SYNTHESIZING DYSARTHRIC SPEECH USING MULTI-SPEAKER TTS FOR DSYARTHRIC SPEECH RECOGNITION

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Engineering
at the University of Kentucky

By
Mohammad Soleymanpour
Lexington, Kentucky
Director: Dr. Michael T. Johnson, Professor of Electrical and Computer Engineering
Lexington, Kentucky
2022

ABSTRACT OF DISSERTATION

SYNTHESIZING DYSARTHRIC SPEECH USING MULTI-SPEAKER TTS FOR DSYARTHRIC SPEECH RECOGNITION

Dysarthria is a motor speech disorder often characterized by reduced speech intelligibility through slow, uncoordinated control of speech production muscles. Automatic Speech recognition (ASR) systems may help dysarthric talkers communicate more effectively. However, robust dysarthria-specific ASR requires a significant amount of training speech is required, which is not readily available for dysarthric talkers.

In this dissertation, we investigate dysarthric speech augmentation and synthesis methods. To better understand differences in prosodic and acoustic characteristics of dysarthric spontaneous speech at varying severity levels, a comparative study between typical and dysarthric speech was conducted. These characteristics are important components for dysarthric speech modeling, synthesis, and augmentation. For augmentation, prosodic transformation and time-feature masking have been proposed. For dysarthric speech synthesis, this dissertation has introduced a modified neural multi-talker TTS by adding a dysarthria severity level coefficient and a pause insertion model to synthesize dysarthric speech for varying severity levels. In addition, we have extended this work by using a label propagation technique to create more meaningful control variables such as a continuous Respiration, Laryngeal and Tongue (RLT) parameter, even for datasets that only provide discrete dysarthria severity level information. This approach increases the controllability of the system, so we are able to generate more dysarthric speech with broader range.

To evaluate their effectiveness for synthesis of training data, dysarthria-specific speech recognition was used. Results show that a DNN-HMM model trained on additional synthetic dysarthric speech achieves WER improvement of 12.2% compared to the baseline, and that the addition of the severity level and pause insertion controls decrease WER by 6.5%, showing the effectiveness of adding these parameters. Overall results on the TORGO database demonstrate that using dysarthric synthetic speech to increase the amount of dysarthric-patterned speech for training has significant impact on the dysarthric ASR systems.

KEYWORDs: Dysarthria, speech recognition, Speech-To-Text, Synthesized speech, Data augmentation.

Mohammad Soleymanpour
(Name of Student)

August 7, 2022
Date

SYNTHESIZING DYSARTHRIC SPEECH USING MULTI-SPEAKER TTS FOR DSYARTHRIC SPEECH RECOGNITION

By
Mohammad Soleymanpour

Michael T. Johnson
Director of Dissertation

Daniel Lau
Director of Graduate Studies

Date

ACKNOWLEDGMENTS

I wish to express my thanks to Dr Michael T. Johnson who kindly supervised this work and supported me throughout my Ph.D. study at the University of Kentucky. He always gives me the space to think and grow not only in my research but also in the other aspects of my education.

I would like to thank Dr. Jeffrey Berry for his insightful research help and suggestions, and I also would like to thank all my dissertation committee members Dr. Donohue, Dr. Sanchez, Dr. Jacobs and Dr. Kavuluru. Without their support, suggestion, and comments I could not have made it.

I would like to thank my all of family, particularly my parents. Without their unending love and support, this research work would not be possible to come to fruition. My appreciation also goes out to my friends for their encouragement and support all through my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

1.1. Statement of The Problem and Motivation

Dysarthria is a motor speech disorder, often caused by traumatic injury or neurological dysfunction, that decreases speech intelligibility through slow or uncoordinated control of speech production muscles [1]. People with moderate and severe levels of dysarthria may be less able to communicate with others through speech due to poor intelligibility [2]. Although individuals with dysarthria may have the cognitive and language abilities to formulate communication, they may not be able to reliably plan and execute the muscle control needed for sufficiently intelligible speech. Statistics shows that non-progressive dysarthria affects approximately 480,000 new people per year due to stroke and traumatic brain injury. Cerebral palsy is among the most common sources of dysarthria, including 0.26 percent of all seven-year-old children in the United States have moderate or severe cerebral palsy and 0.2 percent are involved in mild severity. In addition, there are other causes of dysarthria including Parkinson's disease, amyotrophic lateral sclerosis (ALS), and multiple sclerosis [2].

To provide dysarthric talkers with better communication or better tools for diagnosis and treatment, speech technologies can be effective. Technologies such as Automatic Speech Recognition (ASR) have the potential to significantly increase the quality of dysarthric speakers' communication. The use of ASR is now widespread, with systems such as Siri, Alexa, and Google assistant in common use. Although these systems work reasonably well with typical speech, and are slowly improving for accented speech, they have difficulty understanding dysarthric speech. Having a dysarthria-specific ASR can potentially help dysarthric talkers to be understood better and ameliorate their communication struggles. Different methods have been used to increase the performance of such systems for dysarthric speech, allowing dysarthric individuals to have a robust and reliable aids for communication and improving quality of life.

Another important application of speech technology is automatic assessment of dysarthria severity level, to analyze speech and estimate dysarthria severity level and speech intelligibility for clinical purposes. Such technology is not yet commonplace but could help speech pathologists and physicians in the early-stage dysarthria diagnosis or during treatment. Dysarthria severity level is conventionally assessed clinically using subjective assessments of neuromuscular function during both speech and non-speech tasks. These tests are often time-consuming to implement clinically, and some approaches suffer from a lack of intra-rater reliability, due to the subjective nature of these tools [3]. Automated assessment of dysarthria severity level and speech intelligibility could improve both the efficiency and reliability of clinical assessment. This need has led researchers to investigate systems to assess various clinical characteristics of dysarthric speech.

However, to have reliable and robust dysarthria-related speech applications, there is an essential need to have access to a substantial amount of recorded dysarthric speech for training. Current datasets containing dysarthric speech are insufficient for automatic speech recognition, severity assessment and dysarthric speech intelligibility enhancement tasks. Although there are a few publicly available dysarthric speech datasets, including TORGO [4], UASpeech [5] and Nemours [6], each of these have significant limitations in both size and diversity. TORGO is a popular dysarthric speech dataset of aligned acoustic and articulatory recordings from 15 speakers with eight dysarthric speakers. [4]. UASpeech includes 19 speakers with cerebral palsy. All participants utter the same 765 isolated words, 455 of them unique. [5]. The Nemours dataset contains 814 short nonsense sentences, 74 sentences spoken by each of 11 male speakers with various levels of dysarthria. [6]. Most of the utterances from these datasets consist of single words which do not capture cross-word co-articulation or allow for accurate modeling of prosody and pause characteristics in continuous dysarthric speech. None of these datasets are designed or sufficient for speech recognition and using them to support ASR is challenging. Because there are not an adequate number of conversational sentences, ASR systems trained with these types of datasets are often less robust. Modern ASR methods assume that training data includes a sufficiently large set of speakers, often hundreds to thousands hours of speech data, to adequately capture enough inter-speaker variability. For example, the LibriSpeech and TED_LIUM datasets used for ASR training contain about 1000 and 450 hours of data, respectively, hundreds of times more data than the dysarthric datasets described above. Because of the limited size, dysarthric datasets also have a relatively small number of speakers and are not sufficient to capture speaker variability. As will be shown later in this dissertation, current dysarthric speech datasets lack enough male and female individuals within the same group of severity to learn to distinguish even broad categories of dysarthric severity. To have a robust and generalized model for dysarthria severity level assessment, we need to have a much wider diversity of training speech, including diversity across gender and speaking styles. Including additional speakers with the same categories of dysarthria severity level can address this problem and improve assessment of dysarthria for pathologists and physicians and impact millions of patients suffering from dysarthria.

To address the data insufficiency issues described above for these and other dysarthric speech technology applications, this dissertation focuses on the development of data augmentation approaches for dysarthric speech applications, mainly for Speech recognition. The core idea proposed here is a combination of domain-based and deep-learning based speech synthesis models that are able to generate accurate speech with variability across the dimensions most important to dysarthric speech technologies, including speaking styles as well as prosodic characteristics like speaking rate and intonation patterns, and pause models that correlate with dysarthric severity level.

1.2. Dysarthric speech augmentation and synthesis

Data augmentation is a machine learning technique to generate additional supplemental training data. Augmentation has been widely applied to many different domains, including both image and speech processing. Image data augmentation is typically divided into basic image manipulations and deep learning approaches. While basic image manipulations contains methods such as kernel filters, geometric transformation, random erasing, mixing image and color space transformations, newer deep learning approaches introduce adversarial training, neural style transfer and GAN data augmentation [7-12]. These methods are used in image applications such as facial recognition, handwritten digits, medical image diagnosis, content reconstruction, and supper-resolution [7, 8, 10]. For speech applications, augmentation methods have been used to improve speech recognition [13-18], clinical speech applications [19-22], voice scene classification [23-25], children’s speech technologies [26-29], and speaker identification and verification [30-33]. Techniques such as Vocal Tract Length Perturbation (VTLP) and Statistical Feature Mapping approaches have been implemented for Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) for acoustic modeling [14]. For ASR data augmentation, there have been successful methods for tasks such as simulated Room Impulse Resonances (RIRs) [15], adding source-point noises [15], a voice conversion data augmentation [34, 35], and pitch shifting and speech perturbation [34].

For dysarthric research, temporal and speed modification have been applied on normal speech to simulate artificially dysarthric speech [20] and there has also been augmentation work using transformation methods to convert healthy speech to dysarthric speech.

The approach proposed in this dissertation focuses on the synthesis of dysarthric speech using neural multi-talker speech synthesis. To synthesize dysarthric speech, there is a need to build a system controlling different characteristics of dysarthric speech for generating variant dysarthric speech. As will be discussed later in this dissertation, and according to a number of studies [4, 36-39], such a system should have the following capabilities in order to support generation of authentic and diverse speech: 1) ability to control the speaking rate (duration), pitch, energy for a variety of dysarthria severity levels, 2) ability to learn and model pause behavior of dysarthric speakers (e.g., duration of pause and pause occurrence) and control pause insertion locations and durations 3) ability to learn and model individual voice characteristics of speakers and use these to generate new speaking styles 4) ability to learn and model these characteristics from a small amount of dysarthric speech data.

1.3. Contributions of the work

This dissertation first presents a comparative study between typical and dysarthric speech, to better understand differences in prosodic and acoustic characteristics of dysarthric spontaneous speech at varying severity levels. These characteristics are an important component for dysarthric speech modeling, synthesis, and enhancement, which are themselves important to tasks such as data augmentation for improving dysarthric speech assessment and recognition. To compare typical and dysarthric speech timing, we analyze the mean duration of vowels and consonants to find the speaking rate difference between dysarthric and typical speech. This timing information is essential to model speaking rate across severity levels. The mean pauses duration and the occurrence of pause between words are essential parameters to model the pause rate and duration for various severity levels. Two other important prosody characteristics of speech, pitch and intensity, are also evaluated for each speaker.

The second contribution of this work is a voice conversion-based data augmentation method using GAN and CycleGAN to convert typical speech to dysarthric speech. This method is effective at generating dysarthric speech, but the quality and variability of the speech is not sufficient to improve performance of speech technologies such as ASR when used to generate additional training data for augmentation. Although the method is not sufficient for effective data augmentation, the experimental work highlights some of the challenges of the augmentation task and led to the development of the next two contributions described below.

The third contribution of this dissertation is to explore a specialized data augmentation approach to enhance the performance of end-to-end dysarthric ASR. The proposed method contains prosodic transformation and time-feature masking. In prosodic transformation, we modify the speaking rate and shift the pitch to alter vocal excitation characteristics and prosodic structure. We also exploit time and feature masking in the spectral domain to alter the MFCCs representing vocal tract acoustics. Experimental results with this approach demonstrate that applying prosodic and time- feature masking on both dysarthric and normal speech represent better performance and underscore the need for speech from various dysarthria severity levels. Overall results indicate that using augmentation to increase the amount of dysarthric-patterned speech for training has significant impact on dysarthric ASR systems, particularly for speech with more severe dysarthria.

The fourth contribution is an innovative approach for synthesizing dysarthric speech using end-to-end multi-talker speech synthesis. The synthesis model generates dysarthric speech based on parameters representing key dysarthric speech characteristics, allowing control of parameters such duration, energy, pitch, dysarthria severity level and the occurrence of pause. These represent the most salient features of realistic dysarthric speech. In addition, this model has an ability to catch the voice characteristics of individuals using a decoder and speaker embedding, making it a multi-talkers TTS [40] capable of generating speech in a wide range of speaking styles. This is a useful capability for speech synthesis for data augmentation because it allows generation of a robust set of training data. Experimental results with this approach demonstrate that using dysarthric synthetic speech to increase the amount of dysarthric-patterned speech for training has significant impact on dysarthric ASR systems.

This chapter has provided an overview of the data sufficiency problem faced by dysarthric speech applications. The aim of this dissertation is to address this problem through synthesis of natural dysarthric speech for the purpose of data augmentation. Speech technologies for dysarthric speakers are of great importance but face a number of challenges because of the limited training data available as well as the great diversity of dysarthric speech. The remaining chapters are organized as following: chapter two represents a background and some useful information. Chapter three discusses a comparative studying to understand the main differences between typical and dysarthric speech will be discussed. Chapter four explains a dysarthric speech augmentation method using prosodic transformation and masking for speech recognition. Chapter five contains the main contribution of this dissertation which is a neural multi-talker TTS with a dysarthria severity level coefficient and a pause insertion model to synthesize dysarthric speech for varying severity levels. The final chapter concludes this work and proposes methods suggested by current results that could provide additional future benefit.

Chapter 2: Background

The background describes the fundamental concepts and algorithms required to accomplish this dissertation. First, we briefly review the required speech background and technologies. Following this, dysarthria, its different types, dataset as well as the speech recognition systems are described. The third section explains augmentation and synthesis technologies such as multi-talkers neural Text-to-Speech architecture and voice-conversion based data augmentation. Then, we finally review some typical and dysarthric speech augmentation studies.

2.1. Speech Background and Technologies

2.1.1. Speech Production

Speech production is a complicated motor task that involves approximately 100 orofacial, laryngeal, pharyngeal, and respiratory muscles [41]. These processes are carried out by the lungs, the larynx, and the upper vocal tract including the jaw, lips, tongue, and mouth walls. Speech production originates when a message is formulated in the brain of a speaker and then mapped to a sequence of intended sound units. The Neuro-Muscular [42] system then plans the required muscular movements to controlled speech articulators such as the tongue, lips, teeth, jaw and velum in way that will produce the desired spoken message with the desired prosodic characteristics, including intonation, loudness, and timing. The vocal tract then physically creates the necessary sound sources and the appropriate vocal tract shapes over time to create the corresponding acoustic waveform [42].

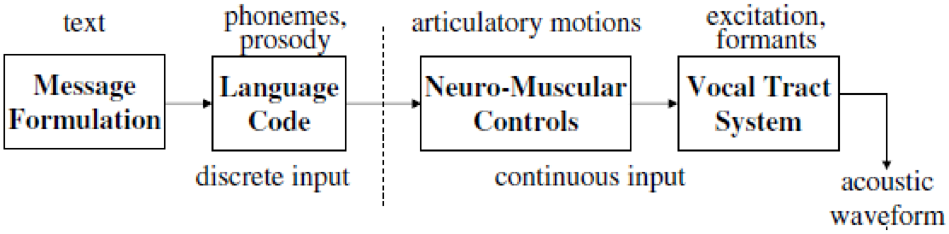


Figure 2.1: An overview of speech production[42]

2.1.2. Acoustic Feature

An initial step in a speech processing system is the computation of a set of acoustic features from sampled speech [43]. Meaningful quantitative features representing acoustic characteristics of speech are important for speech and audio applications. Such features should be robust toward speaker, environmental, pronunciation, accent and other variance which is not relevant to the target speech processing task. For speech recognition systems, a number of different approaches to feature extraction have been used historically, including Mel Frequency Cepstrum Coefficients (MFCC) as well as their the first-and seconded derivatives [44], Linear Predictive Cepstral Coefficients (LPCC)[44], Perceptual Linear prediction (PLP)[45], Filter Bank Analysis Feature Space Maximum Likelihood Linear Regression (FMLLR)[46], and others. Among these features, MFCC and FMLLR are relevant feature for dysarthric speech recognition tasks, either DNN-HMM model or end-to-end model.

2.1.2.1 MFCC

MFCC is the most common feature extraction method used for speech, based on a perceptual model of human hearing. MFCC is the spectral representation of the framed input speech which is obtained by Fast Fourier Transform (FFT). Then, Mel filter banks are applied to perceptually simulate the human auditory system [44]. In order to calculate the MFCC, the first step is to create a frame size of 20 to 40 milliseconds with overlapping of 10 to 25 milliseconds. After that, an FFT is implemented to extract frequency magnitudes, and then Mel Filter Banks are used to integrate the frequency content over perceptually spaced bands. The final step is the Discrete Cosine Transformation (DCT) to convert the frequency information into the cepstral domain with MFCCs representing the shape of the vocal tract spectrum [44]. Figure 2.2 MFCC Derivation [44] shows the main steps of MFCC feature extraction.

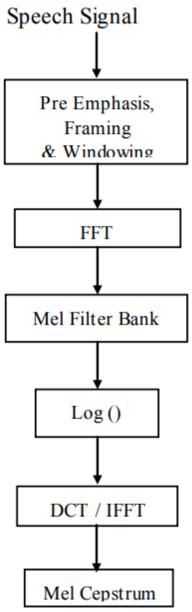


Figure 2.2: MFCC Derivation [44]

2.1.2.2 Dynamic MFCC feature:

In addition to the Cepstral coefficients over each frame, it is common to add dynamic features to represent the rate of change of frame-based characteristics over time. Temporal information of speech is calculated using the first and second derivatives of the Cepstral coefficients. The first-order derivatives, known as delta coefficients, capture the rate of change of the MFCC features, and the second-order derivatives, called delta-delta-coefficients, capture the second derivative or acceleration of those features. The delta coefficients are calculated using a standard linear regression formula as follows:

$$\Delta c_m = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |k_i|}, \tag{1.1}$$

where $C_m(n)$ denotes the m^{th} feature for the nth time frame, k_i is the ith weight, and T is the number of successive frames used for computation. The delta–delta coefficients are computed by taking the first-order derivative of the delta coefficients [47].

2.1.3. Automatic Speech Recognition

Automatic speech recognition (ASR) is a prominent technology that is essential to enabling human–human and human–computer interactions, which has been an active research area for several decades [48]. Conventional ASR systems typically consist of frond-end processing, acoustic modeling, a decoder, and lexicon and language modeling. The front-end performs feature extraction, such as the MFCC calculation described in the previous section. The acoustic model characterizes the feature vectors with respect to the statistical characteristics of underlying acoustic units, typically phonemes, to determine the posterior phoneme probabilities of each frame. The role of the decoder, which historically has been implemented using a Hidden Markov Model (HMM)[49-51], is to search the space of word sequences based on acoustic probabilities as well as simple language models over word sequences to find the most likely sequence of acoustic units for the corresponding feature vectors. A more advanced language model can also be used to re-evaluate the most likely word sequences from the acoustic search and determine the overall most likely word sequence based on a combination of acoustic and linguistic analysis Figure 2.3 shows the main components of a typical ASR system.

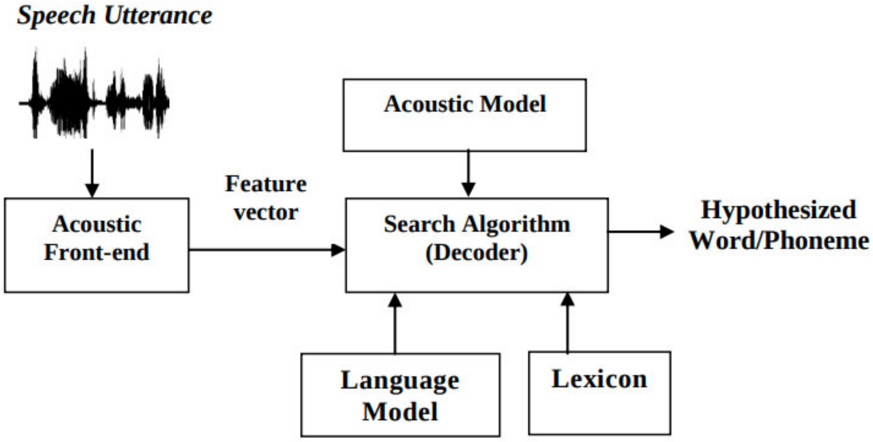


Figure 2.3: A typical ASR architecture [52]

2.1.3.1 Models and algorithm

Conventional ASR systems are based on Bayes' theorem to hypothesize the most likely character/word sequence among all possible character/word sequences given an acoustic feature sequence. The goal is to generate the sequence $W = w_1, w_2, \ldots, w_N$ which maximizes the probability of the acoustic feature sequence *X*:

$$W = \arg\max \frac{P(W)P((X|W)}{P(X)} \qquad 1.2$$

where *P(X|W)* indicates the likelihood of acoustic feature *X* given sequence *W* and *P(w)* plays the role of the language model which determines the prior probability of the given sequence.

The primary historical approach to ASR has been the Hidden Markov Model (HMM) using Gaussian Mixture Model (GMM) which was widely used. In this approach, each state of the HMM represents an acoustic unit, using the GMM to represent the spectrum of speech signal. However, Deep Neural Networks (DNNs) have started to be used instead of GMMs for estimating acoustic posterior probabilities to build a hybrid model.

The main purpose of the hybrid approach is the use of a forced alignment to obtain a frame level labeling for training the neural network [53]. Neural networks architectures such as multi-perceptron with many layers, Deep Belief Neural Network, LSTM, GRU, and CNN were replaced with GMM in statistical speech recognition models to improve the speech recognition performance for different scenarios. The acoustic modelling, language modelling and sequence decoding components of this model are separately trained and then attached together to form a complete system. In the past few years, there has been research in applying Recurrent Neural Networks (RNN) and addressing training problems such as vanishing and exploding gradient. This progress leads people to use a new variant of Recurrent neural networks (RNNs) instead of HMM to encode sequence history in their internal state to predict phonemes based on all the speech features observed up to the current frame [54]. State-of-the-art ASR systems have started to migrate towards End-to-end systems which integrate these components. End-to-end systems directly map the speech signal to word or sub-word sequence and are jointly trained in a single model.

Currently, there are two main architectures for end-to-end speech recognition: Connectionist Temporal Classification (CTC) model and Attention-based modeling [55]. CTC is an approach to train the model without frame-level alignment. The early CTC-based model was not completely end-to-end as it needed a separate language model. The attention model concept was introduced in Machine Translation to solve RNN-based Sequence to Sequence modeling problems. The concept is that an attention model is integrated into the overall architecture which learns how to probabilistically associate each element of the output sequence with an associated region or regions of the input sequence. An attention-based ASR system consists of two components, an encoder and a decoder. The encoder converts the input *X* to a higher feature representation sequence *h* with a fixed length, while the decoder outputs target sequences based on previous outputs, the current hidden state, and attention model probabilities [55]. We explain a hybrid ASR mode and an end-to-end model.

2.1.3.2 Hybrid ASR model: DNN-HMM

Deep Neural Network- Hidden Markov Model (DNN-HMM) based speech recognition systems have become very popular and effective in the last decade. This architecture is able more effectively to obtain underlying nonlinear relationship among data in comparison with GMM-HMM [56]. The following figure shows an overview of DNN-HMM speech recognition architecture. In this architecture, the HMM models the sequential feature of Speech signal and the scaled observation likelihood of all states[53]. It is a finite state structure consisting of three components: transition probability matrix A representing transition probability from state i to state j, prior probability π showing the prior probability of state i and emission probability vector β showing the emission probability of observation x in state j [57].

The main purpose of the hybrid approach is the use of a forced alignment to obtain a frame level labeling for training the neural network [53]. Neural networks architectures such as multi-perceptron with many layers, Deep Belief Neural Network, LSTM, GRU, and CNN were replaced with GMM in statistical speech recognition models to improve the speech recognition performance for different scenarios.

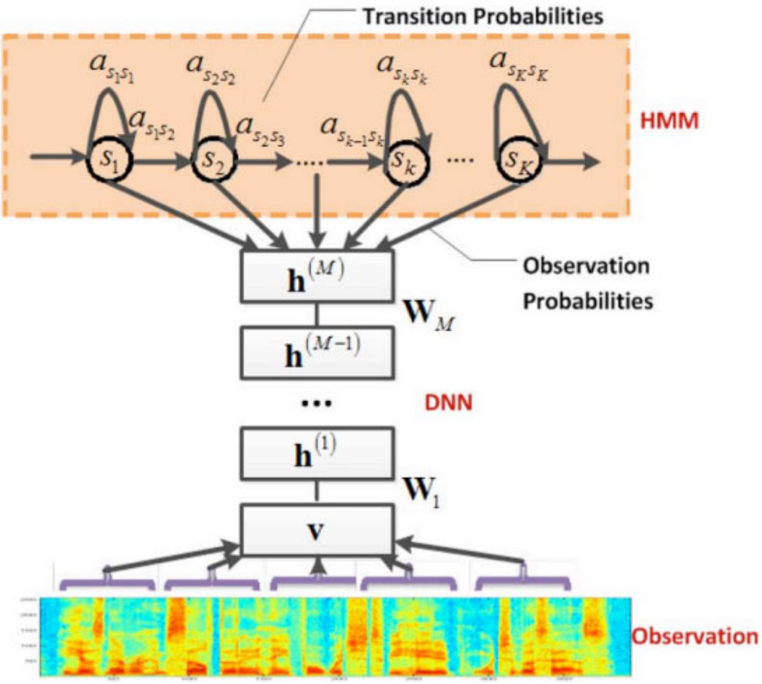


Figure 2.4: an Overview of DNN-HMM models [53]

2.1.3.3 End-to-end ASR: Listen, Attend and Spell

The advent of DNN based ASR has produced a significant improvement in speech recognition system. Using Recurrent Neural Networks for the Language model (LM) have been shown to further improve the performance of such ASR systems. However, these need to be trained separately and then integrated. However, end-to-end ASR systems attempt to map the speech signal to word or sub-word sequence, integrating acoustic and language modeling into a single network. Results of end-to-end systems are rapidly approaching those of state-of-the-art fully tuned DNN recognition systems.

The CTC model and sequence to sequence (seq2seq) model were two of the first end-to-end systems. However, CTC assumes that the outputs are conditionally independent [58]. The Listen, Attend and Spell (LAS) model was one of the first to be proposed to address these limitations. LAS is a neural network that transcribes spoken utterances with character-level labeling based on orthographic transcriptions, i.e. it directly maps acoustics to letter sequences. In this model, no independence assumption is made and HMMs are not needed for initial alignment and labeling as with many DNN systems. LAS contains two main components, the listener and the speller, connected together through an attention vector. The attention vector uses an attention mechanism to estimate the desired alignments in long sequences. The listener is implemented as a hierarchical Bidirectional Long-Short-Term Memory (BLSTM), taking the audio features as an input and converting this into a higher-level representation feature sequence. The speller is an RNN decoder, which takes the high-level representation along with the attention vector to generate the output character sequences. During training, the listener and speller are jointly trained to make a true end-to-end ASR system. Figure 2.5 presents the architecture of the LAS model. In the following sections, the listener and the speller parts are explained in detail.

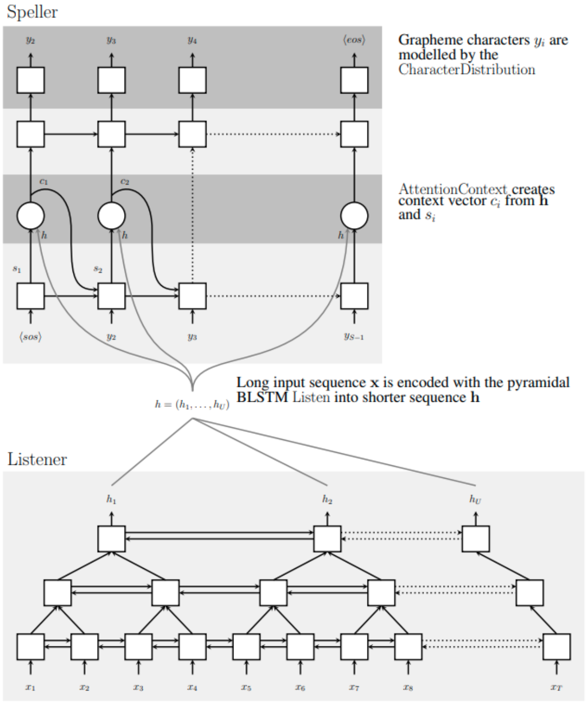

Speller
y_2
y_3
y_4
$\langle eos \rangle$
Grapheme characters y_i are modelled by the CharacterDistribution
c_1
c_2
AttentionContext creates context vector c_i from h and s_i
h
h
h
s_1
s_2
$\langle sos \rangle$
y_2
y_3
y_{S-1}
Long input sequence x is encoded with the pyramidal BLSTM Listen into shorter sequence h
$h = (h_1, \ldots, h_U)$
Listener
h_1
h_2
h_U
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_T

Figure 2.5: Listen, Attend and Spell(LAS) architecture [58]

The first component of LAS is the listener which is shown in the bottom part of Figure 2.5 above. The Speller takes acoustic feature as inputs and generates English characters as outputs. The input is shown by $x = (x_1, ..., x_T)$, which is acoustic features and an output is $y = (<sos>, y_1, .., y_S, <eos>)$ which shows the character sequence. Here <sos> and <eos> are the special start-of-sentence token, and end-of-sentence tokens, respectively [58]. The goal is to model *y* in each time step *i* as a conditional distribution over the previous recognized characters and input speech as follows:

$$P(y \mid x) = \prod_i P(y_i \mid x, y_{<i}). \qquad 1.3$$

- Listen

The main role of the listener is to take the sequence of acoustic feature vectors *X* and transform it into a high-level representation $h = (h_1, ..., h_U)$ where $U <= T$:

$$h = Listen(x). \qquad 1.4$$

To construct the listener, a Bidirectional Long Short-Term Memory (BLSTM) with a pyramid structure is used. The pyramid design is applied to expand the context of the input in an efficient manner [58], enabling the attention model to find the pertinent information as well as reducing the computational cost, particularity during training.

This pyramid structure is particularly beneficial for domains such as dysarthric speech recognition. One of the characteristics of dysarthric speech is a low and inconsistent speaking rate, and people with severe dysarthria may generate somewhat lengthy acoustic output even for a short sentence. The ability of the BLSTM pyramid structure to capture extended context can be very helpful to handle this challenge.

- Attend and Spell

AttendAndSpell function is shown at the top of the block diagram in Figure 2.5, and is based on an attention-model directed LSTM transducer. The transducer provides the model a probability distribution at each output step due to the all previous characters generated[58]. The context vector is defined as:

$$c_i = AttentionContext\left(s_i, h\right), \tag{1.5}$$

where s_i is the current hidden state and h is the high level representation vector from the listener at each time step i, AttentionContext generates a context vector c_i, containing the information of the acoustic signal needed to emit the next character [58].

The context vector itself is one of the parameters needed to calculate probability distribution of the output characters as well as the hidden decoder state Si.

$$s_i = RNN(s_{i-1}, y_{i-1}, c_{i-1}) \tag{1.6}$$

$$P(y_i \mid x, y_{<i}) = CharacterDistribution(s_i, c_i) \tag{1.7}$$

where the decoder state s_i is a function of the previous state s_{i-1}, the previously emitted character y_{i-1} and context c_{i-1}. In addition, the CharacterDistribution is an MLP using softmax activation over entire characters [58].

2.1.3.4 Evaluation

The performance of an ASR system is usually evaluated by accuracy and speed. The accuracy is commonly measured by Word Error Rate (WER) or Character Error Rate (CER) while speed is captured through the real-time-factor of computation on established hardware platforms.

To measure CER or WER, Levenshtein distance is used [52] based on an alignment between the ASR output and the known transcription:

$$WER = \frac{S + D + I}{N}, \qquad 1.8$$

where, $S, D,$ and I are the number of substitutions, deletions and insertions, respectively, and N is the number of words in the reference transcription. Word Recognized Word (WRR) is an alternative metric to WER, calculated as follows:

$$WRR = 1 - WER = \frac{N - (S + D + I)}{N} = \frac{H - I}{N}, \qquad 1.9$$

where $N - (S + D)$ is equal to the number of correctly recognized words.

2.2. Dysarthric Speech background and technologies

2.2.1. Dysarthria

Dysarthria is a motor speech disorder, often caused by traumatic injury or neurological dysfunction, that decreases speech intelligibility through slow or uncoordinated control of speech production muscles[1]. Although individuals with dysarthria may have the cognitive and language abilities to formulate communication, they may not be able to reliably plan and execute the muscle control needed for sufficiently intelligible speech. Dysarthric speech is primarily characterized by slow speaking rate, imprecise phoneme articulation, hypernasality, harsh voice and mono-pitch, and breathiness [19]. For example, disability to control of soft palate movement caused by disruption of the vagus cranial nerve potentially leads to hypernasality. Also, an inadequacy of tongue and lip dexterity often produces heavily slurred speech [1]. Individuals with dysarthria have difficulties controlling the laryngeal muscles that regulate vocal fold tension and glottal airflow, so the fundamental frequency of voiced sounds is unstable. While talkers without dysarthria can maintain their rhythm and energy distribution for a experiments of consonant-vowel repetition sequence, the people with dysarthria are unable to keep these factors steady during the same experiment[37].

Cerebral palsy is among the most common sources of dysarthria, with 0.26 percent of all seven-year-old children in the United States having moderate or severe cerebral palsy and 0.2 percent having mild severity [59]. Non-progressive dysarthria affects approximately 480,000 new people per year due to stroke and traumatic brain injury. In addition, there are other causes of dysarthria including Parkinson's disease, amyotrophic lateral sclerosis (ALS), and multiple sclerosis.

2.2.1.1 Prosody in Dysarthria

In linguistics, prosody refers to aspects of speech occurring over a longer time frame than phonetic segments, at the syllable, word, and sentence level. The main elements of prosody include intonation, loudness, and timing. Acoustic correlates of these prosodic elements are fundamental frequency, amplitude, and duration. Prosodic characteristics contain substantial information, including not only speakers' emotion but also linguistic structures such as differentiation of questions and statements [60].

Many studies have investigated the differences between normal and dysarthric speech. The speaking rate of individuals without voice disorders is between 150 and 250 words per minute. However, the typical speaking rate for dysarthric individuals is less than 15 words per minute, over ten times slower and with a higher degree of variability. Furthermore, different types of dysarthria, described in more detail in the following section, display a wide variation in severity levels and speaking rates. For example, speakers with Amyotrophic Lateral Sclerosis (ALS) talk twice as slow on average compared to healthy speakers. Abnormal speaking rate has multiple acoustic consequences. For instance, if a one-syllable word is prolonged by a long voiced phoneme, it will often be misinterpreted as a multisyllabic word by listeners [1]. Another example is that people may incorrectly understand a single word as two when a voiceless plosive is followed by a lengthy occlusion.

Zhang et al have shown that compared to typical speech, dysarthric speech is indicated by slower speaking rate, imprecise phoneme articulation, hypernasality, harsh voice, mono-pitch, and breathiness [6]. In their study [6] the speech of 10 French normal and dysarthric speakers was analyzed to determine the ability of these speakers to make question-statement contrasts. Findings show that individuals with dysarthria generate smaller intonational differences in comparison with normal speakers and that the overall speaking rate of dysarthric speakers is lower than that of normal speakers [60]. In another study[61], eight speakers with severe dysarthria caused by cerebral palsy were studied to determine the extent of their pitch and duration control. The results indicate that although most of the speakers could produce short, medium and long versions of the vowel /a/, they could only produce two distinct levels of pitch [61].

2.2.1.2 Motor speech disorders of Dysarthria

The most common motor speech disorders are dysarthria and apraxia. Dysarthria is a set of neurogenic speech disorders characterized by "abnormalities in the strength, speed, range, steadiness, tone, or accuracy of movements required for breathing, phonatory, resonatory, articulatory, or prosodic aspects of speech production" [1]. The primary types of dysarthria recognized by perceptual attributes and associated locus of pathophysiology [1] are as follows:

Flaccid: Flaccid dysarthria is usually caused by damage to lower motor neurons, resulting in two common characteristics of this disorder. First, the source of this disorder originates from impairment of the lower motor neurons of cranial or spinal nerves. Secondly, many people with flaccid dysarthria have weak speech and respiratory musculature. Individuals with flaccid dysarthria are often characterized by slow articulation movement, a degree of hypernasal resonance, and hoarse-breath phonation[62]. This type of dysarthria can be caused by anything that disrupts the flow of motor impulses an the cranial or spinal nerves, conditions such as brainstem stroke, tumors and so forth damage lower motor speech [62].

Spastic: Spastic dysarthria is primarily caused by bilateral damage to upper motor neurons. Spastic dysarthria presents through imprecise articulation, monotonous pitch, labored speech and prolonged words [62]. Individuals with spastic dysarthria speak slowly with expanded syllables and longer pauses [63]. The main cause of spastic dysarthria is stroke, amyotrophic lateral sclerosis(ALS), traumatic head injury, multiple sclerosis [62].

Unilateral Upper Motor Neuron: This is a recently recognized form of dysarthria which is associated with damage to the upper motor neurons that support cranial and spinal nerves related to speech production. It presents through imprecise production of consonants due to weakness of the lower face, lips and tongue muscles [62].Stroke, tumors, traumatic brain injury cause unilateral upper motor neuron dysarthria.

Ataxic: The most common feature of ataxic dysarthria is damage to the cerebellum system. Individuals with ataxic dysarthria are characterized by having speech errors associated with timing and identical stress on each syllable. Also, articulation errors occur with mild to severe intermittent, harshness, monotonous pitch and volume, and there is increased and unnatural stress [62]. The main source of Ataxic is stroke, toxic condition, traumatic head injury, tumors and degenerative disease. However, there are other conditions like viral infections and a bacterial abscess that can bring ataxic dysarthria[62].

Hypokinetic: Hypokinetic dysarthria is associated with the basal ganglia control system. It can affect all parts of speech production; however, the primary characteristics of people with hypokinetic dysarthria are weak voice, articulation and altered prosody. Prosodic changes include monopitch and monoloudness, reduced force, inaccurate consonants and irregular silences and harshness voice. Hypokinetic is relatively unique type of dysarthria, in that it is symptomized by an increased rate of speech. Most individuals with hypokinetic dysarthria show the same causative factor [62]. Parkinson's syndrome, traumatic head injury, toxic metal poisoning and stroke are the main causes of hypokinetic dysarthria[62].

Hyperkinetic: The diagnosis of hyperkinetic dysarthria is comparatively difficult due to several causative disorders, the most common of which is Parkinson's disease. One common cause of hyperkinetic disorders is a malfunction of the basal ganglia, which helps to control movement of speech production organs. Patients with hyperkinetic have imprecise articulatory movement, harsh voice, and abnormal prosodic characteristics [62]. Several movement disorders such as chorea, myoclonus, tics, essential tremor and dystonia as well as stroke can lead to hyperkinetic dysarthria [62].

2.2.1.3 Dysarthria evaluation and treatment

There are a number of standard tests for evaluating dysarthric speech, and detailed informal dysarthric speech assessment tools are also available. Following are a few of the most common assessment approaches.

Frenchay Dysarthria Assessment (FDA-2): This standardized test was introduced in 1992 and is used to differentiate the types of dysarthric speech and to assess dysarthric speech intelligibility. The test allows physicians and speech therapists to identify the most impactful factors reducing speech intelligibility and to plan treatments. With FDA-2, subjects are rated on their performance of various tasks along 28 relevant perceptual dimensions of speech grouped into 8 categories, specifically reflex, respiration, lips, jaw, soft palate, laryngeal, tongue, and intelligibility. The time required for the assessment is relatively short[1].

Assessment of Intelligibility of Dysarthric Speech: This is another standardized test which has been broadly used to assess dysarthric speech intelligibility. Subjects are evaluated on single words and sentences based on speaking rate. A set of 50 spoken words are evaluated by native speakers, and intelligibility is measured by a ratio of the number of correctly understood words by native listeners to the total number of words. For sentences subjects are asked to utter 22 sentences with a length ranging from 5 to 15 words, and listeners are asked to transcribe the spoken sentences. This test also provides information regarding severity estimation and communication efficiency[1].

2.2.2. Dysarthric Automatic Speech Recognition (ASR)

Because of the substantial differences between normal and dysarthric speech, standard ASR systems do not work well with dysarthric individuals [64-67]. In order to effectively transcribe dysarthric speech, there is a need to build a robust dysarthric ASR system which is trained on dysarthric speech. The goal of the proposed research work is to build such a system based on advanced methods of data augmentation coupled with a robust end-to-end ASR system using both hybrid (DNN-HMM) and end-to-end (Listen, Attend and Spell) approaches. As will be discussed in chapter 4 and chapter 5, these models are used to evaluate the performance our proposed data augmentation and synthesis in our experiments.

2.2.2.1 Related work

Much less research has been done on dysarthric ASR as opposed to ASR for normal speech. However, there have been several methods used to enhance the performance of dysarthric ASR systems, including improving acoustic and language models, feature engineering, speaker adaptation, and data augmentation. Early work on dysarthric ASR system related to isolated words, computer command or digits. In [68], researchers built a command word recognition system for dysarthric speakers. Initially, the main goal was to build a voice assistance device for people with severe dysarthria. A small vocabulary, speaker-dependent, isolated-word condition was applied for training their HMM-based model. In [69], GMM-HMM and DNN-HMM as well as speaker adaptation methods were compared for dysarthric speech, using the TORGO dataset and Kaldi Speech recognition toolkit to build the ASR system.

DNN- and GMM-HMM based acoustic models have been explored in several ways to improve the Word Error Rate (WER) of previous HMM-based dysarthric ASR machines. In [64], speaker normalized Cepstral features and combined DNN-HMM models were trained on TORGO dataset to evaluate the effect of using normal, dysarthric, and combined speech. In [70], convolutional LSTM (CLSTM) was used to capture the characteristics of dysarthric speech, to take advantage of local features and model temporal dependencies of the features. The model was evaluated on a collected data including 9 dysarthric people.

Some data augmentation techniques have been implemented for dysarthric speech recognition as well. In [20], temporal and speed modification were applied on healthy speech to simulate dysarthric speech, and DNN-HMM based Automatic Speech Recognition (ASR) and Random Forest based classification were used for evaluation. The dataset used to evaluate the data augmentation approach is Universal Access Speech (UASpeech) corpus.

Another approach to improving dysarthric ASR is feature extraction and normalization. One approach has been to enhance MFCC features using deep neural network autoencoders to raise the performance of dysarthric speech recognition [71]. This approach used severity-based adaptation before performing the autoencoder-based feature improvement, with a DNN-HMM model and the UASpeech corpus. Another feature extraction method based on Convolutional Bottleneck Networks (CBN) was implemented for dysarthric ASR systems [72] to decrease the influence of unstable speaking style with Athetoid Cerebral Palsy (ASL) speakers.

Phase-based representations of the dysarthric speech features is presented in [73]. This representation is able to capture properties of vocal tract resonances of the dysarthric speech signal. Speech recognition performance with phase-based representation features was compared to the standard MFCC features, evaluated on UASpeech. In [74], three inter-speaker normalization approaches in acoustic, articulatory, and combined spaces are explored to address the high variation of articulation across dysarthric speakers. The Procrustes matching approach based on physiological modeling in the articulatory space, Vocal Tract Length Normalization (VTLN), and Feature Space Maximum Likelihood Linear Regression (FMLLR) were used [74].

Speaker Adaptation is another approach that can help to improve dysarthric ASR. In [75], a speaker-adaptive recognition system for dysarthric speakers was proposed. Two implementations have been evaluated: 1) MAP adaptation of speaker-independent systems trained on normal speech and, 2) modification of the transition probability matrix that is a linear interpolation between fully ergodic and left-to-right structures. According to their findings, the left-to-right HMMs show slower error rate than transition-interpolated HMMs in speaker-dependent systems. Another finding is that applying both adaptation and transition-interpolation does not enhance the performance of dysarthric ASR system more than applying adaptation alone [75]. In [17], an interpolation-based technique is exploited to capture a prior acoustic model from a speaker trained on healthy speech and then adapt it to the dysarthric speaker. Results demonstrate that the adaptation techniques are an effective approach to robust dysarthric ASR models. In [76], acoustic and lexicon model adaptation were evaluated among people with dysarthric speech, tracking deletions, substitutions, insertions, and distortions of phonemes in each speaker.

Other transformation approaches have been implemented to modify dysarthric features to be more like normal speech [77]. This includes modifying formants and energies from dysarthric speech to approximate desired normal targets, using formant synthesis. The efficiency of their transformed speech was examined through a perceptual test and ASR model based on the HTK HMM Toolkit.

2.2.3. Datasets

There are a few publicly available dysarthric speech datasets, including TORGO [4], UASpeech [64] and Nemours [6]. They are mainly used to analyze dysarthric speech and to understand the difference between dysarthric and typical speech. TORGO is a popular dysarthric speech database of aligned acoustic and articulatory recordings from 15 speakers, containing 8 dysarthric speakers and 7 controls. This dataset includes non-word, short words, restricted and non-restricted sentences. Two types of microphones were used to record the data, an 8-element microphone array and a head-mounted microphone. The number of utterances for each dysarthric talker averages 700; whereas for normal speakers the average is 1560. Dysarthric speakers are categorized into three dysarthria severity levels such as Very Low, Low, and Medium and into two groups of intelligible and non-intelligible. Table 2.1 shows the Speakers' level of dysarthria severity and their corresponding intelligibility categories. The standardized Frenchay Dysarthria Assessment described in 2.2.3 used to assess the motor functions of each subject.

Table 2.1: Properties of various participants in TORGO dataset

Severity Level	Speaker ID	Number of Utterances	Intelligibility Category
Normal	FC01	296	Intelligible
	FC02	2183	
	FC03	1924	
	MC01	2141	
	MC02	1112	
	MC03	1661	
	MC04	1614	
Very low	F04	675	
	M03	806	
Low	F03	1097	Unintelligible
Medium	M05(L/M)	610	
	F01	228	
	M01	739	
	M02	772	
	M04	659	

Another dysarthric dataset is UA-speech collected by the University of Illinois [5]. This dataset includes speech recordings of 15 dysarthric speakers (4 female and 11 male) and 13 control speakers (4 female and 9 male). Each speaker was asked to read isolated works shown on a laptop screen, including utterances containing 10 digits, 26 radio alphabet letters, computer commands, common words from the Brown corpus of written English, and uncommon words from children's novels selected to maximize phone-sequence diversity. All participants produced the same 765 word in citation form, 455 of them unique. Speech was recorded with an eight-channel microphone array at a sampling rate of 48 kHz, but in this experiment only one channel is used to extract features [5].

In the UA-speech dataset, speech intelligibility was assessed by five native English listeners for each dysarthric speaker. The listeners had no experience of transcription and working with a person involved speech disorders. They were asked to orthographically transcribe each work uttered by a given dysarthric speaker and their confidence for the corresponding transcription. The average score among five listeners shows the speech intelligibility of each speaker which is ranged between 0 to 100. Speakers are categorized in four groups defined as very low (0-25%), low (26-50%), middle (51-75%), and high (76-100%)[5]. Table 2.2 shows information of each individuals in the UASpeech dataset.

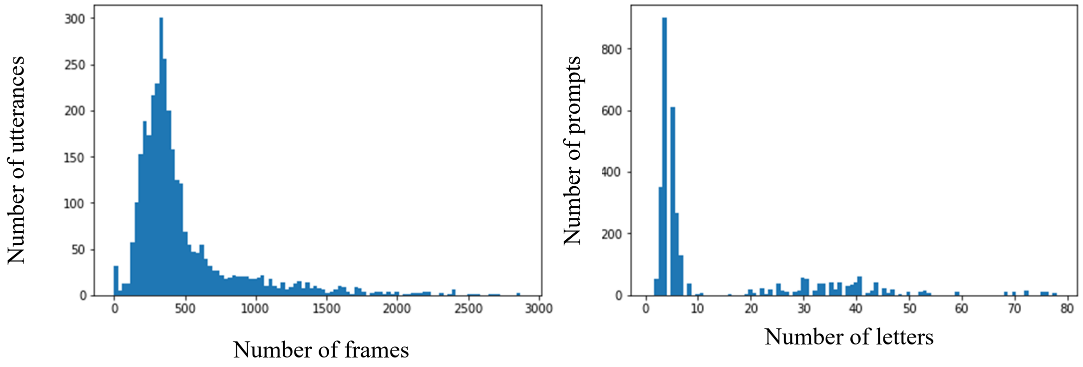
Table 2.2: Properties of various participants in UASpeech dataset [5]

Speaker	Age	Speech Intelligibility	Dysarthria diagnosis
M01	>18	Very Low	Spastic
M04	>18	Very Low	Spastic
M05	21	Mid	Spastic
M06	18	Low	Spastic
M07	58	Low	Spastic
M08	28	currently being rated	Spastic
M09	18	High	Spastic
M10	21	currently being rated	Mixed
M11	48	Mid	Athetoid
M12	19	currently being rated	Mixed
M13	44	currently being rated	Spastic
M14	40	currently being rated	Spastic
F02	30	Low	Spastic
F03	51	Very Low	Spastic
F04	18	Mid	Athetoid
F05	22	High	Spastic
M01	>18	Very Low	Spastic
M02	>18	High	Spastic
M03	>18	Low	Spastic
F01	>18	Low	Spastic

2.2.4. Statistics of TORGO dataset and comparison with a typical speech dataset

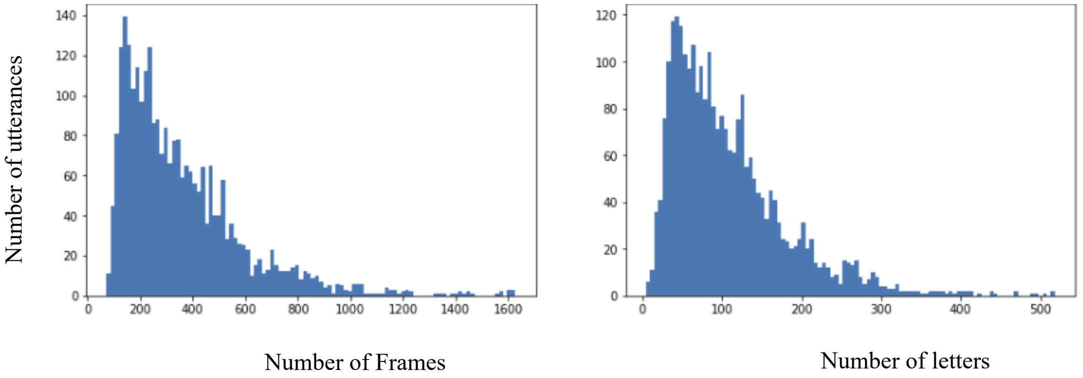
As a preliminary study for this dissertation work, a comparison study of a normal small-scale speech dataset, the Librispeech development set, with TORGO dataset to see some differences between them. The right-hand histogram of Figure 2.6 and Figure 2.7 show the length of prompts in the TORGO and Librispeech datasets, respectively. As shown, most of the prompt's length in letter in the TORGO dataset is shorter than 10 as letter is directly predicted in output of recent ASR.

However, the length of most utterances in Librispeech are between 20 and 150 letters. This highlights the previously mentioned problem of data sufficiency. We do not have access a dysarthric dataset with sufficiently long utterances to train ASR systems. In addition, although the utterances in TORGO are shorter than those in Librispeech, the number of speech frames are much higher than Librispeech because of the slower speaking rate among dysarthric talkers. The histograms on the left in Figure 2.6 and Figure 2.7 depict this information, which indicates the need to handle long-length utterances for dysarthric speech.



(a) (b)

Figure 2.6: Distribution of a) acoustic frame b) text length for TORGO Dataset



(a) (b)

Figure 2.7: Distribution of a) acoustic frame b) text length for Librispeech Dev Set

2.3. Augmentation and synthesis technologies

2.3.1. Generative Adversarial Network (GAN) and Cycle-consistent GAN

Generative Adversarial Networks (GANs) have grown in popularity in the deep learning community due to their ability to generate data based on a training data distribution. The GAN architecture was first introduced by Ian Goodfellow and his colleagues in 2014, and since then it has been widely used in many applications, including domain adaptation, image-to-image transformation, and data augmentation. Furthermore, researchers have successfully worked to improve the initial version of GAN. Consequently, there have been variants of GAN such as Convolutional GAN (CGAN), Adaptive Boosting GAN (AapGAN), and CycleGAN. There are two key problems being addressed in the GAN area. The first problem is that it is hard to train a GAN model, because of what is known as a collapse issue. It is straightforward for the GAN to achieve Nash equilibrium during training, however, there is an imbalance in the convergence characteristics of the two internal network components, which will be described in more detail in the next section. The second problem is that it is hard to measure the similarity and dissimilarity between real data and its corresponding generated data. Currently, research in this area is focused on practical applications that allow for clearer evaluation functions for training [78].

2.3.1.1 GAN architecture

A GAN consists of two components, the generator and the discriminator. While the generator produces a sample from a desired training data distribution, the discriminator is a binary classifier that determines whether the sample came from training set (a real sample) or was generated by the generative model. The adversarial training between the generator and discriminator leads both models to enhance their abilities until generated samples are indistinguishable from actual ones by the discriminator model [79]. In Figure 2.7 shows the general architecture of GAN.

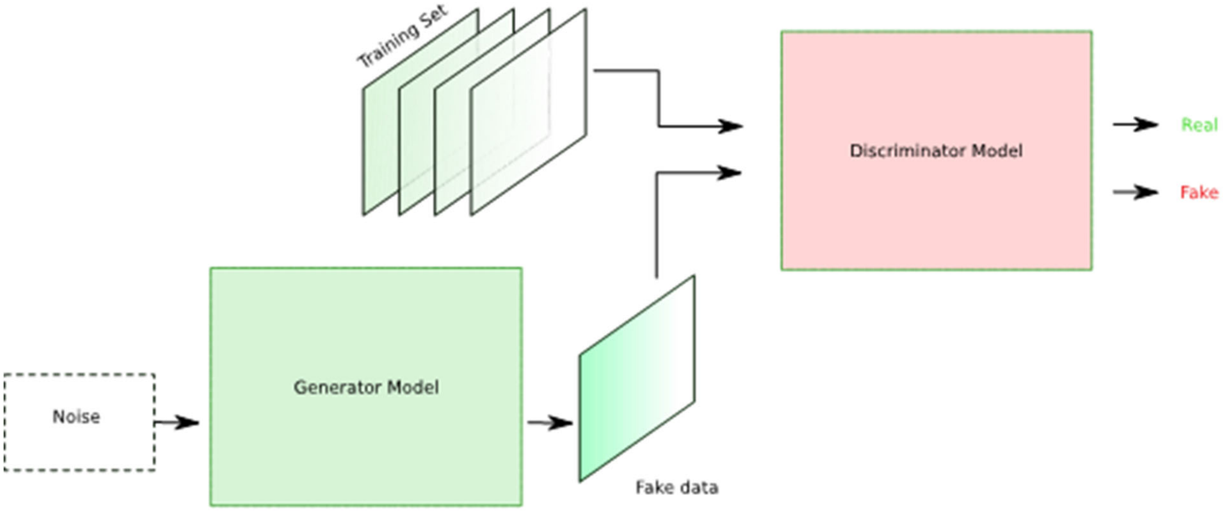


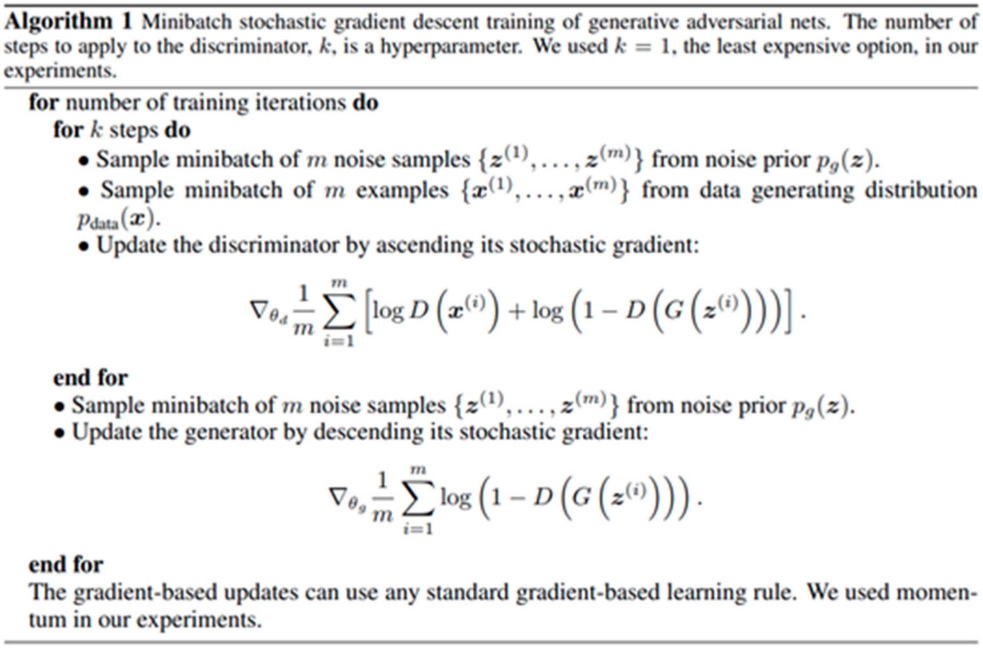
Figure 2.8: Generative Adversarial Networks (GAN) model

To implement a GAN, a prior probability on input noise variables $P_z(z)$ is first defined to specify the generator' distribution P_g over data *x*. Second, a mapping is represented by Generator G which is the based-on input noise variable *z* and the networks parameters *β,* resulting in generating fake data in the output of the generator. Third, to judge the output of generative model, a discriminator *D(x, P_d)* is defined with a scalar output. The discriminator *D* is trained to maximize the probability of correctly labelling both training examples and generated samples, while the generator is simultaneously trained to minimize log(1-D(G(z))), which is equivalent to minimizing the accuracy of the discriminator. Overall, this means that the generator *G* and the discriminator *D* have a minmax relationship as follows [79]:

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim P_{data}(x)}\left[LogD(x)\right] + \mathrm{E}_{z\sim P_z(z)}[\log(1-D(G(z)))] \qquad 1.10$$

Practically, there is an imbalance in the speed of convergence and the amount of training data needed for the generator and the discriminator, and because of this there may be inadequate information for the generator *G* to learn properly. In the early stage of training when the generator is poor, the discriminator can easily detect the generated sample with high confidence. To address this problem and have much stronger gradients early in learning, the generator *G* can be trained to maximize log *D(G(z))* instead of training *G* to minimize log(1-D(G(z))). Table 2.2 shows how GAN works step by step [79].

Table 2.3- GAN algorithm [79]



Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

```
for number of training iterations do
    for $k$ steps do
        • Sample minibatch of $m$ noise samples $\{z^{(1)}, \dots, z^{(m)}\}$ from noise prior $p_g(z)$.
        • Sample minibatch of $m$ examples $\{x^{(1)}, \dots, x^{(m)}\}$ from data generating distribution $p_{\text{data}}(x)$.
        • Update the discriminator by ascending its stochastic gradient:
              $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[ \log D\left(x^{(i)}\right) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right) \right].$
    end for
    • Sample minibatch of $m$ noise samples $\{z^{(1)}, \dots, z^{(m)}\}$ from noise prior $p_g(z)$.
    • Update the generator by descending its stochastic gradient:
          $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right).$
end for
The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.
```

2.3.1.2 Cyclic-GAN

A GAN effectively trains the generator to synthesize entirely new data from a random input, learning the probability distribution of the data itself. In some applications, there is a need to transform data from one domain to another without having pair data for training, rather than synthesizing from scratch. In this situation, two GANs can be connected in an invertible cyclic structure, called a CycleGAN. The main goal of CycleGAN to learn mapping function between two domains *X* and *Y* given the training data X_i and Y_i. As illustrated in the following figure, there exist two cycles with two generators. In the first cycle or forward direction, generator G maps the *X* to *Y* and then return it back to the source domain by generator *F*, the output of which is a predicted $\hat{X}$ that in theory should match the original input. In the backward direction, the generator *F* transforms *Y* to *X* and then Generator G return it back to the target domain Y'. To train these generators, there exist two discriminators D_x and D_y, where D_x aims to distinguish between real samples x and translated samples *F(Y)*; similarly, the D_y tries to discriminate between y and *G(X)*. In order to learn the two mappings simultaneously, adversarial and cycle consistency losses are defined. The adversarial loss matches the distribution of the generated sample to the target domain's data distribution, and a cycle consistency loss enables the model to avoid contradicting the learned generates *G* and *F* [80].

Adversarial losses are applied to both mapping functions *G* and *F*. For mapping function *G* which transforms *X* to *Y* and its discriminator D_y. The adversarial loss is defined as:

$$L(G, D_Y, X, Y) = \mathrm{E}_{y \sim P_{data}(y)}\left[Log D_Y(y)\right] + \mathrm{E}_{x \sim P_{data}(x)}[\log(1 - D_Y(G(x)))], \qquad 1.11$$

where the generator *G* aims to produce a sample similar to the target domain's samples *Y* and the discriminator D_y wants to detect the real sample Y and generated sample by generator G. The adversarial loss works similarly for generator G and D_x. The generator *F* wants to produce a sample to be similar to the samples in source domain *X*, while the discriminator D_x aims to differentiate between the real sample *X* and the generated sample by *F(Y)* [80].

In addition to the adversarial loss, a cycle-consistent loss is used to design better possible mapping functions. For example, the cyclic mapping function *F* should be able to return back the generated sample G(x) = $\hat{X}$ to X in the forward cycle. Similarly, the cyclic mapping function *G* should be able to convert generated sample F(Y)=$\hat{Y}$ back to target domain[80]. To represent this constraint, the following cyclic consistency loss is formulated:

$$L_{cyc}(G, F) = \mathrm{E}_{x \sim P_{data}(x)}\left[\| F(G(x)) - x \|_1\right] + \mathrm{E}_{y \sim P_{data}(y)}\left[\| G(F(y)) - y \|_1\right]. \qquad 1.12$$

When these adversarial and cyclic-consistency losses are combined, the ultimate objective loss will be:

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F), \qquad 1.13$$

Where λ is an parameter to regularize the importance of the two criteria [80]. Given this loss structure, the training objective is to minimize the following equation. Figure 2.9 represents the forward and backward directions of CycleGAN.

$$G^*, F^* = \arg\min_{G,F}\max_{D_X, D_Y} L(G, F, D_X, D_Y). \qquad 1.14$$

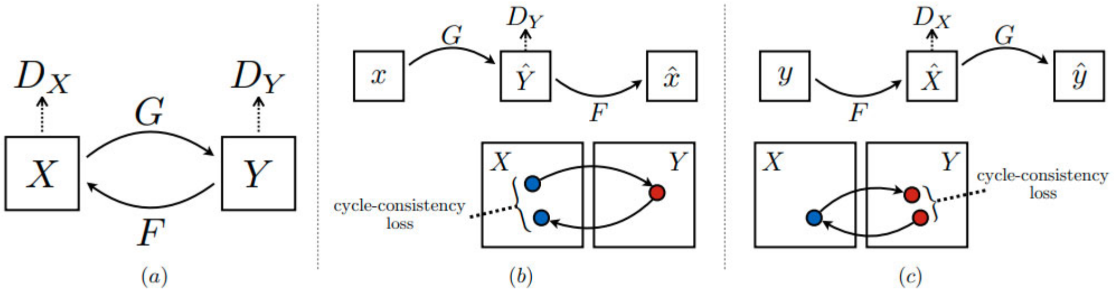


Figure 2.9: Forward and Backward mapping direction of CycleGAN [80]

2.3.2. Voice Conversion

Voice conversion transforms the identity of the target speaker into that of the source speaker without changing the linguistic content. This method is used in speech-related applications such as speech synthesis, animation production, and identity protection [81]. Voice conversion is categorized by their training setups, vocoders, and the other parameter modification applied. As shown in Figure 2.10, a typical voice conversion system consists of speech analysis, feature mapping and speech reconstruction.

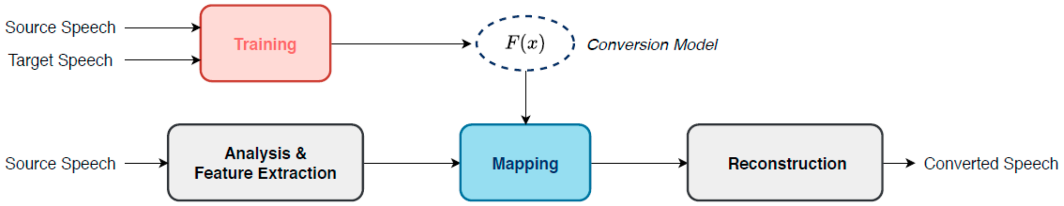


Figure 2.10: A typical block diagram of voice conversion systems [82]

There are two main categories of vocoder used for voice conversion and reconstruction, hand-designed vocoders such as STRAIGHT[83] and WORLD[84], and neural vocoders WaveNet[85] and WaveRNN [86]. A majority of vocoders such as STRAIGHT or and WORLD are designed based on the source-filter model of speech production. In this type of vocoders, speech parameters including spectrum, aperiodicity component, and fundamental frequency are extracted.

2.3.2.1 Related work

Voice conversion studies are mainly divided into parallel and non-parallel approaches, according to the type of training data used. The early studies of the voice conversion have been focused on parallel approach where source and target utterances during training are the same.

Methods like vector quantization [87], Fuzzy vector quantization [88], and dynamic time warping [89] are some of those that have been investigated so far. In addition, non-negative matrix factorization [90] is one of the successful non-parametric methods. After the advent of deep learning, several studies have been conducted parallel and non-parallel approaches. Laskar et al have implemented an artificial neural network to capture the nonlinearity of vocal tract characteristics between the source and target domain [91]. The results demonstrated that the proposed ANN based model can be an alternative for GMM based voice conversion. Wu et al [92]have designed a nonparametric framework for voice conversion, exemplar-based sparse representation with residual compensation. In this method, a spectrogram is reconstructed as weighted linear combination of speech segments. In [93], a DNN has been proposed to convert both timbre and prosodic features. The timbre feature is a high-resolution spectral feature and the prosodic ones are F0, intensity and duration. According to objective and subjective evaluation, the DNN based voice conversion can generate high-quality converted speech.

In recent non-parallel voice conversion, Wu et al [94] have proposed the use of average voice model and i-vectors for LST based voice conversion without a need for parallel training data. Subjective evaluation indicated the effective ness of the proposed method. In [95], a flexible spectral conversion framework based on variational auto-encoder was proposed that facilitates training without aligned data. Both subjective and objective evaluation on VCC2016 speech corpus demonstrated that the results is comparable to the baseline trained on aligned data. More recently, Hsu et al [96] have investigated non-parallel VC framework using a variational autoencoding Wasserstein generative adversarial network (VAW-GAN). The generative adversarial network focus on explaining the observation with latent variables. The results on VCC2016 dataset proved the effectiveness of the proposed framework with an improved conversion quality. In [97], the authors have designed a parallel-data-free voice conversion based on cycleGAN, called CycleGAN-VC. A CycleGAN consists of forward and inverse mappings simultaneously using adversarial and cycle-consistency losses. This enables the model to find an optimal pseudo pair from unpaired training data[97]. Another GAN-based voice conversion approach using cycle-consistent adversarial network for non-parallel has been presented in [98]. Subjective evaluation showed that their results outperformed the baseline using the Merlin open-source neural network speech synthesis system. In [99], the authors have developed the GAN-based voice conversion using a speech enhancement method. This speech enhancement method is used to improve the low-quality pre-existed data and then used them to train voice conversion. Their results represented the enhance models significantly improved the SNR and similarity without degrading the naturalness of the voice. In StarGAN [100], non-parallel manyto-many voice conversion (VC) using a version of GAN. As mentioned in this work, this approach does not require a large number of data, simultaneously learn many-to-many mappings, is parallel-data-free.

2.3.3. Neural Speech Synthesis

Speech synthesis, also known as text-to-speech, aims to generate natural and intelligible speech given input text. The first computer-based speech synthesizer was invented in the mid-20th century [101]. Since then, people have tried to increase the quality in terms of naturalness and intelligibility of synthesized speech. The initial methods of speech synthesis were based on articulatory and formant synthesis, and later concatenative synthesis was introduced.

Formant Synthesis is based on individually controllable formant filters to generate accurate estimations of the vocal-track transfer function. This method was the main speech synthesis until the early 1980's, which was known as a rule-based speech synthesis [102]. The basic assumption of formant synthesis is to model vocal tract transfer function by simulating formant frequencies and formant amplitudes [102]. The synthesis is a source-filter-method that is based on mathematical models of the human speech organ. In the formant synthesis model, the sound is generated from a source, which is periodic for voiced sounds and white noise for obstruent sounds. This basic source signal is then fed into the vocal-tract model. This signal passes into oral cavity and nasal cavity and finally it passes through a radiation component, which simulates the load propagation characteristics to produce speech pressure waveform[103].

Articulatory Synthesis generates speech by modeling of Human articulator motion[102, 104, 105]. Articulatory speech synthesis applies mechanical and acoustic models of speech production to synthesize speech. Articulatory speech synthesis transforms a vector of anatomic or physiologic parameters into a speech signal with predefined acoustic properties. It produces a complete synthetic output, based on mathematical models of lips, teeth, tongue, glottis and velum as well as transit of airflow along the supraglottal cavities. Acoustic models contain number of smaller uniform tubes which generate natural speech. Articulatory synthesis models have an interim stage, in which the motion of the tubes is controlled by some simple process to model the fact that the articulators move with a certain inherent speed. However, there are two challenges in articulatory synthesis. The first is how to generate the control parameters form the specification, and the second is how to find the right balance between highly accurate model[106].

In concatenative synthesis speech is produced by concatenating the segments of recorded speech followed by post-processing. Generally, concatenative synthesis is able to generate a natural sounding synthesized speech [102, 107, 108]. Speech signal processing of natural speech databases plays a key role in the concatenative synthesis. The segmental database is built to reflect the major phonological features of a language. Concatenation techniques take small units of speech, either waveform data or acoustically parameterized data, and concatenate sequences of these small units together, then processing the concatenated waveforms to adjust prosodic characteristics such as intonation and to minimize boundary artifacts between segments. These types of speech synthesizers have their own drawbacks such as less naturalness, artifacts, and noise [101].

Later, statistical synthesis models were developed for speech production. Statistical speech synthesis models generate speech from previously learned statistical models instead of natural speech segments, requiring much less storage than natural segments [109]. This model mainly, consists of a text analysis module, a parameter prediction module, and a vocoder to convert acoustic features to speech. The text analysis module first processes the input text with steps such as text normalization, grapheme to phoneme conversion, and word segmentation [101]. After processing the input text, linguist features such as phonemes, duration POS tags from various granularities are extracted [101]. These linguistic features along with acoustic features are used to train an acoustic model then a vocoder is used to convert the predicted acoustic features to speech. Although the statistical speech synthesis models could improve the synthesized speech in comparison with the previous models, the intelligibility of speech generated by this model is low due to artifacts and the quality of synthesis is still far away from human speech.

Since 2010, neural speech synthesis models have been developed based on the significant advancements in deep neural networks and recurrent neural networks architectures as well as the hardware technologies that allow implementation of these computationally expensive architectures [101]. The paradigm of initial neural models was adopted to replace HMM for acoustic models. However, research later focused on generating directly acoustic feature from phoneme sequence instead of linguist features instead of linguist features. Wang et al [101] have explored the first neural speech synthesizer with directly generating acoustic feature from phoneme sequence. WaveNet was an early successful neural speech synthesis model that directly generated audio from linguistic features. Some end-to-end models like Tacotron 1/2[110, 111] , Deep Voice 3 [112], and FastSpeech 1/2 [56, 113] were introduced to simplify text analysis modules and directly take character/phoneme sequences as input, and simplify acoustic features with mel-spectrograms. For example, Tacotron [111] is a sequence-to-sequence model for producing magnitude spectrograms from a sequence of characters. It simplifies the traditional speech synthesis architecture by replacing the production of these linguistic and acoustic features with a single neural network trained from data alone. Tacotron applies the Griffin-Lim algorithm [114] to convert the mel-spectrogram to waveform [111]. Deep Voice 3 [115], a fully-convolutional attention-based neural text-to-speech (TTS) system. Deep Voice 3 matches state-of-the-art neural speech synthesis systems in naturalness while training an order of magnitude faster.

Later, fully end-to-end TTS systems are developed to directly generate waveform from text, such as ClariNet [116], FastSpeech 2s [56] and EATS [117].

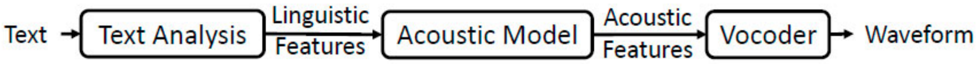


Figure 2.11: Three key components in neural TTS [101]

Since text-to-speech is a one-to-many problem, there are many possible synthesized variants of speech for a given input text[110, 118]. Outputs differ from each other due to pitch, duration, sound volume, speaking style and other prosodic and acoustic characteristics.

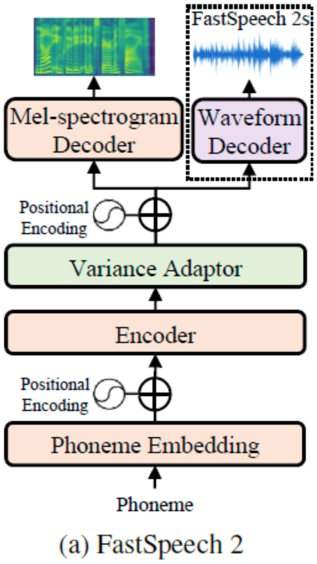


Figure 2.12: The overall architecture for FastSpeech 2 and 2s[56].

One recent neural synthesis architecture is FastSpeech, which is a non-autoregressive model. The FastSpeech architecture is mainly a Transformer, explained in more detail in the next section, consisting of an encoder converting phoneme embedding sequence to phoneme hidden sequence and a Mel-spectrogram decoder that converts the adapted hidden sequence to Mel-spectrogram. It uses a variance adaptor to add additional information like pitch, energy, duration to the phoneme hidden sequence to generate variant speech. In the FastSpeech2 variant of this architecture, [56], there are three predictors of pitch, duration and energy. To better predict variations in pitch contour, a continuous wavelet transform (CWT) is used to decompose the continuous pitch series into a pitch spectrogram and take the pitch spectrogram as the training target for the pitch predictor, which is optimized with MSE loss[56]. The duration predictor takes the phoneme hidden sequence and predicts the duration of each phoneme. Duration of each phoneme determines how many frames in the Mel-spectrogram are corresponded to that phoneme. An Energy predictor computes the L2-norm of the amplitude of each short-time Fourier transform (STFT) frame as the energy [56]. Pitch, duration, and energy predictors have a similar model structure which consists of a 2-layer 1D-convolutional network with ReLU activation, each followed by the layer normalization and a dropout layer, and an extra linear layer to project the hidden states into the output sequence[56].

Multi-speaker variants of speech synthesis systems can learn prosody characteristics, speaker and style variation extracted from the training set, and can use speaker embeddings to generate speech in a variety of speaker styles [40, 119-121]. This synthesis model generates speech based on parameters representing key speech characteristics, allowing control of parameters such duration, energy, pitch, emotion, accent and emotion. Wang et al [122] have proposed a bank of embeddings that are jointly trained within Tacotron, which is called "Global style tokens" or GSTs. This embedding helps the model to learn a large range of acoustic expressiveness and control varying speed and speaking style [122]. Another extension[123] have been presented to the Tacotron speech synthesis model to learn prosodic characteristics by conditioning on the reference acoustic representation. Lee et al [124]have been presented prosody embedding for emotional and expressive speech synthesis architecture. They have proposed temporal structures in the embedding networks, enabling fine-grained control of the speaking style of the synthesized speech [124].In addition, this model has an ability to catch the voice characteristics of individuals using a decoder and speaker embedding, making it a multi-talkers TTS [40] capable of generating speech in a wide range of speaking styles. This allows for generation of relatively large amounts of the high-quality synthesized speech across a range of speaker characteristics and speaking styles.

2.3.3.1 Transformer

The Transformer architecture was first introduced in machine translation to speed up the training process using a self-attention sequence-to-sequence architecture. A Transformer mainly consists of stacked self-attention and point-wise, fully connected, layers for both the encoder and decoder. Figure 2.13 shows the overall architecture of the Transformer [125].

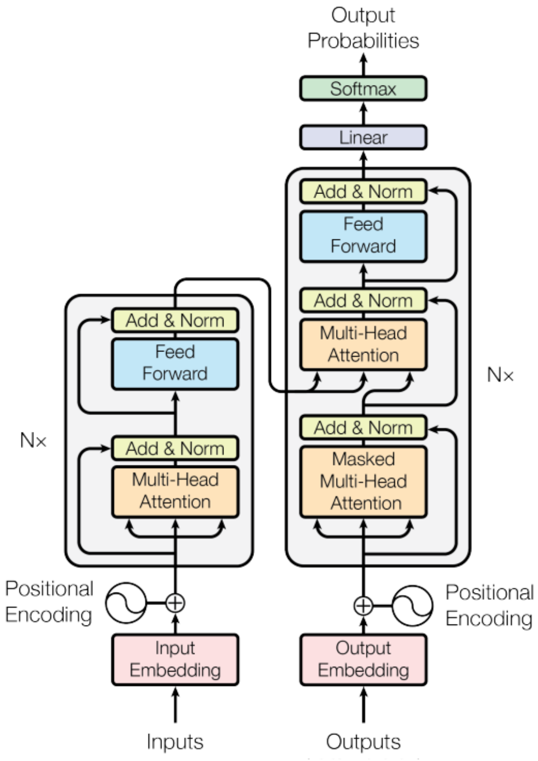


Figure 2.13: The overall architecture for Transformer [125]

- Encoder and decoder

Like the previous sequence-to-sequence model, the Transformer is based on an autoencoder architecture (an encoder-decoder) [126]. The encoder takes an input sequence and maps it to a hidden sequence, containing six identical layers that each layer contains sub-layers of feed-forward neural network and self-attention. Also, a residual connection followed by normalization layer is applied to each of the sub-layers.

The decoder converts the hidden sequences to target representation one element at a time. Similar to encoder, the decoder is composed of a stack of six identical layers. However, each layer here has encoder-decoder attention in addition to feed-forward neural networks and self-attention.

- Attention

Attention is a core component of the Transformer architecture, mapping queries as set of key-value pairs to an output target [125]. There are two types of attention applied in the encoder and decoder, Scaled Dot-Product attention and Multi-head attention. In the former, all of vectors query Q, key K and value V are multiplied by different weight matrices and then every query is compared with every key to find the highly similar keys for each. Practically, the attention function is computed on a set of a queries as following:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad 1.15$$

Multi-head attention enables the model to jointly attend to information from different representation subspaces at different positions [125]. Figure 2.14 represents the two types of attention used in the Transformer architecture.

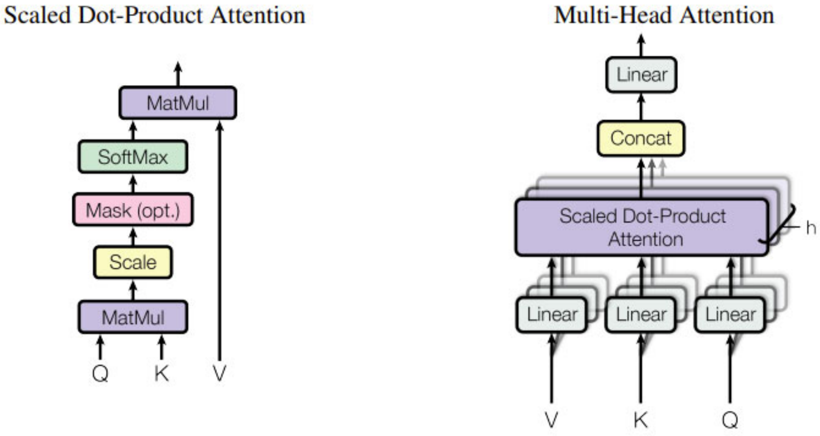


Figure 2.14: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel. [125]

- Positional Encoding

Positional encoding is computed to inject some information about relative or absolute position of tokens/symbols in the sequence as this architecture is lake of recurrence and convolution[125]. To compute the positional encoding, sine and cosine function are used, for more information see section 3.5 in [125].

2.4. Speech data augmentation

Data augmentation is a machine learning technique to generate additional supplemental training data. The purpose of data augmentation is to improve a model's performance or prevent a model from overfitting. Augmentation has been widely applied to many different domains, including both image and speech processing. Image data augmentation is typically divided into basic image manipulations and deep learning approaches. While basic image manipulations contains methods such as kernel filters, geometric transformation, random erasing, mixing image and color space transformations, deep learning approaches introduce adversarial training, neural style transfer and GAN data augmentation [7]. These methods are used in image processing applications such as facial recognition, handwritten digits, medical image diagnosis, content reconstruction, supper-resolution [7].

2.4.1. Typical speech

For speech applications, in addition to manipulation approaches such as adding different noises, pitch modification and speech perturbation, there are methods using voice conversion-based approaches as well as synthesis-based model to generate new speech data. , Augmentation methods have been used to improve typical speech recognition [13-18], clinical speech applications [19-22], voice scene classification [23-25], children speech technologies [26-29], and speaker identification and verification [30-33]. Cui et all [14] have applied Vocal Tract Length Perturbation (VTLP) and Statistical Feature Mapping approaches on Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) for acoustic modeling [14]. Ko et al have [13] proposed the combination of the VTLP and SFM with two stacked architectures with three speed factors to generate new set of data [13], evaluated on 4 LVCSR tasks. The results show an average improvement of 4.3 percent among the four tasks. In [15], data augmentation for far-field ASR was implemented, using a simulated Room Impulse Resonances (RIRs) and evaluating the impact of adding source-point noises as augmentation. Results suggest that the acoustic model trained by simulated RIRs data not only works well in the far-field ASR but also improves the performance in close-talking scenario. In [18], Data Augmentation and ensemble Method (EM) were combined in a single model. First, the VTLP approach and then several feature perturbation methods were carried out to augment the training data. EM techniques were applied to integrate the posterior probabilities of individual DNN acoustic models trained on different sets of data, including voting, averaging, and Linear Logistic Regression (LLR) for Fusion and Calibration. Google DeepMind group has introduced WaveNet which is a simple DNN for generating raw speech waveforms [127]. A single WaveNet can capture the characteristics of each speaker with an equal contribution and can switch between them by conditioning on a given speaker identity. One of the questions addressed by this paper is whether or not WaveNets generates raw speech signals with subjective naturalness in the field of text-to-speech [127]. Wang et al have presented a voice conversion data augmentation using WaveNet, pitch shifting and speech perturbation . The results demonstrate a relative improvement of 10.3% on a speech recognition task. Another approach is the SpecAugment method of Park et al [16], which is a simple method of augmentation [16]. This has been applied directly on the input feature of NNs and consists of the features, masking blocks of frequency channels, and masking of time steps, applied on an end-to-end ASR system. Shahnawazuddin et all [35]have proposed a voice conversion based data augmentation for children speech application using GAN. In this paper [35], the aim is to convert acoustic features of adult into those of children. In another GAN-based data augmentation work [128], data imbalance of training data was addressed for Speech Emotion Recognition (SER). The results indicate that the proposed method relatively improved the performance the tasks by 10% and 5% in two related datasets.

Chapter 3: Comparison of Typical and Dysarthric Suprasegmental Characteristics

Previously only a few studies have looked carefully at prosody across multiple levels of dysarthric severity. In this chapter, we compare timing and acoustic characteristics of dysarthric and typical spontaneous speech at varying severity levels. This information will help us to understand the main differences between the dysarthric and normal talkers with regard tophonemes, pause rhythm and speaking rate as well as overall acoustic analysis, which will provide essential information for synthesizing dysarthric speech.

Dysarthria is a motor speech disorder, often caused by traumatic injury or neurological disfunction, that decreases speech intelligibility through slow or uncoordinated control of speech production muscles [1, 129]. Talkers with moderate and severe levels of dysarthria struggle to communicate with others through speech due to poor intelligibility[2].

Prosodic characteristics are an important component of methods for dysarthric speech modeling, synthesis, and enhancement, which are themselves important to tasks such as data augmentation for improving dysarthric speech assessment and recognition. To create or modify such prosodic characteristics, we need to understand and model the dominant characteristics of dysarthric speech, including pause, speaking rate, pitch and intensity and find the main differences between dysarthric and typical speech. For example, in dysarthric data augmentation, prosodic features and models are needed to control duration, pitch, and intensity profiles. In dysarthric speech enhancement, we need to analyze the difference between the original dysarthric speech and enhanced speech in terms of prosodic match, which is an important part of naturalness.

This chapter investigates suprasegmental characteristics between typical and dysarthric speakers at varying severity levels, with the long-term goal of improving methods for dysarthric speech synthesis/augmentation and enhancement. First, we aim to analyze phonemes, speaking rate and pause characteristics of typical and dysarthric speech using the phoneme- and word-level alignment information extracted by Montreal Forced Aligner (MFA). Then, pitch and intensity declination trends and range analysis are conducted. The pitch and intensity declination are measured by fitting a regression line. These analyses are conducted on dysarthric speech in TORGO, containing 8 dysarthric speakers involved with cerebral palsy or amyotrophic lateral sclerosis and 7 age- and gender-matched typical speakers. These results are important for the development of dysarthric speech synthesis, augmentation to statistically model and evaluate characteristics such as pause, speaking rate, pitch, and intensity.

3.1. Related work

Some prior research has been conducted to analyze and evaluate dysarthric speech characteristics. Bunton et al [130] have analyzed prosody of dysarthric talkers with a perceptual rating. Acoustic measurements such as fundamental frequency f0 and intensity measures in a tone unit, the basic unit of intonation in a language, were computed in conversational speech. Their findings indicated that Amyotrophic Lateral Sclerosis (ALS) subjects with poor intelligibility have greater range than that of ALS subjects with good intelligibility.

Bigi et al [36] have compared speaking style in dysarthric and healthy groups using a syllable-based analysis. Their finding shows mean syllable-based speaking rates in groups for the healthy speakers were higher in comparison with dysarthric speakers. In another study, Zhang et al [37] have analyzed articulatory and acoustic features of dysarthric speech such as the distribution of the duration of repeating 'ah-p-iy', autocorrelation function acoustic signal of 'ah-p-iy' and its corresponding intensity. This work demonstrated that dysarthric talkers do not have full control of source excitation and that the energy of dysarthric speech decays gradually from the beginning.

The effect of sentence length on intelligibility, speaking rate and pause duration in people with amyotrophic lateral sclerosis (ALS) was studied in [131]. Findings showed that pause and speaking rate have direct relationship with utterance length among dysarthric talkers. Yunusova et all in [38] have conducted speech and pause analyses in a reading aloud task for patients with ALS and Frontotemporal Dementia (FTD). Their findings demonstrated differences between patient and healthy groups on the passage reading task. Kuo and Tjaden [39] have examined acoustic variations in a passage reading task for talkers with dysarthria in Slow, Loud and Habitual conditions, with variation in characteristics comparable across the three conditions. Feenaughty et al [132] has assessed features such as speech, articulatory rate, pause type and duration in two tasks of oral reading and narrative speech in Multiple Sclerosis (MS). Results supported the predicted differences in overall speech timing for speech tasks that are different in cognitive-linguistic demand.

Rudzicz et al [4] who collected the TORGO dataset have analyzed the mean duration of vowels and the consonants in dysarthric and healthy groups. The mean duration of each vowel and selected consonants in the dysarthric group was 33% to 63% higher than that of normal speakers.

3.2. Methodology

To compare typical and dysarthric speech timing, we applied Automatic Speech Recognition (ASR) based forced alignment to obtain phoneme- and word-level alignment information on the TORGO dataset, a publicly available dataset containing dysarthric speech described previously in section 2.2.3. Then, we analyzed the mean duration of vowels and the consonants find the speaking rate difference between dysarthric and typical speech. This timing information is essential to model speaking rate across severity levels. Next, the mean pauses duration and the occurrence of pause between words per sentence for each speaker and each dysarthria severity level werecomputed, to model the pause rate and duration for various severity levels. Two other important prosody characteristics of speech, pitch and intensity, werealso evaluated for each speaker. In addition, we computed f0 and intensity declination [133, 134] to understand pitch contour and intensity changes over time. Finally, speaking rate was assessed.

3.2.1. Dataset

The dataset used in this work is TORGO [4]. As described section 2.2.3, TORGO contains 8 dysarthric speakers involved with cerebral palsy or ALS and 7 age- and gender-matched typical speakers. This dataset consists of non-word, short words, restricted and non-restricted sentences. There are an average of 700 utterances for each dysarthric talker; whereas for normal speakers the average is 1560 [129] . Dysarthric speakers in the TORGO data are categorized into three dysarthria severity levels, Very Low, Low, and Medium and into two groups for intelligibility, intelligible and non-intelligible [129]. The standardized Frenchay Dysarthria Assessment by a speech-language pathologist was applied to investigate the motor functions of each subject [4].

3.2.2. Phonetic, pause and speaking rate analyses

The first analysis focused on phonemes, pauses and speaking rate analysis. To investigate this, a noise reduction was applied to reduce white noise. To analyze phoneme and pause differences of dysarthric and normal speech, the Montreal Forced Aligner (MFA) [135] was trained on the dysarthric and normal speech separately. This aligner is an open-source alignment system built on top of Kaldi [136], an open-source automatic speech recognition system. MFA uses Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) ASR systems adapted from Kaldi recipes. This aligner first trains monophone GMMs to generate an initiative alignment and then train triphone GMMs to tackle sparsity in generating the ultimate alignment and predict accurate boundaries. MFA uses 13 Mel Frequency Cepstral Coefficients (MFCC) [137], Delta and Delta-delta on a window size 25 ms with a frame shift of 10 ms [135]. The word- and phoneme-level alignment information were used for the following analyses.

3.2.2.1 Phonemes

Using the phoneme-level alignment information, the mean duration of all vowels and a selected group of consonants were computed. First, each sub-vowel duration in milliseconds was calculated, including /aa/, /ae/, /ah/, /ao/, /aw/, /ay/, /eh/, /er/, /ey/, /ih/, /iy/, /ow/, /oy/, and /uw/. These were then used to calculate the mean duration of individual vowels including both monopthongs and dipthong. Secondly, we obtained the mean duration for selected consonants including /l/, /w/, /y/, /ng/, /n/, and /m/.

3.2.2.2 Pause

To compute accurate pause duration and occurrence rate for various severity levels of dysarthric speech, we calculated the pause duration and counted the pause occurrence between words in all sentences from word-level alignment information. The silences at the beginning and ending of utterances were excluded.

3.2.2.3 Speaking rate

Both words per minute and syllables per second were computed as metrics of speaking rate. For words per minute, the duration of all utterances spoken by each speaker were first calculated, with silence at the beginning and end of each utterance excluded. Then the number of words in all utterances for the corresponding speaker was obtained. Finally, the fraction of the number of words per the overall duration shows the speech rate. For syllables per second the procedure is the same, but speech duration in second was divided by the number of syllables for each speaker, a python function was used to extract syllables of each word.

3.2.3. Acoustic analyses

In this analysis, we computed pitch contour and intensity to figure out how dysarthric and typical speech are different with regard to prosodic characteristics in overall range and long-term trends. To this end, pitch/intensity range and declination analyses were used as indicators of speech variability.

3.2.3.1 Pitch

A python wrapper was implemented to script Praat [138] commands for extracting pitch contour with 10 ms time step, floor 75 Hz and pitch ceiling 500 Hz. For pitch range, the pitch contour between the 25th and 75th percentiles were determined and then the range of the pitch contour was computed. For pitch declination, an ordinary least-squares linear regression was fitted to the f0 contour and the slope coefficients of the regression line was used as a measure of declination [133, 134]. Before applying the linear regression, the zeros at the beginning and ending of pitch contour were trimmed.

3.2.3.2 Intensity

Similarly, a Praat [138] Python script was also used to extract intensity information. To calculate the intensity range, we measured the intensity range as the difference in intensity between the 25th and 75th percentiles. In addition, to calculate the slope coefficients of the intensity, a moving average with a window length of 200 was applied and then a linear regression was fitted for each utterance to model the declination of the intensity envelop.

3.3. Results and Discussion

Figure 3.1 shows the mean duration of all vowels for each speaker. The vowel duration of individuals with moderate dysarthria severity is 28% greater than that of normal talkers in spontaneous speech. However, the duration among the very low and low groups is not significantly different. Figure 3.2 represents the mean duration of the consonants for each speaker as well. While the mean duration of the consonants among moderately dysarthric individuals (the group with highest severity level in the dataset) was longer by 82%, there is not a significant different among two other groups. Table 3.1 shows these results based on the dysarthria severity level.

Rudzicz et al in [4] compared the mean duration of each vowel and the consonant set in two groups of dysarthric and control speakers, and found that the mean duration of each vowel and the consonants in the dysarthric group was 33% to 63% higher than that of normal speakers. Our results are consistent with that finding but are more comprehensive in evaluating across different speakers and severity levels.

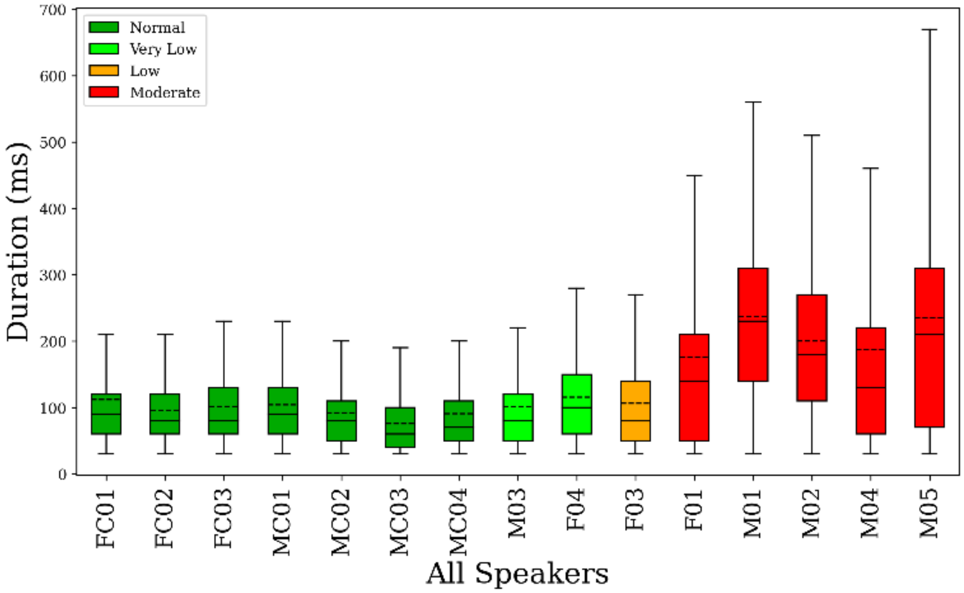


Figure 3.1 Mean duration of vowels for each speaker

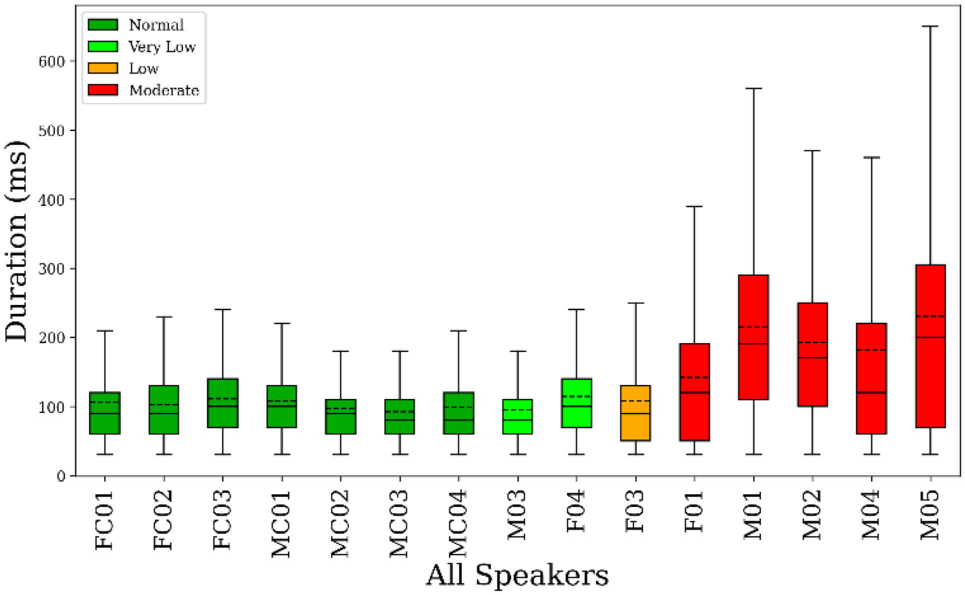


Figure 3.2: Mean duration of the consonants (/l/, /w/, /y/, /ng/, /n/, /m/) for each speaker

Table 3.1: Descriptive statistics across groups—mean(std)

	Normal	Very Low	Low	Moderate
Vowel Mean Dur. (ms)	114(14.1)	136(13.7)	142(0)	259(25)

Consonant Mean Dur. (ms)	101(4.6)	102(9.93)	105(0)	184(29)
Speaking Rate (syll. per sec)	3.56(.34)	3.31(.16)	3.21(0)	1.76(.31)
Speaking Rate (word per min)	147(12)	137(8.2)	130(0)	77(16.4)
Pause Duration	151(68)	246(37)	321(0)	580(171)
Pause Occurrence	0.26(.1)	0.57(.27)	1.21(0)	2.51(.88)

Figure 3.3 depicts the syllables per second speaking rate for each speaker. There is a clear decrease in speaking rate for very low, low, and moderate severity, decreasing by 5.6%, 8.5%, and 49.8% respectively. Table 3.1 also shows that the speaking rate based on word per minute is lower by 6.8%, 11.5% and 47.6%, respectively, for the same three groups, indicating consistency between the two-rate metrics.

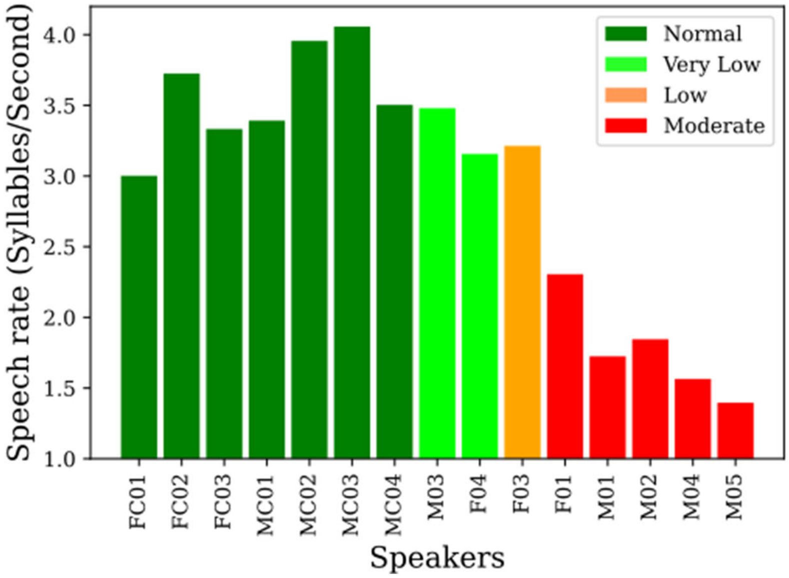


Figure 3.3: Speech rate of each speaker (syllables per second)

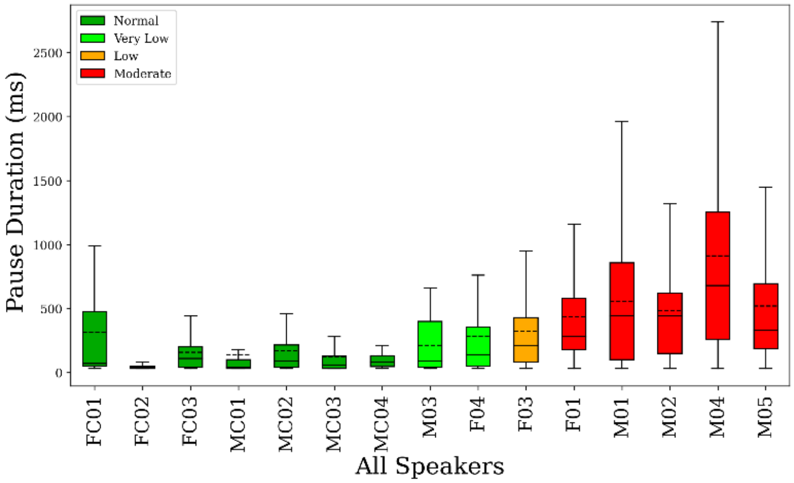


Figure 3.4: Mean duration of pause occurred between words for each speaker

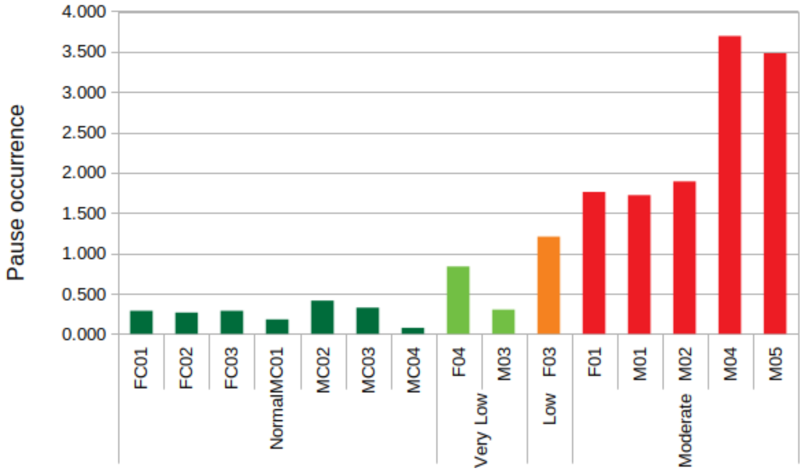


Figure 3.5: Pause occurrence between words per utterance

In [132], the speaking rate for narrative task for multiple sclerosis (MS) was about 3.7, 3.5 and 3.1 syllables per second for control, MS with lower severity, and MS with higher severity groups, respectively. This is similar to our own findings, but our study shows these results for spontaneous rather than read speech and with more differentiation of severity levels, showing more disparity across groups.

Yunusova et al in [38] examined reading aloud in patients with ALS. The speaking rate reported in this study is 156 ± 27.27 words per minute for the mild group and 176±20.93 for the control group. This is a much less significant difference than we found across groups in spontaneous speech.

Figure 3.4 shows mean duration of pause between words for each speaker. Figure 3.5 represents the number of pause occurrences between words per utterance for each speaker. According to Table 3.1, the mean duration of pause among person with normal, very low, low, and moderate severity group is 151, 246, 321, and 580 milliseconds, respectively, indicating that the mean pause duration among the very low, low and moderate groups is 62%, 112%, and 284% longer in comparison with normal talkers, respectively. The pause occurrence per sentence between words among typical, very low, low and moderate groups is about 0.26, 0.57, 1.21 and 2.51, respectively, showing that this value is 120%, 365%, and 865% longer among the very low, low and moderate groups in comparison with typical speaker.

The effect of sentence length on pause duration was investigated in [131] across persons with dysarthria due to ALS . Their results showed that the pause duration over sentence length for the group with higher severity level was increased by higher rate in comparison with the group with lower severity level.

In [132], the mean pause duration for narrative task was about 680, 600 and 560 milliseconds for MS with higher severity, MS with lower severity and Control group, respectively. The number of pauses for the narrative task was 17.5, 17.1, and 14.95 for MS with higher severity, MS with lower severity and Control group for the given passage, respectively. While there a significant difference between the mean pause duration between the typical group and talkers with low and moderate severity for spontaneous speech, this difference is much smaller for narrative task.

Figure 3.6 and Figure 3.7 show the pitch slope coefficient and range distribution, respectively. The range of pitch declination among dysarthric individuals is less than that of the normal talkers. Also, the distribution of pitch range among dysarthric talkers is greater indicating that dysarthric talkers are less able to control their pitch.

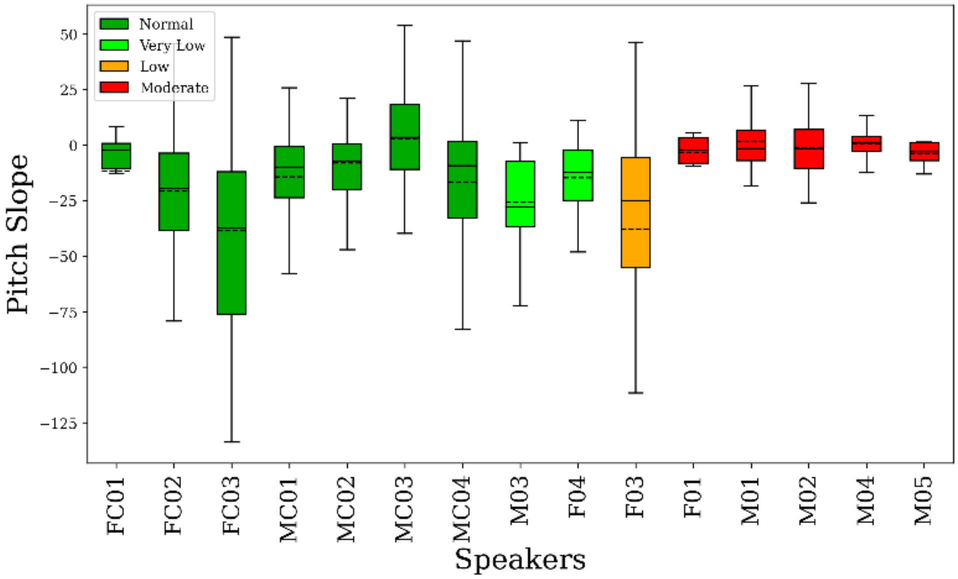


Figure 3.6: Pitch slope for all speakers

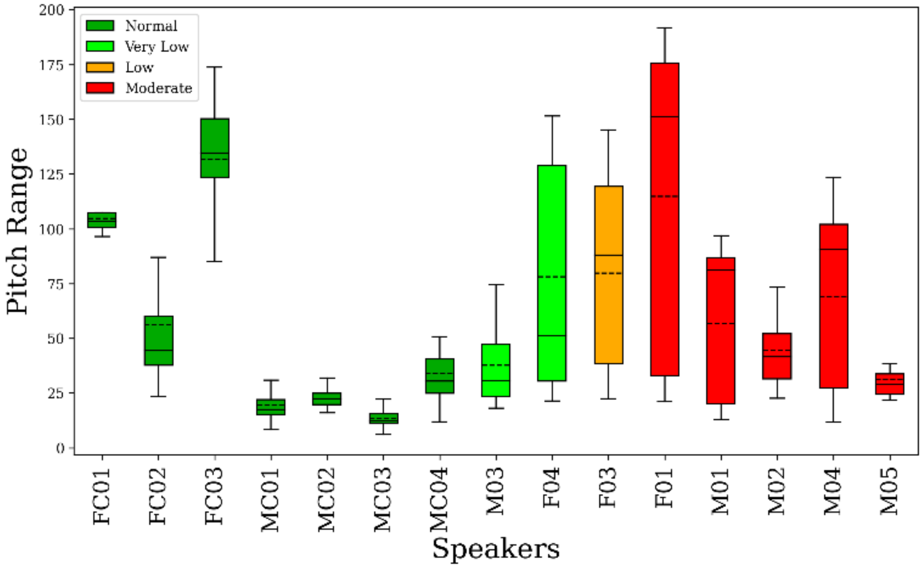


Figure 3.7: Pitch range for all speakers

Bunton et al in [130], conducted an investigation of f0 range for four subject groups. Their findings indicated that ALS subjects with poor intelligibility have greater range than that of ALS subjects with good intelligibility. However, our results indicated that there is not a significant difference between groups with good intelligibility (Very low and Low severity levels) and the group with poor intelligibility (Moderate level). In addition to pitch range, we have considered pitch declination for each individual. This information shows that typical speakers have a greater range of pitch contour changes during speech.

Figure 3.8 and Figure 3.9 show the intensity slope and range distribution, respectively. The intensity results indicate that dysarthric talkers have a wider loudness variation, which can be interpreted as suggesting that dysarthric individuals are probably less able to control their loudness. In addition, the intensity slope shows that loudness decrement among the dysarthric talkers occurs more frequently than within normal talkers during speech.

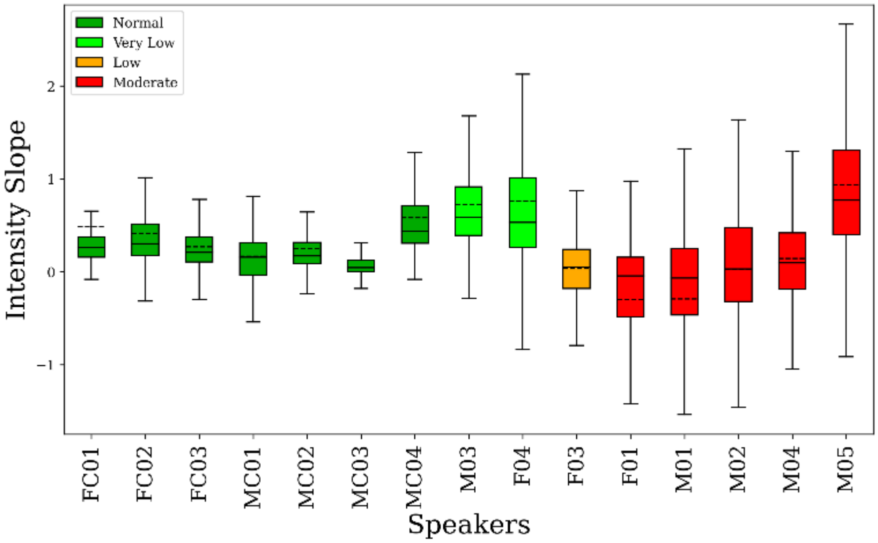


Figure 3.8: Box plot of intensity slope for all speakers

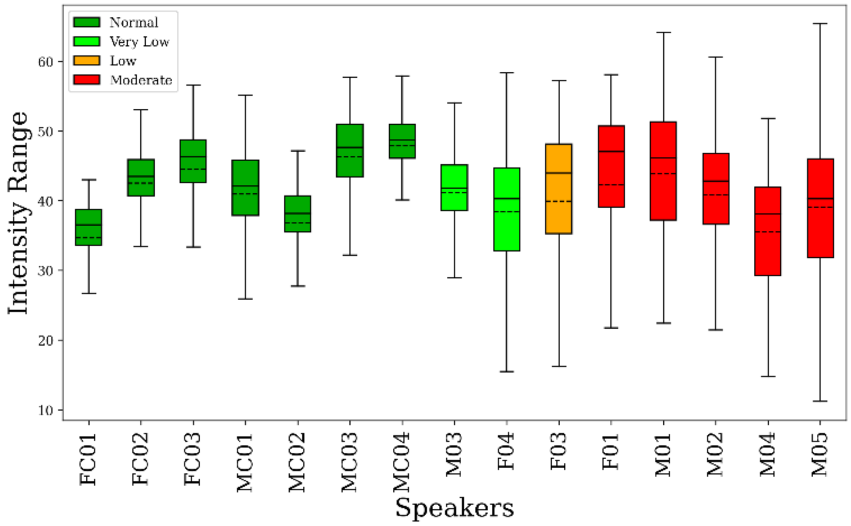


Figure 3.9: Box plot of intensity range for all speakers

In [130], Bunton et al investigated intensity range across syllables in a tone unit for ALS1(less severity), ALS2(higher severity), and control groups. Results showed that the intensity range is 0.75, 0.54, and 0.61, respectively. In contrast, our results suggest that individuals with higher dysarthria severity level have a greater range of intensity. In addition, the intensity slope results in our experiment demonstrate that dysarthric speakers tend to have more negative slope in comparison with non-dysarthric speakers. This may indicate that dysarthric talkers are less able to preserve intensity or loudness over time.

In order to analyze the variation of sound amplitude for dysarthric speech, the short-term energy of the utterance with ‘ah-p-iy’ repetition was calculated in [37]. Their results showed that the amplitude of the peaks has relatively consistent value and no significant envelope deceasing for the speakers without dysarthria, whereas the energy of dysarthric speech decreased gradually from the beginning. In contrast, in our study all sentences of each talkers were used to compute the declination of overall intensity instead of using limited phoneme repetitions. Our findings revealed higherloudness variability among dysarthric talkers than shown in [6].

3.4. Conclusion

In this chapter, we analyzed suprasegmental prosodic characteristics between typical and dysarthric speaker with different severity levels. The phoneme duration, speaking rate and pause characteristics of typical and dysarthric speech were analyzed using the phoneme- and world-level alignment information extracted by MFA. Pitch and intensity declination trends and range analysis were also conducted. Our findings demonstrate that there is a signification difference between the vowel duration between the typical talkers and dysarthric talkers with low and moderate severity. However, the consonant duration differences are less obvious among typical, very low, and low groups. Dysarthric speakers with very low and low severity represent relatively a close speaking rate to that of the typical speakers. Pause duration and occurrence are very distinguishable among various severity levels. In addition, pitch results indicate the variation of pitch among dysarthric speakers is wider in comparison with typical speakers, suggesting that they are less able to control their pitch. Intensity results demonstrate that dysarthric talkers gradually decrease their loudness during speech, and that the intensity variation is more evident among dysarthric speakers. These results are important for the development of dysarthric speech synthesis to statistically model and evaluate characteristics such as pause, speaking rate, pitch, and intensity.

Chapter 4: Dysarthric Speech Augmentation Using Prosodic Transformation and Masking for Subword End-to-end ASR

In this chapter, we explore a specialized data augmentation approach to enhance the results of end-to-end dysarthric ASR. The proposed method contains prosodic transformation and time-feature masking. In prosodic transformation, we modify the speaking rate and shift the pitch to alter vocal excitation characteristics and prosodic structure. Next, we exploit time and feature masking in the spectral domain to alter the MFCCs representing vocal tract acoustics. In addition, we apply sub-word modeling instead of a character-based model because of the high pronunciation variability of the speech. Two experiments are carried out using the proposed approach on the TORGO dataset.

4.1. Introduction

Talkers with dysarthria may exhibit imprecise articulation, irregularities of vocal pitch and quality, atypical nasal resonance, slow and inconsistent speaking rate, inconsistent pauses, as well as altered linguistic stress and speech sound timing [62]. As discussed in section 2.2, speech technologies such as Automatic Speech Recognition (ASR) have the potential to be very beneficial to increase quality of dysarthric speakers' communication.

Early research on dysarthric ASR systems for continuous speech was based on Gaussian Mixture Model-Hidden Markov Models (GMM-HMMs), and later Deep Neural Network-HMM (DNN-HMMs) [64, 139]. In these approaches, there are individual models for acoustics, language, and pronunciation that are separately trained to build an ASR system. Recently, advanced ASR architectures have begun to be applied to the task of building dysarthric ASR systems. Kim et al [70] have investigated the use of Contextual Long Short-Term Memory Recurrent Neural Networks (CLSTM-RNNs) for dysarthric speech recognition. Their experimental evaluation on a dataset collected from nine dysarthric patients showed that their approach provided an improvement over both standard Convolutional Neural Network (CNN) and LSTM-RNN based speech recognizers. Yu et al [140] have considered a range of DNNs such as Time Delay Neural Networks (TDNNS) and LSTM have been developed for dysarthric speech recognition applications. In addition, they trained two out of domain ASR systems and then adapted them to Universal Access Dysarthric Speech (UASpeech) data. Finally, a combined model gave an overall word accuracy of 69.4% on the 16-speaker test set [140]. In another work, to have better feature representation dysarthric speech, Vachhani et al [71] have developed deep autoencoders to improve the dysarthric ASR performance using typical speech. Also, severity based tempo adaptation was analyzed in their work [71]. The results on Universal Access dysarthric speech represented 16 percent improvement. These works have developed their methods on available dysarthric speech datasets which are not designed for speech recognition, such as those previously discussed in section 2.2.3. As discussed, training a robust and reliable speech recognition system requires more dysarthric speech than is currently available.

To address the low number of unique words in publicly available datasets for dysarthria, Harvill et al [141] have proposed a data augmentation method using voice conversion that allows dysarthric ASR systems to accurately recognize words outside of the training set vocabulary. They demonstrated that a voice conversion system can capture the relevant vocal characteristics of a speaker with dysarthria with a small amount of dysarthric speech data. Xiong et al [142] have investigated an improved transfer learning framework to create robust personalized ASR systems for dysarthric talkers. This showed on averaged 11.6% and 7.6% relative recognition improvement in comparison to the conventional speaker-dependent training and data combination, respectively. To further improvement, they analyzed utterance-based data selection of the source domain data based on the entropy of posterior probability. In [143], Shahamiri proposed Speech Vision (SV) systems that cope with challenges like data scarcity and phoneme labeling imprecision. To address the data scarcity problem, the proposed system adopts visual data augmentation techniques, generates synthetic dysarthric acoustic visuals, and leverages transfer learning. Their results on UASpeech dataset showed that the system improved the recognition accuracy 67% of UA-Speech speakers [143].

In addition, other techniques using modern sequence discriminative training like lattice-free maximum mutual information (LF-MMI) [144], have been used for improving dysarthric speech recognition. In [145], Wang et al introduced a reinitialize base model adaptation via meta-learning to obtain better model initialization. Their experimental results on UASpeech corpus showed that the proposed method achieves 54.2% and 7.6% relative word error rate reduction compared with the base model without finetuning and with the model directly fine-tuned from the base model, respectively [145].

End-to-end ASR systems have become a focus of research, showing competitive accuracies with state-of-the-art systems for normal speech[58, 146]. End-to-end systems are jointly trained directly on transcriptions without any need of alignment between the speech waveform and the transcript. End-to-end architectures are robust with respect to different noise backgrounds and speakers[58]. However, to build an end-to-end ASR system, a large amount of training data is required to train the system. While there is a sufficient amount of such data for training an end-to-end ASR system of normal speakers, we do not have access to nearly enough data to effectively train an end-to-end dysarthric ASR system.

There are a few publicly available dysarthric speech datasets, including TORGO [4], UASpeech [64] and Nemours [6]. However, none of these datasets are designed for speech recognition, and using them to support ASR is challenging. Because there is not an adequate amount of conversational speech in these datasets. ASR systems trained with these are often less robust. Furthermore, modern ASR methods assume that training data includes a sufficiently large set of speakers to adequately capture enough inter-speaker variability, but these dysarthric datasets all have a relatively small number of speakers and are not sufficient for an end-to-end ASR system to capture speaker variability[129].

4.2. Methodology

The end-to-end system used for the work presented in this chapter is the Listen, Attend and Spell (LAS) architecture. This method is explained in the section 2.1.3.3 in detailed, so the following section only briefly describes the main components as a review.

4.2.1. Listen, Attend, and Spell (LAS)

LAS is an end-to-end neural network ASR architecture to transcribe spoken utterances to character sequences at each time. It contains two main components, the listener and the speller. The listener takes the audio features as an input and converts it into a higher-level representation feature. The speller is an RNN decoder taking the high level representation from the listener along with the attention vector to generate the output characters. The attention vector uses an attention mechanism to generate probability distribution over character sequences. The goal of the LAS architecture is to model the current output character y at each time step i as a conditional distribution over the previously recognized characters and input speech as follows[58]:

$$P(y \mid x) = \prod_i P(y_i \mid x, y_{<i}) \quad . \qquad (1)$$

4.2.1.1 Listener

The main role of the listener is to take acoustics features X and transform them into a high level representation. To construct the listener, a Bidirectional Long Short-Term Memory (BLSTM) with a pyramid structure is used. The pyramid design is applied to expand the context of the input in an efficient [58], enabling the attention model to find the pertinent information as well as reducing the computational cost, particularity during training.

This pyramid structure is particularly beneficial for domains such as dysarthric speech recognition. One of the characteristics of dysarthric speech is a low and inconsistent speaking rate, and people with severe dysarthria may generate somewhat lengthy acoustic output even for a short sentence. The ability of the BLSTM pyramid structure to capture extended context can be very helpful to handle this challenge.

4.2.1.2 Attend and Spell

The speller component is based on an attention LSTM transducer. The transducer provides the speller a probability distribution at each output step due to all previous characters generated [34]. The context vector is defined as:

$$c_i = AttentionContext\left(s_i, h\right), \tag{3.1}$$

where S_i is the current hidden state and h is the high level representation vector from the listener at each time step i, AttentionContext generates a context vector C_i , containing the information of the acoustic signal needed to emit the next character [58].

The context vector itself is one of the parameters needed to calculate probability distribution of the output characters as well as the hidden decoder state S_i.

$$s_i = RNN(s_{i-1}, y_{i-1}, c_{i-1}) \tag{3.2}$$

$$P(y_i \mid x, y_{<i}) = CharacterDistri \tag{3.4}$$

The decoder state is a function of the previous state, the previously emitted character and context vector [58].

4.2.2. Sub-word Model

Character-based ASR systems generate a character as output, rather than phonetic sequences[146, 147]. However, character-based models face difficulties decoding long utterances because of the extent of the context needed by the attention mechanism. In addition, word-level output is another option that a decoder can generate. World-level ASR is a model that directly learns and generates word-level sequences. However, the model is not able to recognize Out-Of-Vocabulary (OOV) words and requires a large softmax layer to include all vocabulary words, which in turn increases the computational cost [146]. Using sub-words for decoding instead of full words addresses both these issues and has substantial benefit for handling longer utterances and also OOV issues.

In all the experiments included here, we have used Byte Pair Encoding (BPE) to generate the sub-words. Sub-words can be any combination of characters within a word. For example, the decoder's output in our model can generate 'a', 'f', 'th', 'en', 'the', 'is'. Sub-words are particularly applicable to dysarthric speech because of the high pronunciation variability of the speech.

4.2.3. Data Augmentation

The aim here is to use a combination of both prosodic transformation and time-feature masking to generate new speech data. One significant difference between dysarthric and typical speech is that speaking rate may be about twice as slow, an average, for talkers with dysarthria [2, 37]. However, reduced speaking rate may not be consistent and can be quite variable. In addition, dysarthric vocal excitation may be unstable because individuals with dysarthria may not effectively control vocal fold closure and vibration. This may cause inconsistent vocal quality and pitch throughout an utterance [37]. To simulate these characteristics, two functions are designed for lowering speaking rate and pitch-shifting of normal speech. Then, time and feature masking are applied to the MFCC features after prosodic transformation. Figure 4.1 shows the block diagram of the proposed data augmentation. As illustrated here, speaking rate and pitch-shifting modifications are conducted in the prosodic transformation module whereas time-feature masking modifies the MFCC of the speech after prosodic transformation.

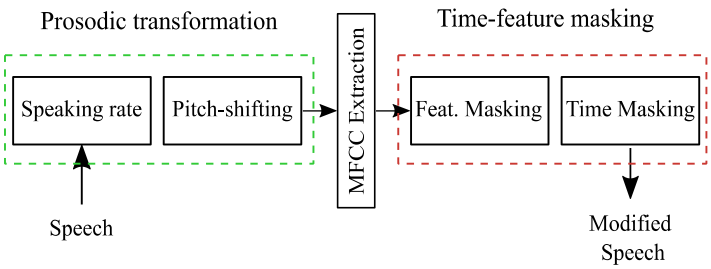


Figure 4.1: Block diagram of data augmentation

The speaking rate of normal speakers is decreased by a multiplicative constant.

Next, in order to simulate variations in sound source and pitch the following equation is used to shift the pitch of each non-overlapping frame as follows:

$$FF_{modified} = FF_{original} * (1 + pitch\ factor) \qquad 3.4$$

where FF indicates the fundamental frequency and the pitch factor is randomly obtain from a uniform distribution between -0.5 to 0.5. and then apply the speaking rate and pitch shifting modification only on speech.

After completing prosodic transformation and extracting MFCCs, time and feature masking are performed to give a new modified MFCC feature matrix. This method is inspired by [16], which applied time and frequency masking on mel spectrogram features. However, here we are performing feature masking instead of frequency. Time masking replaces MFCC coefficients in a selected time range with the mean of all MFCCs for the utterance, while feature masking replaces a selected set of MFCCs with the mean value across the entire utterance. Between 3 and 5 time masks are applied per utterance, chosen randomly, each of duration between 4 and 8 frames, also chosen randomly. Either 2 or 3 feature masks are applied per utterance, each mask of width 1 to 3 coefficients, both parameters chosen randomly. The time-masking is applied in the center 50% of the utterance. The feature masking is performed on the MFCC coefficients or first-derivative of MFCC. Figure 4.2 depicts an example of the augmented speech.

4.3. Experimental setup

We performed two experiments using the proposed approach. In the first, we used normal speech and created augmented speech with dysarthric-like characteristics using prosodic transformation. For each normal speech utterance, the speaking rate was lowered by a multiplicative coefficient of 0.85, 0.7, 0.6, and 0.5, resulting in 4 additional utterances. For each, pitch modification was applied a single time using equation (5). Following this, time and feature masking were implemented on the MFCC features of each new utterance.

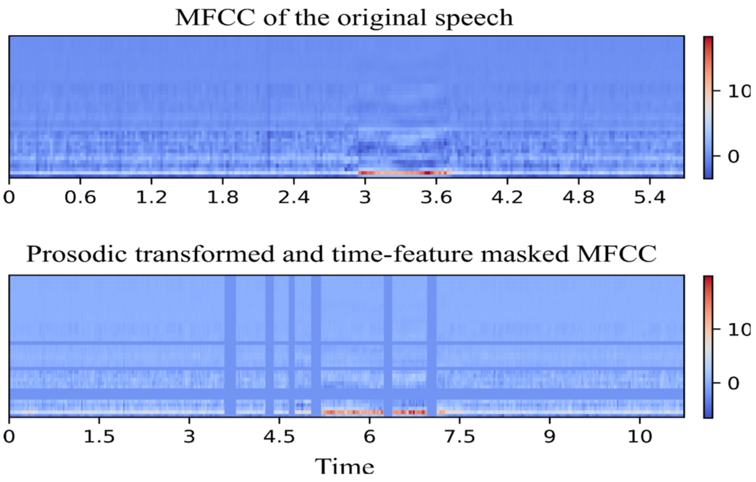


Figure 4.2: MFCC of the original and augmented speech

In the second experiment, both normal and dysarthric speech were augmented. The procedure was the same for normal speakers as in the previous experiment, creating 4 augmented utterances for each original utterance. However, for dysarthric speech only masking was applied. After this, time and feature masking were applied four times on each dysarthric speech utterance to generate the same number of augmented utterances for the dysarthric speech as for the normal speech. This resulted in four times more data than the original data to train the model, excluding the test speaker.

To evaluate the effectiveness of the proposed approach, the methods were implemented on the TORGO dataset. This dataset contains 8 dysarthric speakers and 7 normal speakers, and includes non-words, short words, sentences. The number of utterances for each dysarthric talker averages 700; whereas for normal speakers the average is 1560 [129][4]. Since the data are relatively limited, a leave-one-speaker-out classification method was applied in order to have the maximum data for training and the ability to evaluate the performance of the system on each speaker.

The first 13 MFCCs along with first- and second-order derivatives are extracted as features to represent the input speech. This 39-dimentional feature vector is computed with a window step size of 10 milliseconds and a frame size of 25 milliseconds. To create the sub-word units, the training prompts were used to train the BPE. There are 56 possible sub-word units as outputs of the decoder's softmax layer.

For the Listener, there are 3 layers of pBLSTM with 256 cells (128 for each direction), which in turn reduce the time resolution by a factor of eight. For the Speller function, a single layer of 256 LSTM nodes was used. The loss function for training was Stochastic Gradient Descent (SGD) with learning rate 0.001, epoch 100, and batch size 32. During the decoding, beam search with width of 1, 5, 10 and 15 was used. We chose beam width 10 for all experiments here as it showed the best results in initial experiments.

4.4. Results and discussion

To evaluate the results of the two experiments, Word Error Rate (WER) and Character Error Rate (CER) were calculated for each test speaker. WER of each severity group is compared with a recent work conducted by Yue, Z., et al [139] on the same dataset. Finally, we report CER of the various experiments based on the severity level.

Table 4.1 shows both WER and CER of the two experiments along with the baseline. Using the augmented speech improves the performance of LAS model for each speaker except F01 and M01 in the prosodic transformation plus masking on only normal speech in experiment 1. On average, the first and second experiments reduce CER by 5.3%, 11.3%, respectively. Also, those decrease the WER by 5.6% and 11.4%.

Table 4.1: WER and CER of each test speaker for Prosodic Transformation plus Masking of normal speech (Experiment 1) and Prosodic Transformation plus Masking of both normal and dysarthric speech (Experiment 2)

Severity Level	Spk	Baseline		Exp. 1		Exp. 2	
		CER	WER	CER	WER	CER	WER
Mild	F04	19.0	44.0	16.3	39.3	16.6	35.8
	M03	16.0	37.0	12.9	29.7	14.3	32.8
Moderate	F03	46.0	74.0	45.0	69.0	39.3	62.8
Severe	F01	51.0	76.0	53.2	76.0	49.6	68.1
	M05	59.0	84.0	54.2	78.2	53.2	76.9
	M01	61.0	86.0	59.6	86.5	50.5	74.3
	M02	63.0	88.0	57.2	81.5	54.8	80.1
	M04	64.0	88.0	60.2	84.3	57.7	80.1

In order to represent the effect of the proposed approaches as a function of the level of severity of the dysarthric speech, Table 4.2 compares the average WER for the different dysarthria severity levels.

Table 4.2: Prosodic Transformation plus Masking of normal speech (Experiment 1) and Prosodic Transformation plus Masking of both normal and dysarthric speech (Experiment 2). Both experiments include a combination of isolated word and sentence data.

Severity level	baseline	Exp.1	Exp.2	Results of [139]	
				isolated word	Sent.
Mild	40.5	34.5	34.3	27.0	38.0
Moderate	74.0	69.0	62.8	64.5	65.6
Severe	84.4	81.3	75.9	82.0	86.4

Although both experiments show similar improvement, the second experiment represents better performance among the moderate and severe levels. The WER improvement of the second experiment in comparison with baseline is 15.3%, 15.1%, 10.1% for mild, moderate and severe categories, respectively. Overall, comparison with the augmentation method from [139] indicates that the proposed augmentation method provides more improvement than prior approaches.

Table 3 lists CER based on severity levels. The CER improvement of the first experiments for mild, moderate and severe levels are 16.6%, 2.2% and 4.5, respectively. However, this improvement for the second experiment in comparison with baseline is 12.0%, 14.6 and 10.7%, respectively. These results demonstrate that the second experiment have more effective on the moderate and severe levels whereas only augmenting normal speakers primarily enhances the performance of the mild group.

Table 4.3: CER for different severity levels

Severity level	baseline	Exp. 1	Exp. 2
Mild	17.5	14.6	15.4
Moderate	46.0	45.0	39.3
Severe	59.6	56.9	53.2

Figure 4.3 depicts the training loss for test speaker F03. As shown, extra augmented speech in the first and second experiments prevent the model from early overfitting.

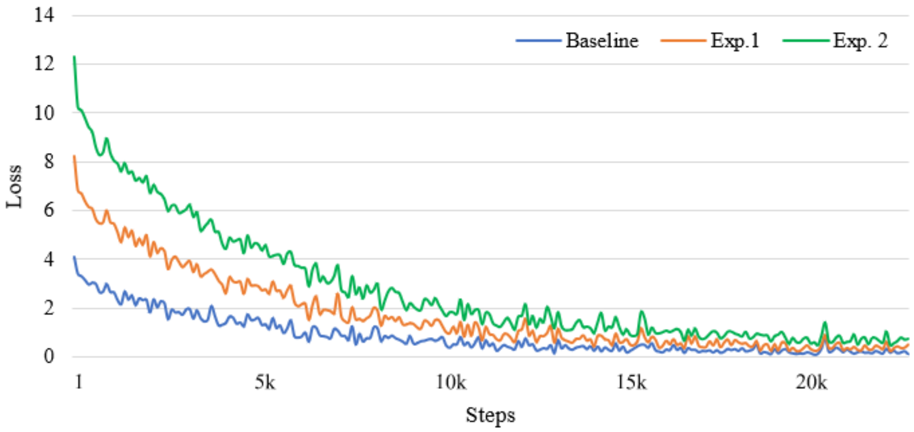


Figure 4.3: Loss of the various experiments for Test speaker F03

4.5. Conclusion

In this chapter, we explored a specialized augmentation approach to exploit an end-to-end ASR system based on sub-word models. The LAS architecture was trained on the TORGO dataset plus augmented speech. The proposed approach contained two methods, prosodic transformation and time-feature masking. The results show that applying prosodic and time- feature masking on both dysarthric and normal speech represent better performance and underscore the need for speech from various dysarthria severity levels. Overall results indicate that using augmentation to increase the amount of dysarthric-patterned speech for training has significant impact on dysarthric ASR systems, particularly for speech with more severe dysarthria.

Chapter 5: Synthesizing dysarthric Speech using end-to-end Text-To-Speech systems

Recent progress in end-to-end TTS systems such as Tacotron [110, 111], FastSpeech [40, 56], Deep-Voice [112] support synthesized speech with high quality and naturalness with varying prosody. These improvements in synthesizing speech inspired us to attempt synthesis of realistic dysarthric speech for ASR training data augmentation. Such neural speech synthesizers have been used to generate new utterances for ASR application for low resource languages [119-121, 148, 149]. Multi-speaker speech synthesis systems can learn prosody characteristics, speaker and style variation extracted from the training set, and can use speaker embeddings to generate speech in a variety of speaker styles [119-121]. This allows for generation of relatively large amounts of the high-quality synthesized speech across a range of speaker characteristics and speaking styles.

In this chapter, we propose a method based on Multi-talker neural TTS to synthesize dysarthric speech to enhance the results of dysarthric ASR. In addition to traditional prosodic variables such as speech rate, energy, and pitch, we add two new variables to control dysarthric severity and extent of pause insertion. These parameters enable us to generate a broad range of synthesized speech to improve the training of dysarthric ASR systems. To assess the effectiveness of the synthetic speech, we evaluate the Deep Neural Network-Hidden Markov Model (DNN-HMM) models with and without augmented speech. Experiments are carried out using the proposed approach on the TORGO dataset.

5.1. Methodology

For the baseline synthesis model, we modified FastSpeech2 [56] and a recent variant [40] to synthesize dysarthric speech. Figure 5.1 shows the main block diagram of the proposed method. In the modified version of the FastSpeech2, the energy, pitch and forced-alignment duration [135] of each speaker's utterances are incorporated into the phoneme hidden sequence through a "variance adaptor" module, resulting in more controllability of these prosodic parameters.

The multi-talker variant of FastSpeech2 decoder works like a voice conversion system, making it a multi-talkers TTS [40] capable of generating speech in a wide range of speaking styles. This is a useful capability for speech synthesis for data augmentation because it allows generation of a robust set of training data.

The prosodic characteristics of dysarthric speech greatly differs from typical speech, specifically at moderate and high severity levels. One significant difference between dysarthric and typical speech is that the speaking rate is often substantially slower for talkers with dysarthria [2, 3]. However, this reduced speaking rate is often not consistent throughout the utterance. In addition, dysarthric vocal excitation may be unstable because many individuals with dysarthria cannot effectively control vocal fold closure and vibration. This may cause inconsistent vocal quality and pitch throughout an utterance.

5.1.1. Synthetic Dysarthric Speech

Differences in speech style and speaking rate significantly depend on the dysarthria severity level of the talkers [129]. To be able to synthesize accurate dysarthric speech, we add a dysarthria severity predictor in the variance adaptor to simulate the characteristics of different severity levels of dysarthric speech. The severity embedding is added as an input to the variance adaptor before the pitch/energy/duration predictors. This allows the system to detect the relative characteristics of different severity groups, especially duration, pause and voice harshness, and variance of pitch and energy. It also allows additional control of the duration of the speech like the duration, pitch, and energy predictors, the dysarthria severity level predictor has a similar model structure which consists of a 2-layer 1D-convolutional network with ReLU activation, each followed by the layer normalization and a dropout layer, and an extra linear layer to project the hidden states into the output sequence [56].

Based on the structure of the TORGO dataset and the amount of data available, speakers are categorized into three dysarthria severity levels: normal, very low/low, and medium and into two intelligibility groups: intelligible and non-intelligible [129]. During training, the TORGO non-dysarthric speakers are used for the Normal category, label 0. Because there is only one speaker in the Low category, Very Low and Low are combined together to form the middle category, coefficient 1. The highest severity level in the dataset, labeled Medium, is used for the third category, coefficient 2.

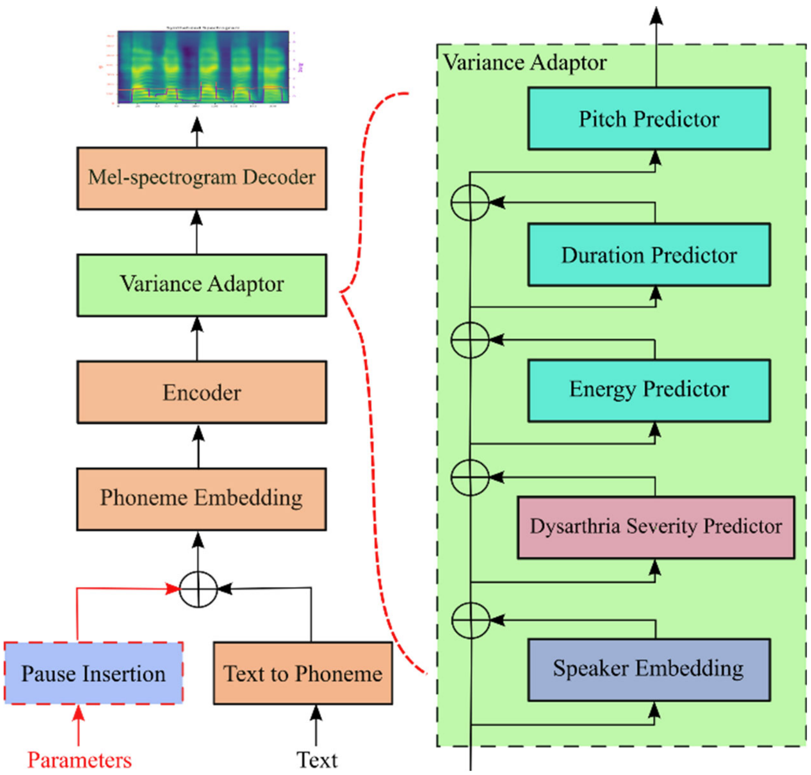


Figure 5.1: An overview of the proposed architecture

5.1.2. Pause Insertion

Pause is another important indicator of dysarthric speech. Analysis of the TORGO data set indicates that the number of between-word pauses per sentence among typical, very low, low and moderate groups is about 0.26, 0.57, 1.21 and 2.51, respectively. As a ratio to normal speakers, this means that the number of pauses is 120%, 365%, and 865% more frequent among the very low, low and moderate groups in comparison with typical speaker in TORGO dataset. The effect of sentence length on pause duration has also been previously investigated in persons with dysarthria due to Amyotrophic Lateral Sclerosis (ALS) [4]. Their results show that the pause duration over sentence length for the group with higher severity level is increased by a higher rate in comparison with the group with lower severity level.

Although FastSpeech2 can already synthesize normal pause patterns for a given text, it is not sufficient to represent the patterns in dysarthric speech. To address this issue, we add a binary parameter to control insertion of additional pauses. Although pauses in dysarthric speech sometimes occur between phonemes within a word as well, the current version supports insertion of pauses only between words. To implement this, possible inter-word positions are identified, and then the maximum number of pauses is determined based on the severity level and length of the given sentence. For longer texts or for speakers with a higher dysarthria severity level, the model inserts more pauses. Since many of the sentences in the TORGO dataset are relatively short, there is not enough data to learn a complex model for pause insertion, so a simple model is used. The model uses the number of words in the sentence and the dysarthric severity level to determine the number of pauses to be inserted. Once this is set, the locations of the pauses are chosen randomly at inter-word locations in the sentence. The pause insertion model is shown in the bottom left of the architecture in Figure 5.1.

For ASR, Pytorch-kaldi [150] was used to train DNN ASR models . A light Gated Recurrent Unit (liGRU) architecture was implemented, trained on fMLLR transformed features Baseline configuration files provided in the Pytorch-kaldi repository for common speech databases like TIMIT, Librispeech were used as reference and the final architecture was based on experimental results using a small number of training set speakers [151].

5.1.3. Frame and phoneme level of Masking

There are two options for pitch and energy modifications in the Variance adaptor, phoneme, and frame levels. In the frame level, the target duration is applied and then pitch and energy modifications are implemented while the modification of pitch and energy are carried out before the adjusting the target mel spectrogram duration in the phoneme level. The mel mask is used in the frame level modification, the source mask is applied in the phoneme level modification to modify pitch and energy. Masking is a method of padding to the maximum length of the input sequence which are phonemes or the maximum length of the output sequence which is here the mel spectrogram length.

The main paper [56] was implemented based on the frame level feature for pitch and energy modifications. However, a recent variant of the paper was found the phoneme level feature is more effective and their synthesized speech is more natural [40].

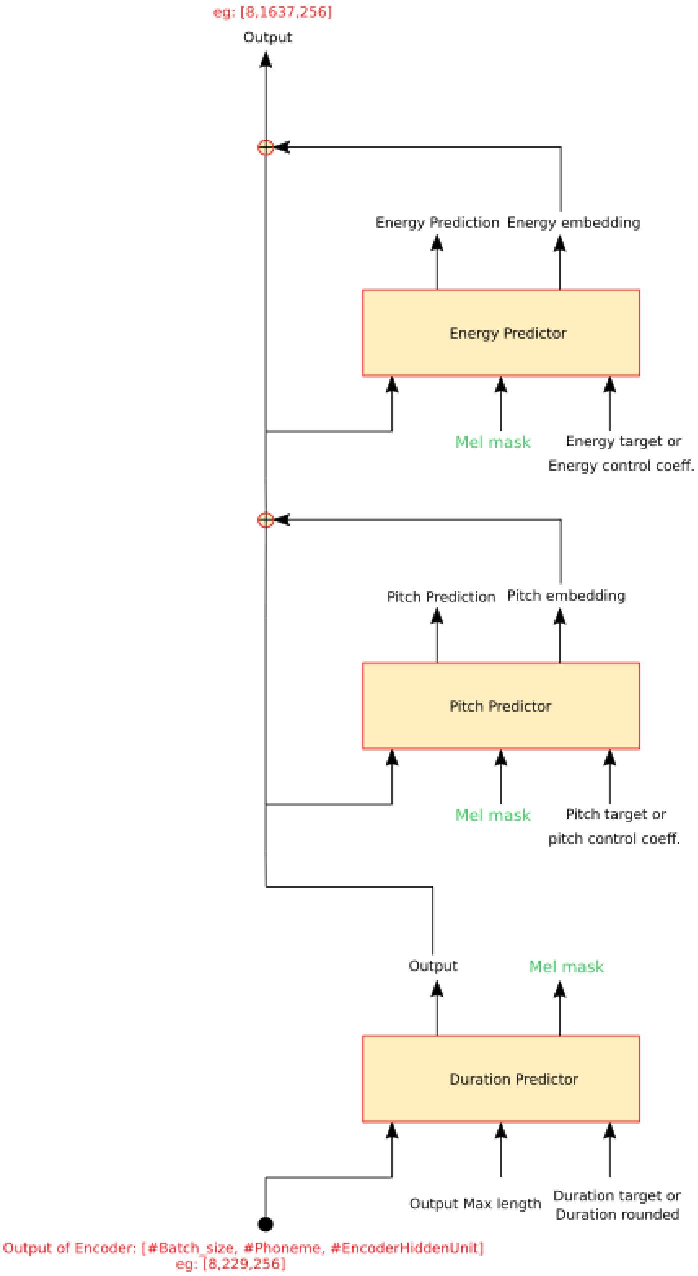


Figure 5.2: Frame level masking

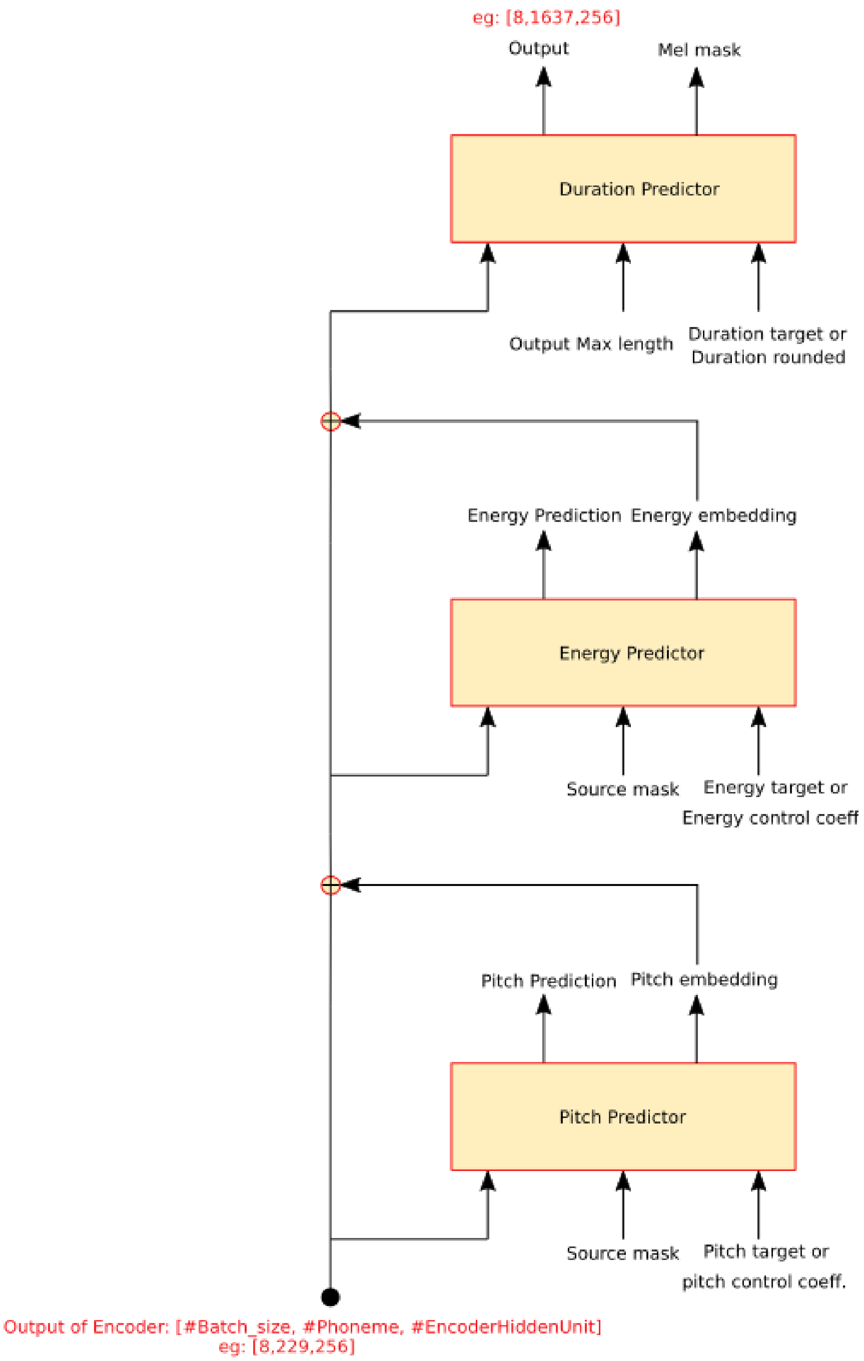


Figure 5.3: Phone level masking

5.2. Experimental setup

FastSpeech2 contains 4 feed-forward transformer blocks in the encoder and mel-spectrogram decoder. The decoder generates an 80-dimensional mel-spectrogram from hidden state. The size of phoneme embedding is 256 in our implementation. The adjusted model was trained with a GeForce RTX 2080 Ti on the TORGO [4] dataset, containing 8 dysarthric speakers and 7 normal speakers. This dataset consists of non-words (which are excluded in this experiment), short words, restricted and non-restricted sentences. Dysarthric speakers are categorized into three dysarthria severity levels, very low, low, and medium and into two groups for intelligibility, intelligible and non-intelligible [129]. The number of utterances for each dysarthric talker averages 700; whereas for normal speakers the average is 1560 [64].

After training the TTS models, the text in TORGO was used to synthesize additional dysarthric speech. The effect of the synthesized speech was evaluated by implementing two experiments on speech recognition application.

In the first experiment, the focus is on the effect of the severity predictor and pause insertion. Synthesized speech for augmentation was synthesized with three different severity coefficients of 0.0, 1.0, and 2, with the pause insertion turned on. Pitch, energy and duration coefficients were fixed at 1.0. The number of augmented sentences was three times that of the original TORGO dataset.

For the second experiment, a wider range of dysarthric speech was synthesized for augmentation across all controllable parameters. Parameters for pitch, energy, duration, as well as severity level were varied across a range with pause insertion activated as shown in Table 5.1 below. The number of augmented sentences was ten times that of the original TORGO dataset.

Table 5.1: The prosody coefficients for synthesizing dysarthric speech in the two experiments

Coef.	Baseline	Exp. 1	Exp. 2
Pitch	-	1.0	[0.1, 0.6, 1.2, 1.75]
Energy	-	1.0	[0.1, 1.0, 2.0]
Duration	-	1.0	[1.0, 1.3, 1.6, 1.8]
Severity level	-	[0.0, 1.0, 2.0]	[0.0, 1.0, 2.0]
Pause insertion	-	Yes	Yes

Total utterance	~ 16000	~ ×3	~ ×10

The synthesized speech was applied for training the DNN-HMM model with light bidirectional GRU [150] architecture, with five layers containing 1024 cells each, activated by Relu activation function and dropout of 0.2. The number of epochs was 10 to 12 to achieve the best result of each experiment. The architecture applies monophone regularization [152]. A multi-task learning procedure was applied using two SoftMax classifiers, one estimating context-dependent states and the second one predicting monophone targets [151].

For testing, a leave-one-speaker-out cross-validation procedure was applied across the original TORGO dataset.

5.3. Results and discussion

Before evaluating the performance of the synthetic data augmentation on dysarthria-specific DNN-HMM speech recognition, we first review and assess the quality of the synthesized dysarthric speech itself. Figure 5.3 shows the synthesized speech of speaker MC04 for the input text “We would like to play volleyball” for severity level of 0, 1 and 2, respectively. Pitch, energy and duration coefficients are the same across the various severity levels shown here. As indicated in Figure 5.4, synthesized speech duration increases with increasing severity level. Duration is one of the key indicators of different levels of severity. However, the rate of change is variant across the different phonemes and depends on the speaker’s speech characteristics and utterance. Unlike the fix-rated speaking rate changes of speech augmentation method in 4.2, this capability of the proposed system is one of the key factors to synthesize of dysarthric speech. This capability which is based on speakers’ speech characteristics allows us to expand the existing speech by adding new dysarthric talkers to the system.

Other parameters, including harshness, blurred quality, and unintelligibility have also been synthesized and can be heard and evaluated on the provided demo web page[1]. To analyze the quality of synthesized dysarthric speech, we have provided speech and spectrograms generated for M02, M04 (dysarthric speaker) and MC02 and MC04 (control talker) in the section “Dysarthria

[1] https://mohammadelc.github.io/SpeechGroupUKY/

Severity Level" [2] of the demo page. To investigate the effect of changing the dysarthria severity level parameters on synthesizing dysarthric speech, all other parameters are kept fixed. For each speaker there are three speech utterances synthesized by three different coefficients of 0, 1 and 2, corresponding to severity levels of normal, very-low-and-low and moderate, respectively. By increasing the severity coefficient, more severe dysarthric speech was generated, especially for the highest level, "moderate". The dysarthric speech characteristics such as harshness, blurred and unintelligibility are more obvious in synthesized speech at that level.

One of the other metrics for evaluation was evaluated is whether or not the "dysarthric-ness" quality of the synthesized speech is different for typical and dysarthric speakers when changing the severity level coefficients. The term "dysarthric-ness" is used to refer to the authenticity/accuracy of the synthesis engine in generating speech that sounds genuinely dysarthric to a human listener. Comparing the dysarthric-ness quality of synthesized speech for dysarthric and typical speakers shows us that the synthesized speech for dysarthric speakers is more naturally similar to real dysarthric speech than the speech synthesized using a typical non-dysarthric speaker as the synthesis target.

For example, for Speaker MC02 when saying "This is the pad" for different severity level coefficient, the phoneme /p/ is heard /c/, which represents a specific common mispronunciation that sometime happens for dysarthric speech. In addition, we observed that the duration of the synthesized phoneme /æ/ in words such as bad, sad, dad and pad varies significantly, and most notably for the word "pad", the sound of this phoneme is much longer than that of the rest. This is also representative of typical dysarthric characteristics

For pause insertion, we aimed to build a system to learn the pause patterns including both duration and frequency of pauses as the two main factors for each given speaker. To learn the pause length, the model considers this intrinsically as it would with other phonemes; thus, pause length can be learned during training which is dependent on each speaker. As discussed in Section 5.1.2 regarding pause insertion, for longer texts or for speakers with a higher dysarthria severity level, the model inserts more pauses. The Pause Insertion section[3] of the demo page shows the synthesized dysarthric speech with pauses for two dysarthric target speakers M02 and M05 for the given input text "How we can synthesize better dysarthric speech?". To illustrate that we can

[2] https://mohammadelc.github.io/SpeechGroupUKY/#:~:text=adding%20these%20parameters.-,Dysarthria%20Severity%20Level,-Abbreviation%3A%20Pitch

[3] https://mohammadelc.github.io/SpeechGroupUKY/#:~:text=to%20play%20volleyball%22-,Pause%20Insertion,-Number%20of%20pause

control pause insertion, speech was synthesized with three different numbers of pauses, 0, 1 and 2. The result demonstrates that with an increasing number of pauses the length of utterances will change correspondingly, and the pauses can be observed in the audio and spectrograms. In addition, the pause length for different speakers is different. For example, for M05 the pause length is much longer than that of speaker M02.

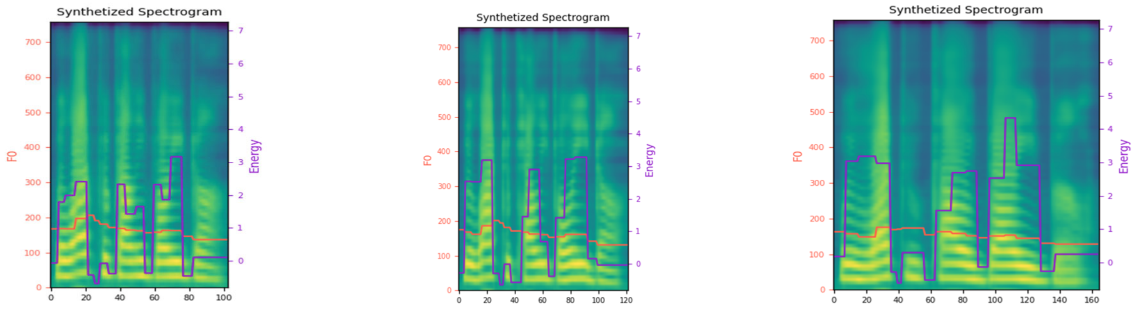
To evaluation the effect of other parameters like pitch, energy and duration controllability, different examples of synthesized speech for target speaker M05 with the input text "Bad and good" are presented. In the section "Duration, Pitch and Duration controls on a fixed severity level"[4] at the demo page, the first row shows the results for changing duration coefficients. With changing this coefficient from 1.0 to 1.3 and then 1.6, when the other factors are fixed, the model generated correspondingly longer speech as expected. For energy and pitch, three speech utterances were synthesized with coefficients of 0.5, 1.0 and 2.0. The purple line in the spectrogram of the second row in this section shows the change in energy of the synthesized speech caused by changing the energy coefficient.

To see the effect of the pitch coefficient, the third row in this same section of the demo page plots the synthesized speech for three pitch coefficients. The orange line in these spectrograms indicates that with increasing or decreasing the pitch coefficient, the pitch in synthesized speech was changed, again as expected.

Other dysarthric characteristics are also learned by the system itself, rather than being controlled by a specific parameter, as part of the speaker embedding process that models the target speaker. For example, it can be seen that in the synthesized speech[5], the synthesized speech based on target speaker M05 has a stutter at phoneme /b/ before "best", something that might be considered one of the characteristics of dysarthric speech. However, we do not explicitly control this parameter, it is instead totally learned and generated by the system itself as part of speaker modelling and training.

[4] https://mohammadelc.github.io/SpeechGroupUKY/#:~:text=Duration%2C%20Pitch%20and%20Duration%20controls%20on%20a%20fixed%20severity%20level

[5] https://mohammadelc.github.io/SpeechGroupUKY/#:~:text=Bad%20and%20good%22-,Other%20Observation,-We%20have%20noticed



Severity level: 0.0 (Normal)

Severity level: 1.0 (Combined Levels of very low and low)

Severity level: 2.0 (Moderate)

Figure 5.4: Effect of dysarthria severity coefficients in synthesizing dysarthric speech for speaker MC04

To evaluate the results of the two data augmentation ASR experiments as mentioned Section 5.2, the Word Error Rate (WER) was calculated for each test speaker, with varying amounts and types of training data coming from the synthesized speech for those target speakers. Table 5.2 shows the WER of the two experiments along with the baseline and compares them with the results of the best models of two other published works preformed on the same TORGO dataset using hybrid speech recognition models [69, 153].

Results show that the WER performance of the baseline is similar to that of the two comparison methods for the lowest few severity levels, and slightly better for the highest ("medium") severity. The average WER across all speakers is 44.5%, 56.2% and 43.3% for our baseline, [153] and [69], respectively.

In the first experiment that only used severity synthesis and pause insertion, the synthesized speech used for augmenting ASR training improved the performance of the DNN-HMM model for each speaker except M03, which declined slightly. Average WER performance across all speakers improves, with WER dropping from 44.5% to 41.6%. The second experiment with additional prosody variance and data augmentation shows further performance improvement, with individual improvement for all 8 speakers in the dataset.

Average WER performance across all speakers improves, going from 44.5% to 39.2%. On average, the first and second experiments reduces WER by 6.5 %, 12.2% with the respect to the baseline, respectively.

Table 5.2: WER of each test speaker for the two augmentation experiments: Exp.1 included augmented speech across 3 severities with pause insertion, and Exp. 2 included augmented speech across severity, pause, pitch, energy, and duration.

Severity Level	Test Spk	WER (%)				
		Baseline	**Exp. 1**	**Exp. 2**	**[153]**	**[69]**
Very low	F04	16.8	16.3	14.5	18.3	13.1
	M03	10.9	12.7	10.7	18.2	17.7
Low	F03	46.6	39.3	36.8	44.2	39.1
Moderate	F01	58.3	52.4	50.4	71.5	39.6
	M01	55.4	51.3	50.3	69.3	62.2
	M02	44	43.1	38.4	70.9	42.9
	M04	65.8	64.2	62	79.9	69.0
	M05	58.2	53.6	49.6	77.2	62.6
Overall Average		44.5	41.6	39.2	56.2	43.3

To summarize the effect of the proposed approaches as a function of the level of severity of the dysarthric speech, Table 5.3 shows the average WER for speakers at the different dysarthria severity levels. This shows that augmentation using synthetic speech at three dysarthria levels with pause insertion improves the WER of each severity level on average except for the group with the low severity. Augmentation using synthetic speech at three severity levels plus pause insertion, further varying energy, pitch, and duration improved WER across all severity levels.

Table 5.3: WER of each severity level for the two augmentation experiments.

Severity level	baseline	Exp. 1	Exp. 2	Improvement	
				Exp.1	**Exp.2**
Very Low	13.8	14.5	12.6	-4.7%	**9**%
Low	46.6	39.3	36.8	**7.3**%	**21**%
Moderate	56.3	52.9	50.1	**6**%	**11**%
All	44.5	41.6	39.2	**6.5%**	**12.2%**

5.4. Robustness and an extension of the proposed method:

In the current version of the dysarthric speech synthesis described in the section Methodology 5.1, we have used a single dataset, TORGO, with three discrete dysarthria categories, e.g., normal, combination of very low and low, and moderate, as measures related to dysarthric characteristics. Neither of these are ideal. It would be preferable to be able to include data from multiple datasets, and to have more meaningful indicators or measures or dysarthric characteristics. However, because as discussed previously in Dataset 2.2.3, there are very few datasets of dysarthric speech available, and each of these have their own unique labels and measures of dysarthria that are not common or standardized, this is not directly possible.

To make the method more robust and synthesize a broader range of dysarthric speech, in this section we present a preliminary study based on the idea of selecting more explanatory dysarthric measures with broader scales, and then connecting those measures into the previously presented synthesis model, enabling generation of synthesized dysarthric speech from controllable explanatory parameters rather than discrete severity level variables.

This approach has the potential to allow integration of multiple datasets for training the synthesizer, since a trained prediction model on the selected parameters can be built for each dataset based on whatever information and labels each has available. For example, if there is well-defined label for a primary dataset like TORGO, semi-supervised approaches like label propagation [154] can be used to expand it. With this approach, it is possible to combine labeled data with abundant unlabeled data to train deep neural network. Thus, any additional data allows to train the current synthesis model better as the model itself is a relatively data-driven approach. It also allows more intuitive control of the synthesized speech using control parameters that have more perceptual meaning associated with them.

For the preliminary study shown here, we have used for a parameter a combination of dimensions taken from the Frenchay dysarthria assessment information available in TORGO. The Frenchay assessment measures 28 relevant perceptual factors of speech grouped into 8 dimensions, including reflex, respiration, lips, jaw, soft palate, laryngeal, tongue, and intelligibility. A 9-point scale was used to rate each dimension. For example, for the cough reflex dimension, a talker would receive a grade of 'a' (8) for no difficulty, 'b' (6) for occasional choking, 'c' (4) if the patient requires particular care in breathing, 'd' (2) if the patient chokes frequently, and 'e' (0) if they are unable to have a cough reflex[4].These dimensions reflect the severity level of dysarthria each talker in a range between 1 to 8, where 1 indicates the highest severity and 8 the lowest severity(normal). Table 5.4 shows these dimensions along with the variance of each of the dimensions across speakers (rightmost column). Some of these dimensions have very little variability across speaker, such as “Reflex” which within the range of 1 to 8 has a minimum value of 6.67. Those with the largest variability across speakers include Respiration (range 5, variance 5.05), Laryngeal (range 5.5, variance 5.73), Tongue (range 5.83, variance 4.69) and Intelligibility (range 6.33, variance 8.44).

Among these dimensions, we have selected three, Respiration, Laryngeal and Tongue, because of the highest variance and being relatively discriminative across different severity levels. The average of these three dimensions were used to create a single dysarthria indicator, referred to as “RLT”, also shown in Table 5.5. If there was enough amount of speech data for each of these three labels, it would be ideal to use them as three separate parameters since we could generate more variant speech data. Given the small amount of data here, using a combination allows the system to converge better, similarly to the method described 5.1.1 on combination of groups very low and low. In that method, since there is only one speaker in the Low category, Very Low and Low are combined together to form the middle category. In practical, we need to consider this issue and find out if there is a fairly enough data for each category helping the model to learn from. Thus, this is the main reason we convert the three selected dimensions of respiration, laryngeal, tongue into a single numeric indicator.

Table 5.4: Average score of different dimensions of Frenchay dysarthria assessment for each speaker.

Dimension	Normal	**F04**	**M03**	**F03**	**F01**	**M01**	**M02**	**M04**	**M05**	**Var**
Reflex	8.0	6.67	7.67	6.67	8.00	8.00	8.00	6.33	7.33	0.48
Respiration	8.0	8.00	7.50	8.00	5.00	3.00	3.00	3.00	5.50	**5.05**
Lips	8.0	8.00	7.80	8.00	5.60	5.00	5.00	3.00	3.60	3.95
Jaw	8.0	8.00	8.00	8.00	5.50	8.00	8.00	5.00	8.00	1.25
Velum	8.0	8.00	8.00	8.00	5.33	6.67	6.67	7.33	7.33	0.86
Laryngeal	8.0	8.00	7.00	8.00	5.00	2.50	2.50	2.75	4.50	**5.73**
Tongue	8.0	6.67	7.50	6.67	2.83	2.33	2.33	3.33	2.17	**4.69**
Intelligibility	8.0	8.00	8.00	8.00	2.33	2.33	2.33	1.67	5.33	**8.44**

Table 5.5: RLT combination score and its corresponding label coefficient

Dimension	Normal	**F04**	**M03**	**F03**	**F01**	**M01**	**M02**	**M04**	**M05**	**Var**

Average Score	8.0	7.67	7.68	7.67	4.95	4.73	4.73	4.05	5.47	2.38
Respiration + Laryngeal +Tongue	8.0	7.56	7.33	7.56	5.14	2.61	2.61	3.03	4.06	
	8.0	7.00	7.00	7.00	5.00	2.00	2.00	3.00	4.00	
Coeffs in range (0,6)	6.0	5.0	5.0	5.00	3.00	0.00	0.00	1.00	2.00	

To build a predictor, we translated the RLT to a new scale and used it as a target variable for prediction in a machine learning model. Since the embedding vectors begins from 0, we converted these to 0 to 6 and then injected this information as an input coefficient to a predictor. Figure 5.5 shows the variance adaptor that replaces dysarthria severity levels predictor with RLT predictor.

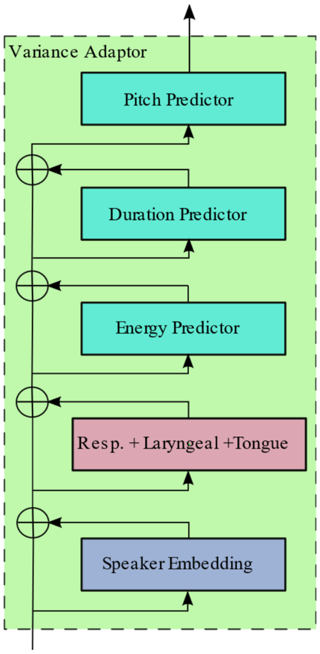


Figure 5.5- Variance adaptor with RLT (Respiration-Laryngeal-Tongue)

Like the previous procedure described in Section 5.1 , we have an encoder-decoder model containing 4 feed-forward transformer blocks with a module called variance adaptor between them as shown in Figure 5.1. This variance adaptor contains different predictors, and the main responsibility of this module is to train the predictors and control their predictions. The model receives the input text and converts it to the corresponding phonemes. Phoneme embedding sequences generated by phoneme embedding module are used as input to the encoder to output hidden state sequence, which is the input of our variance adaptor.

In the variance adaptor, after adding the speaker embedding sequence to the hidden state sequence of the encoder, an RLT embedding sequence will be added to hidden state sequence. During training, the RLT true label, which is a number between (0 to 6), is used to train the RLT predictor so that the corresponded embedding vector is added to the hidden state sequence. Then, embedding sequence of pitch, energy and duration predictors is added to the hidden state sequence came from the previous step to form the output of variance adaptor. Finally, the decoder takes the output of the variance adaptor and generates 80 Mel-spectrogram. The vocoder High-Fidelity GAN (HiFi-GAN) [155] is applied to convert the Mel-spectrogram to audio file.

The RLT predictor has a model structure consisting of a 2-layer 1D-convolutional network with ReLU activation, each followed by the layer normalization and a dropout layer, and an extra linear layer to project the hidden states into the output sequence.

After having the entire model trained, the model is ready to synthesize new speech for a given input text. During synthesizing (inference), the variance adaptor is the main part to control parameters like dysarthria severity level, pitch, energy and duration as well as speakers' identification (ID). By changing RLT labels from 0 to 6, the mode generates speech corresponding to that RLT level. In addition, pause is also controlled using the same method as described in the section Section 5.1.2. Generally, the number of pauses in an utterance is determined based on the dysarthria severity and length of the given sentence. For longer texts or for speakers with a higher dysarthria severity level, the model inserts more pauses.

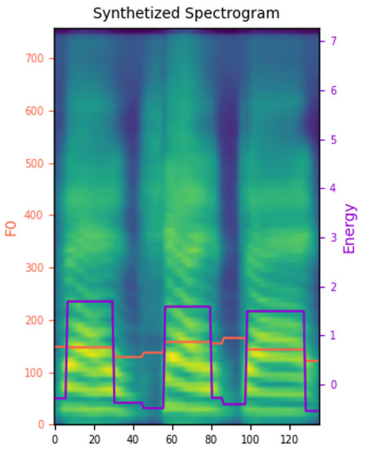
.

Figure 5.6 and Figure 5.7 demonstrate the two examples of synthesized dysarthric speech using RLT predictor for input text of "bad sad dad" and "we are in the classroom", respectively. You can find their audio file here[6]. In these examples, all other coefficients are fixed to the prediction of the model itself and are identical in two examples. These figures indicate that speaking rate would change if we changed the RLT coefficient. For RLT 0 as it is expected to see longer synthesized speech, the model with RLT predictor generated longer speech, which is one of the indicators of dysarthric speech. However, we can see the same length of generated speech for some

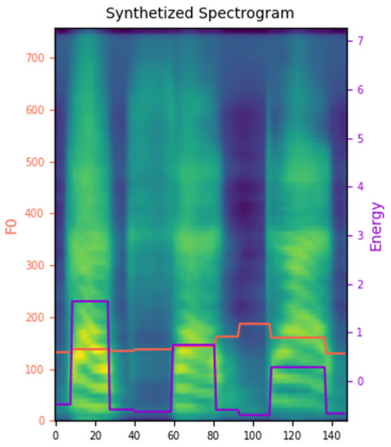
[6] https://github.com/Mohammadelc/SpeechGroupUKY/tree/main/rtlAudioFiles

close RLT coefficients, such as RLT 6 and 5, which are the normal speaker group and talkers with very low severity levels, respectively. To differentiate normal speech and speech with a very low severity level dysarthria is quite challenging, even for experienced listeners who are listening to authentic (not synthesized) speech .

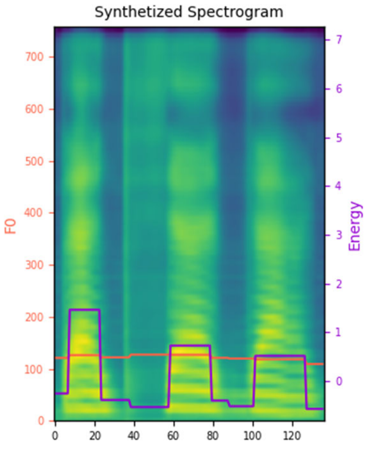
Another observation from this plot is that even for the same phonetic sound like /æ/ in bad, sad and dad, the duration of the synthesized phoneme in these three words is different for a specific target speaker. From this we can observe that this model can learn each sound characteristics based on the other following and proceeding sound context and generate prosodic characteristics such as duration appropriately.



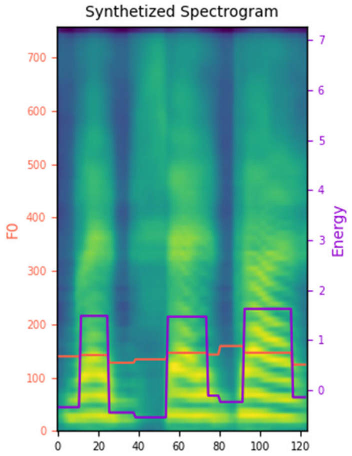
RLT0



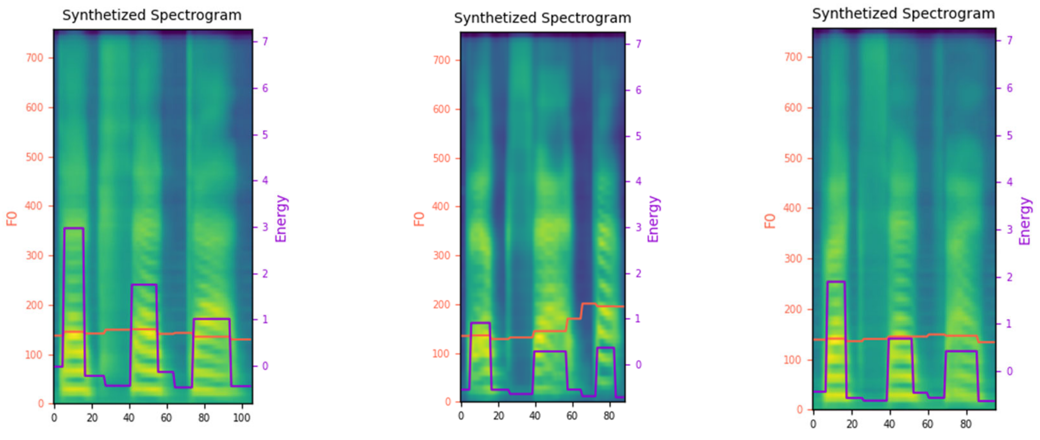
RLT1



RLT2

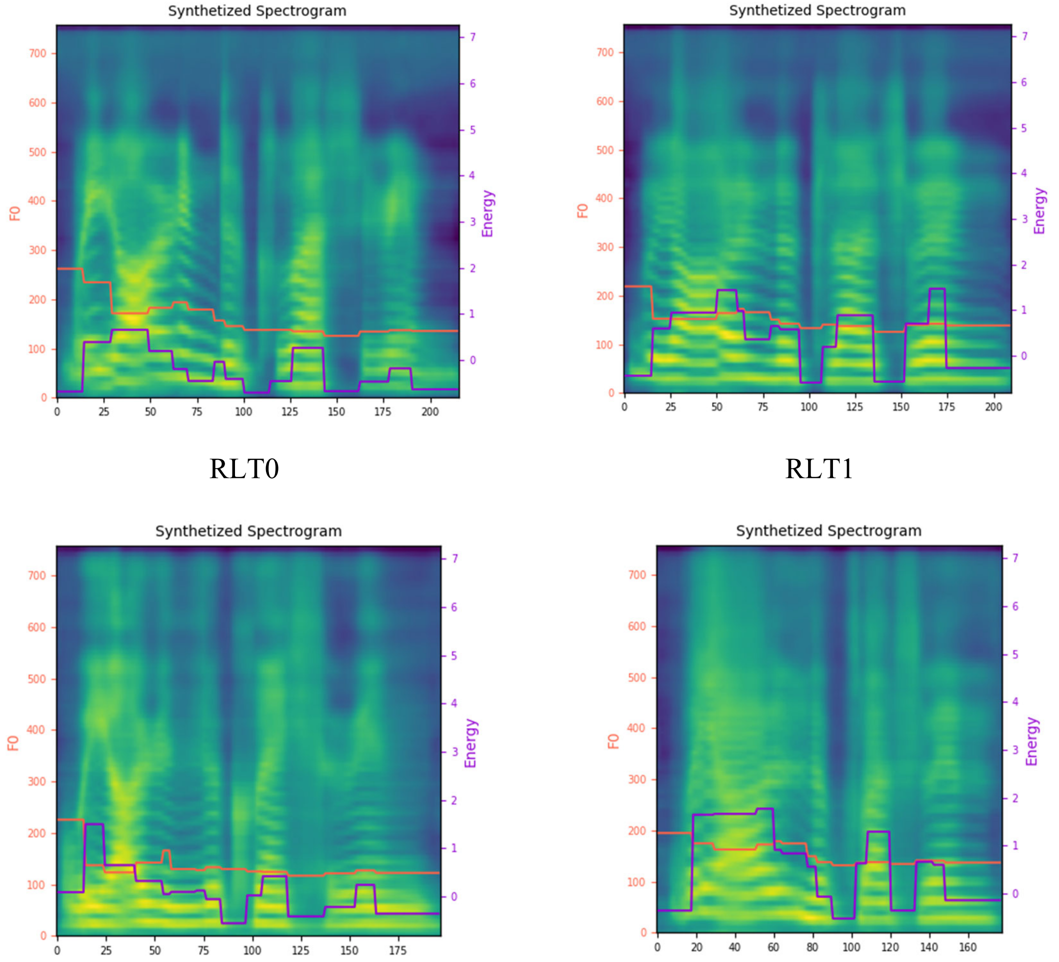


RLT3



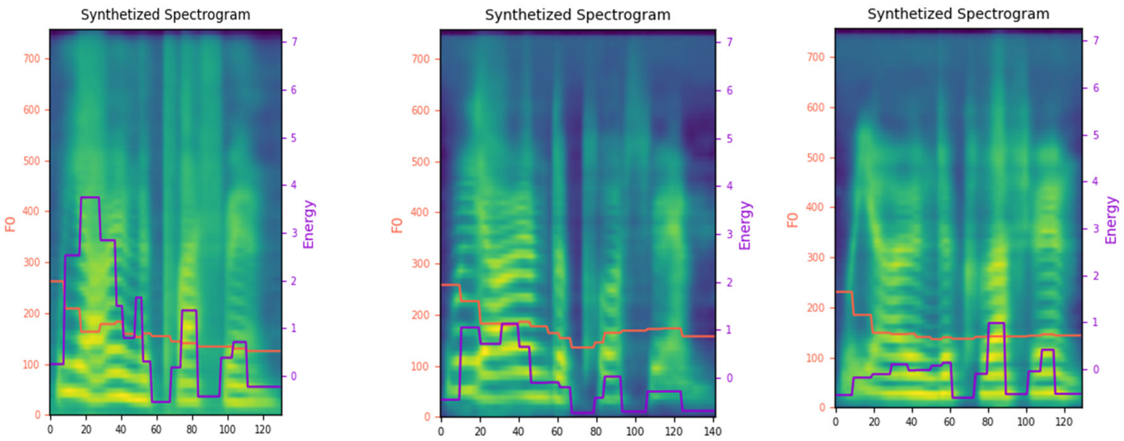
RLT4 RLT5 RLT6

Figure 5.6: Effect of various RLT on synthesized dysarthric speech for an input text: "*Bad sad dad*"



RLT0 RLT1

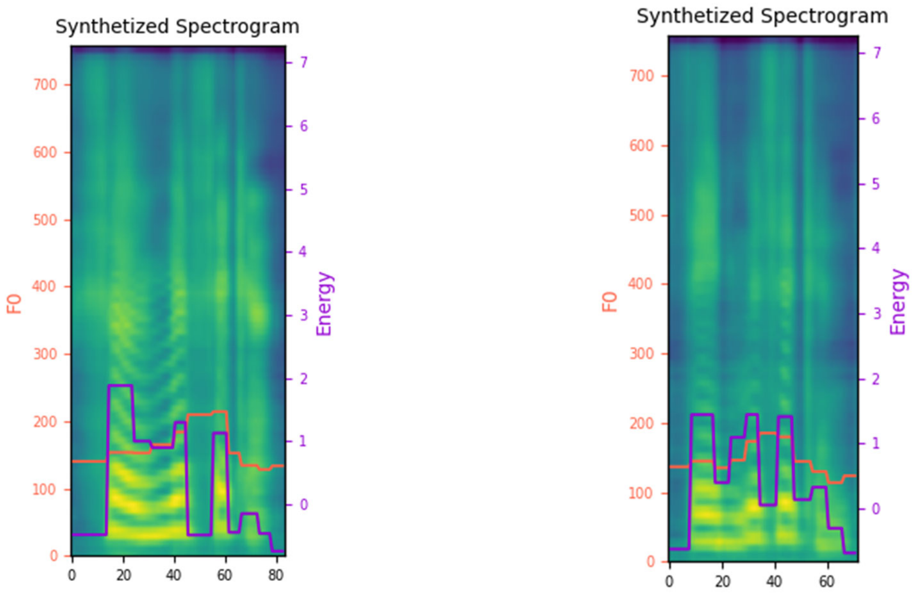
RLT2 RLT3

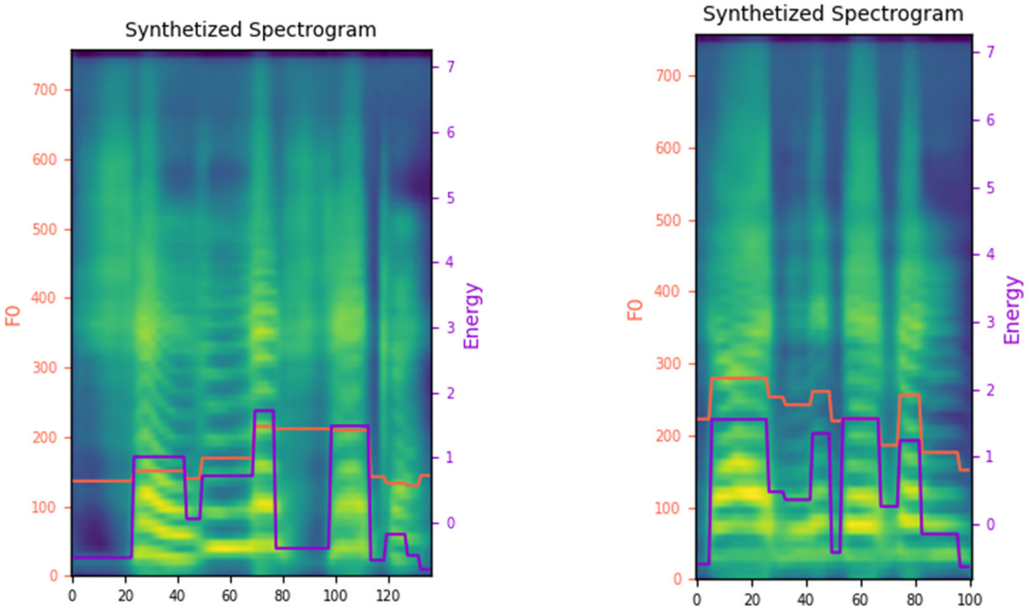


RLT4 RLT5 RLT6

Figure 5.7: Effect of various RLT on synthesized dysarthric speech for an input text: "*We are in the classroom*"

Note that even for the same input text, the length of the result is quite different across different RLT levels, as well as across different target speakers. For example, the length of the utterance for the input text "significant" is different for speakers M02 and M04. For RLT 0 and 4, the length of generated dysarthric speech for M02 is about 84 and 130 frames, respectively. However, this value for M04 is about 100 and 70, respectively. This shows that the model differentiates speech characteristics of each individual.





M02 for RLT 4 M02 for RLT 4

M02 for RLT 0 M02 for RLT0

Figure 5.8: Comparing the length of generated audio for the input "Significant" for M02 and M04

5.5. Conclusion

In this chapter, we have modified a neural multi-talker TTS by adding a dysarthria severity level coefficient and a pause insertion model to synthesize dysarthric speech for varying severity levels. We evaluate its effectiveness for data augmentation of training data for dysarthria-specific speech recognition. Results are shown for two different experiments: the first includes augmented speech across 3 severities with pause insertion, and the second includes augmented speech across severity, pause, pitch, energy, and duration. Overall results on the TORGO database demonstrate that using dysarthric synthetic speech to increase the amount of dysarthric-patterned speech for training has significant impact on the dysarthric ASR systems. A demonstration web page with audio results of the synthesis is available at https://mohammadelc.github.io/SpeechGroupUKY/.

In addition, we have introduced an extension to make more robust dysarthric synthesized speech and increase the controllability of the system by adding the dimensions of Respiration, Laryngeal and Tongue (RLT). These dimensions are selected because of the highest variance and being relatively discriminative across different severity levels. This extension allows the model to generate dysarthric speech with broader range.

Chapter 6: Conclusions and Future Work

This chapter summarizes the original contributions and conclusions presented in this dissertation. Some future research subjects are also suggested that could improve and facilitate the progress of the important topics discussed in this work.

6.1. Original Contributions

This dissertation first presents a comparative study between typical and dysarthric speech, to better understand differences in prosodic and acoustic characteristics of dysarthric spontaneous speech at varying severity levels. These characteristics are important components for dysarthric speech modeling, synthesis, and enhancement, which are themselves important to tasks such as data augmentation for improving dysarthric speech assessment and recognition. To compare typical and dysarthric speech timing, we analyze the mean duration of vowels and consonants to find the speaking rate difference between dysarthric and typical speech. This timing information is essential to model speaking rate across severity levels. The mean pauses duration and the occurrence of pause between words are essential parameters to model the pause rate and duration for various severity levels. Two other important prosody characteristics of speech, pitch and intensity, are also evaluated for each speaker.

The second contribution of this work is an investigation of a voice conversion-based data augmentation method using GAN and CycleGAN to convert typical speech to dysarthric speech. This method is effective at generating dysarthric speech, but the quality and variability of the speech is not sufficient to improve performance of speech technologies such as ASR when used to generate additional training data for augmentation. Although the method is not sufficient for effective data augmentation, the experimental work highlights some of the challenges of the augmentation task and led to the development of the next two contributions described below.

The third contribution of this dissertation is an exploration of a specialized data augmentation approach to enhance the performance of end-to-end dysarthric ASR. The proposed method contains prosodic transformation and time-feature masking. In prosodic transformation, we modify the speaking rate and shift the pitch to alter vocal excitation characteristics and prosodic structure. We also exploit time and feature masking in the spectral domain to alter the MFCCs representing vocal tract acoustics. Experimental results with this approach demonstrate that applying prosodic and time- feature masking on both dysarthric and normal speech represent better performance and underscore the need for speech from various dysarthria severity levels. Overall results indicate that using augmentation to increase the amount of dysarthric-patterned speech for training has significant impact on dysarthric ASR systems, particularly for speech with more severe dysarthria.

The fourth contribution is an innovative approach for synthesizing dysarthric speech using end-to-end multi-talker speech synthesis. The synthesis model generates dysarthric speech based on parameters representing key dysarthric speech characteristics, allowing control of parameters such duration, energy, pitch, dysarthria severity level and the occurrence of pause. These represent the most salient features of realistic dysarthric speech. In addition, this model has an ability to catch the voice characteristics of individuals using a decoder and speaker embedding, making it a multi-talkers TTS capable of generating speech in a wide range of speaking styles. This is a useful capability for speech synthesis for data augmentation because it allows generation of a robust set of training data. Experimental results with this approach demonstrate that using dysarthric synthetic speech to increase the amount of dysarthric-patterned speech for training has significant impact on dysarthric ASR systems.

The fifth contribution is an RLT predictor that replaces with dysarthria severity level predictor with a continuous and more perceptually meaningful metric that can be utilized across multiple datasets. This predictor is a combination of Respiration, Laryngeal and Tongue (RLT) that have the highest variance and being relatively discriminative across different severity levels. This method has the potential to allow the new dysarthric synthesis model to be trained from data across datasets with different labeling mechanisms and adds the benefit of supporting one or more control parameters that are based on perceptually meaningful categories rather than the more generic severity level indicator.

6.2. Recommendation for Future work

To expand the current model, possibilities include applying a semi-supervised approach such as label propagation to extend the amount of speech data and number of speakers, adding additional features like articulatory positions, and using a continuous scale to define dysarthria severity level. In addition, to increase the benefit of this model for augmentation, Zero-shot learning could be used to add a new dysarthric talker with only a few speech utterances after training the main model and using out-of-domain text to synthesize dysarthric speech. In the following subsections, more information on each of these directions is provided.

6.2.1. Applying out-of-domain text on dysarthric speech

In this dissertation, we applied in-domain dysarthric speech domain. However, to increase the number of unique utterances, out-of-domain text can be used to enrich the existing utterances. To accomplish that, text can be collected from other speech recognition related datasets such as Librispeech, VCTK, LJ Speech and use them as input text to synthesize speech. In this way, a dysarthria-specific ASR is trained on a larger variety of utterances that is more robust and more effective for practical applications. It is anticipated that an ASR model trained with this additional data will further increase performance only systems trained on a limited unique word lexicon.

6.2.2. Zero-shot method

The current version of the model can capture voice characteristics of the speaker used for training. To expand the number of target speakers, it is possible to incorporate a Zero-shot learning procedure. Zero-shot learning is a well-known method that has recently found significant use in the speech domain, especially in the voice conversion subdomain. The procedure involves learning the voice characteristics of a new speaker with only a few speech samples. That would be valuable to accomplish for the dysarthric task, in order to increase the size and distribution of speakers in the existing data set. For example, the current version of the speech synthesis model generates male speech better than female speech as there are not enough dysarthric female speakers across the different dysarthria severity levels in the training dataset. However, zero-shot learning might improve this so that the dysarthric models are strengthened and closer to the robustness of typical speech models. Therefore, even a few minutes of speech training for a given talker whom was not in the original training set may help the robustness of the dysarthric speech model.

6.2.3. Continuous scaling

Instead of using discrete label of dysarthria severity level or RLT to train the TTS model, it would be possible to integrate real-continuous values for each severity level or to increase to a broader range of severity levels. This would increase the flexibility to synthesize new dysarthric speech. However, to represent dysarthric speech in a continuous way to the synthesis model, there should be enough data to train the model. However, a sufficient amount of data is not currently publicly available in a single dysarthric dataset. One of the ways we can address this problem is to incorporate label propagation as described in Section 6.2.5 below to integrate all of the exiting datasets and use these to train the dysarthric speech synthesis model.

6.2.4. Adding articulation feature

Articulatory features, including Reflex, Respiration, Lips, Jaw, Velum, Laryngeal, Tongue, Intelligibility can be added as additional information for each input to the Variance adaptor in the main model. Accurate values of these feature can help the model do differentiate speech characteristics across all dysarthria severity levels. However, there is a need of a reliable acoustic-to-articulatory inversion model to accomplish this, as articulatory features are not available in all exiting dysarthric datasets. To this end, we recommend using TORGO dataset to train an acoustic-to-articulatory model first since this dataset was designed for this purpose. Then, the acoustic-to-articulatory model can be applied to other datasets like UASpeech to generate articulatory features. The generated articulatory and acoustic features together can be used to train the dysarthric speech synthesis model. One of the key features of using variance adaptor module is that giving more information allows the system to build a better predictor and then better dysarthric speech.

6.2.5. Label propagation

It is a semi-supervised learning that can combine data carefully labeled by humans with abundant unlabeled data to train deep neural networks. Figure 6.1 is an example to understand a high-level idea and is relatively similar to k-Nearest Neighbor (KNN).

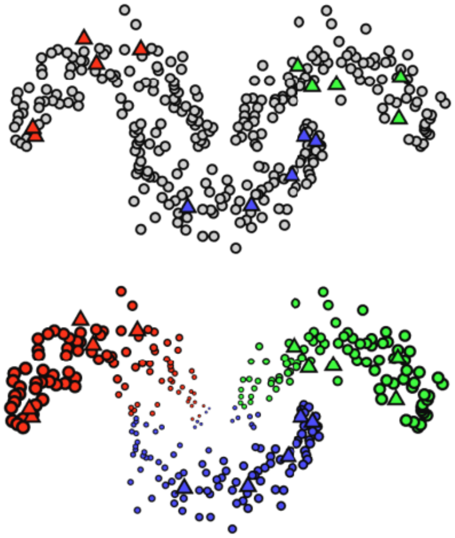


Figure 6.1: Label propagation on a toy example [154]

For example, if we want to align two main dysarthric dataset, UASpeech dataset is mainly used as labeled dataset and TORGO as unlabeled dataset because of two reasons. First, UASpeech contains severity level From Low to High while the highest severity level in TORGO is moderate. Second, UASpeech is perceptually evaluated; however, the TORGO dataset was evaluated by Frenchay dysarthria assessment.

In another case, if there is a dataset with subjective assessment of dysarthria, even with few data, it would be useful to be used as a labelled data and the other available datasets like TORGO and UASpeech could be used as the unlabeled data.

The proposed method can also be used in other low-resource domains such as accented speech and child speech recognition systems, other tasks which suffer from lack of robust data for training. For accented speech, the model can train based on the different accented of a language to generate synthetic speech for each accent. For children scenario, the model can be extended to this area of research to generate new speech for different age category. Therefore, the synthetic data can be used as training data to make a more robust accented/children speech recognition system.

6.3. Conclusion

In this dissertation, we have investigated dysarthric speech and methods to synthesize dysarthric speech. This work has analyzed suprasegmental prosodic characteristics between typical and dysarthric speaker with different severity levels. The phoneme duration, speaking rate and pause characteristics of typical and dysarthric speech as well as energy and pitch were analyze. For augmentation, we started with prosodic transformation and time-feature masking. However, to synthesize dysarthric speech, we have used an end-to-end multi-talker TTS model to have better controllability on the parameters such as pitch, energy, duration, severity level and pause insertion for varying severity levels. In addition, we have extended this work by adding Respiration, Laryngeal and Tongue (RLT) instead of dysarthria severity level. This increases the controllability of the system, so we are able to generate more dysarthric speech with broader range.

With the synthetic dysarthric speech generated with this model, we can address some of problems with existing dysarthric datasets. As discussed in section Datasets 2.2.3, one of the problems of the existing dysarthric dataset is lack of utterances with different length e.g., most utterances in TORGO are a single word. The new method proposed here can synthesize text input with different lengths and enrich the training dataset for ASR. In the term of speaker variability, the new approach allows us to synthesize dysarthric speech with variant voice characteristic of the speakers in training. Generally, by using this system, more dysarthric speech will be available for dysarthria-specific tasks like speech recognition and dysarthria severity and intelligibility assessment.

This dissertation has proposed a new method to generate dysarthric speech with controllability on parameters that can generate the main identifying characteristics of dysarthric speech. This methodology supports dysarthria-related speech applications such as speech recognition to be trained on more data with more robust models. Our overall results demonstrate that using dysarthric synthetic speech to increase the amount of dysarthric-patterned speech for training has significant impact on the dysarthric ASR systems and suggests the possibility of using this same approach for other applications impacted by lack of training data.

References:

1. Duffy, J.R., Motor Speech Disorders E-Book: Substrates, Differential Diagnosis, and Management. 2019: Elsevier Health Sciences.
2. Mitchell, C., et al., Interventions for dysarthria due to stroke and other adult-acquired, non-progressive brain injury. Cochrane Database of Systematic Reviews, 2017(1).
3. Kent, R.D., Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. American Journal of Speech-Language Pathology, 1996. **5**(3): p. 7-23.
4. Rudzicz, F., A.K. Namasivayam, and T. Wolff, The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation, 2012. **46**(4): p. 523-541.
5. Kim, H., et al. Dysarthric speech database for universal access research. in Ninth Annual Conference of the International Speech Communication Association. 2008.
6. Menendez-Pidal, X., et al. The Nemours database of dysarthric speech. in Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96. 1996. IEEE.
7. Shorten, C. and T.M. Khoshgoftaar, A survey on image data augmentation for deep learning. Journal of big data, 2019. **6**(1): p. 1-48.
8. Perez, L. and J. Wang, The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621, 2017.
9. Bagherinezhad, H., et al., Label refinery: Improving imagenet classification through label progression. arXiv preprint arXiv:1805.02641, 2018.
10. Yi, X., E. Walia, and P. Babyn, Generative adversarial network in medical imaging: A review. Medical image analysis, 2019. **58**: p. 101552.
11. Taylor, L. and G. Nitschke. Improving deep learning with generic data augmentation. in 2018 IEEE Symposium Series on Computational Intelligence (SSCI). 2018. IEEE.
12. Moosavi-Dezfooli, S.-M., A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
13. Ko, T., et al. Audio augmentation for speech recognition. in Sixteenth Annual Conference of the International Speech Communication Association. 2015.
14. Cui, X., V. Goel, and B. Kingsbury, Data augmentation for deep neural network acoustic modeling. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2015. **23**(9): p. 1469-1477.

15. Ko, T., et al. A study on data augmentation of reverberant speech for robust speech recognition. in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. IEEE.

16. Park, D.S., et al., Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779, 2019.

17. Sharma, H.V. and M. Hasegawa-Johnson, Acoustic model adaptation using in-domain background models for dysarthric speech recognition. Computer Speech & Language, 2013. **27**(6): p. 1147-1162.

18. Rebai, I., et al., Improving speech recognition using data augmentation and acoustic model fusion. Procedia Computer Science, 2017. **112**: p. 316-322.

19. Jiao, Y., et al. Simulating dysarthric speech for training data augmentation in clinical speech applications. in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. IEEE.

20. Vachhani, B., C. Bhat, and S.K. Kopparapu. Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. in Interspeech. 2018.

21. Mirheidari, B., et al., Data augmentation using generative networks to identify dementia. arXiv preprint arXiv:2004.05989, 2020.

22. Geng, M., et al., Investigation of Data Augmentation Techniques for Disordered Speech Recognition.

23. Mun, S., et al., Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane. Proc. DCASE, 2017: p. 93-97.

24. Zhang, T., K. Zhang, and J. Wu. Data Independent Sequence Augmentation Method for Acoustic Scene Classification. in INTERSPEECH. 2018.

25. Han, Y. and K. Lee, Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation. arXiv preprint arXiv:1607.02383, 2016.

26. Nagano, T., et al. Data Augmentation Based on Vowel Stretch for Improving Children's Speech Recognition. in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2019. IEEE.

27. Kathania, H., et al., Data augmentation using prosody and false starts to recognize non-native children's speech. arXiv preprint arXiv:2008.12914, 2020.

28. Sheng, P., Z. Yang, and Y. Qian. GANs for Children: A Generative Data Augmentation Strategy for Children Speech Recognition. in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2019. IEEE.

29. Fainberg, J., et al. Improving Children's Speech Recognition Through Out-of-Domain Data Augmentation. in Interspeech. 2016.

30. Rituerto-González, E., et al., Data augmentation for speaker identification under stress conditions to combat gender-based violence. Applied Sciences, 2019. **9**(11): p. 2298.

31. Cai, W., et al. Countermeasures for Automatic Speaker Verification Replay Spoofing Attack: On Data Augmentation, Feature Representation, Classification and Fusion. in INTERSPEECH. 2017.

32. Wu, Z., et al. Data Augmentation Using Variational Autoencoder for Embedding Based Speaker Verification. in INTERSPEECH. 2019.

33. Qin, X., D. Cai, and M. Li. Far-Field End-to-End Text-Dependent Speaker Verification Based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation. in INTERSPEECH. 2019.

34. Wang, J., S. Kim, and Y. Lee. Speech Augmentation Using Wavenet in Speech Recognition. in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. IEEE.

35. Shahnawazuddin, S., et al., Voice Conversion Based Data Augmentation to Improve Children's Speech Recognition in Limited Data Scenario. Proc. Interspeech 2020, 2020: p. 4382-4386.

36. Bigi, B., et al. A syllable-based analysis of speech temporal organization: a comparison between speaking styles in dysarthric and healthy populations. in Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH 2015). 2015. International Speech Communications Association.

37. Zhang, C., et al. Investigation on articulatory and acoustic characteristics of dysarthria. in The 9th International Symposium on Chinese Spoken Language Processing. 2014. IEEE.

38. Yunusova, Y., et al., Profiling speech and pausing in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). PloS one, 2016. **11**(1): p. e0147573.

39. Kuo, C. and K. Tjaden, Acoustic variation during passage reading for speakers with dysarthria and healthy controls. Journal of communication disorders, 2016. **62**: p. 30-44.

40. Chien, C.-M., et al. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. IEEE.

41. Simonyan, K. and B. Horwitz, Laryngeal motor cortex and control of speech in humans. The Neuroscientist, 2011. **17**(2): p. 197-208.

42. Rabiner, L.R. and R.W. Schafer, Introduction to digital speech processing. Vol. 1. 2007: Now Publishers Inc.

43. Ramaswamy, G.N. and P.S. Gopalakrishnan. Compression of acoustic features for speech recognition in network environments. in Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181). 1998. IEEE.

44. Dave, N., Feature extraction methods LPC, PLP and MFCC in speech recognition. International journal for advance research in engineering and technology, 2013. **1**(6): p. 1-4.

45. Hermansky, H., Perceptual linear predictive (PLP) analysis of speech. the Journal of the Acoustical Society of America, 1990. **87**(4): p. 1738-1752.

46. Suh, Y. and H. Kim. Data-driven filter-bank-based feature extraction for speech recognition. in SPECOM2004. 2004.

47. Lawrence, R., Fundamentals of speech recognition. 2008: Pearson Education India.

48. Yu, D. and L. Deng, AUTOMATIC SPEECH RECOGNITION. 2016: Springer.

49. Yu, S.-Z., Hidden semi-Markov models. Artificial intelligence, 2010. **174**(2): p. 215-243.

50. Al-Qatab, B.A. and R.N. Ainon. Arabic speech recognition using hidden Markov model toolkit (HTK). in 2010 International Symposium on Information Technology. 2010. IEEE.

51. Sha, F. and L.K. Saul. Large margin hidden Markov models for automatic speech recognition. in Advances in neural information processing systems. 2007.

52. Karpagavalli, S. and E. Chandra, A review on automatic speech recognition architecture and approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition, 2016. **9**(4): p. 393-404.

53. Dahl, G.E., et al., Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on audio, speech, and language processing, 2011. **20**(1): p. 30-42.

54. Li, J., et al. LSTM time and frequency recurrence for automatic speech recognition. in 2015 IEEE workshop on automatic speech recognition and understanding (ASRU). 2015. IEEE.

55. Wang, D., X. Wang, and S. Lv, An Overview of End-to-End Automatic Speech Recognition. Symmetry, 2019. **11**(8): p. 1018.

56. Ren, Y., et al., Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558, 2020.

57. Fohr, D., O. Mella, and I. Illina. New paradigm in speech recognition: deep neural networks. in IEEE international conference on information systems and economic intelligence. 2017.

58. Chan, W., et al., Listen, attend and spell. arXiv preprint arXiv:1508.01211, 2015.

59. Hasegawa-Johnson, M., et al. Audiovisual phonologic-feature-based recognition of dysarthric speech. 2006. Citeseer.

60. Patel, R., Prosodic Control in Severe Dysarthria. Journal of Speech, Language, and Hearing Research, 2002.

61. Patel, R., Phonatory control in adults with cerebral palsy and severe dysarthria. Augmentative and Alternative Communication, 2002. **18**(1): p. 2-10.

62. Freed, D., Motor speech disorders: diagnosis & treatment. 2011: Nelson Education.

63. Rudzicz, F., Production knowledge in the recognition of dysarthric speech. 2011, Citeseer.

64. Joy, N.M. and S. Umesh, Improving acoustic models in torgo dysarthric speech database. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2018. **26**(3): p. 637-645.

65. Sanders, E., et al., Automatic Recognition Of Dutch Dysarthric Speech, A Pilot Study. 2002.

66. Hawley, M.S., et al., A speech-controlled environmental control system for people with severe dysarthria. Medical Engineering & Physics, 2007. **29**(5): p. 586-593.

67. Hawley, M., et al. STARDUST; speech training and recognition for dysarthric users of assistive technology. in 7th European Conference for the Advancement of Assistive Technology (AAATE 2003). 2003.

68. Green, P., et al. Automatic speech recognition with sparse training data for dysarthric speakers. in Eighth European Conference on Speech Communication and Technology. 2003.

69. Espana-Bonet, C. and J.A. Fonollosa. Automatic speech recognition with deep neural networks for impaired speech. in International Conference on Advances in Speech and Language Technologies for Iberian Languages. 2016. Springer.

70. Kim, M.J., et al. Dysarthric Speech Recognition Using Convolutional LSTM Neural Network. in INTERSPEECH. 2018.

71. Vachhani, B., et al. Deep Autoencoder Based Speech Features for Improved Dysarthric Speech Recognition. in Interspeech. 2017.

72. Nakashika, T., et al. Dysarthric speech recognition using a convolutive bottleneck network. in 2014 12th International Conference on Signal Processing (ICSP). 2014. IEEE.

73. Sehgal, S., S. Cunningham, and P. Green. Phase-based feature representations for improving recognition of dysarthric speech. in 2018 IEEE Spoken Language Technology Workshop (SLT). 2018. IEEE.

74. Hahm, S., D. Heitzman, and J. Wang. Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization. in Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies. 2015.

75. Sharma, H.V. and M. hasegawa Johnson. State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition. in Proceedings of the NAACL HLT 2010 workshop on speech and language processing for assistive technologies. 2010.

76. Mengistu, K.T. and F. Rudzicz. Adapting acoustic and lexical models to dysarthric speech. in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2011. IEEE.

77. Tolba, H. and A.S. El_Torgoman. Towards the improvement of automatic recognition of dysarthric speech. in 2009 2nd IEEE International Conference on Computer Science and Information Technology. 2009. IEEE.

78. Wang, Z., Q. She, and T.E. Ward, Generative adversarial networks in computer vision: A survey and taxonomy. arXiv preprint arXiv:1906.01529, 2019.

79. Goodfellow, I., et al. Generative adversarial nets. in Advances in neural information processing systems. 2014.

80. Zhu, J.-Y., et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. in Proceedings of the IEEE international conference on computer vision. 2017.

81. Qian, K., et al. F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder. in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. IEEE.

82. Sisman, B., et al., An overview of voice conversion and its challenges: From statistical modeling to deep learning. arXiv preprint arXiv:2008.03648, 2020.

83. Kawahara, H., I. Masuda-Katsuse, and A. De Cheveigne, Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech communication, 1999. **27**(3-4): p. 187-207.

84. Morise, M., F. Yokomori, and K. Ozawa, WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE TRANSACTIONS on Information and Systems, 2016. **99**(7): p. 1877-1884.

85. Oord, A.v.d., et al., Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

86. Kalchbrenner, N., et al., Efficient neural audio synthesis. arXiv preprint arXiv:1802.08435, 2018.

87. Abe, M., et al., Voice conversion through vector quantization. Journal of the Acoustical Society of Japan (E), 1990. **11**(2): p. 71-76.

88. Shikano, K., S. Nakamura, and M. Abe. Speaker adaptation and voice conversion by codebook mapping. in 1991., IEEE International Sympoisum on Circuits and Systems. 1991. IEEE.

89. Helander, E., et al. On the impact of alignment on voice conversion performance. in Ninth Annual Conference of the International Speech Communication Association. 2008.

90. Luan, Y., et al. Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition. in 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). 2014. IEEE.

91. Laskar, R.H., et al., Comparing ANN and GMM in a voice conversion framework. Applied Soft Computing, 2012. **12**(11): p. 3332-3342.

92. Wu, Z., et al., Exemplar-based sparse representation with residual compensation for voice conversion. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014. **22**(10): p. 1506-1521.

93. Nguyen, H.Q., et al., High quality voice conversion using prosodic and high-resolution spectral features. Multimedia Tools and Applications, 2016. **75**(9): p. 5265-5285.

94. Wu, J., Z. Wu, and L. Xie. On the use of i-vectors and average voice model for voice conversion without parallel data. in 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). 2016. IEEE.

95. Hsu, C.-C., et al. Voice conversion from non-parallel corpora using variational auto-encoder. in 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). 2016. IEEE.

96. Hsu, C.-C., et al., Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. arXiv preprint arXiv:1704.00849, 2017.

97. Kaneko, T. and H. Kameoka, Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv preprint arXiv:1711.11293, 2017.

98. Fang, F., et al. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. IEEE.

99. Lorenzo-Trueba, J., et al., Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data. arXiv preprint arXiv:1803.00860, 2018.

100. Kameoka, H., et al. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. in 2018 IEEE Spoken Language Technology Workshop (SLT). 2018. IEEE.

101. Tan, X., et al., A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561, 2021.

102. Indumathi, A. and E. Chandra, Survey on speech synthesis. Signal Processing: An International Journal (SPIJ), 2012. **6**(5): p. 140.

103. Klatt, D.H., Review of text-to-speech conversion for English. The Journal of the Acoustical Society of America, 1987. **82**(3): p. 737-793.

104. Lukose, S. and S.S. Upadhya. Text to speech synthesizer-formant synthesis. in 2017 International Conference on Nascent Technologies in Engineering (ICNTE). 2017. IEEE.

105. Kröger, B.J. and P. Birkholz, A gesture-based concept for speech movement control in articulatory speech synthesis, in Verbal and Nonverbal Communication Behaviours. 2007, Springer. p. 174-189.

106. Hill, D., L. Manzara, and C. Schock. Real-time articulatory speech-synthesis-by-rules. in Proceedings of AVIOS. 1995. Citeseer.

107. Tabet, Y. and M. Boughazi. Speech synthesis techniques. A survey. in International Workshop on Systems, Signal Processing and their Applications, WOSSPA. 2011. IEEE.

108. Wouters, J. and M.W. Macon, Control of spectral dynamics in concatenative speech synthesis. IEEE Transactions on Speech and Audio Processing, 2001. **9**(1): p. 30-38.

109. Tiomkin, S., et al., A hybrid text-to-speech system that combines concatenative and statistical synthesis units. IEEE transactions on audio, speech, and language processing, 2010. **19**(5): p. 1278-1288.

110. Wang, Y., et al., Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135, 2017.

111. Shen, J., et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. IEEE.

112. Ping, W., et al., Deep Voice 3: 2000-Speaker Neural Text-to-Speech. 2017.

113. Ren, Y., et al. Almost unsupervised text to speech and automatic speech recognition. in International Conference on Machine Learning. 2019. PMLR.

114. Griffin, D. and J. Lim, Signal estimation from modified short-time Fourier transform. IEEE Transactions on acoustics, speech, and signal processing, 1984. **32**(2): p. 236-243.

115. Ping, W., et al., Deep voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654, 2017.

116. Ping, W., K. Peng, and J. Chen, Clarinet: Parallel wave generation in end-to-end text-to-speech. arXiv preprint arXiv:1807.07281, 2018.

117. Dinh, L., J. Sohl-Dickstein, and S. Bengio, Density estimation using real nvp. arXiv preprint arXiv:1605.08803, 2016.

118. Gadermayr, M., et al. An asymmetric cycle-consistency loss for dealing with many-to-one mappings in image translation: a study on thigh MR scans. in 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). 2021. IEEE.

119. Rosenberg, A., et al. Speech recognition with augmented synthesized speech. in 2019 IEEE automatic speech recognition and understanding workshop (ASRU). 2019. IEEE.

120. Li, J., et al., Training neural speech recognition systems with synthetic speech augmentation. arXiv preprint arXiv:1811.00707, 2018.

121. Chen, Z., et al. Improving Speech Recognition Using GAN-Based Speech Synthesis and Contrastive Unspoken Text Selection. in INTERSPEECH. 2020.

122. Wang, Y., et al. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. in International Conference on Machine Learning. 2018. PMLR.

123. Skerry-Ryan, R., et al. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. in international conference on machine learning. 2018. PMLR.

124. Lee, Y. and T. Kim. Robust and fine-grained prosody control of end-to-end speech synthesis. in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. IEEE.

125. Vaswani, A., et al. Attention is all you need. in Advances in neural information processing systems. 2017.

126. Dong, L., S. Xu, and B. Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. IEEE.

127. Hu, H., T. Tan, and Y. Qian. Generative adversarial networks based data augmentation for noise robust speech recognition. in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. IEEE.

128. Chatziagapi, A., et al. Data Augmentation Using GANs for Speech Emotion Recognition. in INTERSPEECH. 2019.

129. Soleymanpour, M., M.T. Johnson, and J. Berry. Increasing the Precision of Dysarthric Speech Intelligibility and Severity Level Estimate. 2021. Cham: Springer International Publishing.

130. Bunton, K., et al., Perceptuo-acoustic assessment of prosodic impairment in dysarthria. Clinical Linguistics & Phonetics, 2000. **14**(1): p. 13-24.

131. Allison, K.M., Y. Yunusova, and J.R. Green, Shorter sentence length maximizes intelligibility and speech motor performance in persons with dysarthria due to amyotrophic lateral sclerosis. American journal of speech-language pathology, 2019. **28**(1): p. 96-107.

132. Feenaughty, L., et al., Speech and pause characteristics in multiple sclerosis: A preliminary study of speakers with high and low neuropsychological test performance. Clinical linguistics & phonetics, 2013. **27**(2): p. 134-151.

133. Yuan, J. and M. Liberman, F0 declination in English and Mandarin broadcast news speech. Speech Communication, 2014. **65**: p. 67-74.

134. Looze, C.D., et al. Pitch declination and reset as a function of utterance duration in conversational speech data. in Sixteenth Annual Conference of the International Speech Communication Association. 2015.

135. McAuliffe, M., et al. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. in Interspeech. 2017.

136. Povey, D., et al. The Kaldi speech recognition toolkit. in IEEE 2011 workshop on automatic speech recognition and understanding. 2011. IEEE Signal Processing Society.

137. Soleymanpour, M. and H. Marvi, Text-independent speaker identification based on selection of the most similar feature vectors. International Journal of Speech Technology, 2017. **20**(1): p. 99-108.

138. Jadoul, Y., B. Thompson, and B. De Boer, Introducing parselmouth: A python interface to praat. Journal of Phonetics, 2018. **71**: p. 1-15.

139. Yue, Z., et al. Exploring Appropriate Acoustic and Language Modelling Choices for Continuous Dysarthric Speech Recognition. in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. IEEE.

140. Yu, J., et al. Development of the CUHK Dysarthric Speech Recognition System for the UA Speech Corpus. in Interspeech. 2018.

141. Harvill, J., et al. Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary. in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. IEEE.

142. Xiong, F., et al. Source domain data selection for improved transfer learning targeting dysarthric speech recognition. in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. IEEE.

143. Shahamiri, S.R., Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2021. **29**: p. 852-861.

144. Hermann, E. and M.M.-. Doss. Dysarthric speech recognition with lattice-free MMI. in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. IEEE.

145. Wang, D., et al. Improved End-to-End Dysarthric Speech Recognition via Meta-learning Based Model Re-initialization. in 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). 2021. IEEE.

146. Zeyer, A., et al., Improved training of end-to-end attention models for speech recognition. arXiv preprint arXiv:1805.03294, 2018.

147. Chiu, C.-C., et al. State-of-the-art speech recognition with sequence-to-sequence models. in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. IEEE.

148. Rossenbach, N., et al. Generating synthetic audio data for attention-based speech recognition systems. in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. IEEE.

149. Mimura, M., et al. Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition. in 2018 IEEE Spoken Language Technology Workshop (SLT). 2018. IEEE.

150. Ravanelli, M., T. Parcollet, and Y. Bengio. The pytorch-kaldi speech recognition toolkit. in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. IEEE.

151. Khanal, S., M.T. Johnson, and N. Bozorg. Articulatory Comparison of L1 and L2 Speech for Mispronunciation Diagnosis. in 2021 IEEE Spoken Language Technology Workshop (SLT). 2021. IEEE.

152. Ravanelli, M., et al., Light gated recurrent units for speech recognition. IEEE Transactions on Emerging Topics in Computational Intelligence, 2018. **2**(2): p. 92-102.

153. Yue, Z., H. Christensen, and J. Barker. Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition. in Proceedings of the

Annual Conference of the International Speech Communication Association, INTERSPEECH 2020. 2020. International Speech Communication Association (ISCA).

154. Zhao, Y., et al., Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. arXiv preprint arXiv:2008.12527, 2020.

155. Yang, J., et al., VocGAN: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network. arXiv preprint arXiv:2007.15256, 2020.