazon Alexa, and Microsoft Cortana demonstrate impressive performance under normal speech conditions, allowing users to interact seamlessly with technology. These state-of-the-art systems owe their success to advanced technology design and the quality, diversity, and scale of the speech data used for training, as well as speaker characteristics and the surrounding environment [17].

However, these commercial ASR systems typically perform well when trained on extensive datasets that represent a wide range of voices. When groups are underrepresented in training data such as individuals with speech disorders, these systems struggle to achieve accurate recognition. For instance, ASR models trained on thousands of hours of "normal" speech perform well for typical users but often fail to recognize individuals with medical conditions that affect speech [18]. While people with mild dysarthria, whose speech acoustically resembles that of normal speakers, may benefit from these systems, the moderate, moderate-to-severe, and severe dysarthric speakers—who need such systems the most—derive limited benefit. This is due to the greater acoustic variability in their speech, and the challenge of training ASR systems with reasonable accuracy given the limited availability of dysarthric speech data.

A major challenge in improving ASR for dysarthric speech is the scarcity of high-quality dysarthric speech datasets. Unlike typical speech data, which is widely available, dysarthric speech corpora are relatively limited, making it difficult to train robust models that generalize well across different severity levels. Additionally, the high variability in dysarthric speech—caused by differences in articulation, phonation, and prosody—further complicates recognition.

To address these challenges, research efforts have focused on augmenting dysarthric speech data through various techniques, including noise-based data augmentation, transfer learning, and model adaptation. By introducing augmented data into ASR training, researchers aim to improve model robustness and enhance recognition accuracy for dysarthric speakers. Just like normal speech data that is available in abundance, the noisy data is also a source that is in abundance. Hence, the solution involves using noise as a source for augmentation, where transformations are applied to the original dysarthric speech rather than normal speech data. This method allows for the creation of diverse training samples that can work effectively for both isolated word and sentence-level ASR tasks.However, it is essential to carefully select which portions of the noise will be used for augmentation, as dysarthric speech samples are inherently distorted. Adding further degradation could significantly hinder recognition performance. Therefore, this research focuses on analyzing the specific aspects of noise that are beneficial for our augmentation efforts. The detailed process of employing noise as a source for data augmentation will be discussed in the following sections and chapters.

1.4 Understanding Noise as a factor for data augmentation

Noisy data has emerged as a significant source for data augmentation in various speech processing applications, with substantial support in the literature [1]. The core idea of this research is to inject noise into existing training examples to create new speech samples. This process can be particularly advantageous for enhancing the performance of Automatic Speech Recognition (ASR) systems for dysarthric speakers, as it can effectively increase the size and diversity of the training dataset.

Noisy data is typically characterized by disordered or corrupted signals, which can impede accurate interpretation. When such noise is introduced into already distorted dysarthric speech, there is a risk of further degrading intelligibility and recognition performance. Studies have shown that excessive noise can obscure phonetic features critical for speech recognition, ultimately leading to a deterioration in performance metrics [5].

To counteract these challenges, it is essential to conduct a thorough analysis of the frequency characteristics of various noise types. Previous research has demonstrated that certain types of noise, such as white noise or colored noise, can have varying effects on speech intelligibility, depending on their frequency content [14]. By selectively incorporating noise that complements the frequency range of dysarthric speech, we can create a more balanced training set that maintains intelligibility. The use of noise for data augmentation not only increases the quantity of available training data but also introduces variability that can enhance the generalization capabilities of ASR models. This is particularly important for dysarthric speakers, as their speech can vary significantly due to the nature of their condition.

1.5 Objective of the Research

The primary objective of this research is to synthesize new dysarthric speech samples by introducing controlled noise into the original dysarthric speech, thereby expanding the dataset while preserving the unique characteristics of dysarthric speech. This approach seeks to effectively address the issue of data sparsity without degrading the recognition

performance, particularly for moderate-to-severe dysarthric speakers, where conventional methods often fail. By carefully analyzing and selecting noise types that align with the frequency characteristics of dysarthric speech, the research aims to develop a robust data augmentation framework. Further, the proposed approach has employed to check its effectiveness at both the isolated word and continuous dysarthric speech recognition aspects.

1.6 Research Contribution

The aim of this research is to develop a noise-based data augmentation approach for dysarthric speech signals for developing efficient dysarthric speech recognition systems. The key contributions of this research work are as follows:

- **Comprehensive Literature Review on Data Sparsity in Dysarthric Speech Recognition:** A thorough review of existing research was conducted to examine the challenges associated with dysarthric speech recognition, particularly in low-resource settings. The study explores various data augmentation techniques, transfer learning approaches, and other methodologies aimed at mitigating data sparsity. Additionally, different evaluation measures and frameworks for dysarthric speech recognition are analyzed to identify gaps and limitations in the field.

- **Utilizing Noise as a Data Augmentation Source:** This research uses noise as an abundant and versatile source for data augmentation. By carefully selecting noise based on its frequency characteristics, noise is used as a transformation tool to generate new dysarthric speech samples without degrading the speech quality.

- **Category-Based Learning Approach:** To handle low-resource data settings at continuous dysarthric speech acenario, dysarthric speakers are grouped into categories based on the severity of their condition. A transfer learning framework is applied, where ASR systems are trained on each category and then adapted to individual speakers, effectively addressing the data sparsity challenge.

- **Targeting the demand category of Dysarthria:** Unlike conventional data augmentation methods, which struggle with moderate-to-severe and severe dysarthric speakers, this research specifically addresses the needs of these categories. The

proposed approach ensures that new augmented data samples are effectively synthesized for severe cases, where data augmentation is crucial.

- **Creation of an Augmented Dysarthric Speech Database:** A new augmented dataset has been developed using noise-based augmentation techniques applied to a standard dysarthric speech corpus. This dataset preserves both the speech quality and the speaker's identity, ensuring that augmented speech remains representative of actual dysarthric speech while offering increased diversity for model training.

- **Comparative Analysis with Current Literature Trends:** To establish the novelty and effectiveness of the proposed approach, this research includes a critical comparative analysis with existing literature in dysarthric speech recognition. By evaluating state-of-the-art data augmentation and transfer learning techniques, this study highlights how the proposed method aligns with or surpasses current trends, providing a structured discussion on its strengths and limitations.

- **Unlimited Data Synthesis:** Leveraging noise as a primary source, the proposed method allows for the creation of an unlimited number of augmented speech examples. With proper noise analysis, this method ensures a consistent supply of new data, making it scalable for various use cases.

- **Adaptable to Any Speech Unit:** The proposed data augmentation technique is flexible enough to be applied to different speech units, including sentences, words, and syllables. This versatility leads to improved ASR performance across a range of dysarthric speech tasks, surpassing state-of-the-art augmentation methods that primarily focus on isolated words.

  These contributions collectively address the challenges of data sparsity in dysarthric speech recognition while ensuring a robust, scalable solution for improving ASR systems.

Chapter 2

SPEECH CORPORA USED FOR THE RESEARCH

2.1 Introduction

This chapter provides a detailed discussion of the dysarthric speech and noise corpora used in this research work. The selection of these databases plays a crucial role in determining the effectiveness of the proposed data augmentation methods. First, the chapter introduces the dysarthric speech datasets used, highlighting their characteristics, speaker demographics, speech unit types (isolated words, sentences, continuous speech), and availability. Since dysarthric speech datasets are inherently scarce, selecting appropriate corpora with diverse dysarthria severity levels is essential to ensure robust model training and evaluation.

2.2 Dysarthric Speech Corpora

In this research, two primary dysarthric speech corpora were utilized to facilitate comprehensive analyses of both isolated words and continuous speech patterns: the Universal Access (UA) Dysarthric Speech Corpus [4] and the Nemours Dysarthric Speech Corpus [10]. The UA corpus was employed for studies involving isolated word recognition, while the Nemours corpus provided data for continuous speech analysis. A detailed examination of each corpus is presented below, highlighting their development, structure, and significance in advancing dysarthric speech recognition research.

2.2.1 Universal Access (UA) Dysarthric Speech Corpus

The UA Dysarthric Speech Corpus [4] comprises speech data from 19 individuals diagnosed with cerebral palsy-induced dysarthria and 13 control speakers without speech impairments. Each dysarthric participant contributed a total of 765 isolated word recordings, encompassing:

- **300 Uncommon Words:** A single utterance of each word, selected to introduce variability and challenge to the dataset.
- **26 Radio Alphabet Letters:** Each letter was repeated three times, facilitating analysis of phonetic consistency and articulation.
- **19 Computer Command Words:** Commonly used commands, each repeated thrice, to simulate practical ASR applications.
- **10 Digits:** Numbers zero through nine, each repeated three times, essential for numerical data entry tasks.
- **100 Common Words:** Frequently used words, each repeated three times, to assess everyday speech recognition.

This structured approach resulted in a rich dataset of 765 words per speaker, meticulously recorded in a controlled laboratory environment to ensure high audio quality. The recordings were captured using an 8-element microphone array, with one element dedicated to synchronization purposes, and sampled at a rate of 48 kHz. For the current research analysis in the upcoming chapter 3, the uncommon words are not considered as they have only a single utterance. This setup was chosen to accurately capture the nuances of dysarthric speech, providing a reliable basis for subsequent analyses and ASR system training.

2.2.2 Nemours Dysarthric Speech Corpus

The Nemours Dysarthric Speech Corpus [10] contains recordings from 11 adult male speakers with varying degrees of dysarthria. However, for this research, only 10 speakers were considered, as phonetic transcriptions were unavailable for one speaker. The corpus

is designed to capture continuous speech, making it particularly useful for studying the challenges of recognizing dysarthric speech beyond isolated words.

Each dysarthric speaker in the corpus uttered 74 short nonsense sentences in English. These sentences were constructed using a controlled template:

- The sentence structure follows the pattern:
  "The **X** is **Y**ing the **Z**", where:
  X and **Z** are monosyllabic nouns **Y** is a disyllabic verb
- There are 37 unique sentences, and for the second set of 37 sentences, the nouns (X and Z) are swapped, while the verbs remain unchanged.
- The total set contains 37 verbs, which are used twice across the 74 sentences.

This structured approach ensures linguistic consistency while allowing researchers to analyze phoneme and word-level variations in dysarthric speech.

2.3 Noisy Data

For the noisy data the NOISEX-92 database is used. The NOISEX-92 database [15] is one of the most widely recognized noise datasets in speech processing research. Each noise sample in the NOISEX-92 database is sampled at 16 kHz, which is a standard sampling rate for speech processing applications. The dataset includes both real-world and synthetic noise sources, each representing different acoustic challenges.

For the current research, noise from NOISEX-92 was used as a source for dysarthric speech augmentation. Several distinct noise types were selected, each contributing unique acoustic properties that influence speech intelligibility and ASR performance. The following are the noise types available in the database:

- Factory noise
- Pink Noise
- Benz Noise
- Car Noise

- Bus Noise
- Volvo Noise

Beyond NOISEX-92, additional noise types were included from research sources, specifically golf noise and bus noise [3]. These noises were selected based on their distinct frequency characteristics and their potential impact on dysarthric speech augmentation. These include golf noise and bus noise.

2.4 Summary

This chapter provided a detailed overview of the speech corpora and noise datasets used in this research for dysarthric speech augmentation and recognition enhancement. The chapter discussed the two dysarthric speech corpora—UA Dysarthric Speech Corpus and Nemours Dysarthric Speech Corpus—which were used to analyze both isolated word-level and continuous speech-level dysarthria, respectively. Additionally, the chapter elaborated on the NOISEX-92 noise database and other external noise sources, which were utilized for noise-based data augmentation in dysarthric speech recognition.

Chapter 3

NOISE ANALYSIS - A SOURCE OF DATA AUGMENTATION FOR DYSARTHRIC SPEECH DATA

3.1 Introduction

Data augmentation is crucial for addressing data sparsity in dysarthric speech recognition, where obtaining sufficient training samples is challenging. While traditional augmentation methods focus on increasing data volume, maintaining speaker identity while ensuring diversity remains complex. This section explores an underutilized approach—using noisy data for augmentation.

However, noise inherently distorts speech and, if not carefully managed, can degrade recognition performance, especially for dysarthric speech, which is already impaired. The key challenge is to harness noise effectively without further compromising intelligibility.

This chapter examines how different noise types and frequency ranges can enhance ASR robustness while preserving dysarthric speech characteristics. Unlike conventional noise augmentation aimed at improving environmental robustness, this approach strategically injects noise to enrich the training set while retaining essential dysarthric features.

We systematically analyze noise sources, their frequency properties, and their suitability for augmentation. The next section details the dysarthric speech and noise database used in this study.

3.2 Noise Analysis for being a source for data augmentation

The analysis of using noise for data augmentation in dysarthric speech recognition systems, involves understanding how noise affects the speech signal. One crucial aspect of this analysis is the study of the dominant frequency components present in different types of noise. This analysis helps identify the noises that have the potential to either corrupt or enhance the training data based on their frequency characteristics. To analyze the frequency components of different noise types, Linear Prediction (LP) analysis [6] is employed. The analysis proceeds by plotting the histogram of dominant frequencies for various categories of noise as shown in Figure 3.1. This helps to understand the energy distribution across different frequency ranges for each noise type, which is important because noise that overlaps with speech frequencies has the potential to degrade the intelligibility of the speech signal.

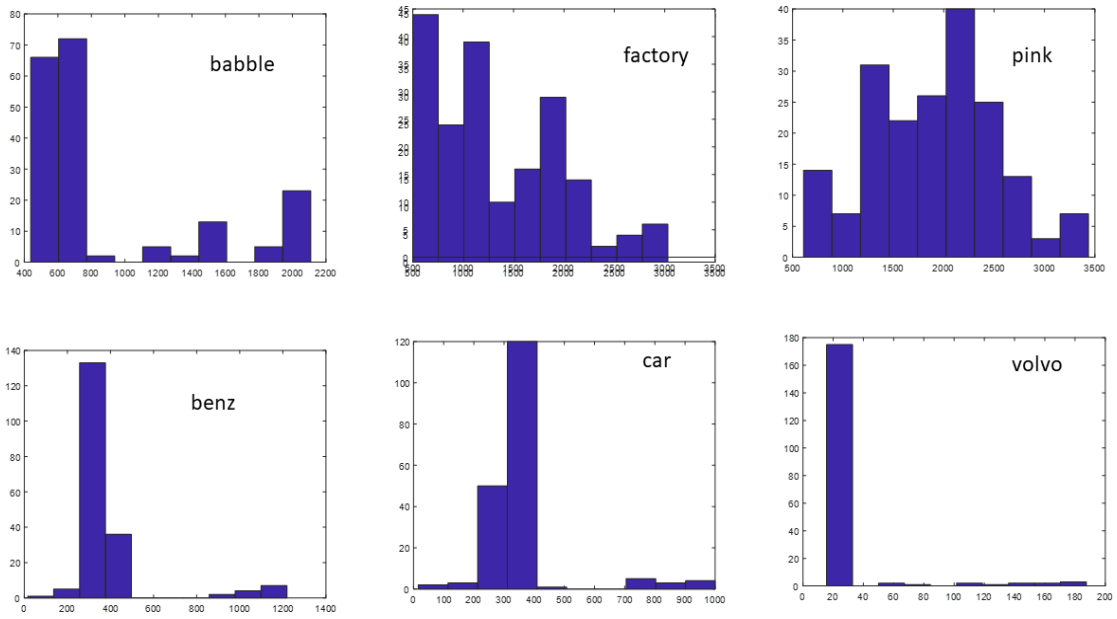

Fig. 3.1 Histogram plot of dominant frequency spectral components extracted for babble, factory, pink, benz, car and Volvo noise

From the frequency analysis in Figure 3.1, it is evident that low-frequency noises such as car noise, Volvo noise, golf noise, bus noise and Benz noise have dominant frequency components outside the primary speech frequency range (500 Hz to 4000 Hz). These noises, with lower energy in the critical speech regions, are less likely to mask the

speech signal and therefore may be more appropriate for augmenting dysarthric speech data. On the other hand, noises like pink noise, babble noise, and factory noise, which have dominant energy within the speech frequency range, are more likely to degrade the quality of the dysarthric speech during augmentation by masking key phonetic features. As a result, these noises might be less effective for this purpose unless carefully managed in terms of intensity and SNR.

By carefully selecting and analyzing noise types, it is possible to enhance dysarthric speech recognition systems through data augmentation, ensuring robustness without sacrificing the quality and intelligibility of the speech data.

3.3 Augmenting Noise to Dysarthric Speech

The analysis conducted on noise-augmented dysarthric speech highlights the impact of various noise types on different categories of dysarthric speakers. The aim of this analysis was to understand how noise, when introduced at a specific signal-to-noise ratio (SNR) of +5 dB [1], influences the quality of the dysarthric speech during augmentation and to evaluate the most suitable types of noise for creating new training samples for Automatic Speech Recognition (ASR) systems.

The noise types selected in the previous section were used to augment dysarthric speech data from the UA Speech Corpus discussed in Section 2.2.1, covering a range of dysarthric speakers from high, mid, low, to very low intelligibility. As an example, Figures 3.3 and 3.2 compare the effect of noise augmentation for a low-intelligibility dysarthric speaker (F02) when car noise and factory noise were introduced.

In this analysis, noise was added to the dysarthric speech at +5 dB SNR. This choice of SNR ensures that the augmentation focuses on transforming the data to create additional training samples rather than improving the intelligibility of the speech. The dominant frequency range of each noise type was used to analyze how different noises affect the original dysarthric speech data. The following are the steps involved in adding noise to the dysarthric speech data [7]:

1. **Selection of Clean Dysarthric Speech Signal and Noise Source:**

    - A clean dysarthric speech signal is selected from the available dysarthric speech dataset.

- A pure noise signal is extracted from a Noisex-92 corpus, which may includes the car, golf, train, and pink noises. The type of noise chosen depends on the augmentation strategy.

2. **Determining the Desired Signal-to-Noise Ratio (SNR):**

   - The level of noise added to the speech is controlled by setting a specific Signal-to-Noise Ratio (SNR), which determines how much the noise affects intelligibility.
   - A lower SNR (e.g., 0 dB) results in a heavily distorted signal, whereas a higher SNR (e.g., 20 dB) retains most of the original speech clarity.
   - The ITU-T P.56 standard is employed to measure the active speech level and assess the power of the clean speech signal, ensuring uniformity in the augmentation process.

3. **Random Selection of a Noise Segment:**

   - Since, we have already analysed the noises, a segment of noise can be randomly selected from the analysed noise data.
   - The duration of the noise segment is adjusted to match the duration of the clean speech signal.
   - Random selection ensures variability in the noise patterns, improving model generalization.

4. **Noise Power Calculation and Adjustment:**

   - The power (energy) of the selected noise segment is calculated.
   - The noise power is adjusted to achieve the predefined SNR.
   - The adjustment ensures that the noise level is appropriately scaled before mixing with the speech signal.

5. **Mixing Noise with the Speech Signal:**

   - The scaled noise signal is added to the clean speech signal to synthesize a noisy, augmented dysarthric speech sample.

- The final waveform is carefully checked to ensure that speech intelligibility is not overly degraded.

6. **Dataset Expansion and Storage:**

   - Multiple noisy versions of each dysarthric speech sample are generated using different SNR levels and various types of noise.
   - The augmented dataset is stored in a structured format, ready for use in ASR training and evaluation.

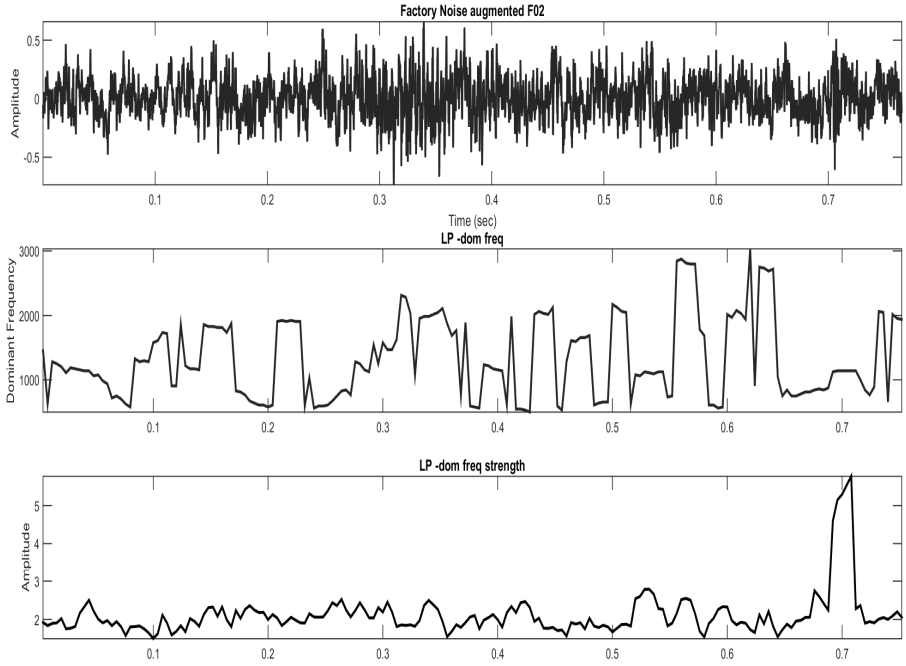

Fig. 3.2 Augmentation of factory noise on F02 - Low-intelligibility category of dysarthric speakers

This suggests that low-frequency noises such as car noise, volvo and benz are better suited for augmenting dysarthric speech, as they preserve more of the original speech characteristics, especially in the critical speech frequency range. On the other hand others are still a choice of transformation source by using them at the appropriate SNR levels which will be discussed in the upcoming sections.

To assess the impact and preservation of quality and speaker identity in noise-augmented dysarthric speech, a correlation analysis and speaker verification analysis are conducted.

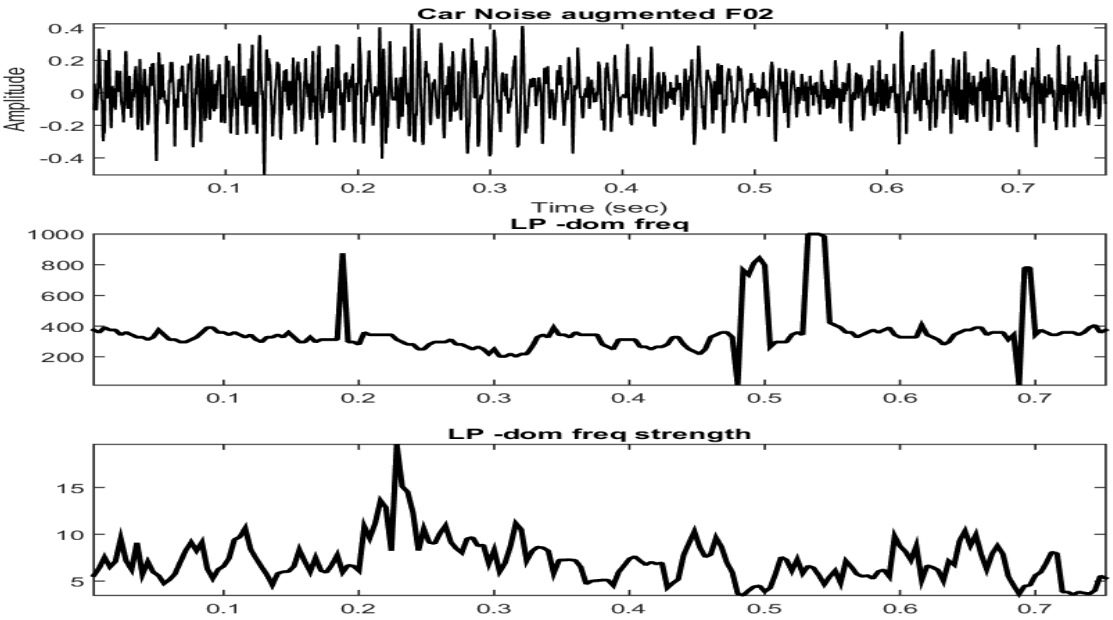

Fig. 3.3 Augmentation of car noise on F02 - Low-intelligibility category of dysarthric speakers

3.3.1 Correlation Analysis of Clean and Augmented Speech

Correlation analysis is a statistical technique used to measure the similarity between two signals. For dysarthric speech augmentation, it is essential to evaluate whether the augmented speech data retains the key acoustic features of the original dysarthric speech.

Correlation values were computed for each dysarthric speech sample before and after augmentation across different noise categories and SNR levels. The correlation coefficients were averaged for each noise type to obtain generalized trends. The correlation values are presented in Table 3.1 . It is seen that while high-frequency noises can still be useful for augmentation, they should be carefully managed by selecting their low-frequency segments to avoid excessive interference with speech components in continuous speech scenarios. The second level of analysis peroformed to check the speaker identity is to performa a speaker verification task that is discussed below.

3.3.2 Speaker verification analysis

To evaluate whether the noise-augmented dysarthric speech samples retain speaker identity, a speaker verification analysis was conducted. This analysis ensures that the augmented speech maintains speaker-specific characteristics and does not deviate significantly from the original dysarthric speech samples. For the evaluation, four

Table 3.1 Correlation Between Dysarthric Speech Data and Noise Augmented Dysarthric Speech Data

Category of Noise	**Correlation coefficient**
Pink	0.61
Factory	0.6
Babble	0.81
Car	0.88
Volvo	0.9
Benz	0.87

speakers from the UA Dysarthric Speech Corpus were selected, each representing different levels of speech intelligibility:

- High intelligibility: Speaker M09
- Mid intelligibility: Speaker F04
- Low intelligibility: Speaker M07
- Very low intelligibility: Speaker F03

A Gaussian Mixture Model (GMM)-based speaker verification system was trained to assess whether the noise-augmented samples still represent the respective speakers. 70% of the original training data was used for training the GMM model, ensuring that the system had sufficient speaker-specific information for verification. The remaining 30% of the data was reserved for testing.

Instead of using the original speech samples for testing, only noise-augmented versions of the test samples were provided to the trained speaker model. This classification approach allowed for an objective evaluation of whether the noise-augmented speech samples were still identifiable as belonging to the original dysarthric speaker.

The speaker verification results indicate that across high, mid, and low-intelligibility categories, the augmented speech samples remained close to the original in terms of speaker identity.

Based on the analysis, the following inferences can be drawn:

- Low-frequency noises such as car noise, Volvo noise, and benz noise are more suitable for augmenting dysarthric speech data. These noises do not overlap sig-

nificantly with the speech frequency range, thus allowing the speech components to remain relatively intact.

- High-frequency noises like factory noise, babble noise, and pink noise can still be used for data augmentation, but careful selection of their low-frequency components is necessary to prevent the masking of important speech characteristics. This is particularly important when working with continuous speech data, where high-frequency noise can have a more pronounced negative impact.

- The goal of noise augmentation in this context is not to improve the intelligibility of the dysarthric speech but rather to create transformed training samples that retain the identity and characteristics of the dysarthric speaker. This helps ASR systems to generalize better without losing the phonetic and prosodic features of the speaker.

- Since the augmented speech data is derived directly from dysarthric speech, the speaker identity and speech errors are still retained in the augmented data. This ensures that the augmented samples remain representative of the original speech patterns, which is crucial for training ASR systems tailored to dysarthric speakers.

3.4 Summary

This chapter analyzes the use of noise for data augmentation in dysarthric speech recognition. The approach was evaluated through correlation analysis and speaker identification tests, ensuring that augmented data preserves essential speech characteristics while introducing controlled variability. Results emphasize the importance of selecting low-frequency noise to maintain intelligibility and speaker identity.

A newly created augmented database, derived from a standard dysarthric speech dataset, is introduced. This enriched dataset helps address data sparsity while supporting robust ASR training. The findings set the stage for further evaluations, including analyzing WER in ASR systems trained on noise-augmented speech. This research ultimately aims to enhance dysarthric speech recognition for real-world applications.

Chapter 4

SNR-SELECTION-BASED-DATA AUGMENTATION FOR DYSARTHRIC SPEECH RECOGNITION

4.1 Introduction

Building on the previous chapter, this chapter explores augmenting dysarthric speech data with selected low-frequency noise at appropriate SNR levels. These noise types, which fall outside the critical speech frequency range (500 Hz–4000 Hz), minimally distort phonetic content, making them suitable for augmentation without compromising intelligibility.

SNR levels from +5 to +20 dB (in 5 dB increments) were chosen to balance noise and speech, ensuring intelligibility while introducing controlled variability. Additionally, this chapter evaluates augmentation effectiveness using key metrics and compares noise-based augmentation with other methods in the literature.

4.1.1 Augmentation Process

Based on the noise analysis performed in the previous chapter, the following low-frequency noise types were chosen for augmenting dysarthric speech data:

- Volvo noise
- Car noise

- Benz noise
- Bus-I noise
- Bus-J noise
- Golf noise

These noise types were selected because of their frequency distribution, which predominantly lies below the critical speech frequency range, making them ideal for data augmentation. The addition of these noise types simulates real-world acoustic environments without overwhelming the dysarthric speech characteristics.

The noise augmentation was carried out by adding each selected noise type to the dysarthric speech data at the frame level across the four different SNR levels. This method ensures that the noise is incorporated at every time frame of the speech signal, introducing variations that emulate real-world conditions where speech is often contaminated with background noise. Each dysarthric speech example was therefore augmented 24 times (6 noise types $\times$ 4 SNR levels), producing a diverse set of noisy speech samples for training.

To ensure that the augmented data are truly distinct and useful for training purposes, the features extracted from the augmented speech were analyzed and compared with the original speech features. The results, as shown in Figure 4.1, confirm that the feature sets from the noisy data are not mere copies of the original data but are instead sufficiently different to serve as unique training examples. The line graph in Figure 4.1 demonstrates that the features from the original speech and the augmented data do not overlap, indicating that the augmentation process successfully created new and diverse training samples.

One of the key advantages of this approach to data augmentation is its scalability. The number of augmented examples is not limited by the available noise types or SNR levels; it can be expanded further by introducing additional low-frequency noise types or by exploring a broader range of SNR levels. This flexibility allows for the creation of as many augmented speech samples as necessary to address the data sparsity issue in dysarthric speech recognition.

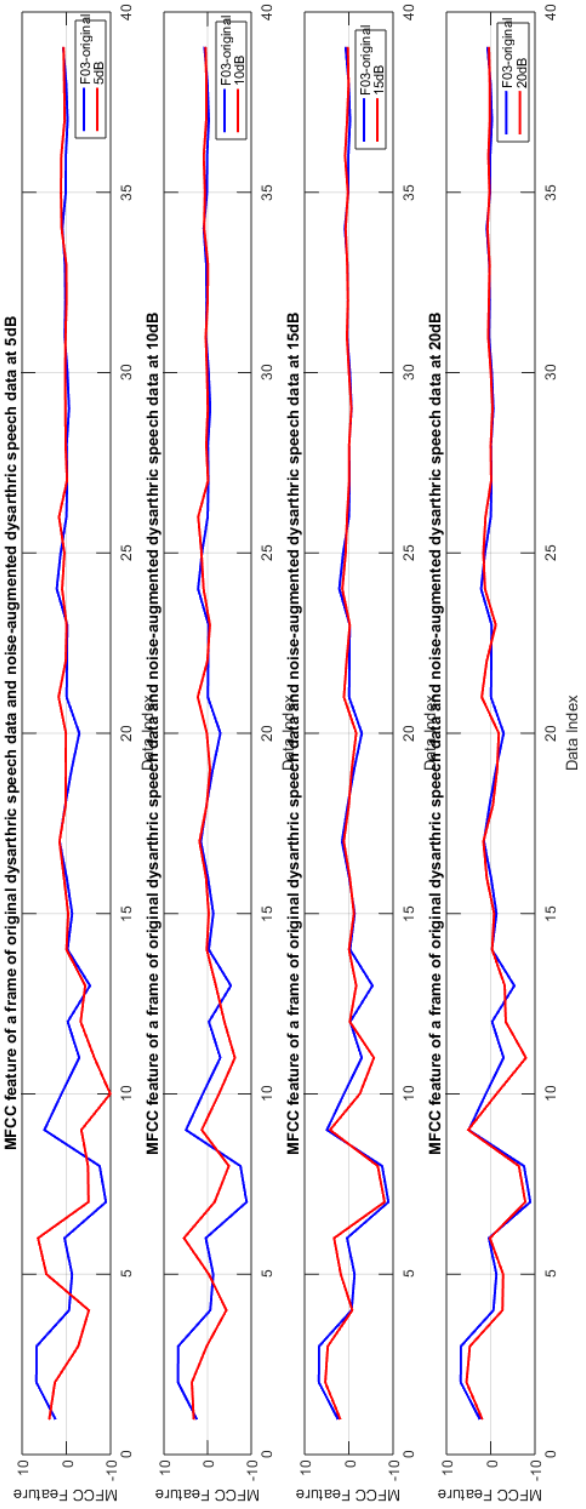

Fig. 4.1 Line Graph comparing the MFCC features of original dysarthric speech data with the features of noise augmented versions

4.2 DNN-HMM-based Automatic Dysarthric Speech Recognition System with Augmented Speech Data

In this section, we delve into the creation and performance evaluation of a DNN-HMM-based automatic dysarthric speech recognition system using augmented speech data. This system is built specifically for the UA dysarthric speech corpus discussed in Section 2.2.1 of Chapter 2, leveraging noise-augmented examples to address the inherent data sparsity in dysarthric speech datasets.

The augmentation process begins by using the noise selection strategy discussed in previous sections. Each dysarthric speech example is augmented 24 times, covering a variety of low-frequency noises and SNR levels from +5 dB to +20 dB. For the UA dysarthric speech corpus, which consists of three words per iteration, the third word and its augmented examples are reserved for testing, while the first two words and their corresponding augmented samples are used for training.

For each dysarthric speaker, the system trains with 50 dysarthric speech datasets, comprising 48 augmented examples and 2 original examples. For testing, the system uses 25 dysarthric speech datasets, with 24 augmented examples and 1 original example. This substantial increase in training examples allows the system to better learn and generalize from the diverse speech patterns and noise conditions, thus improving its robustness.

The feature extraction step employs 39 Mel-frequency cepstral coefficients (MFCC) feature vectors, including 13 static, 13 delta, and 13 acceleration coefficients. These MFCCs capture the spectral characteristics of dysarthric speech and are crucial for modeling phonetic variability across different levels of speech intelligibility. To normalize these features, Cepstral Mean Variance Normalization (CMVN) is applied, which removes variations caused by noise and channel distortions.

Following feature extraction, HMMs are trained for each word using the training examples. The number of HMM states is set according to the number of phones in a word, using two mixture components per state to capture acoustic variability. The HMMs serve as the foundational model for recognizing dysarthric speech, providing temporal modeling of the speech signal.

Once the HMMs are trained, a Deep Neural Network (DNN) is trained using the Kaldi Toolkit [12] to further improve the system's recognition capabilities. The DNN

is used in a hybrid DNN-HMM architecture, where the DNN estimates the posterior probabilities of HMM states. The input to the DNN consists of fMLLR-transformed features (Feature-Space Maximum Likelihood Linear Regression), which are commonly used to adapt the features to better match the speaker and the acoustic conditions.

The DNN architecture comprises five hidden layers, each employing the tanh activation function. The DNN is trained over 15 epochs for each dysarthric speaker.

4.2.1 Performance Evaluation (Word Error Rate - WER)

The performance of the DNN-HMM-based dysarthric speech recognition system is evaluated in terms of WER as given in equation 1.1, of Chapter 1. Table 4.1 shows the WER for each noise type used in the data augmentation process, as well as the WER for the baseline system (without augmentation).

The results indicate that noise augmentation significantly improves the recognition performance across all dysarthric speakers.

Table 4.1 WER performance of UA dysarthric speech recognition systems augmented with the SNR-based noise selection

Word Error Rate (WER) UA Dysarthric Speech Corpus								
Speaker ID	**Baseline without data augmentation**	**Car Noise**	**Benz Noise**	**Golf**	**Volvo**	**Bus-i**	**Bus-j**	**WER Including all 6 noises**
M08 (H)	5.97	5.19	5.45	4.21	5.1	5.89	5.12	4.21
M09 (H)	6.49	4.68	5.32	5.32	5.45	4.18	4.21	3.87
M10 (H)	2.34	1.17	0.97	0.95	0.69	1.9	2.01	0.98
M14 (H)	7.2	3.16	2.06	2.48	6.43	5.64	3.42	2.11
F05 (H)	4.42	2.4	2.08	3.46	2.14	3.55	2.32	3.12
F04 (Mod)	16.25	13.21	12.02	16.06	12.02	13.85	12.9	14.21
M05 (Mod)	18.7	12.12	16.82	12.21	16.82	13.46	10.35	14.12
M11 (Mod)	18.79	17.58	15.32	9.7	17.88	8.4	9.35	12.33
F02 (L)	5.45	3.33	4.29	5.12	5.12	4.05	5.41	3.12
M07 (L)	9.09	6.88	8.57	2.95	7.43	4.65	5.08	3.19
F03 (VL)	44.44	40.99	43.12	40.8	31.21	18.23	19.19	32.15
M04 (VL)	98.18	65.18	68.14	43.31	43.12	49.65	43.21	45.33
M12 (VL)	44.85	39.56	40.79	40.61	39.23	30.37	31.64	36.44

4.3 Summary

This chapter provides an in-depth exploration of noise augmentation for dysarthric speech, investigating the impact of six different noise types and four SNR levels. The findings demonstrate that noise augmentation serves as a valuable technique for increasing the diversity of training data while preserving speaker identity. However, the proposed method is not constrained to the selected noise types; additional low-frequency noises can be incorporated to further enhance the robustness of augmented speech data.

Furthermore, while the study considers SNR values up to 20 dB, it is observed that increasing SNR beyond this threshold offers negligible improvements. This is because, at higher SNRs, the differences between the augmented and original speech features diminish, reducing the effectiveness of augmentation in improving recognition performance. This highlights the importance of selecting an optimal SNR range to balance augmentation effectiveness and data quality.

Additionally, the analysis in this chapter is conducted exclusively on isolated words, providing valuable insights into how noise augmentation affects word-level dysarthric speech recognition. However, real-world applications of dysarthric speech recognition often involve continuous speech, where data sparsity and variability present more significant challenges. The next chapter expands on this work by addressing the complexities of continuous dysarthric speech, exploring methods to mitigate data sparsity and enhance recognition performance in more naturalistic speech scenarios.

Chapter 5

CATEGORY-BASED AND TARGET-BASED DATA AUGMENTATION FOR DYSARTHRIC SPEECH RECOGNITION USING TRANSFER LEARNING

5.1 Introduction

This chapter focuses on handling data sparsity problems in continuous dysarthric speech scenarios. The intuitive concept for this chapter is to combine successful and the noise-based data augmentation at the dysarthric category level to cater the low data sparsity conditions in continuous dysarthric speech recognition systems. This section uses category-wise transfer learning technique. The following section describes the continuous dysarthric speech database used for the analysis.

5.2 Data Augmentation Techniques for the Two-Stage Transfer Learning Approach

This study validated the experiments using the Nemours dysarthric speech corpus [10] discussed in Chapter 2 section 2.2.2. This work uses two data augmentation approaches

to train the source and target TL model. Both approaches utilize the dysarthric speech data itself to synthesize new examples. Data augmentation through noise is performed at the first level, and data augmentation through the VM-MRFE approach is performed at the second level for the target model. For any target dysarthric speaker, the source model is decided based on the category of the dysarthric speaker. If the dysarthric speaker is from a severe category, then the data for source model training includes all the dysarthric speech data from the severe category and its noise augmented examples. The corresponding target data for training the target mode is the dysarthric speech of the target speaker itself augmented using VM-MRFE approach as shown in Figure 5.1. Hence, initially category-wise models are trained leaving the target dysarthric speaker.

For the training data in both the cases (source & target) 37 utterances from the Nemours dysarthric speech corpus are used, with the remaining 37 reserved for the testing. Data augmentation methods are implemented on the training dataset. The source model is trained category-wise (mild, moderate, and severe) using noise-based data augmentation. This choice is rooted in the fact that it pre-trains the initial layers of the neural network based on the common characteristics inherent to the noise-based data-augmented examples. The shared characteristics specific to a category (mild, moderate, and severe) in the noise-based data augmentation method are primarily centered on the acoustic information of each category rather than the diverse noise characteristics. This is because, in this approach, low-frequency analysis-based noises are specifically added to the dysarthric speaker, as opposed to introducing generic noise. Consequently, the impact and nature of influence on each original example vary, precluding it from being a generalizable feature. The following sub-section discusses dysarthric speech augmentation using noise as a source.

5.2.1 Stage 1: Noise-based data augmentation approach for source model training

Noise categories such as "Volvo," "Golf," "Car," "Benz," "Babble," "Train," "Bus-i," and "Bus-j" are chosen for noise-based data augmentation. In contrast to the previous chapter, noises such as "train" and "babble" are also being considered in this work. This is because the source model is simply a pre-trained version specific to the category of the target dysarthric speaker rather than a speaker-dependent model used for dysarthric

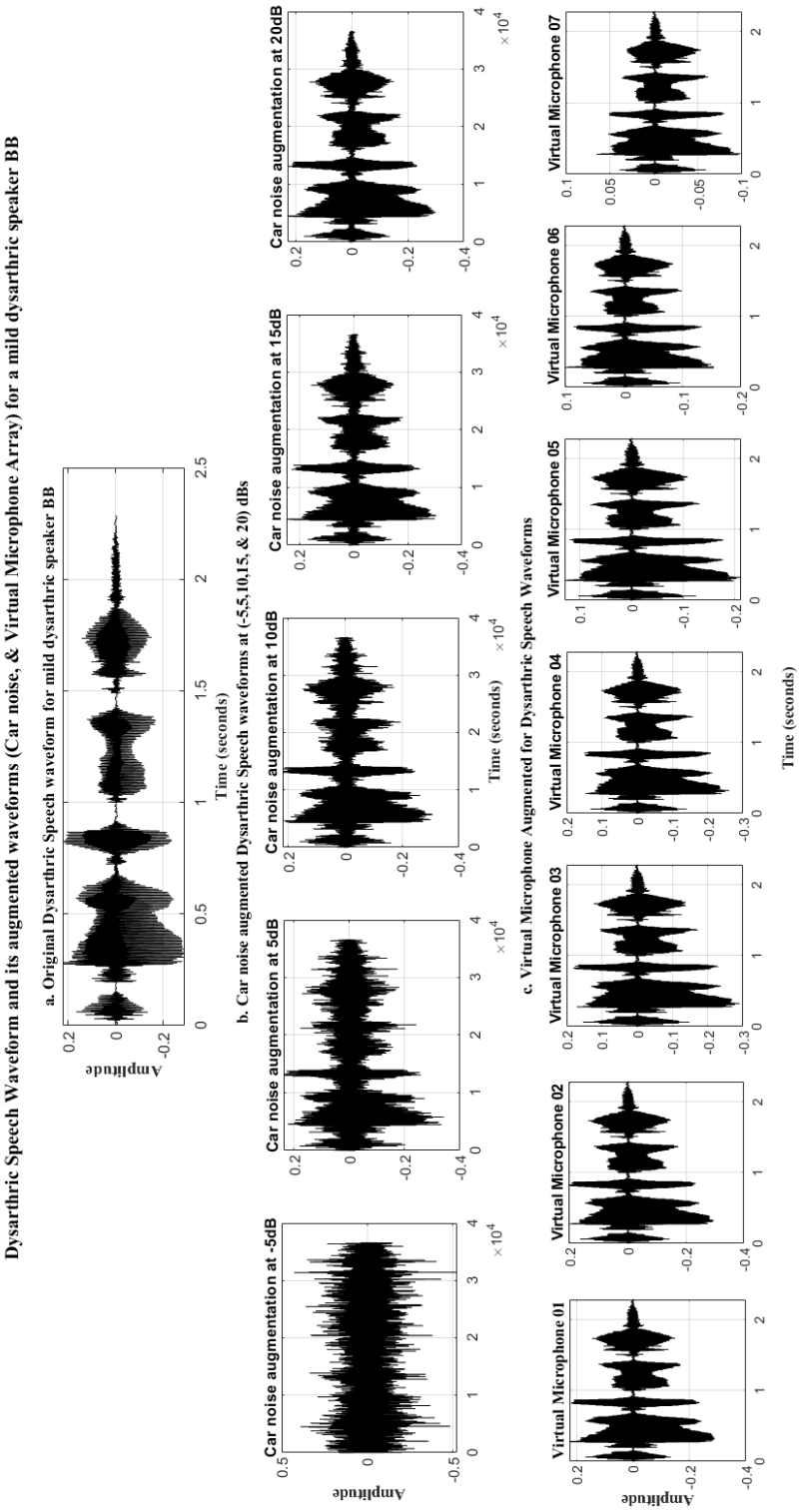

Fig. 5.1 Dysarthric speech signal for a mild dysarthric speaker BB and its augmented versions

speech recognition. SNR dB levels ranging from -5dB to +20dB in steps of 5 were used for augmenting noise data to dysarthric speech data. Eight different noise conditions across five SNR dB levels (-5, 5, 10, 15, and 20dB) were applied, and the noise data was integrated with dysarthric speech data at a frame-by-frame level. Figure 5.1(a) shows the original dysarthric speech signal for the speaker BB with mild dysarthria, and 5.1(b) shows the augmented version with noise data for the same speaker. It's apparent that the augmented example is not simply a copy of the original but also represents a new instance of speech by the dysarthric speaker. A line graph is generated in Figure 5.2 to compare the characteristics of the original dysarthric speech data with its noise-augmented versions at different SNR levels. This graph illustrates the variations in speech data across different Signal-to-Noise Ratio (SNR) levels, highlighting how noise impacts speech characteristics. This confirmed that the two sets of data are distinct from each other. Figure 5.2 clearly shows that the original speech data and the noise-augmented speech data have unique features, meaning each example can be treated as a new training example. Each dysarthric speech example was augmented with noise data across various SNR ranges, resulting in 40 additional unique examples (8 types of noise and 5 dB levels). Therefore, after augmentation, each word has a minimum of 80 examples. The noise-based augmentation approach is quite flexible regarding the number of augmented examples, as it is not limited to a specific number. In contrast to the previous chapter, this study demonstrates the flexibility of incorporating additional low-frequency noises for augmentation by utilizing different noises. This noise-based dysarthric speech augmentation technique is applied to train the source model. The research incorporates data from three categories of dysarthric speakers: individuals with mild, moderate, and severe symptoms. Training the source model for a specific dysarthric speaker follows a category-based approach. For example, when training the source model for a mild dysarthric speaker "X," we use all the examples from the mild category except "X." As a result, new source models need to be trained for each dysarthric speaker with a “leave 1-out” approach and further details on training

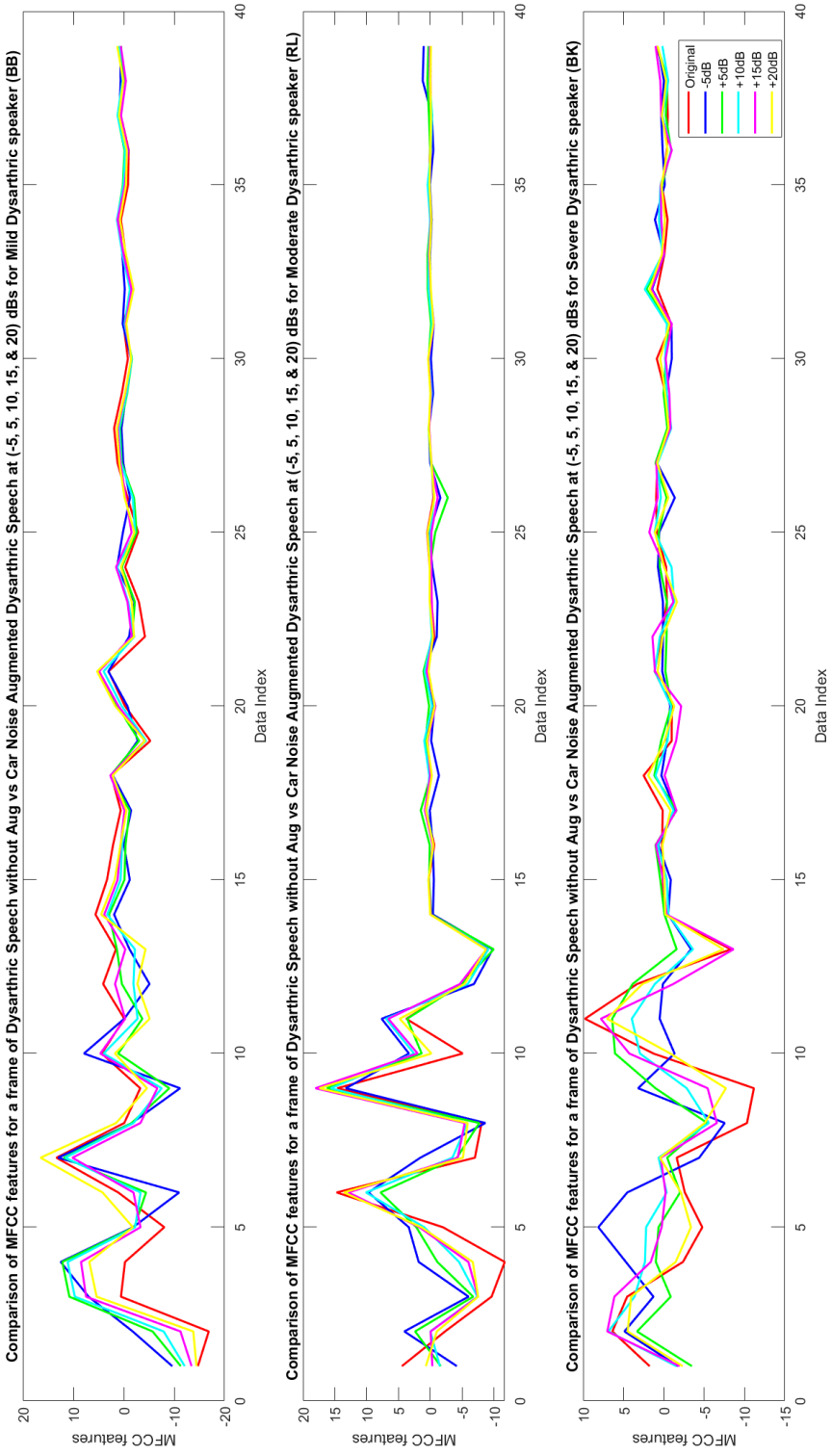

Fig. 5.2 Line graph comparing the features of original dysarthric speech data of mild, moderate, and severe dysarthric speakers

5.2.2 Stage 2: VM-MRFE-based data augmentation approach for target model training

In stage 2, virtual microphone array (MA) signals are synthesized using the original sample from the dysarthric speaker, followed by the application of multi-resolution feature extraction (MRFE) to these synthesized MA signals [7], [9]. A linear array configuration with 7 microphones generates the virtual microphone array (MA) signals.

To prevent spatial aliasing, the microphones are positioned d = 0.02 m apart. The virtual MA signals are produced using the phase spectrum, incorporating a phase shift of e^{jkd} [2], corresponding to a time delay of d. In this context, $k = 2\pi f/c$, where c represents the speed of sound (343 m/s). Let M be the spectrum of the source signal and M_n be the spectrum of the n^{th} element which is given by,

$$M_n = Me^{jk(n-1)d}; n = 1, 2, ..., N \tag{5.1}$$

where, N is the total number of elements in the array.

In Figure 5.1(c), the virtual MA signals and their corresponding original dysarthric speech signal is shown. These virtual MA signals are used for multi-resolution feature extraction (MRFE) [13]. The 39-dimensional MFCC features are extracted from these signals, including 13-dimensional static MFCCs, 13-dimensional delta coefficients, and 13-dimensional acceleration coefficients. These features are derived by subtracting the cepstral mean from noise-augmented and virtual-augmented versions. The characteristics are derived at various resolutions by employing numerous window sizes ranging from 10 ms to 20 ms in 2 ms intervals, and each window size has a 50% frame rate. This yields 5 unique resolutions. This research uses a starting window size of 10 ms for MRFE to guarantee that the frame size exceeds one pitch period for every speaker. Consequently, by using the VM-MRFE method, each source example is expanded by a factor 35 (7 (virtual MA examples $*$ 5 (MRFE)) via this two-tier data augmentation technique. An interesting aspect of these two data augmentation methods is that they use the original dysarthric speech samples to create augmented data while preserving the identity and errors of the dysarthric speaker. The techniques are used with continuous dysarthric speech data, and we consider approaches for expanding the pool of augmented dysarthric speech samples. As discussed in the previous section, a two-stage transfer learning

approach is then trained using a specific dysarthric speech example and its corresponding source model.

5.3 Training Two-Stage Transfer-Learning approach for dysarthric speech recognition system

The concept of transfer learning entails commencing with a pre-trained model using a vast dataset and subsequently adjusting the parameters with a smaller dataset in shown in Figure 5.3. The model trained is based on dysarthric speech data using a VM-MRFE approach. To do this, a pre-trained model is trained for any specific dysarthric speaker by choosing other speakers of the same category and their corresponding noise-augmented examples for pre-training, leaving the target dysarthric speaker out of this pre-training process. Thus, each target dysarthric speaker has a separate source model involved. For Nemours corpus, as shown in Table 5.1 (sourced from [8]), dysarthric speakers are classified into mild, moderate, and severe categories. In Table 5.1, the distribution example for pre-training the TL model while leaving out the target is shown. For the source model's initial training, 13-dimensional MFCC features are extracted and then transformed into a 40-dimensional vector using linear discriminant analysis and maximum likelihood linear transform, as illustrated in Figure 5.3. Additionally, we implement speaker adaptive training along with GMM-HMM training, using feature space maximum likelihood linear regression as described in [16]. Next, a DNN architecture, referred to as factored time delay neural networks (TDNNF), is employed, combining convolutional neural networks (CNNs). The core of this architecture consists of five CNN layers that receive input from 40-dimensional log-Mel spectrogram features. Following these layers are nine TDNN-F layers and one linear layer, culminating in the output layer. The TDNN-F training process utilizes the lattice-free maximum mutual information criterion. The linear layer is an additional hidden layer incorporated specifically for speaker adaptation purposes. The learning rates initially set at 0.0002 for training the source model over five epochs gradually decrease to 0.0005. Since the source model pre-trains the initial layers of the neural network based on the common characteristics inherent to the noise-based data-augmented examples. The common characteristics specific to a dysarthric category (mild, moderate, and severe) in the noise-based data augmentation method

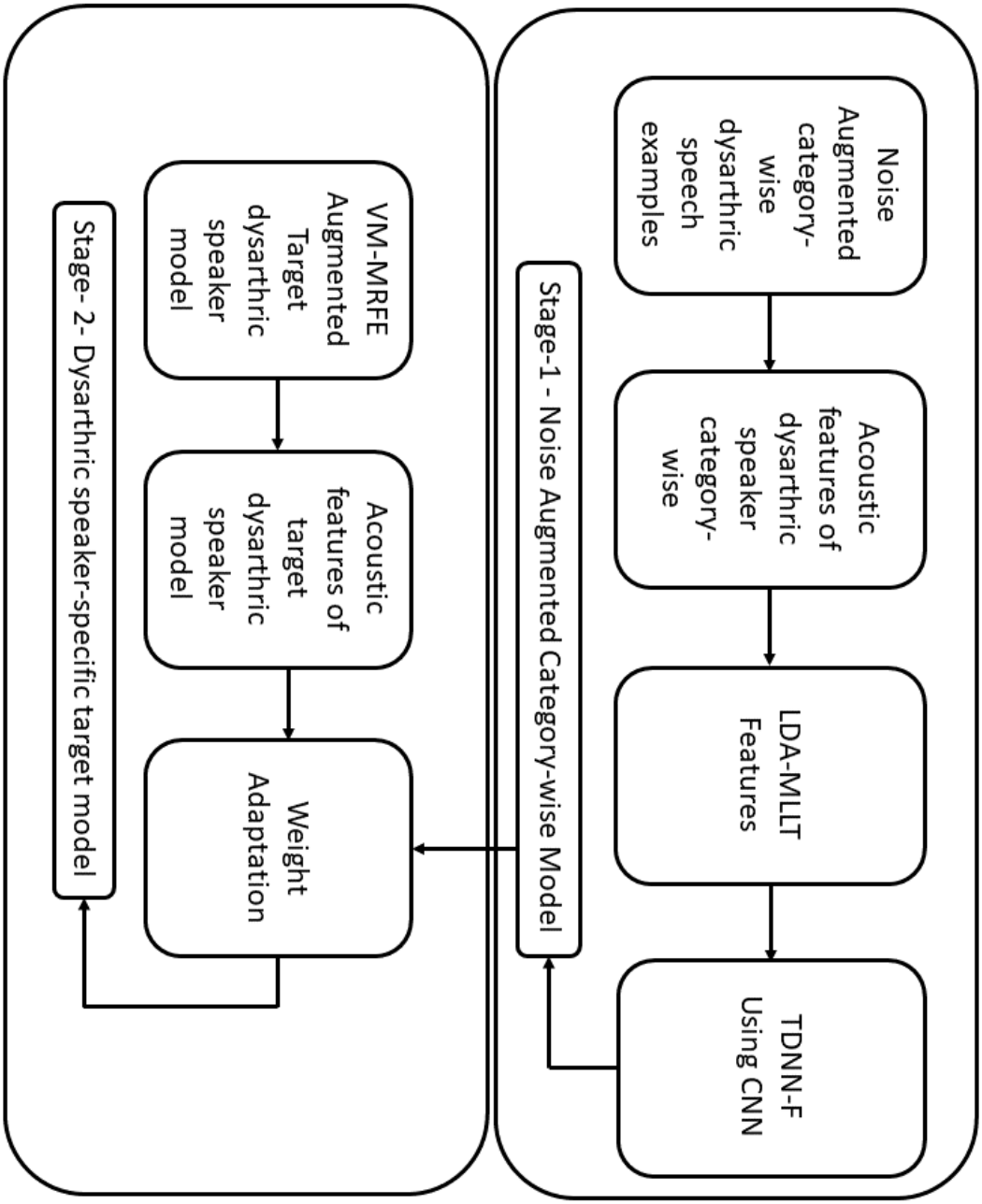

Fig. 5.3 Block Diagram of the two-stage transfer learning approach

Table 5.1 Number of dysarthric speakers in each category and Number of examples for pre-training the stage 1 source model

Dysarthric Speaker ID	**Category**	**No. of Examples for Pre-training Source Model**
FB	Mild	40 (8 noises × 5dB levels) × 37 (training utterances) × 3 (speakers leaving the target) = 4440 examples
BB		
MH		
LL		
RL	Moderate	40 (8 noises × 5dB levels) × 37 (training utterances) × 2 (speakers leaving the target) = 2960 examples
JF		
BV		
SC	Severe	40 (8 noises × 5dB levels) × 37 (training utterances) × 2 (speakers leaving the target) = 2960 examples
BK		
RK		

which is centered on the acoustic information of each category rather than the diverse noise characteristics is captured in the source model. This is because, in this approach, low-frequency analysis-based noises are specifically added to the dysarthric speaker, as opposed to introducing generic noise, hence the speaker characteristics are inherited over the noise characteristics which has a very poor influence.

With the pre-trained source model, the target model with VM-MRFE examples is fine-tuned at the second stage of the TL approach. The examples of each dysarthric speaker, as given in Table I, are used to fine-tune each speaker individually, making it a speaker-dependent system. For fine-tuning, each dysarthric speaker has 1295 dysarthric speech examples (35 * 37), where 35 is from VM-MRFE-based data augmentation, and 37 are the examples allocated for training. In stage two, we opt for three epochs to be used, and the learning rate is set at 0.0005, which is half of what the original model used. Learning is transferred from the original noisy dysarthric speech data to the VM-MRFE data by linearly adjusting the weights of the original model within the final TDNN-F layer. This adjustment is tailored to the characteristics of the augmented target dysarthric speech data. This two-stage transfer learning process is performed separately for ten dysarthric speakers. The original model is selected for each speaker based on the category of the target dysarthric speaker.

Table 5.2 Performance of Two-Stage TL-based dysarthric speech recognition system

Category	**Dysarthric Speaker ID**	**WER of dysarthric speech recognition without data augmentation**	**WER of dysarthric speech recognition using two-stage TL-based data augmentation**
Mild	FB	14.31	4.89
	BB	15.87	4.18
	MH	10.98	6.9
	LL	22.11	5.64
Moderate	RL	23.12	8.55
	JF	24.21	10.85
	BV	34.12	9.46
Severe	SC	52.33	22.4
	BK	58.12	26.05
	RK	63.19	38.65

As mentioned previously, we used the remaining 37 utterances from each dysarthric speaker as test data without applying data augmentation. Table 5.2 shows the performance evaluation of the dysarthric speech recognition systems based on two-stage transfer learning using this test data. It can be observed from the table that under the mild category of dysarthric speakers, the reduction in WER is almost 16.47%, and for the moderate category, it is 24.66%, and for severe, it is 34.54%. A greater reduction of WER is observed as the severity increases, which could also be attributed to the flexibility of the utterance syntax in the corpus.

5.4 Summary

This chapter presents a two-stage transfer learning approach for continuous dysarthric speech recognition. In the first stage, dysarthric speakers are grouped into mild, moderate, and severe categories, with ASR models trained separately using noise-augmented data to improve generalization. Low-frequency noise augmentation enhances robustness without distorting speech characteristics.

In the second stage, the VM-MRFE augmentation approach adapts each speaker using their category-specific model, efficiently updating weights without training independent models. This method significantly benefits severe dysarthric speakers by leveraging pre-learned patterns.

The two-level augmentation—category and speaker-based—ensures data diversity while preserving speaker identity. Experimental results show this structured transfer learning approach outperforms traditional methods, achieving better recognition, particularly for severe dysarthria.

Chapter 6

CONCLUSIONS AND FUTURE SCOPE

6.1 Conclusions

This thesis addresses data sparsity in dysarthric speech recognition through noise-based data augmentation. Unlike conventional methods, it explores noise augmentation as a means to enhance ASR performance while preserving dysarthric speech characteristics.

The study identifies low-frequency noise sources (e.g., car, bus, Volvo) that minimally interfere with speech intelligibility. Two augmentation strategies were proposed: (1) noise augmentation at specific SNR levels for isolated words, significantly reducing WER, and (2) a two-stage transfer learning approach for continuous speech, categorizing speakers by severity and adapting models through VM-MRFE-based augmentation.

Results show substantial improvements, particularly for severe dysarthria, surpassing prior augmentation methods. By relying solely on dysarthric speech data, this approach ensures speaker identity preservation and enables scalable dataset expansion without additional recordings.

Overall, this research introduces a flexible augmentation framework that enhances dysarthric speech recognition across various speech units. Improved ASR performance can facilitate better e-health services and communication accessibility, promoting greater independence and social inclusion for individuals with speech impairments.

6.2 Future Scope

This research advances dysarthric speech recognition through noise-based data augmentation, particularly benefiting moderate-to-severe dysarthric speakers. Future work can explore:

- **Extension to Other Speech Disorders:** Applying noise augmentation to conditions like apraxia, stuttering, and Parkinson's-related dysfluencies to improve ASR performance across diverse impairments.
- **Integration with Speech Therapy:** Developing ASR-driven therapy tools for real-time feedback, personalized training, and interactive speech exercises.
- **E-Health Applications:** Implementing ASR for medical consultations, speech-to-text documentation, and telemedicine to improve healthcare accessibility.
- **Personalized ASR Models:** Developing adaptive ASR systems that learn user-specific speech patterns through transfer learning and speaker adaptation.

References

[1] Actlin Jeeva, Muthu Philominal and Nagarajan, Thangavelu and Vijayalakshmi, Parthasarathy (2016). 'Discrete cosine transform-derived spectrum-based speech enhancement algorithm using temporal-domain multiband filtering'. *IET Signal Processing*, 10(8):965–980.

[2] Arcienega, Mijail and Drygajlo, Andrzej and Malsano, Joseph (2000). Robust phase shift estimation in noise for microphone arrays with virtual sensors. *in Proc. 10th IEEE Eur. Signal Process*, 1(1):1 – 4.

[3] Hirsch, Hans-Hans-Günter and Pearce, David (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *ISCA Tutorial and Research Workshop (ITRW) on Automatic Speech Recognition*, 1(1):1–10.

[4] Kim, Heejin and Hasegawa-Johnson, Mark and Perlman, Adrienne and Gunderson, Jon R and Huang, Thomas S and Watkin, Kenneth L and Frame, Simone and others (2008). 'Dysarthric speech database for universal access research'. *in Proceedings of* 9^{th} *Annual Conference on International Speech Communication Association*, 1(1):1741–1744.

[5] Ko, Tom, Peddinti, Vijayaditya, Povey, Daniel, Khudanpur, Sanjeev (2015). 'Audio augmentation for speech recognition'. *Interspeech*, 10(8):3586–3589.

[6] Makhoul, John (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(5):561–580.

[7] Mariya Celin TA and Nagarajan, T and Vijayalakshmi, P (2020). Data Augmentation Using Virtual Microphone Array Synthesis and Multi-Resolution Feature Extraction

for Isolated Word Dysarthric Speech Recognition. *IEEE Journal of Selected Topics in Signal Processing,*, 14(2):346–354.

[8] Mariya Celin TA and Rachel, G Anushiya and Nagarajan, T and Vijayalakshmi, P (2018). A weighted speaker-specific confusion transducer-based augmentative and alternative speech communication aid for dysarthric speakers. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 27(2):187–197.

[9] Mariya Celin, TA and Vijayalakshmi, P and Nagarajan, T (2023). Data Augmentation Techniques for Transfer Learning-Based Continuous Dysarthric Speech Recognition. *Circuits, Systems, and Signal Processing*, 42(1):601 – 622.

[10] Menendez-Pidal, Xavier and Polikoff, James B and Peters, Shirley M and Leonzio, Jennie E and Bunnell, H Timothy (1996). 'The Nemours database of dysarthric speech'. *in Proceedings of* 4^{th} *Int. Conf. Spoken Lang. Process.,*, 3(1):1962–1965.

[11] Pennington, Lindsay and Parker, Naomi K and Kelly, Helen and Miller, Nick (2016). 'Speech therapy for children with dysarthria acquired before three years of age'. *in Cochrane Database of Systematic Reviews*, 7(1):1–37.

[12] Povey, Daniel and Ghoshal, Arnab and Boulianne, Gilles and Burget, Lukas and Glembek, Ondrej and Goel, Nagendra and Hannemann, Mirko and Motlicek, Petr and Qian, Yanmin and Schwarz, Petr and others (2011). The Kaldi speech recognition toolkit. *In: Automatic Speech Recognition and Understanding Workshop*, 1(1):1 – 4.

[13] Priyanka, M Anbu Swarna and Solomi, V Sherlin and Vijayalakshmi, P and Nagarajan, T (2013). Multiresolution feature extraction (MRFE) based speech recognition system. *in Proceedings of IEEE Int. Conf. Recent Trends Inf. Technology*, 1(1):152–156.

[14] Salamon, Justin, Bello, Juan Pablo (2017). 'Deep convolutional neural networks and data augmentation for environmental sound classification'. *IEEE Signal Processing Letters*, 24(3):279–283.

[15] Varga, Andrew and Steeneken, Herman JM (1993). Assessment for automatic speech recognition: NOISEX-92. *NATO Science Technology Organization*, 1(1):1–10.

[16] Xiong, Feifei and Barker, Jon and Yue, Zhengjun and Christensen, Heidi (2020). Source domain data selection for improved transfer learning targeting dysarthric speech recognition. *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1(1):7424–7428.

[17] Young, Victoria and Mihailidis, Alex (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.

[18] Young, Victoria and Mihailidis, Alex (2018). 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.

LIST OF PUBLICATIONS

International Journals

1. Sarkhell Sirwan Nawroly, Decebal Popescu, Mariya Celin TA, M. P. Actlin Jeeva, "SNR-Selection-Based-Data Augmentation for Dysarthric Speech Recognition", Studies in Informatics and Control, Vol. 32, Issue. 4, pp. 129-140 December 2023. (Impact Factor: 1.2) DOI: 10.24846/v32i4y202312 Accession Number: WOS: 001164650200012 Indexed: 2024-03-18 ISSN: 1220-1766 IDS Number: IE4Y7

2. Sarkhell Sirwan Nawroly, Decebal Popescu, Mariya Celin TA, M. P. Actlin Jeeva, "Analysis for using Noise as a source of data augmentation for Dysarthric Speech Recognition", Circuits, Systems, and Signal Processing, 2025. Springer Nature: www.springernature.com (Impact Factor: 1.8) DOI: 10.1007/s00034-025-03054-4

3. Sarkhell Sirwan Nawroly, Decebal Popescu, Mariya Celin TA, "Category-based and Target-based Data Augmentation for Dysarthric Speech Recognition using Transfer Learning", Studies in Informatics and Control, Vol. 33, Issue 4, pp. 83-93, 2024. (Impact Factor: 1.2) DOI: 10.24846/v33i4y202408 Accession Number: WOS: 001390868500008 ISSN: 1220-1766 IDS Number: R3Z4K