Continuous speech recognition for people with dysarthria

Zhengjun Yue
Supervisor: Prof Heidi Christensen, Prof Jon Barker

Department of Computer Science

University of Sheffield

This dissertation is submitted for the degree of

Doctor of Philosophy

I would like to dedicate this thesis to my loving parents, who have been the biggest and continuous source of motivation throughout my life ...

Acknowledgements

First and foremost, I would like to thank Prof Heidi Christensen and Prof Jon Barker for their encouraging and careful supervision throughout this great 3-year white-knuckle ride. I am deeply grateful to the TAPAS (Horizon 2020 Marie Skłodowska-Curie Actions Innovative Training Network European Training Network (MSCA-ITN-ETN)) project for having funded my PhD. Without their support, I would not have been able to do it. Special thanks to Dr Neil Walkinshaw and Dr Stefan Goetze not only for their insightful comments in our PhD panel meetings but also for having encouraged me to pursue a PhD. I would like to express my gratitude for their support and valuable feedback.

I would like to thank all my past and present colleagues at work for all the encouragement and tea-time discussions. I want to thank Jack Deadman and Gerardo Roa Dabike who are always willing to help me out of the problems I have met, and the lovely girls, Fatimah A Alzahrani and Dalia Attas, who always make me feel I am in a lovely lab. I want to particularly thank Meg Thomas for proofreading my thesis.

Finally, I thank my family: my father Xiujie Yue and my mother Xinmei Zhang, for their constant love, support and encouragement. They are the source of my strength, inspiration and happiness. They are my reasons to keep going. They are someone who I can never thank enough. I will use the rest of my life to love them.

Abstract

Dysarthria is a motor speech disorder caused by damage to the nervous system. People with dysarthria often have poorer motor control of their speech articulators resulting in atypical speech. Consequently, the intelligibility of dysarthric speech is often affected and this could affect the communication ability of people with dysarthria, which may cause social exclusion. Dysarthria is also often associated with physical disabilities. This group of people, therefore, have a higher need for automation and voice-enabled interfaces that could improve their daily life. Automatic speech recognition (ASR) technology is becoming ever more ubiquitous. However, the performance on dysarthric speech still lags far behind the mainstream ASR systems designed for typical speech. The large systematic dysarthric and typical speech mismatch, the high speaker variability and the data scarcity make the task challenging. Moreover, the focus on dysarthric speech recognition research has not moved from isolated word to more challenging connected speech scenarios yet. There is a clear need to improve continuous dysarthric speech recognition.

This thesis is the first to systematically investigate various methods for continuous dysarthric speech recognition. Experimental work conducted on the TORGO dysarthric corpus shows that the deployed approaches and the developed systems effectively improve the recognition performance. The key findings are as follows. Firstly, applying an out-of-domain language model allows for a more reasonable decoding space for continuous dysarthric speech, leading to fairer performance. Secondly, incorporating features extracted from an autoencoder-bottleneck feature extractor which is jointly optimised with a speech recogniser is shown to effectively lead to better recognition performance. Employing the monophone regularisation as an auxiliary task can further benefit the performance. Thirdly, by incorporating real articulatory information alongside acoustic features, a multi-modal acoustic-articulatory system is demonstrated to achieve encouraging performance. The best feature fusion scheme is explored and shown to achieve better results. In conclusion, this thesis makes promising progress in improving continuous dysarthric speech recognition.

List of Acronyms and Abbreviations

ADSR automatic dysarthric speech recognition

ACDSR automatic continuous dysarthric speech recognition

MND motor neurone disease

PLP perceptual linear prediction

MFCC Mel-frequency cepstrum

OOD out-of-domain

MTL multi-task learning

EMA electromagnetic midsagittal articulography

CE cross-entropy

WER word error rate

MAMR maximum articulator motion range

OOV out-of-vocabulary

GMM Gaussian mixture model

HMM hidden Markov model

AM acoustic model

LM language model

SI speaker-independent

SD speaker-dependent

CI context-independent

CD context-dependent

MLP multi-layer perceptron

DNN deep neural network

CNN convolutional neural network

RNN recurrent neural network

LSTM long short-term memory

LiGRU light gated recurrent unit

TDNN time-delay neural network

TDNN-F time-delay neural networks

GAN generative adversarial network

AE autoencoder

BN bottleneck

CBN convolutive bottleneck network

AE-BN autoencoder bottleneck

PD Parkinson's disease

CP cerebral palsy

DFT discrete Fourier transform

DCT discrete cosine transform

CMVN cepstral mean and variance normalisation

LOSO leave-one-speaker-out

SAT speaker adaption training

fMLLR feature-space MLLR

MAP maximum a posteriori

MLLT maximum likelihood linear transform

LDA linear discriminant analysis

MMI maximum mutual information

LF-MMI lattice-free maximum mutual information

EPG electropalatog raphy

ER enrolment data

ID interaction data

VOT voice onset time

VD vowel duration

FD fricative duration

VF vowel formant

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

Dysarthria is the most commonly acquired speech disorder, accounting for 53% of all speech disorders [Morris et al., 2016]. It is caused by the damage to the nervous system. Stroke, cerebral palsy (CP) and Parkinson's disease (PD) are the most prevalent causes of dysarthric speech in the UK [Therapists, 2006]. This neurological damage causes the weakening or uncoordination of the muscles used for speaking. Therefore, people with dysarthria often have poorer motor control of their speech articulators, producing speech characterised as being heavily slurred, having a slower speaking rate, abnormal pauses, false starts, and repetitions [Darley et al., 1969]. As a result, this group of people, especially those with severe dysarthria, are difficult to understand for people who are not familiar with their way of speaking. The reduced intelligibility of dysarthric speech can cause problems in human-human communication which in turn can affect social interaction, employment and even education.

People with dysarthria also often have physical motor disabilities such as those associated with stroke or cerebral palsy. They often have more limited or involuntary body movements. Consequently, they may experience difficulty carrying out simple daily tasks such as using handles or switches. It is also difficult for them to interact with physical devices such as keyboards and touch-screens. Especially nowadays, human-machine interaction is increasingly needed by the general public as computer systems are now part of everyday life (e.g., social media, email). People with dysarthria are increasingly left behind due to the barrier in human-machine interaction caused by physical disabilities.

These problems demonstrate an increasing need to establish systems that can facilitate human-human communication and human-machine high-performance interaction for people with dysarthria, which may improve their wellbeing and independence [Holmes et al., 2010].

1.1 Motivation

Automatic speech recognition (ASR) technology underpins today's voice-enabled interfaces by converting spoken utterances into text transcriptions. The speech-to-text translation process enables us to use voice with artificial technology (e.g., remote controls)

whether it is in our homes, cars, offices or elsewhere. In this case, speech provides an attractive interface for hand-free human-machine interaction. As a replacement for more 'traditional' physical interface elements like keys, knobs and handles, ASR-enabled technology has the potential to help people with dysarthria carry out simple daily tasks with a better and more natural interaction with machines, without the need for fine motor skills. In addition, a high-performance ASR system can provide an accurate text transcription for a dysarthric utterance which helps people with dysarthria better communicate with others (e.g., via automatic captioning or text-to-speech synthesis). Eventually, this technology would greatly improve the quality of life of people with dysarthria.

Speech recognition technology clearly has a lot of potential for people with dysarthria. However, currently available commercial ASR systems do not work reliably for dysarthric speech. Mainstream ASR systems are trained on large amounts of *typical* speech and therefore fail to capture the specific characteristics of a person's dysarthria. There is therefore a clear need for a high-performance ASR system for dysarthric speech. Although some progress has been made in recent years, the ASR performance on dysarthric speech is still considerably lower than that achieved with typical speech.

There are several reasons why establishing high-performance ASR systems for people with dysarthria is challenging. The large systematic dysarthric and typical speech mismatch, coupled with the high speaker variability (usually depending on the severity and type of dysarthria) means that large amounts of training data matched to a person's dysarthria is typically required to train adequate acoustic models to learn the mismatch information and normalise different speakers well. However, there are very few dysarthric datasets available and those that do exist contain a limited amount of data. The data scarcity problem makes dysarthric speech difficult to model with recent data-hungry ASR approaches designed for typical speech, for which there are hundreds and thousands of hours of training data available. Moreover, most dysarthric datasets were not collected to be used for ASR training but instead for purposes such as speech assessment. For these reasons, great care needs to be taken when designing the ASR systems for dysarthric speech.

It is also noticeable that most of the previous studies [Kim et al., 2008a, 2018; Xiong

et al., 2018, 2020; Yu et al., 2018] on automatic dysarthric speech recognition (ADSR) have focused on isolated words, since the commonly used dysarthric speech datasets mainly contain single word utterances. Although a single word recognition system can be a great help for people with dysarthria, it would not be adequate to cover most of their needs. Systems that can only recognise isolated words severely limit the range of activities that people with dysarthria can carry out. These systems do little to reduce the exclusion caused by the communication problems experienced by people with dysarthria. Instead, continuous speech (phrases and sentences) is much more natural to use. In order to improve the communication ability of people with dysarthria, similar to the development of typical speech recognition [Makhoul and Schwartz, 1995; Wachter et al., 2003; Young, 1996], the research focus of ADSR needs to move from the single word on to continuous speech.

However, achieving an acceptable performance for a continuous dysarthric speech recogniser is much more challenging than building a desirable isolated word recogniser for dysarthric speech. Firstly, continuous dysarthric speech has more variability. The reduced motor control of articulators makes people with dysarthria experience difficulty moving their articulators from one target pronunciation position to another. As a result, speaking multiple words is more challenging for people with dysarthria and more mispronunciations will occur. Due to the high variability in dysarthric speech, locating the word boundaries in a sentence and tackling the potential effects of coarticulation [Makhoul and Schwartz, 1995] become more challenging. These problems do not need to be considered in the isolated word recognition task. Secondly, and related, decoding is particularly difficult for continuous dysarthric speech where there is increased uncertainty. When decoding single words, the output will be one of the words in a closed-set vocabulary. Even if the words vary a lot in pronunciation by a given speaker, it is a much easier task as opposed to free continuous speech recognition using a probabilistic language model. Finally, compared with isolated words, there is much less continuous dysarthric speech data available which is insufficient to train an adequate model applicable for continuous dysarthric speech.

The purpose of this research is to explore various approaches to building reliable and

high-performance ASR systems for *continuous dysarthric speech*. This will improve the ability of people with dysarthria to communicate in a more natural manner with both humans and machines, and eventually, increase their social participation and improve their independence in life. Difficulties in recognising continuous dysarthric speech mentioned above will be considered when designing the systems, including data scarcity issues and high degree of speech variation. It is essential to exploit ways of maximising the usage of the currently available data and searching for additional sources of information to capture the specific characteristics of dysarthria. Research questions and the main contributions will be presented in the next two sections.

1.2 Research Questions

In order to achieve the above goals, the following research questions will be addressed:

RQ1: To progress the ADSR task from isolated word to continuous speech, it is important to identify and consider the difference between these two tasks. Most previous ADSR studies have focused on isolated word recognition, using closed-set (usually small-sized) vocabularies, where the training and test vocabularies were the same. However, this is not applicable to how continuous speech ASR systems should be evaluated. When it comes to continuous speech, a language model is needed and greatly affects the results. ADSR is a low-resource task. The collected corpora normally have a limited number of sentences with limited vocabulary sizes. This leads to problems when using a language model only trained on the limited prompts within such dysarthric datasets. In this case, the training and test sets are usually non-disjoint leading to unfair decoding (the language model is tuned very highly to the sentences present in the corpus). In addition, the limited vocabulary size results in a large out-of-vocabulary rate, which will also influence the recognition performance. This leads us to the question: **what is an appropriate evaluation framework for continuous dysarthric speech recognition, given current data limitations?**

RQ2: With the popularity of deep learning approaches, there is a growing interest in applying deep learning methods for speech representation learning, e.g., bottleneck and

autoencoder-based features. These approaches may also be applicable to dysarthric speech representation learning. Compared with hand-crafted features, neural networks can learn richer representations which can capture the dysarthric and typical speech mismatch and represent dysarthric speech better. As noted in **RQ1**, data scarcity is a major issue in the ADSR task. The lack of dysarthric data limits the performance achieved by data-hungry deep learning approaches designed for typical speech where large datasets are readily available. Exploiting out-of-domain data is a good way to address this issue [Christensen et al., 2013; Xiong et al., 2020; Yılmaz et al., 2019]. This leads us to the question: **What is a good way to leverage typical speech, which is out-of-domain, to learn more robust representations for continuous dysarthric speech?**

RQ3: All produced spoken sounds are the result of muscles contracting, which is reflected by the movement of different articulators of the speaker – called articulatory information. Due to speakers with dysarthria having poorer control of their articulators, the unintended and involuntary movements make it hard to reproduce their speech (i.e., a typical phonetic token can be pronounced differently), which lacks consistency. Consequently, the produced speech perceived by the listeners always exhibits high variability, and there is often no robust acoustic cues for a specific phoneme. Compared with acoustic representations, articulatory space is simpler to model. The latter may help shape each produced sound which helps in recognising the sound. When it comes to continuous speech, it is also more suitable to better capture coarticulation. Articulatory information may therefore hold complementary information that could be explored in ADSR systems. Detailed analysis of the articulatory space of dysarthric speech is essential to support this assumption. It can provide evidence for the difference between dysarthric and typical speech in the articulatory space. Therefore, is it necessary to address the questions of **how can articulatory information characterise continuous dysarthric speech, and what are the advantages of incorporating articulatory information?**

RQ4: Previous studies have demonstrated that deploying articulatory information alongside acoustic features results in better performance for dysarthric speech recognition. However, the multimodal ADSR task has been limited by the lack of parallel acoustic-articulatory data, therefore most of them use synthetic articulatory features learnt from

an acoustic-articulatory mapping. The learnt articulation can be an inaccurate representation of the real dysarthric articulatory space as the mapping is normally trained on typical speech. Incorporating real articulatory data with acoustic features has not been widely explored in the dysarthric speech community. Whether the real articulatory information is still beneficial in the recent more advanced acoustic models should be explored. So the final research question is **how can articulatory information be incorporated effectively to build multimodal automatic continuous dysarthric speech recognition (ACDSR) systems using recent acoustic models?**

1.3 Contributions

This section identifies the main contributions of this thesis, along with the motivation behind the work carried out.

Contribution 1: Out-of-domain language models.

It is observed that most of the commonly used dysarthric datasets contain much overlap in the prompts read by the speakers. This needs to be considered when these resources are re-purposed and used for ASR evaluation in low-resource data scenarios. Using language models trained on non-disjoint training and test data when recognising continuous speech will most likely result in an unfair design, which potentially produces unrealistically optimistic results. Previous studies have often neglected this problem. This contribution addresses the research question **RQ1**. The work introduces language models trained on out-of-domain (OOD) data [1] and performs evaluation separately on the word and sentence tasks. It provides a fair benchmark for the current state-of-the-art for dysarthric read speech ASR. The work demonstrates that employing OOD language models is fairer and allows for a more meaningful decoding space. The Kaldi recipes of this evaluation framework are available at `https://github.com/zhengjunyue/CADSR-LM`. The work will be introduced in Chapter 5 and will be used as the basis for evaluations in the subsequent chapters. The experimental work has been published as a conference paper [Yue et al.,

[1] Librispeech [Panayotov et al., 2015] is used as the OOD dataset. Compared with other large-size typical speech datasets (e.g., Wall Street Journal [Paul and Baker, 1992]), Librispeech has the most similar recorded speech style with the dysarthric speech dataset TORGO [Rudzicz et al., 2012b] used in this thesis. Librispeech also has public and prunable pre-trained language models.

2020b] in ICASSP 2020:

- Yue, Z., Xiong, F., Christensen, H., & Barker, J. (2020, May). Exploring Appropriate Acoustic and Language Modelling Choices for Continuous Dysarthric Speech Recognition. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020.*

Contribution 2: Autoencoder-based bottleneck features with multi-task optimisation for ACDSR.

The second contribution addresses the research question **RQ2**. It is the first work to demonstrate the effectiveness of learning deep speech representations (i.e., the autoencoder-based bottleneck features) using multi-task optimisation techniques for continuous dysarthric speech recognition. Specifically, the work establishes a framework which allows for jointly optimising the autoencoder bottleneck (AE-BN) feature extractor and the speech recogniser. This enables the speech recogniser to engage in and influence the feature extraction process resulting in speech representations benefits to the phoneme classification specifically. The work also increases the robustness of the dysarthric speech representation learning process by pretraining the feature extractor on the OOD typical data, which addresses the data scarcity problem. Monophone regularisation is applied as a multi-task learning strategy to provide further improvement. The results are state-of-the-art for continuous dysarthric speech recognition in the acoustic domain. The experimental results are fully reproducible and the recipes are available at `https://github.com/zhengjunyue/bntg`. The details will be presented in Chapter 6. The experimental work has been published as a conference paper [Yue et al., 2020a] in INTERSPEECH 2020.

- Yue, Z., Christensen, H., & Barker, J. (2020, July). 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. International Speech Communication Association (ISCA).*

Contribution 3: Acoustic-articulatory speech recognition frameworks.
The current limitation of acoustic-articulatory acoustic modelling for ADSR has two aspects. Firstly, previous work modelled dysarthric articulation parameters from acoustic signals using the acoustic-articulatory mapping knowledge learnt from typical speech, while assuming the articulatory-acoustic mapping remains invariant between typical and dysarthric speech. This leads to uncertainty as to whether the synthesised articulatory data conformed to actual dysarthric speech properties. In contrast, the real recorded dysarthric articulatory data (e.g., electromagnetic midsagittal articulography (EMA) data) can better reflect the dysarthric articulatory space. Secondly, although previous studies showed benefits in incorporating the real articulatory data for the ADSR task, more recent state-of-the-art acoustic models need to be considered in order to test whether such fusion of acoustic and articulatory information is still beneficial. This contribution addresses research questions **RQ3** and **RQ4**, evaluating the contribution of the dysarthric speech articulatory data in combination with acoustic features for automatic dysarthric speech recognition systems. It is the first work to address and analyse the dysarthric and typical speech mismatch in the articulatory space and extends the previous acoustic-articulatory ADSR work with more recent acoustic modelling architectures. It also demonstrates that using the appropriate measure of real articulatory information is beneficial for continuous dysarthric speech recognition. The details will be presented in Chapter 7 and 8. A journal paper covering this work is under preparation:

- Yue, Z., Barker, J., Loweimi, E., Cvetkovic, Z., Christensen, H. Acoustic-articulatory Multimodal Speech Recognition for Dysarthric Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (In preparation).

Contribution 4: Multi-stream acoustic-articulatory feature fusion for dysarthric speech recognition.
Few studies have explored the optimal information fusion scheme for combining acoustic and articulatory representations for dysarthric speech recognition. The acoustic and articulatory representations encode different information, in various formats and with different levels of importance to the task. Consequently, the optimal set of filters to

process each stream will differ. Direct fusion at the input level does not permit such per stream pre-processing. Therefore, it is essential to pre-process each stream individually and fuse the processed streams at a higher level. This work is the first to propose multi-stream acoustic-articulatory architectures which allow for different levels of fusion for dysarthric speech acoustic modelling. It evaluates various levels to uncover the level at which the fusion between both kinds of features occurring during processing is optimal (at the input level, at the medium level or before the output layer). Experimental results demonstrate that fusion at a later level achieves better performance as it underlines the independence of both parameter groups. The optimal fusion level should be high enough to effectively pre-process each information stream for the given task and low enough to leave sufficient capacity after fusion for post-processing the fused streams. This contribution further addresses the research question **RQ4**. The details will be presented in Chapter 8. The experimental work was accepted as a conference paper [Yue et al., 2022] in ICASSP 2022:

- Yue, Z., Loweimi, E., Cvetkovic, Z., Christensen, H., & Barker, J. (2022, May). Multi-modal Acoustic-articulatory Feature Fusion for Dysarthric Speech Recognition. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022.*

1.4 Thesis Overview

A diagram of the organisation of this thesis is shown in Figure 1.1. The remainder of the thesis is presented in Chapters 2 to 9. The content of these chapters are summarised as follows:

Chapter 2: Dysarthric Speech.

This chapter provides a background of what dysarthric speech is in order to give readers a deeper insight. First, it starts with introducing two essential components in communication – speech production and human speech perception which help to understand how to support people with dysarthria using this knowledge of speech perception. Then, a brief description of what dysarthria is and how various types of dysarthria are medically

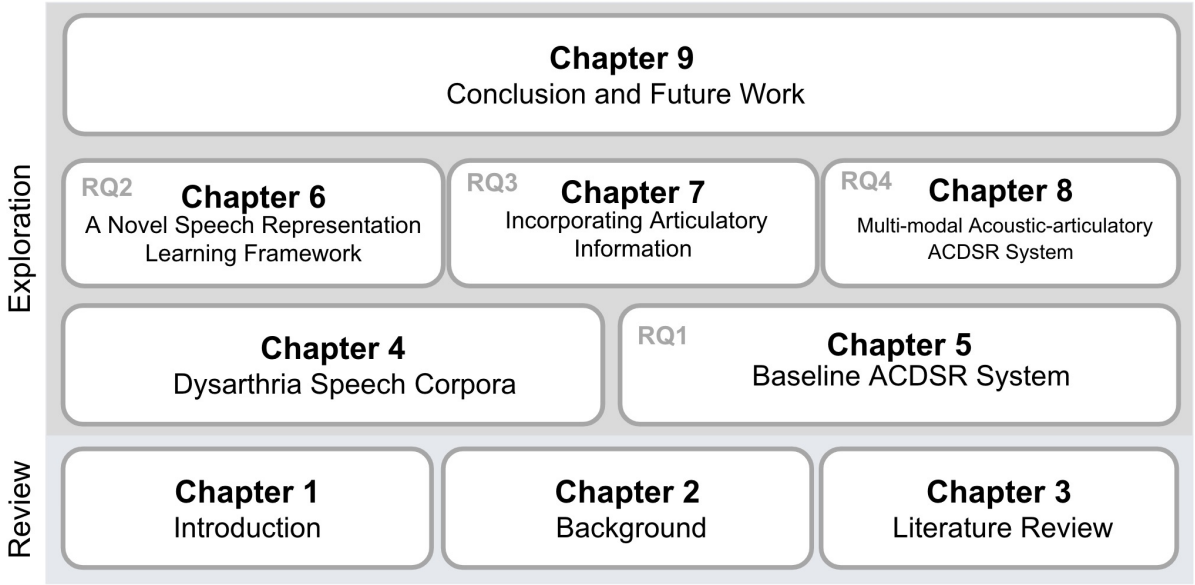


Figure 1.1: Organisation of the thesis indicated by the addressed research questions.

classified is included. The definition of intelligibility and the techniques for assessing the intelligibility (i.e., the severity) for dysarthric speech is discussed. This chapter also gives an analysis of the acoustic characteristics of dysarthric speech, demonstrating the dysarthric and typical speech mismatch, some speech variations and the challenges of developing ASR systems on dysarthric speech.

Chapter 3: A Review of Automatic Recognition for Dysarthric Speech. This chapter reviews the existing research literature concerning ADSR. It begins with a brief introduction of a typical ASR system and its components. A review of the recent progress on ADSR is then presented, including speech representation learning, acoustic modelling, data augmentation and multimodal acoustic modelling. Finally, the current research gaps in dysarthric speech recognition systems and the relevance to the research questions are summarised.

Chapter 4: Dysarthric Speech Corpora and Comparison. This chapter compares the commonly used English dysarthric speech corpora. The previous ADSR studies using the datasets are reviewed. Several dysarthric corpora in other languages are also briefly described. Finally, the most appropriate dataset for this research is selected and will be used in the following experimental chapters.

Chapter 5: Baseline Continuous Dysarthric Speech Recognition System.

In this chapter, a baseline ACDSR system is established. Firstly, a pilot exploration of state-of-the-art ASR system for a widely used dysarthric dataset – TORGO [Rudzicz et al., 2012b] – based on Espana-Bonet and Fonollosa [2016]'s work is presented. The work separates isolated word and sentence recognition into two tasks and reports the results for each task separately. Then, the various designs of language models are discussed and a novel fair evaluation approach employing an OOD language model is proposed. The OOD language model will be used as the basis for evaluations in the subsequent chapters.

Chapter 6: Novel Speech Representation Leaning for Continuous Dysarthric Speech Recognition

The work in this chapter proposes a novel ACDSR framework, which applies autoencoder bottleneck features trained using multi-task optimisation techniques. First, the theory of autoencoders and multi-task optimisation techniques are explained. The system architecture is then presented. A five-fold cross training setup applied in this framework is discussed. The training setup and the acoustic model will be used in the subsequent chapters.

Chapter 7: Incorporating Articulatory Information.

In this chapter the articulatory data recordings available in TORGO are explored. The articulatory motion patterns including 2-D trajectories and 3-D point cloud plots are visualised. The statistical articulatory space distribution regarding maximum articulator motion range (MAMR) is analysed for dysarthric and typical speech. These provide evidence of the articulation mismatch between typical and dysarthric speech.

Chapter 8: Multimodal Acoustic-articulatory Speech Recognition Systems for Continuous Dysarthric Speech.

This chapter demonstrates the effectiveness of incorporating articulatory information for ACDSR. The procedure of processing the articulatory data in TORGO is covered. Different acoustic-articulatory ACDSR systems are explored using various measures of the articulatory data and the subsets of training data. The optimal fusion level and scheme is investigated to further improve the recognition performance for the acoustic-articulatory ACDSR system.

Chapter 9: Conclusion and Future Work.

This chapter summarises the main outcomes of the contributions in this thesis and outlines some potential areas of future research.

The content of thesis is on the website `https://zhengjunyue.github.io/ZJ-thesis/`.

Chapter 2

Dysarthric Speech

Contents

Human speech production is the process of uttering articulated sounds or words, where the speech articulator muscles are controlled by the neuro-motor interface. Neural damage can lead to uncoordinated speech musculature movement, resulting in, for example, poor articulation which causes a motor speech disorder. Dysarthria is the most commonly acquired speech disorder.

Dysarthria was defined by Duffy [2013] as:

> *"A collective name for a group of neurologic speech disorders resulting from abnormalities in the strength, speed, range, steadiness, tone, or accuracy of movements required for control of the respiratory, phonatory, resonatory, articulatory, and prosodic aspects of speech production."*

How common is dysarthria? Researchers don't know exactly how many people have dysarthria. However, it is known to be more common in people who have certain neurological conditions. For example, up to 30% of people with motor neurone disease (MND), 70% to 100% of people with Parkinson's disease, and about 8% to 60% of people with stroke have dysarthria [clinic, 2021]. This shows the necessity of understanding dysarthria in order to support people with dysarthria in ways such as designing assistive technologies.

This chapter starts by introducing the essence of speech production and human speech perception. It then discusses what causes dysarthria, what types of dysarthria there are, and how dysarthric speech intelligibility is assessed. Towards the end of the chapter, the acoustic characteristics of dysarthric speech are analysed to illustrate the dysarthric and typical speech mismatch.

2.1 Human Speech Production and Perception

Speech production and perception are the two essential components in the process of communication. The speech sound waves are produced by the speakers and then travel from the speakers to the listeners. Spoken information is transmitted during this process. Listeners receive the acoustic signals through their ears and perceive the spoken information in an attempt to understand what the speaker intended to express. Understanding the mechanism of human speech production and perception is a prerequisite for knowing

how dysarthric speech is produced and how to support people with dysarthria using this knowledge of speech perception.

2.1.1 Speech Production

The production of spoken language involves three levels of processing: conceptualisation, formulation, and articulation [Levelt, 1999]. Since this thesis will not cover natural language production but focus on the acoustic side of speech, articulation will be explained in detail for speech production. Articulation is how people physically produce speech sounds from the highly coordinated functioning of several components (e.g., lungs, glottis, larynx, tongue, lips and jaw) in the human vocal tract. The vocal apparatus can be divided into three levels according to the relative position to the larynx: the parts below the larynx (i.e., subglottal structure), the larynx and its surrounding vocal folds components (i.e., glottal), and the parts above the larynx (i.e., supraglottal structure). Figure 2.1 shows a diagram showing the human speech production at these three levels.

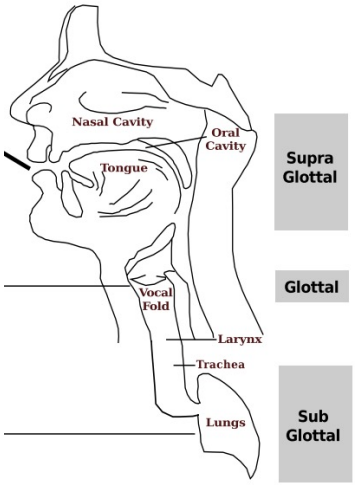


Figure 2.1: Diagram showing human speech production at levels based on proximity to the glottis.

The physical production of speech could be described as an air-oriented process. First, a stream of air flows from the subglottal level as the source of sound energy (i.e., the pulmonary pressure provided by the lungs). The air then moves up to the larynx where it passes through the glottis. The sound is generated by phonation through the glottis. The

movement of the vocal tract enables this process. In particular, the vocal tract vibrates to produce the fundamental frequency and its corresponding harmonics of the voice. Here, the air pressure is transformed according to the movements of the supra glottal articulators, resulting in the production of different vowels and consonants. Articulators within the vocal tract are divided into two types when involved in speech production: *fixed* and *movable* [Tatham and Morton, 2006]. These two types of articulators together create places of constriction or narrowing within the vocal tract. The fixed articulators, such as teeth and palate, are located in a particular region of the vocal tract on a repeatable basis. In contrast, the movable articulators (e.g., lips, jaw and tongue) are dynamic. They move, relocate or change their shapes while an utterance is produced.

The neurological way of describing speech production is brain-oriented, where the speech production originates in various centres of the brain. A timed and coordinated message is delivered to the musculoskeletal structure responsible for speech production directing a neural activation. Speech production is a complex process where various components are involved. Even small problems with any of these systems can cause speech impairment. For instance, damage in the coordination of the motor commands necessary for speech can cause difficulties in the manner of articulation. Damage existing in the central or peripheral nervous system can cause weakness and incoordination in the speech musculature.

2.1.2 Perception

"Speech is produced to be perceived" [Tatham and Morton, 2006]. Speakers convey their thoughts by speaking. At the other end of the information transmission channel is the process of the sounds of language being heard, interpreted and understood, which is known as speech perception. In the process of communication, both correctly speaking or perceiving a message is important. Perception is a learnable process. Listeners cannot begin to understand a language unless they have acquired some knowledge of its phonological, grammatical and semantic systems. Familiarisation with unfamiliar or ambiguous speech signals can facilitate perceptual learning of that same speech signal [Borrie et al., 2012]. Samuel and Kraljic [2009] and McClelland and Elman [1986] have suggested that

the individual's perceptual system is flexible and dynamically adjusts to match the information provided in the incoming signal. When listeners are familiarised with a speech signal that is at first unfamiliar or ambiguous, they are able to modify their perceptual strategies for subsequent processing of the speech. This is called perception learning. The process could explain why people with speech disorders are often intelligible to their close friends and family but not to others. Knowing about the perception learning could help developing assistive techniques to support people with dysarthria. Perception learning is an inherent characteristic of human beings; how might this be applied to machines? Can automatic speech recognition (ASR) systems perform such a process to recognise (dysarthric) speech better? Speech perception research has applications in building computer systems that can recognise speech, in improving speech recognition for hearing- and language-impaired listeners, and in foreign-language teaching. It is also a gold standard for clinical differential diagnosis and judgments of the severity of atypical speech. This motivates the exploration of the adaptation techniques to the target speech for the ASR systems. For instance, Lally et al. [2019] have investigated the speech perception effects on dysarthric speech. As familiarised human listeners better understand dysarthric speech, the supervised (i.e., prior) information could facilitate familiarised ASR models.

2.2 Dysarthria

2.2.1 What is Dysarthria

Dysarthria is defined as a speech disorder caused by neurologic impairments affecting the planning, programming, control or execution of speech. The neurologic impairments disrupt the motor system which controls the physical production of speech [Gowers, 2001]. Different forms of disruption may result in different types of disorder. There are three major causes of dysarthria. Firstly, a death of dopamine cells can cause deficiencies and imbalance in the neurochemical system. This type of disorder is called *neurochemical dysarthria*. Parkinson's disease (PD) is such an example. Speakers with PD usually have problems with body movements accompanied by tremors or shaking. This impaired posture and balance results in changes in speech. Secondly, the gradual decline and

death of neuronal activity can cause MND. This is called *degenerative dysarthria*. The reduced neuronal activity results in reduced motor control and sometimes even in the wrong commands being sent to the articulators and other parts of the body. Thirdly, some external or internal injuries may cause neurologic impairments. For instance, a head injury can lead to the obstruction of blood supply to neurons. Likewise, stroke can cause sudden damage to the brain. This can cause *traumatic dysarthria*. People with stroke or other internal brain injuries usually have slurred or garbled speech.

2.2.2 Types of Dysarthria

Three criterion can be used to categorise dysarthria: the lesion site, the degree of dysarthria severity and the developmental pattern. The motor neurone systems which control speech articulation rely on different parts of the brain. Therefore, the damage to any site may influence speech production. This section focuses on the types of dysarthria classified by the site of the lesion and the expressed speech characteristics. According to the site of the neurological damage, dysarthria can be classified into six types: Flaccid dysarthria, Spastic dysarthria, Ataxic dysarthria, Hypokinetic dysarthria, Hyperkinetic dysarthria and Mixed dysarthria. They are summarised in Table 2.1. In particular,

Flaccid dysarthria is caused by the damage to the *lower* motor neurone system, mostly caused by stroke. An example of this type of dysarthria is bulbar palsy, characterised by weakness in muscle movement and poor reflexes (so-called hypotonia). The produced speech often has too much sound resonation (vibration) in the nose, a breathy voice quality, mono-pitch and imprecise consonant production.

Spastic dysarthria is caused by damage to the *upper* motor neurons, caused mainly by MND. An example of this type of dysarthria is pseudobulbar palsy. People with pseudobulbar palsy cannot control their facial movements, and certain muscles are continuously contracted followed by overactive reflexes. The produced speech is characterised by imprecise consonants, mono-pitch, reduced stress, a harsh and strained voice quality, slow speaking rate and hypernasality.

Ataxic dysarthria is caused by the damage to the *cerebellum*. Ataxia is defined as the lack of order. This causes mistakes in range, force, timing and direction of the speech

articulators. People with ataxic dysarthria often move slowly and have decreased muscle tension. They have less control of fast movements. The respiratory, phonatory and articulatory aspects of speech production are affected. The produced speech is characterised by excess and equal stress, irregular articulatory breakdown, vowel distortion, harsh voice quality, and imprecise consonants.

Hypokinetic dysarthria is caused by the *extrapyramidal* system, which describes a number of centers in the brain and their associated tracts used to coordinate and process motor commands. Damage to the extrapyramidal system can cause involuntary actions and reduced movement. PD is one example of Hypokinetic dysarthria. People with PD often have rigidity in muscles, slowness and limited range in speech movements. The produced speech is characterised by reduced stress, mono-pitch, mono loudness, imprecise consonants, inappropriate silence and a continuous breathy voice quality.

Hyperkinetic dysarthria is also caused by the *extrapyramidal* system. This kind of damage to the extrapyramidal system causes an increase in the movement. Dystonia is one example of Hyperkinetic dysarthria, which is a movement disorder in which a person's muscles contract uncontrollably. The produced speech is characterised by imprecise consonants, distorted vowels, irregular articulatory breakdown, mono-pitch and mono-loudness.

Mixed dysarthria refers to a combination of any of the above types of dysarthria. The causes of the disorders are usually complicated. The most common conditions of mixed dysarthria are degenerative (e.g., MND, PD) and vascular disorders (e.g., stroke).

2.2.3 Dysarthric Speech Intelligibility

The level of dysarthria is assessed by two subjective criteria: Human listener perceptual measures of articulation, and speech intelligibility [Whurr, 1988]. Speech intelligibility is defined as the accuracy with which a message is conveyed by a speaker and recovered by a listener [Yorkston et al., 1999]. It could indicate how well a speaker is understood despite speech impairments [Beukelman and Yorkston, 1979]. Speech intelligibility has frequently been used as the main indicator of the severity level of dysarthria [Kent et al., 1989], which is ranging from *mild*, *moderate* and *severe* or any condition within, such as,

Table 2.1: Types of dysarthria

Types	Damages	Speech Characteristics	Common Disorder
Flaccid dysarthria	Lower motor neurons	Breathy voice, mono-pitch and imprecise consonant	Stroke, degenerative disease and muscular dystrophy
Spastic dysarthria	Upper motor neurons	Mono-pitch, reduced stress, imprecise consonant, slow rate, harsh and strained voice	MND, multiple stroke
Ataxic dysarthria	Cerebellum	Affects the respiratory, phonatory and articulatory	Cerebellar degeneration
Hypokinetic dysarthria	The extrapyramidal system's basal ganglia circuitry	reduced stress, imprecise consonant, mono-pitch, mono loudness and phases of inappropriate silences	Parkinson's disease and parkinsonism
Hyperkinetic dysarthria	The extrapyramidal system's basal ganglia component or portions of the cerebellar circuitry	Abnormal and unexpected involuntary movements	Huntington's disease
Mixed dysarthria	A combination of any of the above forms	Comprehensive characteristics of each type	MND

mild-moderate [Klasner and Yorkston, 2005]. The higher the severity of dysarthria, the lower the speech intelligibility score. Three popular clinical dysarthria assessment tools are commonly used: the Frenchay Dysarthria Assessment [Enderby, 1980], Computerized Assessment of Intelligibility of Dysarthric Speech [Yorkston et al., 1984b] and the Swedish Dysarthria Test [Lillvik et al., 1999]. A significant correlation exists between speech intelligibility and the accuracy achieved by the ASR system [Thomas-Stonell et al., 1998; Wilson and Blaney, 2000]. In general, speakers with higher intelligibility scores (speakers with mild dysarthria and typical speakers) tends to obtain better ASR performance on their speech. In contrast, speech produced by speakers with lower intelligibility scores (speakers with moderate and severe dysarthria) tends to get worse ASR performance. The recognition abilities of humans and ASR systems on dysarthric speech in different severity levels do vary, though. Sy and Horowitz [1993] suggested that, in general, for moderately and mildly impaired speakers, the ASR performance is lower than the human perception performance when considering listeners who were unfamiliar with the speaker. Ferrier et al. [1995] hypothesised that at low and moderately low intelligibility levels, applying speaker adaptation can improve the ASR performance to outperform human

perception in recognising dysarthric speech. The reason might be that computers deal more readily with altered but consistent category boundaries than humans.

2.3 Acoustic Characteristics of Dysarthric Speech

As mentioned above, dysarthria affects the way a person speaks. It interferes with articulation, respiration, phonation, and resonance, causing reduced intelligibility of the produced dysarthric speech. These atypical variations pose a great challenge for mainstream ASR systems to achieve desirable performance on dysarthric speech and they affect recognition accuracy. This section will present the acoustic characteristic of dysarthric speech and how these differ from those of typical speech.

2.3.1 The Mismatch to Typical Speech

Regardless of the large variability, in general, there are three common types of mismatch between dysarthric and typical speech: reduced speaking rate, less distinctive phone classes and boundary position shift.

The **reduced speaking rate** has been shown to be the typical characteristic of severely dysarthric speech [Raghavendra et al., 2001; Turner et al., 1995]. Due to the damage in the neural-motor system, speakers with dysarthria often have difficulty moving their articulators from the position of one pronunciation to another. As a result, they will need more time to produce an utterance with a slower speaking rate than typical speakers. Many previous studies have done experiments to support this view. The mean phoneme duration on several words and sentences for dysarthric versus typical speech was calculated to show that, in general, the duration of dysarthric speech is longer than that of typical speech [Kent et al., 1979]. The authors also found that the duration increased with the increasing severity degree of the dysarthria. In addition, Brown and Aronson [1970] found dysarthria caused prosodic impairments characterised by prolonged intervals and phonemes. This also provides evidence of the slow speaking rate of dysarthric speech. The slow speaking rate can lead to poorer performance of the speech recogniser as non-existing words are inserted into the transcript [Turner et al., 1995].

Less distinct phone classes is also considered to be one of the main characteristics of dysarthric speech. By visualising the space of different vowels, Wilson and Blaney [2000] found that the vowel space is more centralised/overlapping in dysarthric speech while more distinct in typical speech. The less differentiable vowel target space might be due to the reduced flexibility of articulators for people with dysarthria [Kent et al., 2004]. Consequently, it is more challenging for ASR systems to identify the phonemic categories of dysarthric speech. This usually leads to substitution recognition errors.

The **boundary position shifts** refers to the additional shift in the boundaries between voice and voiceless contrast, e.g., 'b' and 'p'. The standard category boundary positions of dysarthric speech were found to shift to a higher value than that of typical speech by Wilson and Blaney [2000]. The new boundaries are consistent for mildly dysarthric speech while inconsistent for more severely dysarthric speech. As a result, the contrast between minimal pairs (e.g., voice and voiceless) is maintained for mildly dysarthric speech but could not be always identified for speech at higher-level severity.

The **word boundary ambiguity** refers to the difficulty of detecting word boundaries in a sentence. In continuous speech, there are no real breaks between individual words. Isolating the words within a sentence is usually conducted by the listener [Maciuszek, 2018]. For example, strings "ice cream" and "I scream" sound indistinguishable but have totally different meanings. This is known as word boundary ambiguity. Coarticulation, which is when the acoustic realisation of a particular phoneme is affected by neighbouring sounds, can also cause word boundary ambiguity [Crowley and Bowern, 2010]. In typical speech, the word boundary ambiguity is often regarded as a lexical or syntactic ambiguity of a specific language. The solution is often the concept of a word or a sentence. However, the word boundary ambiguity becomes bigger because the phonemic ambiguity is sufficiently high in dysarthric speech [Liss et al., 1998]. Wilson and Blaney [2000] found the prevalence of merging acoustic boundaries in dysarthric speech. In addition, the voice and voiceless contrasts and different phone classes are often indistinguishable, which often occur at the word boundaries. Consequently, dysarthric speech sounds slurred and blurred. Compared with isolated words, continuous dysarthric speech is more difficult to recognise, with even harder word boundaries to be detected.

2.3.2 High Degree of Inter- and Intra-speaker Variability

Apart from the main common acoustic characteristic mismatched with typical speech, large variability is also a characteristic of dysarthric speech, affecting ASR performance. The speech variability usually refers to linguistic variability, speaker variability and channel variability [Makhoul and Schwartz, 1995]. Speaker variability is a more related factor in dysarthric speech, including inter- and intra speaker variability.

Inter-speaker variability means that the speech varies from speaker to speaker. As mentioned in Section 2.2.3, dysarthria is assessed into different severity levels, ranging from mild to severe. The acoustic characteristics produced by speakers in different severity levels are different. This is one type of inter-speaker variability. Studies have shown that the greater speech variability often correlates with increasing severity of dysarthria [Ferrier et al., 1995; Wilson and Blaney, 2000]. In addition, speakers with the same level of dysarthria also produce speech with different characteristics. This might be because dysarthria is often caused by various conditions leading to different acoustic characteristics.

All speakers are different to each other but speakers with dysarthria have an even higher level of variability. The mainstream ASR systems for typical speech are usually speaker-independent (SI) models, i.e., the models trained on any other speakers in the dataset except for the test speaker, which is inclusive of various speakers. However, the SI dysarthric speech recognition systems still achieve poor performance due to the large inter-speaker variability. The speech pattern learnt in the SI models are usually not representative of the target dysarthric test speaker. For instance, Mengistu and Rudzicz [2011] has trained SI systems using only dysarthric speech data. The results showed that the model trained on dysarthric speech performs worse (24.85% vs 30.41% word recognition rate) than the model trained on typical speech. This demonstrates that the impact of speaker-variability is even larger than the dysarthric and typical speech mismatch, and more data is required for training a well-performed SI model. Christensen et al. [2012b] explored the dependence between the severity level of dysarthria and the accuracy of the ASR and showed that each speaker with dysarthria has a personalised optimal system. There have been studies exploring ways to build systems robust to inter-

speaker variability, such as speaker adaption training (SAT) and deep learning training; however, a large amount of speech data is always required for training. The data scarcity problem makes these techniques more challenging to address the inter-speaker variability efficiently well for dysarthric speech [Doyle et al., 1997].

Intra-speaker variability refers to the variety of speech produced by a single speaker with dysarthria. Previous studies have analysed various measurements and provided evidence of the intra-speaker variability. [Kent et al., 1979] analysed the duration of the phone segments for dysarthric and typical speech. They found that speakers with dysarthria are more variable in segment duration than typical speakers. Wilson and Blaney [2000] analysed the parameters of voice including voice onset time (VOT), vowel duration (VD), fricative duration (FD) and vowel formant (VF) in each of the utterances. They found that the model adaptation ability to the speakers of dysarthria is slow and it is hard to build a stable model for dysarthric speech. They demonstrated that, in general, speakers with moderate and severe dysarthria exhibit greater intra-speaker variability than the mild speakers and typical speakers. When the intra-speaker variation is too wide to be modelled by the available training data (especially in small size), the speaker-dependent models could perform even worse than the speaker adaptation models [Mengistu and Rudzicz, 2011]. The intra-speaker variability is challenging for ASR systems due to the limited amount of dysarthric speech data resources. And there are always limited number of speakers in the datasets.

As an example of showing the speaker variability, Figure 2.2 and 2.3 plots the waveform amplitude envelope and spectrogram for dysarthric and typical speech of speakers with varying severity levels: M01 (a severe male speaker), F03 (a moderate female speaker), F04 (a mild female speaker) and FC01/FC02 (typical female speakers). The figures are plotted for the word 'Jacket' and the sentence 'Just one side got wet' from the TORGO dataset [Rudzicz et al., 2012b] (one of the main databases with dysarthric speech) consisting of both isolated words and sentences. Figure 2.2 shows that the speaking rate of the speaker with the highest dysarthria severity level is approximately twice that of the typical speakers. Generally, the higher the severity, the lower the speaking rate. It is also observed that for the mildly dysarthric speech, the shapes of the waveform and the spectrogram are

very similar to the typical speaker. In terms of speech intelligibility, vowels, stops, and fricatives are depicted in temporal and spectral structures for typical and mildly dysarthric speech. However, the disfluencies and vague transition are observed between the different phonemes in moderately and especially in severely dysarthric speech. A big difference in the vowel shapes can also be seen in the two figures. The envelope of the vowel is less variant in severely dysarthric speech compared with the typical speech. This might be because of the lack of muscle coordination. For the sentence illustrated in Figure 2.3, the higher the intelligibility, the clearer the word boundaries are. The distortion can be seen at the word boundaries in severely dysarthric speech.

2.4 Summary

This chapter gave an overview of speech production and perception, the background of dysarthria, and an analysis of acoustic characteristics of dysarthric speech. The dysarthric and typical speech mismatch and the inter- and intra-speaker variabilities in dysarthric speech were illustrated. These pose great challenges to the performance of mainstream ASR systems on dysarthric speech, especially on sentence utterances. The next chapter will introduce the basic ASR technologies and review the progress of dysarthric speech recognition.

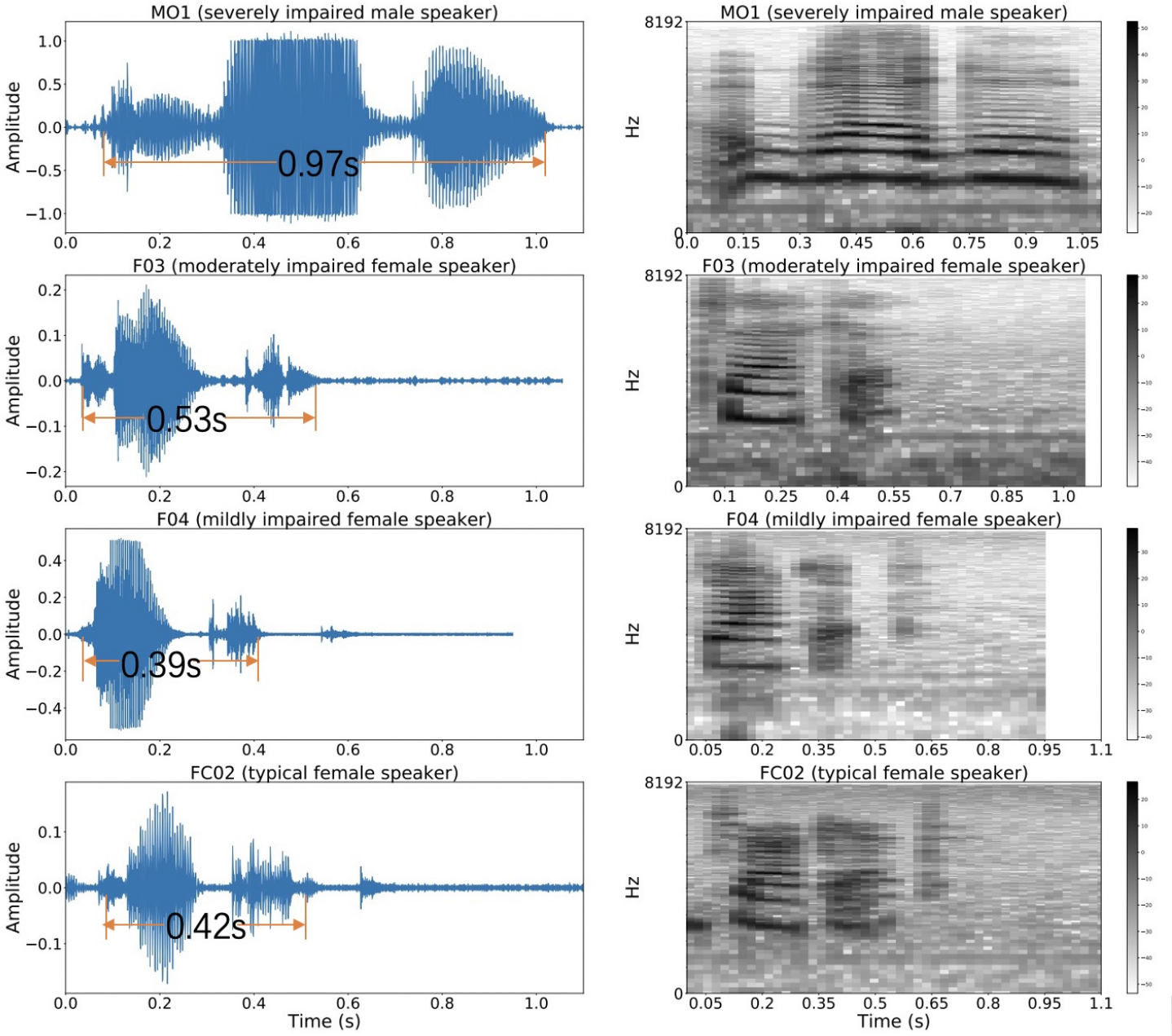


Figure 2.2: Waveform and spectrogram of the word 'Jacket' for speakers with different dysarthria severity.

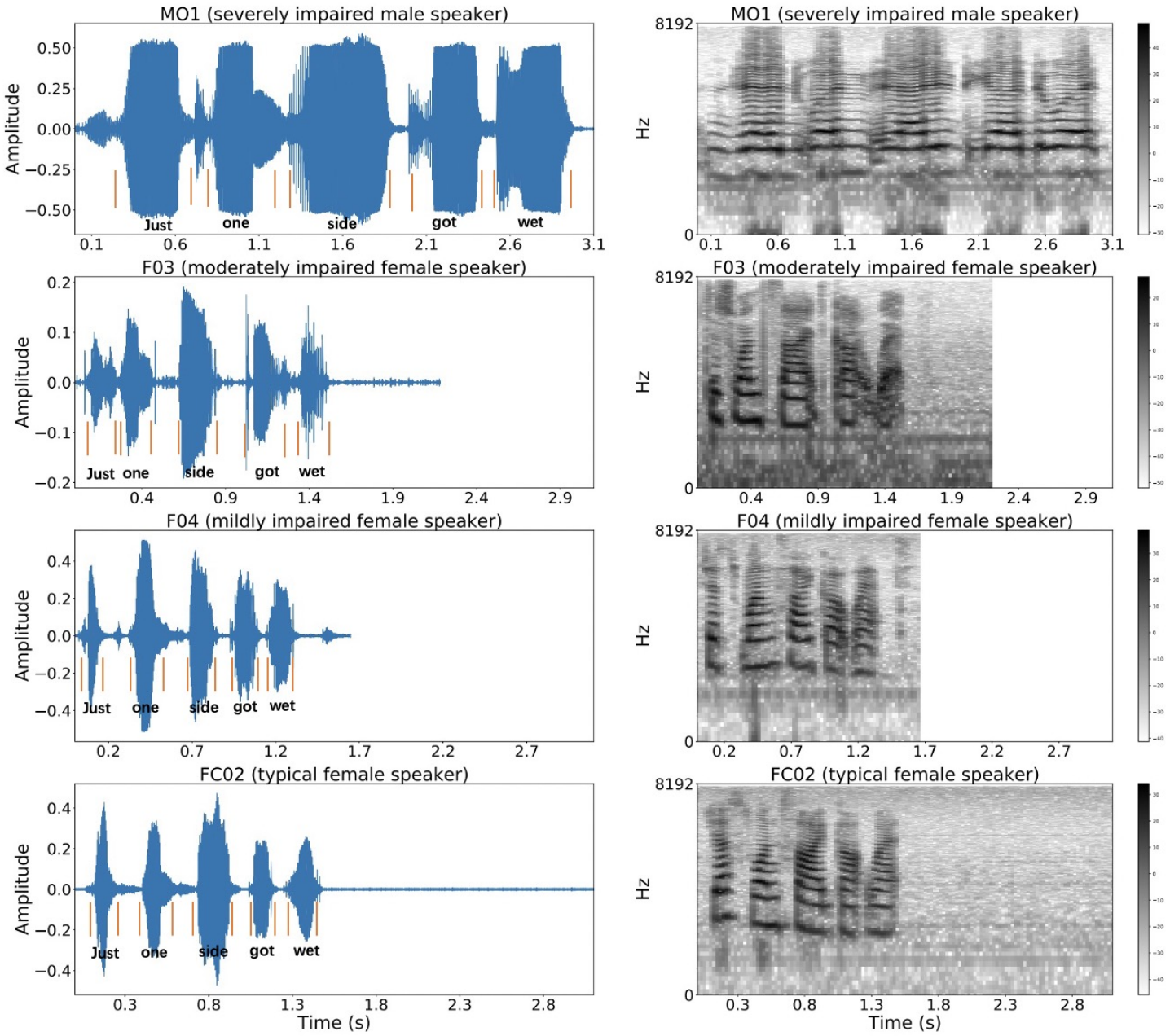


Figure 2.3: Waveform and spectrogram of the sentence 'Just one side got wet' for speakers with different dysarthria severity.

Chapter 3

A Review of Automatic Recognition for Dysarthric Speech

Contents

3.1 Introduction

Speech is an attractive communication interface. It offers potential for people with dysarthria, who find keyboards and touchscreens difficult to use, to effectively interact with machines. It is also an effective input medium that enables people with dysarthria to use speech commands for hands-free interaction with smart devices in their homes or elsewhere. This speech-driven user interface can be provided by automatic speech recognition (ASR) – a speech technology that translates input speech into text transcriptions. ASR technology can also help people with dysarthria better engage in communicating with non-familiar people by accurately transcribing what they have said. Therefore, a high-performance ASR system has clear potential enhancing human-human and human-machine interaction for people with dysarthria.

However, the current performance of ASR technology when applied to dysarthric speech does not give rise to optimism. Dysarthric speech usually show a significant mismatch to typical speech and has a high degree of variability. In addition, there is a lack of suitable training data which limits the effectiveness of applying some deep learning approaches on dysarthric speech. Consequently, mainstream ASR systems designed for typical speech do not work reliably on dysarthric speech. Achieving an acceptable performance for dysarthric speech is challenging. Carefully designed customised automatic dysarthric speech recognition (ADSR) systems are required.

This chapter will review the recent progress and help identify the current research gaps in dysarthric speech recognition. It will also relate the key research questions that the thesis aims to address. The chapter is organised as follows. Section 3.2 will briefly introduce the main components of a typical ASR framework focusing on those that are most relevant in the context of this thesis. Then, Section 3.3 will review how research has progressed in the field of dysarthric speech recognition within some broad application domains. Finally, Section 3.4 will highlight the current research gaps in ADSR systems and summarise the chapter.

3.2 A Typical Automatic Speech Recognition System

Various components of an ASR system need to be adapted in order for the technology to work well with dysarthric speech. Before reviewing the recent ADSR research, this section briefly presents an overview of the generic architecture of a typical ASR system. The main components relevant in the context of this thesis are introduced in detail. Figure 3.1 depicts the generic workflow of a typical ASR system.

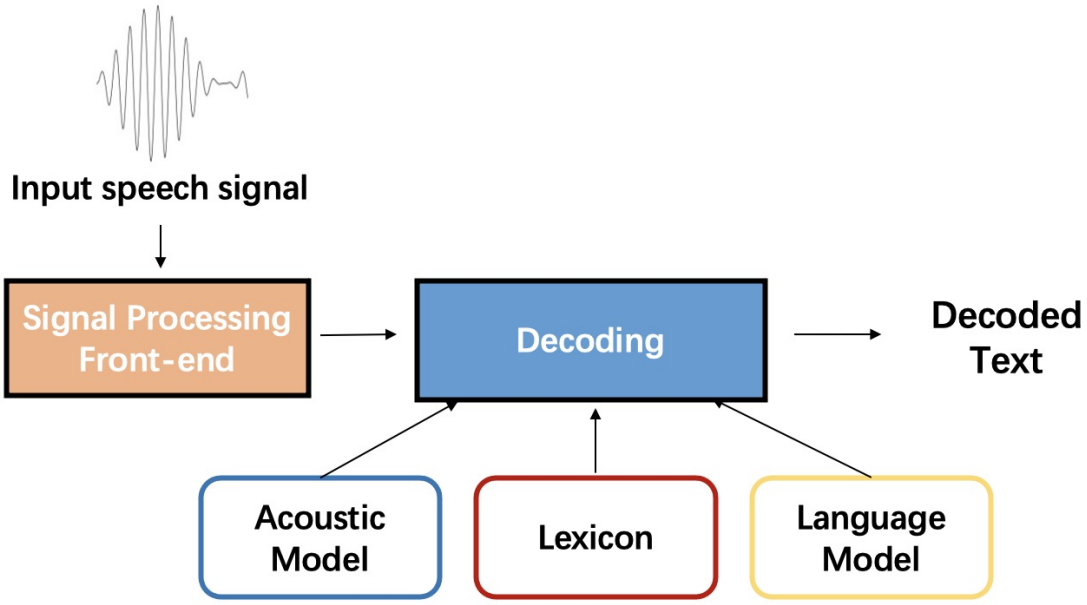


Figure 3.1: A typical ASR system.

- The **signal processing front-end** converts the input speech waveform into frame-level acoustic feature vectors.
- **Decoding** is an implementation of a search algorithm which finds the optimal hypotheses for an input speech signal.
- The **acoustic model** is responsible for matching acoustic feature vectors to individual words as defined in a vocabulary.
- The **lexicon** defines a dictionary of pronunciations of each word.

- The **language model** represents syntactical, semantical and discourse constraints on the word sequence from the acoustic model, which gives the probability of a sequence of words.

As shown in Figure 3.1, a ASR system contains two main stages: feature extraction (front-end processing) and decoding. The former will be discussed later in Section 3.2.1. Decoding is also known as a search/inference process, which searches for the optimal sequence of words $W = w_1, w_2, ..., w_n$ given the acoustic observations $X = x_1, x_2, ..., x_n$ (the sequence of frame-based acoustic feature vectors). In particular, the decoder attempts to determine the following:

$$W^* = \arg\max_W P(W|X) \tag{3.1}$$

where W^* denotes a word sequence.

Applying the Bayes' Theorem, the above equation can be written as:

$$W^* = \arg\max_W \frac{P(X|W)P(W)}{P(X)} \tag{3.2}$$

Since $P(X)$ remains constant for each word sequence W, the search problem can be then defined as two main parts:

$$W^* = \arg\max_W P(X|W)P(W) \tag{3.3}$$

where $P(X|W)$ is estimated using the acoustic model and $P(W)$ is estimated using the language model.

To summarise, the ASR process searches for the most probable word/word sequence W^* given the observed acoustics X, using all knowledge sources (i.e., the acoustic model, lexicon and language model). The algorithm searches for the best path with the highest probability by expanding words, phonemes and hidden Markov model (HMM) states, and assigning scores from the different components. Then a trackback process of the highest probable path produces the best hypothesis transcription. The rest of this section will describe front-end processing and each of the source knowledge components used in the decoding process.

3.2.1 Front-end Processing

Feature extraction is the first step in speech processing which transforms the input waveform into a sequence of acoustic feature vectors. This is also known as front-end processing. Desirable acoustic features used for ASR should be compressed and encode the most relevant information of the speech signal to distinguish between phonemes. The features are also expected to be robust against speaker variability and any noise. The front-end processing removes less significant information such as the intensity and background noise of the speech signal as well as the speaker attributes, which helps the acoustic model become robust to irrelevant data, and therefore, useful for the phoneme classification.

Although there are many types of feature representations, the Mel-frequency cepstrum (MFCC) [Davis and Mermelstein, 1980] is the most commonly used feature in speech recognition which is based on the cepstrum. MFCC features are used in all experiments in this thesis as baselines. This section takes MFCC as an example to explain how handcrafted features are extracted. Figure 3.2 presents a diagram for the process of extracting MFCC features.

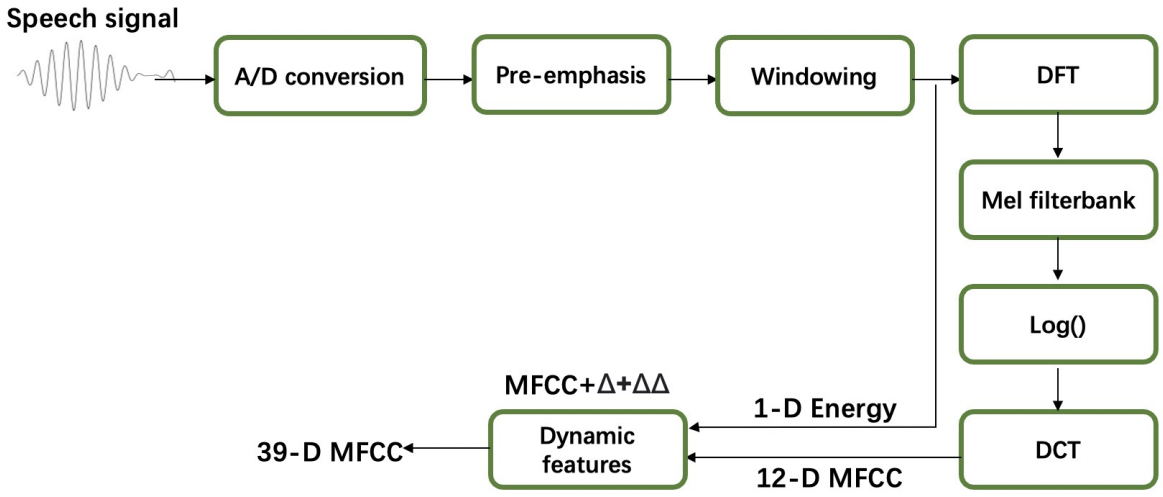


Figure 3.2: MFCC feature extraction (modified).

The speech waveform is a continuous signal. First, the analogue signal is converted into a digital signal by sampling and quantisation, so-called analogue-to-digital (A/D) conversion. The extraction of MFCC starts by pre-emphasis which boosts the amount

of energy (magnitude) of the high frequency components. Then, under the assumption that the speech signal is stationary over short periods of time [Kwong and He, 2001], the pre-emphasised signal is split into overlapping frames[1]. In the context of dysarthric speech, given that the speaking rate is normally slower than typical speech, increasing the frame shift (e.g., from 10 ms for typical speech to 15 ms for severely dysarthric speech) may help compensate for the reduced speaking rate.

The discrete Fourier transform (DFT) extracts spectral information (e.g., the power spectrum) from each windowed frame to compute how much energy is at each frequency band. The design of the feature extraction is motivated by the models of the human auditory system. Given that the human hearing is less sensitive to higher frequencies (above 1000 Hz) and that the human response to signal level is logarithmic, the power spectrum is passed through the Mel filterbank and logarithm. Finally, the discrete cosine transform (DCT) is applied to transform the log magnitude spectrum to the cepstral domain and remove redundant information. The resulting cepstral coefficients (12-D) tend to be less correlated. After adding the energy to the resulting cepstral coefficients, the dynamic coefficients (the first and second order derivatives of the features: $\Delta + \Delta\Delta$) are calculated to supplement the MFCC features. MFCC is highly effective in ASR, but things could be better. Feature transformations have been applied to maximize the separability between phonemes. For instance, linear discriminant analysis (LDA) [Fisher, 1936]) which is a dimensionality reduction technique that is commonly used for supervised classification problems. It takes the speech feature vectors and builds HMM states with a reduced feature space for all data. It models the differences in the phoneme classes and separates the classes. Normalizations have been applied to reduce the mismatch between training and test data as well as to normalize the speaker information. cepstral mean and variance normalisation (CMVN) [Naik, 1995] is a computationally efficient normalization technique for robust speech recognition. It linearly transforms the cepstral coefficients to have the same segmental statistics and hence minimizes speaker mismatches.

[1]Usually, the input signal is analysed in terms of overlapping windows. Each resulting feature vector represents the information in a frame [Jurafsky, 2000]. In ASR systems, 25 or 30 ms are typically used as the frame length [Jurafsky, 2000]. To avoid losing information at the frame boundaries and to obtain a better temporal resolution [Rao and Vuppala, 2014], a frame shift (e.g., 10 ms) which is shorter than the frame length is commonly used.

By looking at the source-filter theory [Fant, 1970] of speech production, the speech signal is the response of the vocal tract *filter* (i.e., the result of the position of the articulators) to a sound *source*. The sound source is mainly characterised by the fundamental frequency (F0)/pitch, and it does not carry relevant information for phoneme identification. As a result, the filter information is more commonly used in ASR. Cepstral analysis, the last step of the MFCC feature extraction process, is one way of separating the source and filter in a speech signal. In ASR systems, the redundant F0 information can be removed through the separation and only the filter information is used. However, this source-filter feature engine might not be optimal for dysarthric speech recognition. In the case of dysarthric speech, the neurological impairment causes atrophy of the musculoskeletal structure, therefore the articulators are poorly controlled. This variation of the articulator positions affects the response of the vocal tract filter, which in turn affects the effectiveness of MFCC. Other important pathological voice parameters (i.e., jitter, shimmer and F0) could be appended when extracting the features.

Recently, with the popularity of deep learning techniques, features extracted from artificial neural networks have been introduced in ASR systems. The features are usually extracted from one of the intermediate layers of the neural network. Compared with handcrafted approaches, the neural networks can learn information relevant for the phoneme classification from the speech signal automatically without the need of tuning the feature parameters manually, which helps avoid losing relevant information to the task. To handle dysarthric variabilities and the mismatch to typical speech, it is necessary but complex to adjust feature parameters, such as the frame shift, filter bandwidth and vocal tract length to particular speakers accordingly. Neural network-based representation learning therefore can be particularly useful for dysarthric speech. However, the lack of dysarthric data always limits the effectiveness of the deep representation learning techniques. Research needs to be conducted to maximise the usage of available data. This thesis will explore approaches to employing deep representations that benefit recognising dysarthric speech.

3.2.2 Lexicon

Lexicons are lists of words with pronunciation for each word expressed as a phone sequence [Jurafsky and Martin, 2009], which is also known as pronunciation dictionaries. The commonly used publicly available English pronunciation dictionaries are Carnegie Mellon University Pronouncing Dictionary (CMU) [Weide et al., 1998], CELEX [Baayen et al., 1996] and PRONLEX [Kingsbury et al., 1997] lexicons, which are widely used for both speech recognition and speech synthesis tasks. The following shows several examples of lexicon entries:

ACTON AE K T AH N

FAMILY F AE M AH L IY

DESERT D EH Z ER T

DESERT D IH Z ER T

... ...

As seen, each element of the lexicon list consists of a word from the vocabulary (e.g., action) and a phone sequence (e.g., AE K T AH N). Most of the words have a single pronunciation while some words may have more (e.g., desert: D EH Z ERT and desert: D IH Z ER T).

Different English speaking accents and speaking styles can cause different recognition results. How to choose a proper pronunciation dictionary for a dataset needs to be considered when designing a speech recognition system. TORGO [Rudzicz et al., 2012b], the dysarthric speech corpus used in this thesis, is a Canadian English dataset. However, only American and British English pronunciation dictionaries are publicly available. One of them needs to be chosen for the TORGO task. Canadian English may best be described as a product of the country's history: born out of treaties and settlement negotiations and migrations between the British and the Americans. Researchers found that although Canadian English is more similar to British English when it comes to spelling and grammar, it is more similar in pronunciation to American English. Both Canadian and American English are considered phonologically North American English and mostly indistinguishable as America has always been Canada's closest neighbour. One obvious difference example between the way Canadians speak and the way the British speak is

the letter "r". The British tend to omit the "r" sound in words when speaking while Americans and Canadians don't. For instance, the pronunciation of the word "far" in the British dictionary is "F AH". Canadians pronounce it as "F AA R", which is the same as in the American pronunciation dictionaries.

The lexicon can also be personalised. However, due to the large speaker variability and the limited amount of dysarthric data, it is hard to create a representative personalised lexicon for TORGO. In conclusion, the CMU American lexicon was used for the TORGO speech recognition task.

3.2.3 Acoustic Modelling

The acoustic model provides the probability $P(X|W)$ for a sequence of feature vectors X given a sequence of words W. HMM is widely used in acoustic modelling which models sub-word units (i.e., monophone or triphone models). Then, the sub-word units are accumulated to produce word-HMMs based on the rules defined by the lexicon. HMM is a stochastic finite-state automaton. Gaussian mixture model (GMM)-HMM has been the most popular acoustic model in ASR systems. Each state of and HMM is associated with weighted mixtures of Gaussian distributions, and includes transition probabilities and observation probability distribution which is represented using a GMM. The probability of generating a sequence of feature vectors is inferred by a set of HMM states, the probabilities of each state and the transition probabilities between the states.

With the development of deep learning techniques, hybrid deep neural network (DNN)-HMM systems are extensively used in ASR research. They replace the GMMs in the HMM with the outputs of a DNN as a phone probability estimator. The DNN learns the appropriate network parameters: weights and bias, for assigning the probability of each possible phonetic label for a given frame of input sequence data using the cross-entropy loss function and softmax outputs. Compared with GMMs, DNNs can directly use multiple frames in the input simultaneously to incorporate acoustic context.

Throughout this thesis, DNNs have clear potential for various tasks such as representation learning and acoustic modelling. However, deep layers result in a large amount of parameters, and training adequate DNNs usually requires a large amount of data.

Dysarthric speech recognition is a domain with very sparse data. How to maximise the usage of existing data to train adequate DNN models is an essential question to explore. The problems associated with wide variabilities in dysarthric speech and different speakers also need to be dealt with when designing acoustic models. This thesis will investigate ways of employing in-domain/out-of-domain (OOD) data for acoustic modelling as well as various deep acoustic models for dysarthric speech.

3.2.4 Language Modelling

The language model covers syntactical, semantic and discourse constraints of the language and assigns a prior probability $P(W)$ for any hypothesised word sequence $W = w_1 w_2 ... w_n$ [Jelinek, 1976]. When recognising isolated words, the output is restricted to a single word within a closed-set vocabulary. A much simpler language model is used to store the prior probability of each single word in the vocabulary. When it comes to large vocabulary continuous speech recognition tasks, the N-gram model [Bahl et al., 1989] is the most commonly used language model which follows an (N-1)-th order Markov assumption and the probability of the current word only depends on the (N-1) predecessor words. $P(W)$ is computed as:

$$P(W) = P(w_1)P(w_2|w_1)\prod_{k=3}^{n} P(w_k|w_{k-N+1}...w_{k-1}) \tag{3.4}$$

where n is the total number of words and N is the language model parameter. The first two items refer to the unigram and bigram. The N-gram language model parameters are estimated using a maximum likelihood. Take tri-gram as an example:

$$P(w_i|w_{i-1}, w_{i-2}) = \frac{count(w_{i-2}, w_{i-1}, w_i)}{count(w_{i-2}, w_{i-1})} \tag{3.5}$$

where $count(w_{i-2}, w_{i-1}, w_i)$ is the number of times that the particular word sequence $(w_{i-2}w_{i-1}w_i)$ occurs in the training data.

In practice, due to data sparsity, the above equation can fail to estimate any missing N-gram sequences ($\text{count}(w_{i-2}, w_{i-1}, w_i) = 0$). This is also called the zero probability

problem. Estimating the parameters for an N-gram model is especially challenging for larger values of N. As N grows larger, the data grows sparser and more zero counts will occur. Such an issue can be addressed by applying smoothing methods that involve discounting and back-off techniques. The Discounting methods (e.g., Witten-Bell discounting [Witten, 1991] and Good-Turning discounting [Good, 1953]) handles the problem by shifting the probability mass from the non-zero count N-grams to the zero or low count N-grams, while the back-off technique (e.g., Katz back-off smoothing [Katz, 1987]) assigns a zero-count n-gram with a scaled factor of its corresponding lower order n-gram counts.

In the past few years, with the advancements of deep neural networks, neural language models (e.g., recurrent neural network (RNN)-based language model (LM)s (e.g., ELMo [Peters et al., 1802], sequence-to-sequence attention-based LMs and transformer-based LMs (e.g., BERT [Devlin et al., 2018])) have been increasingly applied. RNN-based language models are the basic ones. An RNN does not need to follow the Markov assumption; therefore, it takes into account long-term dependencies. It appears to obtain lower perplexities (i.e., an evaluation indicator for language models where a lower value is better) than the N-grams. However, N-gram language models have lower latency and lower computational costs in evaluation.

The N-gram language models are commonly used for dysarthric speech recognition, typically $N = 2, 3, 4$. The 3-gram language model will be used throughout the thesis. Dysarthric speech is a low-resource data domain, where there is much less continuous dysarthric speech data available to train good continuous speech systems. Care needs to be taken when employing in-domain language models to recognise continuous speech. The small vocabulary size may lead to a large out-of-vocabulary rate and an increased word error rate. In addition, the training and test sets tend to be non-disjoint which can lead to overly optimistic evaluation results on sentence utterances.

3.3 Recent Progress in Automatic Recognition for Dysarthric Speech

3.3.1 Introduction

State-of-the-art mainstream ASR technology has obtained significant progress in many real-world scenarios by employing deep learning approaches trained on thousands of hours of speech data. However, improving ASR robustness against the distortions in dysarthric speech is still a big challenge. One reason is that the data-driven deep learning approaches cannot work efficiently using the limited dysarthric speech data. In addition, the significant mismatch to typical speech prevents mainstream ASR, designed for typical speech, from effectively recognising dysarthric speech, particularly for severely affected speakers. Therefore, ASR systems that are dedicated to dysarthric speech need to be explored. There has been increasing interest in the development of ADSR during the last few decades. Having introduced various components in a typical ASR system in Section 3.2, this section will review related studies for dysarthric speech recognition under speech representation learning, acoustic modelling, data augmentation and multimodal modelling.

3.3.2 Representation Learning

Speech representations carry information that the subsequent models will learn, and various representation learning approaches contain or lose different parts of the original signal's information. To build ADSR systems, it is expected that the learnt representation has the ability to normalise the high speaker variability and capture information associated with the dysarthria. For this reason, the speech representation learning approaches designed for typical speech might not be suitable for dysarthric speech. A comparative study was conducted on various conventional acoustic features for dysarthric speech recognition on TORGO by Mathew et al. [2018]. They compared the recognition performances of systems using MFCC, perceptual linear prediction (PLP), filter bank and reflection coefficients feature sets. It was found that the MFCC and PLP perform better than the

filter bank and reflection coefficient for dysarthric speech. Considering the slower speaking rate in dysarthric speech, Selouani et al. [2012] investigated the effect the window size of speech frames has for dysarthric speech. It was observed that a window greater than 25 ms with an extended 15 ms frame shift leads to 8% - 10% improvement on average. However, the scope of these two studies is limited in using the GMM-based acoustic model while deep learning-based acoustic models have been more widely used recently.

Deep speech representations that are learnt through deep learning approaches have drawn more and more attention in the ADSR task, and previous studies have demonstrated the effectiveness of employing deep speech representations. The bottleneck (BN) features refer to compressed features extracted from a neural network bottleneck layer using a supervised criterion such as phoneme prediction accuracy [Grezl and Fousek, 2008]. The bottleneck reduces the feature dimension, forcing the network to discard the redundant information irrelevant to the task. The convolutive bottleneck network (CBN) was proposed to extract disorder-dependent features employing convolutional neural network (CNN)s in Nakashika et al. [2014]. In the dysarthric speech, the key points in local time-frequency regions of an input feature map are often shifted slightly due to the fluctuation of the speech uttered by a person with dysarthria. The CNN was expected to deal with the small local fluctuations by capturing the temporal information while the frame-wise features (e.g., MFCCs) cannot, unless delta features are used. Employing features extracted from CBNs was shown to outperform conventional features.

Deep learning approaches require a large amount of data for training, however, insufficient data was used in this study. As a consequence, OOD data is usually used for pretraining. Takashima et al. [2015] expanded Nakashika et al. [2014]'s work using a pretrained CBN to prevent overfitting, which improved the acoustic modelling of the dysarthric speech. Tandem features were introduced by Hermansky et al. [2000], which fused the conventional features and the BN features extracted from a pretrained DNN model. The BN features were demonstrated to capture complementary information for dysarthric speech that can be beneficially fused with standard short-time spectral input features [Yılmaz et al., 2019]. Multi-level Adaptive Networks features were proposed in Christensen et al. [2013], which is an extension to the standard tandem features, exploiting

OOD data for cross-domain adaptation.

There has been a growth of interest in applying autoencoder (AE) to extracting BN features for dysarthric speech. [Chorowski et al., 2019; Sainath et al., 2012]. In contrast to conventional BN features, AE-BN features are learnt by reconstructing the input features in an unsupervised manner [Sainath et al., 2012]. This is more applicable in the context of low-resource dysarthric speech. Bhat et al. [2018] applied a denoising dysarthric speech feature enhancement framework using an AE. The system learnt non-linear mappings from the dysarthric speech to the typical speech. The enhanced features thereby improved the dysarthric speech recognition performance. This approach is limited to corpora with parallel recordings for both typical and dysarthric speech. These studies have been performed using only isolated-word dysarthric corpora such as UASpeech [Kim et al., 2008a]. The approaches applicable to a broader range of datasets and tasks is under-explored.

In addition to the features extracted from the approaches mentioned above, there have also been some other novel speech representation learning approaches specifically for dysarthric speech. For instance, a Speech Vision model was proposed in Shahamiri [2021] which extracted speech features visually by seeing the shape of the words pronounced by people with dysarthria. The Speech Vision system achieved 67% improved recognition accuracy for speakers in the UASpeech dataset.

3.3.3 Acoustic Modelling

As discussed in Section 2.3, dysarthric speech shows large speech variability associated with the articulation disorder. Research has been conducted to improve the ability of acoustic model to handle this variability. Conventional GMM-HMM acoustic models were employed in ADSR systems at the early stage of the research [Hasegawa-Johnson et al., 2006; Rudzicz, 2010c]. Then owing to the advance of deep learning techniques, deep learning acoustic models started to be widely deployed in ADSR. There have been two comparative studies exploring various deep learning architectures for instance DNNs, CNNs, time-delay neural network (TDNN) and long short-term memory (LSTM)s using dysarthric databases [Espana-Bonet and Fonollosa, 2016; Joy and Umesh, 2018]. The

results demonstrated that the hybrid DNN-HMM models outperform the classical GMM-HMM as well as other variants. Kim et al. [2018] modelled the spectral and temporal characteristics associated with dysarthria using Convolutional LSTM-RNNs [Han et al., 2017; Sainath et al., 2015] on the UASpeech dysarthric dataset. In this framework, the CNNs extract effective local features and LSTM-RNNs model the temporal dependencies of the features. As a result, this Convolutional LSTM-RNNs combined framework handles the local fluctuation caused by articulation disorder by capturing the local temporal-dimensional characteristics.

Although deep learning technologies have been successfully applied to typical ASR systems, the performance on dysarthric speech is still far behind that achieved on typical speech. The main reason is that the lack of sufficient dysarthric speech data limits the generalisation of current state-of-the-art deep learning-based ASR systems [Tu et al., 2017]. The amount of data is insufficient for the networks to capture the speech and speaker variability to train a generalised model. The details of the data scarcity problem will be explained in Section 3.3.4. There are two main approaches for training a model on low-resource data: model adaptation and data augmentation. This section will focus on the model adaptation approach while the latter will be reviewed in Section 3.3.4. Model adaptation is adapting a source model to a target domain. Usually, a source model is first trained on other data resources (e.g., the dysarthric speech data in other datasets or the OOD typical speech data). Then the source model is fine-tuned on a small set of data of the target speaker (speaker adaptation) or of the target dysarthric speech data.

Mengistu and Rudzicz [2011] adapted speaker-independent acoustic and lexicon models to the target speakers. The adaptation resulted in a significant WER reduction on the target speakers. This is known as speaker adaptation. A variety of ADSR systems using maximum likelihood and maximum a posteriori adaptation strategies are built in Christensen et al. [2012b]. They concluded that the model trained on typical speech adapted to the domain of dysarthric speech is a viable way of achieving good performance despite the inherent mismatch. The maximum a posteriori (MAP) estimation can deal with the large mismatch to a large extent. Xiong et al. [2020] investigated the use of transfer learning to adapt DNN models towards target speakers in personalised dysarthric speech

recognition systems. An utterance-based data selection of the source domain data was proposed in this work to improve the transferability towards the target domain further. The selection is based on the entropy of posterior probability, which is seen to obey a Gaussian distribution statistically. A two-step acoustic model adaptation approach was proposed in Takashima et al. [2020], aiming to tackle the large mismatch when adapting the pretrained ASR model trained on typical speech to dysarthric speech. In their architecture, an ASR model was first adapted to multiple speakers with dysarthria in order to learn the speaking style of dysarthric speech. Then the adapted model was further adapted for the target speaker. This adaptation scheme transfers the common knowledge learnt from a speaker-independent dysarthria model into the target speaker-dependent dysarthria model. More recently, a novel Bayesian parametric and neural architectural domain adaptation approach was proposed by Deng et al. [2021]. The model rapidly ports lattice-free maximum mutual information (LF-MMI) trained TDNNs ASR systems developed using a large amount of typical speech data to elderly and disordered speech task domains of more limited quantities.

The typical speech used in the dysarthric speech recognition adaptation system is regarded as OOD data. Exploiting OOD data has been shown to be beneficial for dysarthric speech [Christensen et al., 2013; Yılmaz et al., 2019]. Pretraining a model with OOD data can be especially crucial when little in-domain training data is available. The OOD typical-data pretraining framework was introduced in Christensen et al. [2013] to increase the robustness of the dysarthric speech representation learning process. Different model training setups using different subsets of data (typical, dysarthric or both typical and dysarthric data) were further investigated in Yılmaz et al. [2019] using a deep BN network. They concluded that the best performance is achieved by training the BN feature extractor on a large amount of OOD typical speech while the acoustic model is trained on the extracted dysarthric BN features.

3.3.4 Data Augmentation

Compared with typical speech, dysarthric speech is much more difficult to collect. There are always a limited number of dysarthric speakers involved in the recordings. Some

speakers with dysarthria are afraid of exposing their privacy by recording, and some find it difficult to come to the studio to complete the recording independently. In addition, speakers with severe dysarthria tend to tire quickly and speak slowly. As a result, the number of utterances recorded in a session is limited [Doyle et al., 1997]. The state-of-the-art data-driven ASR systems do not work well when feeding with small datasets. So far, the most commonly used English dysarthria datasets are Nemours (3 hours; American English [Menendez-Pidal et al., 1996]), TORGO (35 hours; Canadian English [Rudzicz et al., 2012b]), UASpeech (64.7 hours; American English [Kim et al., 2008a]) and the homeService (9.5 hours; British English [Nicolao et al., 2016]). More details about each database will be presented in Chapter 4. All of those dysarthric datasets are much smaller than the typical modern speech ASR databases containing thousands of hours of speech data such as LibriSpeech [Panayotov et al., 2015]. Hence, a major current limitation is a need for a large dataset to train better ADSR systems.

In recent years, increasing interest has been attached to the data augmentation approaches to handle the low-resource dysarthric speech data. Data augmentation refers to the process of artificially generating new synthetic samples for training from the original training data. There have been studies of various audio data augmentation approaches, e.g., speed perturbation, time-stretching, pitch shifting, dynamic range compression and adding noise, successfully applied to ASR for typical speech [Ko et al., 2015; Parascandolo et al., 2016; Piczak, 2015; Salamon and Bello, 2017]. Research has also demonstrated the benefit of employing data augmentation approaches on the ADSR task. Motivated by the spectral-temporal level differences of dysarthric speech from typical speech such as slower speaking rates, recent studies in data augmentation for dysarthric speech have been mainly focused on tempo-adjustments [Xiong et al., 2019], speed perturbation [Vachhani et al., 2018] and vocal tract length perturbation [Geng et al., 2020] from typical speech. In this way, not only is the speech generated but also new speakers are simulated. The generated "dysarthric like" speech is characterised by a slower speaking rate and modified vocal tract spectral shape. Then it is used to augment the original limited dysarthric speech training data. For instance, Vachhani et al. [2018] employed temporal and speed modifications to simulate extra dysarthric speech. They analysed phone durations for dysarthric

speech for speakers with different degrees of severity and generated corresponding synthetic dysarthric speech that matches the duration.

Motivated by work in Kaneko et al. [2017a,b], Jiao et al. [2018] applied adversarial training following the voice conversion function to transform typical speech toward dysarthric speech. The approach used listeners' judgement and a classifier to evaluate whether the simulated dysarthric speech matches actual dysarthric speech. The generated dysarthric speech using a convolutional generative adversarial network (GAN) [Jin et al., 2021] has been applied to the ADSR task and obtained the lowest word error rate (WER) on the UASpeech test set.

3.3.5 Multimodal Acoustic Modelling

If one data modality is not enough to capture the information of dysarthric speech, jointly modelling multimodal data may benefit the task. Data from other modalities may hold complementary information to improve the performance on ADSR.

There has been growing interest in building multimodal speech recognition systems recently. The most commonly used data modalities are visual and articulatory data. The articulatory data refers to the articulator movements collected by sensors attaching articulators. Compared with acoustic representations, articulatory information has been shown to be less speaker-variant [Fujimura, 1986] and more suitable to model the coarticulation [Wrench and Richmond, 2000] (will be discussed in Chapter 7). The visual data is usually images of video frames including the speaker's face. These two features have been successfully applied in many recent audio-visual and acoustic-articulatory ASR systems for typical speech [Afouras et al., 2018; Badino et al., 2016; Estellers and Thiran, 2012; Mitra et al., 2017; Wrench and Richmond, 2000]. Incorporating features from other data modalities is also promising for low-resource data ASR tasks. For instance, Abraham et al. [2017] employed articulatory features on low-resource languages ASR by jointly estimating the articulatory features and the acoustic model. The approach achieved relative 23% and 10% improvement on two low-resource Indian language datasets.

The visual features were employed on dysarthric speech in Salama et al. [2014] using the UASpeech dataset. It was demonstrated that the visual features extracted from the

video recordings were highly effective for recognition performance. In particular, they increased the recognition accuracy by around 3%. Incorporating articulatory information has also been shown to benefit ADSR tasks [Rudzicz, 2009; Xiong et al., 2018]. Since speakers with dysarthria differ from the typical speakers in the manner of their articulation, measuring the articulation empirically is beneficial [Rudzicz, 2011; Yılmaz et al., 2018]. It has the potential to outperform the acoustic-features-only frameworks. Due to the limited amount of dysarthric articulatory data, the synthetic articulatory data has been used to support acoustic features to improve acoustic modelling of dysarthric speech [Xiong et al., 2018; Yılmaz et al., 2018]. The synthesiser estimates the articulatory data from the acoustic representations by learning the acoustic-articulatory mapping.

For instance, Xiong et al. [2018] employed LSTM-RNNs to estimate articulatory information from acoustic features by learning the acoustic-to-articulatory mapping. The estimated articulatory features were then augmented with the conventional acoustic features achieving consistent improvement on dysarthric speech. Yılmaz et al. [2018] suggested that jointly using the articulatory and the acoustic features has potential against the spectro-temporal deviations in the dysarthric speech. The synthesisers are normally trained on typical speech and then applied to generate dysarthric speech. Given the mismatch between dysarthric and typical speech, the synthetic articulatory features might not reflect the real dysarthric articulatory space effectively. There have been several studies deploying the real dysarthric articulatory data [Rudzicz, 2010a,c; Rudzicz et al., 2012a]. However, most of these studies are based on GMM-HMM or simple DNN acoustic models. Acoustic-articulatory dysarthric speech recognition systems applying the real articulatory data with recent advanced acoustic models will be established in Chapter 8.

3.4 Summary

3.4.1 Current Research Gaps in Dysarthric Speech Recognition Systems

Although some progress has been made on dysarthric speech recognition, the performance is still unsatisfactory. There are still gaps in the research in ADSR.

1. Most of the previous ADSR studies were conducted on the isolated word dysarthric datasets (e.g., UASpeech). Although a few of them use the dysarthric datasets containing continuous speech, the difference between the isolated word and the sentence tasks has not been investigated.

2. The current acoustic-articulatory ADSR frameworks employed synthesised articulatory data are not optimal. The dysarthric articulation parameters were usually derived from acoustic signals using knowledge about typical speech, making it uncertain whether the synthesised articulatory data is in line with the actual dysarthric speech properties. Actual real recorded dysarthric articulatory data, instead, should be more reliable to use. However, the real dysarthric articulatory information is currently under-analysed. Whether it is still beneficial to incorporate it with more recent acoustic modelling architectures is also under-explored.

As a result, the following chapters will fill the gaps by investigating the difference between the isolated word and sentence tasks and moving the research focus to continuous dysarthric speech recognition.

3.4.2 Summary

There has been much interest in building ASR systems for people with dysarthria. This chapter reviewed various techniques that have been used to address the major challenges in ADSR: the mismatch to typical speech, the high intra- and inter-speaker variability and data scarcity. Although progress has been made, the recognition performance on dysarthric speech is still far behind that for typical speech. It is also noticed that most of the previous studies on dysarthric speech have focused on isolated word recognition with

limited vocabulary sizes. Although the single word recognition system can benefit people with dysarthria, systems that only recognise isolated words limit the range of activities that people with dysarthria can carry out. It is believed that people with dysarthria need help with a more natural way of communication (i.e., producing phrases and sentences in a large vocabulary size). In the following chapters, a series of studies using various speech recognition technologies to build robust ASR systems on continuous dysarthric speech will be carried out. Research questions will be explored and addressed in Chapter 5 (**RQ1**), Chapter 6 (**RQ2**), Chapter 7 (**RQ3**) and Chapter 8 (**RQ4**).

Chapter 4

Dysarthric Speech Datasets and Comparison

Contents

It is important to find an appropriate dataset before conducting the experiments. This chapter reviews some widely used dysarthric datasets used in previous studies. Section 4.1 presents several English dysarthric databases in detail (as this research focuses on English) and briefly introduces several non-English dysarthric datasets. The commonly used English dysarthric speech datasets are listed and compared in Section 4.2 to find the most appropriate dataset for this research.

4.1 Dysarthric Speech Corpora

As based in an English speaking country, this work is particularly interested in English dysarthric corpora. There are five commonly used English dysarthric speech corpora available: the Whitaker [Deller Jr et al., 1993], the Nemours [Menendez-Pidal et al., 1996], the UASpeech [Kim et al., 2008a], the TORGO [Rudzicz et al., 2012b] and the homeService [Nicolao et al., 2016].

4.1.1 The Whitaker Database

The Whitaker Database is an American English corpus that comprises the speech of 6 speakers with cerebral palsy (CP) at different severity levels and one typical speaker. The prompting items comprise 46 words: 26 alphabet letters, 10 single digits and 10 control words ('start', 'stop', 'yes', 'no', 'go', 'help', 'erase', 'rubout', 'repeat', and 'enter'), and other 35 words from the "Grandfather" passage [Johnson et al., 1963]. The participants were asked to repeat each word at least 30 times. The total number of dysarthric utterances is 19275. The audio recordings were sampled at 10 kHz.

The Whitaker dataset has been referenced in papers describing the collection of many other dysarthric speech datasets. However, it has not been commonly used by recent automatic dysarthric speech recognition (ADSR) studies since the vocabulary size is too small to train a practical ADSR system.

4.1.2 The Nemours Corpus

The collection Nemours corpus was motivated by intelligibility assessment and the investigation of general characteristics of dysarthric speech, such as production error patterns. The Nemours database comprises the speech of 11 male speakers with various degrees of CP assessed by a speech therapist using the Frenchay Dysarthria Assessment (FDA) tool. Each speaker spoke 74 short nonsense sentences and two paragraphs of connected speech. Each nonsense sentence in the database has the structure of 'The X is Ying the Z', where X and Z are monosyllabic nouns, and Y is disyllabic verb. During the recording, participants randomly selected X and Z (X $\neq$ Z) without replacement from a set of 74 monosyllabic nouns and selected Y from a set of 37 disyllabic verbs. Note that all of the words within a set differ in a single phoneme. The vocabulary size of the nonsense sentence recordings is 111. It includes two paragraphs taken from the "Grandfather" and the "Rainbow" passage. The audio recordings are sampled at 16 kHz.

Rudzicz [2007] has implemented experiments on the Nemours database by comparing the performance of a speaker-dependent and a speaker-adaptive Gaussian mixture model (GMM)–hidden Markov model (HMM) system. Caballero Morales and Cox [2009] modelled and attempted to correct the errors made by the speaker. The 'The X is Ying the Z' sentence structure in the Nemours datasets enables the phonetic comparison between different utterances. However, Nemours has been used in fewer studies as the dataset is not publicly available now.

4.1.3 The HomeService Corpus

The homeService database is a British English corpus of dysarthric speech data recorded in the home environment. It is motivated by developing a system that works in real environments (e.g., the home) which helps people with dysarthria interact with devices using the commands in a single word. The homeService corpus comprises the speech of five (three males and two females) speakers with severe dysarthria (three of them were with CP, one speaker was with motor neurone disease (MND), and the rest has not had the speaker's condition noted). The speech data was collected in two approaches: enrolment

data (ER) and interaction data (ID). The ER data was recorded as participants read lists of the words that they had chosen as commands in their system. The ID data was recorded as the participants operated the electronic devices in the house with the homeService speech-enabled interface. Three speakers were recorded with both ER and ID data, while the other two speakers only have the ER data due to personal reasons. The total recording duration is approximately 10 hours and the vocabulary size is 131. The audio recordings were sampled at 48 kHz. Only the 16 kHz-sampled, single-channel version of the audio was released. The recordings were collected over several months, allowing longitudinal studies on voice variations which are caused by degenerative speech impairment.

The homeService has been used for dysarthric speech recognition in Nicolao et al. [2016] and for dysarthria severity classification in Purohit et al. [2021]. The additional dysarthric dataset – UASpeech were used in both studies to train background models. The number of speakers with dysarthria in the dataset and the vocabulary size are both small, making it challenging to train an adequate automatic speech recognition (ASR) system for dysarthric speech solely on homeService. In addition, the recordings collected from the home environment is usually noisier than the recordings collected in the lab, so additional processing (e.g., denoising) is needed.

4.1.4 The UASpeech Corpus

UASpeech is an American English dysarthric speech database for Universal Access Research created by the University of Illinois. It is a collection of recordings from 16 speakers with CP and 13 age-matched typical speakers. The prompting items are all isolated words. The audio recordings were recorded by an eight-microphone array sampled at 48 kHz. The aligned visual features of speech were also captured by one video camera along with the audio recordings. The speakers with dysarthria in the database ranged from four severity levels (namely *Severe*, *Moderate-Severe*, *Moderate* and *Mild*) based on a subjective estimate of perceptual speech intelligibility assessment [Kim et al., 2008b]. Each speaker in the datasets read 455 distinct words, composed of 155 common words from 4 groups: 10 digits, 29 Nato alphabet letters, 19 command words ('up', 'down' etc.), 100 common

words ('the', 'this' etc.) and 300 uncommon words from the "Grandfather" passage and PBS (TIMIT sentences). The 455 words are split into three blocks for each speaker, and each block contains the common words and one-third of uncommon words. Blocks 1 and 3 were used as training data while block 2 as test data in previous UASpeech-based studies (e.g., [Christensen et al., 2012a]), and approximately 126,000 utterances were involved in the training and test sets.

UASpeech has been widely used in ADSR studies [Kim et al., 2018; Xiong et al., 2018, 2019; Yu et al., 2018] as it is the largest dysarthric speech corpus in American English with well pre-defined training and test partition. Researchers have also published standard scripts for the ADSR task (e.g., [Xiong et al., 2019, 2020]) on UASpeech, making it easier for others to explore models based on the well-presented baselines. Various models have been implemented and improvements are made on UASpeech. For instance, Xiong et al. [2019] applied the speech tempo adjustments to the acoustic features which reduced dysarthric and typical speech mismatch and achieved consistent recognition performance improvements on UASpeech. Xiong et al. [2018] also made improvements by employing estimated articulatory-based representations to better model dysarthric speech variability. The only limitation of the application of this dataset is that it only contains isolated words, which constraints the ADSR systems being applied to continuous speech.

4.1.5 The TORGO Corpus

The TORGO dataset is a Canadian English corpus created by the University of Toronto. It is a collection of 21 hours of aligned acoustic and articulatory recordings from 15 speakers. Eight of them (five males and three females) are with different degrees of dysarthria (*Severe*, *Moderate to severe*, *Moderate* and *Mild*) with CP or amyotrophic lateral sclerosis. The others are age- and gender-matched typical speakers (four males and three females). On average, 415 and 800 utterances were recorded from each speaker with dysarthria and typical speaker, respectively. The audio recordings were obtained by one head-mounted and one array microphone sampled at 16 kHz in four types of stimulation: non-word, isolated words, sentences and photograph descriptions. Usually, only isolated words and sentences are used in experiments. The non-word recordings are only used to control the

baseline speaker abilities in the recording stage, and the photograph descriptions do not have text transcription. The set of single words consist of English digits, international radio alphabets, twenty most frequent words in British National Corpus [Landow, 1993], and some phonetically contrasting pairs of words selected by Kent et al. [1989]. Most of the restricted sentences were selected from Yorkston-Beukelman assessment of intelligibility [Yorkston et al., 1984a] and the TIMIT database [Lamel et al., 1989]. The dataset consists of 615 unique words and 354 unique sentences. The total vocabulary size is 1573, of which the vocabulary size for the sentence prompts on their own is 1083. Apart from acoustic speech data, aligned articulatory data is also recorded for some of the utterances using a 3-D AG500 electromagnetic midsagittal articulography (EMA) system [Kroos, 2008].

TORGO has also been widely used since it was published [Espana-Bonet and Fonollosa, 2016; Joy and Umesh, 2018; Rudzicz, 2011]. Rudzicz [2011] achieved significant recognition improvements on the dysarthric speech by deploying the production knowledge – articulatory information using the conventional acoustic models. Joy and Umesh [2018] explored various ways and tuned various parameters to improve GMM and deep neural network (DNN) acoustic models for dysarthric speech recognition. Most of the previous TORGO-based work evaluate the whole dataset without considering the difference between the isolated word and sentence. This thesis will investigate a proper evaluation approach that works well for both isolated word and sentence tasks.

4.1.6 Non-English Dysarthric Corpora

There also exists some dysarthric corpora collected for other languages. For instance, the Mexican Spanish corpora [Deller Jr et al., 1993], the Korean dysarthric corpora [Choi et al., 2012], the Cantonese dysarthric corpus in Cantonese [Wong et al., 2015], the TYPALOC corpus in French [Meunier et al., 2016] and the EasyCall corpus in Italian [Turrisi et al., 2021].

Database	utt type	Num of Spk	Num of utt/spk	vocab size
Whitaker	isolated words	7 (6 Dys, 1 Typ)	3200	81
Nemours	sentences	11 (11 Dys, 1 Typ)	74	111
UASpeech	isolated words	29 (16 Dys, 13 Typ)	455	455
TORGO	isolated words and sentences	15 (8 Dys, 7 Typ)	951-969	1573
homeService	isolated words	5 (5 Dys)	various	131

Table 4.1: Details of five popular English dysarthric corpora

4.2 English Dysarthric Corpora Comparison

What should a desirable dataset for this research be? First, as the main target of this research is to improve recognition performance on continuous speech, the dataset should consist of not only isolated words but also sentences. Second, the dataset should be phonetically rich – the bigger the vocabulary size, the more utterances, the more speakers, the better. Moreover, the dataset should be appropriate for the ASR task. For instance, unlike datasets designed for speaker assessment, the one used for ASR should be able to split into non-disjoint training and test sets. It is an asset that the dataset contains data from other modalities (e.g., video or articulatory).

Different corpora are designed for various purposes. Nemours is motivated by the intelligibility assessment and homeService is motivated by developing a system that works in real environments. UASpeech is developed for ADSR research to design assistive technologies for people with dysarthria. TORGO is designed for the comparative study of dysarthric and typical speech. Table 4.1 summarises the details (the utterance type, the number of speakers, the number of utterances per speaker and the vocabulary size) of each English dysarthric speech dataset. UASpeech comprises the most speakers (both dysarthric and typical) and TORGO has the most utterances per speaker in addition to the largest vocabulary size. TORGO appears to be the only dataset consisting of both word and sentence prompting items. Containing a large number of speakers and a medium-sized vocabulary, UASpeech is a well-designed corpus for ADSR with a disjoint set of training and test sets. However, it could only work with isolated word recognition. Both Nemours and TORGO contain sentence utterances, therefore, are able to use for continuous dysarthric speech recognition. Compared with Nemours, the TORGO dataset

has a larger vocabulary size and more recording samples. The utterances in Nemours have a fixed structure of 'The X is Ying the Z', while the utterances in TORGO are more flexible and natural. In addition, the aligned acoustic and articulatory recordings make the TORGO dataset particularly interesting in continuous speech scenarios. The articulatory information could be exploited to build more complicated but robust ADSR systems. Consequently, TORGO is selected for the baseline experiments and further study of the research in this thesis.

However, TORGO is still not a perfect corpus for this ADSR task. For instance, the linguistic overlap among speakers (will be explained in Section 5.2.3) and the unwell explored articulatory recordings for each speaker (will be explained in Section 7.1.2). These limitations can largely influence the recognition performance on this dataset. A strong baseline considering limitations need to build for more fair results.

Chapter 5

Baseline Continuous Dysarthric Speech Recognition System

Contents

This chapter presents a pilot study of the recent automatic dysarthric speech recognition (ADSR) systems using the TORGO dysarthric database, which has been concluded as the best suitable corpus for this research in Chapter 4. A continuous dysarthric speech recognition baseline is then built on TORGO. Unlike previous TORGO-based studies as reviewed in Section 3.3, which evaluated the recognition performance averaged for the whole dataset, the performance is evaluated by prompt types, i.e., considering word and sentence recognition as two separate tasks, in this chapter. Based on the observed results, the problems of the evaluation framework of previous ADSR systems are noticed and discussed. As shown in Chapter 4, very few datasets exist that allow researchers to develop speaker-specific continuous speech recognition systems for people with dysarthria, and they are mostly not designed for automatic speech recognition (ASR), meaning great care has to be taken to choose an appropriate experimental setup. Then, the impact of the language model (LM) design is investigated and the out-of-domain (OOD) LMs trained with data originating from other datasets are proposed.

The work presented in this chapter explores two essential questions: *How to design the baseline experimental framework on TORGO*? and *What is a fair way to evaluate the continuous dysarthric speech recognition system with a limited amount of data and the issues with lack of variability in the prompts*? Following the above research questions, a reproducible benchmark for continuous dysarthric speech recognition is developed by using the recent acoustic models and OOD LMs for further research.

5.1 Baseline Experiment

5.1.1 Data Cleaning

Before building the baseline system, data cleaning needs to be done such as removing the noise data in the dataset. First, the audio recordings annotated with 'xxx', which indicates spurious noise, were discarded. Then, recordings shorter than 15 ms were removed given that the 15 ms frame shift is the smallest unit in an ASR system. Finally, any wrongly annotated recordings were also removed, most of which were accidentally recorded without any acoustic signal but were not annotated as noise in the dataset. Table 5.1 summarises

the remaining number of utterances in both word and sentence types per speaker in TORGO after cleaning the data.

Severity	**Severe**				**M/S**	**Moderate**	**Mild**	
Speaker	**F01**	**M01**	**M02**	**M04**	**M05**	**F03**	**F04**	**M03**
# Utterances in both types	228	739	766	651	572	1072	667	800
# Word utterances	188	561	582	500	440	797	498	610
# Sentence utterances	40	178	184	151	132	275	169	190
# Unique utterances	94	319	333	228	382	422	351	345

Table 5.1: The number of utterances per (F)emale and (M)ale speaker with dysarthria in TORGO. 'M/S': moderate to severe intelligibility. '#': the number of.

5.1.2 System Overview

The open-source Kaldi Speech Recognition toolkit [Povey et al., 2011] created by Johns Hopkins University was used in the experiments. Kaldi is a flexible software that is intended for building speech recognition systems. Four modules in an ASR pipeline are described for the system architecture: feature extraction, acoustic modelling, language modelling and pronunciation lexicon.

Feature extraction: 39-D MFCC+Δ+$\Delta\Delta$ with a spliced context window of length 9 frames is used. The Mel-frequency cepstrum (MFCC)s are then subsequently transformed to a 40-D vector via linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) to get more evolved speaker-independent features. Afterwards, the feature-space MLLR (fMLLR) is employed for speaker adaption training (SAT). 100-D i-vectors are also added to gather specific speaker information during deep neural network (DNN) training.

Language Modelling: The LM used for the baseline is reproduced from previous TORGO-based studies [Espana-Bonet and Fonollosa, 2016; Joy and Umesh, 2018] using the *SRILM* Toolkit [Stolcke, 2002]. In particular, it is a standard trigram LM built on prompts of the training data in TORGO. Interpolated Kneser-Ney discounting is applied to the LM for smoothing. From hereon, this LM used in the baseline is referred to as the *TORGO LM*.

Pronunciation Lexicon: The Carnegie Mellon University (CMU) Pronouncing Dictionary is used as the lexicon. It consists of over 134,000 words and their pronunciations in the ARPAbet phoneme set with 39 phonemes. The standard three-state context-dependent triphone models are used for acoustic modelling. Compared with the monophone model, the triphone model takes into account the position of phones within a word. For instance, for a phone AA, the extended phone list will include AA_B, AA_E, AA_I and AA_S, indicating whether the phone occurs in the beginning, end, internal of a word or as a singleton phone, respectively. This word position-dependent phone set includes 167 phones in total.

Acoustic Modelling: The classical Gaussian mixture model (GMM)-hidden Markov model (HMM) architecture and hybrid DNN-HMM architectures have been explored for acoustic modelling. The GMM-HMM employs a triphone model with speaker adapted transformation. In the GMM-HMM system, first, a monophone model is trained with 13-D MFCC features. Then the obtained alignments are used for basic triphone model training with 39-D MFCC+Δ+$\Delta\Delta$ features. Afterwards, speaker-independent transformations (LDA and MLLT) are applied to get the triphone speaker-independent alignments. Finally, fMLLR is applied for speaker-adaptive training as the speaker-dependent transformation. Discriminative training such as maximum mutual information (MMI) and feature-space MMI training are also employed to fit the HMM parameters.

In the hybrid system, the alignment for DNN senones (i.e., context-dependent phonemic states) is obtained with an additional GMM-HMM training using MFCC+LDA+MLLT+fMLLR features. The factored form of time-delay neural networks (TDNN-F) [Povey et al., 2018] incorporating convolutional neural network (CNN)s is used as a state-of-the-art DNN architecture. The trick of factorising matrices with a semi-orthogonal constraint of TDNN-F has been shown beneficial to ASR tasks [Xiong et al., 2020]. The inputs are 40-D log-Mel spectrogram features. The TDNN-F-CNN architecture used in the experiment comprises two CNN layers at the bottom, followed by nine TDNN-F layers. A linear layer, similar to linear hidden network (LHN) [Gemello et al., 2007], is added for speaker adaption before the output layer.

Severity	Severe				M/S	Moderate	Mild	
Speaker	**F01**	**M01**	**M02**	**M04**	**M05**	**F03**	**F04**	**M03**
# Utterances in train	16158	15647	15620	15735	15814	15314	15719	15586
# Utterances in test	228	739	766	651	572	1072	667	800
# Word utterances in test	188	561	582	500	440	797	498	610
# Sentence utterances in test	40	178	184	151	132	275	169	190
# Unique utterances in train	969	966	963	965	965	951	964	968
# Unique utterances in test	94	319	333	228	382	422	351	345
# Common utterances in train and test	94	316	327	224	378	404	346	344
# Different utterances in train and test	0	3	6	4	4	18	5	1
% Prompt overlap	100%	99.1%	98.2%	98.2%	98.9%	95.7%	98.6%	99.7%

Table 5.2: The number of utterances per (F)emale and (M)ale speaker of the leave-one-speaker-out models in TORGO. 'M/S': moderate to severe intelligibility.

Experimental Setup: As reviewed in Section 4.1 TORGO does not come with a pre-defined training and test partition. In previous studies, researchers used the leave-one-speaker-out approach to maximise the use of the available training data and trained speaker-independent models. In particular, for each split, 14 speakers were used for training and the 15^{th} held out speaker was used for testing. In this way, 15 models were trained and each of the 15 speakers was evaluated separately. The leave-one-speaker-out strategy is employed for training and testing in the baseline experiment in this research. Table 5.2 presents the number of utterances in the training and test sets of TORGO per speaker when applying leave-one-speaker-out. The acoustic models (GMM-HMMs or DNN-HMMs) are trained using both dysarthric and typical speech [Espana-Bonet and Fonollosa, 2016; Joy and Umesh, 2018; Mengistu and Rudzicz, 2011]. The optimal number of HMM states and Gaussians for the task is also explored. It is found that using 6000 HMM states and 6000 Gaussians yields the best result. The training data is augmented using speed perturbation with factors 0.9, 1.0 and 1.1.

5.2 Baseline Results Discussion

The performance of the ASR systems is usually measured by word error rate (WER). The WER is described as follows:

$$WER = \frac{Del + Ins + Subs}{N_{utt}} * 100\% \tag{5.1}$$

where *Del*, *Ins* and *Subs* are the number of deletions, insertions and substitutions, respectively. N_{utt} represents the total number of words in a reference utterance. WER will be used as the main evaluation measurement throughout the thesis.

5.2.1 Results on the Full Test Set

Before looking into the best choice of LM and splitting off the test set, it is worth looking at results in the usual setup first. The baseline results are reported as the WERs averaged for each severity group (i.e., severe, moderate and mild)[1]. The first row of Table 5.3 presents the results of the full test set. It shows that the DNN performs better than the GMM for speakers in all dysarthria severity levels. In particular, 12.0% (69.4% vs. 57.6%), 2.9% (35.9% vs. 33.0%) and 0.8% (15.1% vs. 14.3%) absolute WER has been reduced for the severe, moderate and mild groups by applying the DNN model compared with GMM. This demonstrates that the DNN has a good ability in learning latent information of dysarthric speech and modelling dysarthric speech by being than the GMM.

TORGO LM						
	Severe		Moderate		Mild	
Task	GMM	DNN	GMM	DNN	GMM	DNN
Full Test set	69.6	**57.6**	35.9	**33.0**	15.1	**14.3**
Isolated words	**79.8**	82.0	66.3	**65.5**	22.4	**19.5**
Sentences	62.0	**48.3**	23.3	**22.7**	**11.2**	12.2

Table 5.3: WER using different acoustic models and the TORGO LM for full, isolated words and sentences tasks, averaged for speakers with different dysarthria severity levels.

[1]It is noticed that the WERs for speaker M05 with moderate to severe dysarthria are exceptionally high. After further exploration by splitting the test set by microphone types, it was found that the WER of the head microphone subset is 14.60% higher than that of the array microphone subset, on account of the unexceptionally loud noise in the audio files recorded by the head microphone for speaker M05. Therefore, the results exclude the moderate-to-severe (i.e., speaker M05) group.

5.2.2 Results on Different Prompt Types

Considering the search space of recognising sentences is more complicated than recognising isolated words, the impact of LM on the isolated words and sentences should be different. In order to explore the difference, the results are reported on the word and sentence test sets separately in the second and third rows in Table 5.3. Significantly different performance is observed on the two prompt types for the GMM and DNN systems. And it is surprising that much better performance is achieved on the sentences than on the isolated words for each severity level when using the ***TORGO LM*** (i.e., a trigram LM built on all prompts in the TORGO training data similar to LMs used in Espana-Bonet and Fonollosa [2016]; Joy and Umesh [2018]), with a 26.0% higher performance on average.

When comparing GMM and DNN acoustic model (AM)s, the DNN provides varied performance gains across the tasks. In general, the higher the severity level, the more the DNN is able to improve the performance on the sentence subset (13.7%, 0.6% and -1% [1] reduced WER for the severe, moderate and mild group). The opposite effect is observed for the word subset. Note that for severely dysarthric speech, although the DNN decreases the overall WER achieved by the GMM on the full test set by 12%, it is evident that this overall decrease is the result of a (modest) increase in the word task (2.2%) and a large decrease for the sentence task (13.7%) when looking at the results separately on different prompt types. The DNN even has a negative influence (WER increases by 2.2%) on the word subset compared to the GMM. The task-specific (isolated words and sentences) results give a more nuanced picture of the performance. This inspires further exploration of the essence of various evaluation effects on the word and sentence tasks.

5.2.3 Discussion

Based on the observations in the baseline results, the data in TORGO is thoroughly analysed. Looking back to Table 5.2, although the corpus contains from 15,314 to 16,158 recorded utterances for training for each speaker, only a fraction of these (between 951

[1] The decreased performance on the sentence subset in the mild condition shows evidence of the existence of the bias caused by the in-domain LM. DNN is able to learn more useful acoustic information than the GMM and achieve better performance. However, the GMM acoustic model leads to a better result since it relies more on the LM.

and 969) are in fact unique. There is a significant overlap (as shown in the row of 'Prompt overlap' in Table 5.2) between any given speaker's utterances (in response to word and sentence prompts) and those seen in their training set (provided by the remaining 14 speakers). The high overlapping percentages demonstrate the high degree of repetition within and across speakers. This is sensible for assessment or across speaker comparisons, but not convenient for ASR. In fact, the standard approach of using a leave-one-speaker-out cross-validation setup with this dataset has encouraged previous researchers to train LMs on training sets that are almost completely overlapping with the test set. Therefore, the *TORGO LM* trained on any speaker's training data is highly tuned to the test set.

LMs impose a syntactic and semantic constraint on the ASR decoding process by assigning probabilistic estimates to the occurrence of short word sequences ('n-grams'). The LM is typically trained using large amounts of natural language text data [Tsujii, 2011]. When it comes to low resource data, care has to be taken to not unfairly design the LM so as to give over-optimistic results by training it on within-corpora data. Especially for dysarthric datasets, which have usually not been collected for the purpose of training ASR systems but instead for purposes such as diagnosis and impairment severity assessment, the prompts are largely overlapping across speakers. However, LMs in previous TORGO-based studies have typically been trained on the within-corpus training prompts (i.e., the *TORGO LM*), which means the LMs were unfairly designed – trained on non-disjoint training and test data. This has potentially produced misleading, unrealistically optimistic results for continuous speech recognition. This could be verified by the baseline results shown in Table 5.3 that the recognition results on sentences are much better than the isolated words. Researchers are faced with challenging choices when attempting to set up experimental frameworks aimed at facilitating meaningful research on improving continuous dysarthric speech recognition. When setting up an evaluation framework in an ASR system, it is essential that the chosen LM reflects a realistic scenario as best as possible. The small vocabulary size and the limited amount of continuous utterances limit the system search space and make the system less practical for real application. Training LMs on more text in a larger vocabulary using OOD data is one way of establishing an evaluation framework which allows for a more mearningful decoding space (in terms of

WER). Note this will evidently result in worse baseline performance. However, the result is more meaningful in terms of evaluating the success of acoustic modelling strategies in general, not just fitting the (non-ASR) database available for research.

5.3 Language Model Design

The baseline results (as shown in Subsection 5.2.1 and 5.2.2) present the weaknesses of the *TORGO LM* used in previous TORGO-based ASR studies. To explore the effect of different LMs, two *task-specific* TORGO LMs are trained for the word (*TORGO unigram LM*) and sentence (*TORGO trigram LM*) recognition tasks separately. This enables us to explore how the two distinct word and sentence recognition tasks are affected by the choice of the LM. Then a series OOD LMs originating from LibriSpeech are built to explore the optimal complexity of the LM.

5.3.1 In-domain Task-specific TORGO Language Models

The ***TORGO unigram LM*** for the isolated word utterances is built as a standard unigram LM, whereas the *TORGO trigram LM* is specific to the sentence utterances which is trigram. In particular, the *TORGO unigram LM* is constructed on TORGO's 615 unique isolated words. It restricts one-word output per utterance by a uniform word grammar network. The network contains silence models at the start and the end while all possible test words are in the middle. All words in the corpus are in parallel and they are assigned with the same log probability of -log(1/N), where N is the number of words. This follows the method used in Christensen et al. [2012b]. The grammar in Finite State Transducer (FST) format contains lines like the following:

$$0 \quad 1 \quad !SIL \quad !SIL \quad 0$$
$$1 \quad 2 \quad A \quad A \quad -log(1/N)$$
$$1 \quad 2 \quad ABBREVIATED \quad ABBREVIATED \quad -log(1/N)$$
$$1 \quad 2 \quad ABLE-BODIED \quad ABLE-BODIED \quad -log(1/N)$$
$$1 \quad 2 \quad ABLUTIONS \quad ABLUTIONS \quad -log(1/N)$$
$$1 \quad 2 \quad ABOUT \quad ABOUT \quad -log(1/N)$$

...

2 0

where the 1^{st} and 2^{nd} columns represent the start and the end grammar states, the 3^{rd} and 4^{th} columns represent the recognised words and the last column represents the log probability of the state transducer. This unigram LM could be regarded as a multi-class classification task where the number of classes is 615 (i.e., the number of unique isolated words in the dataset).

The ***TORGO trigram LM*** is built on 313 to 354 (depending on different speakers) unique training sentence prompts as defined by the speaker-specific TORGO training data split. A Witten-Bell discounting [Chen and Goodman, 1999] is applied to this TORGO trigram LM for smoothing. The task-specific TORGO LMs are less complex than the full TORGO LM due to the reduction of the training corpus size and no out-of-vocabulary (OOV) words or extra confusion in the system.

Table 5.4 shows the results of testing with the task-specific LMs, and Figure 5.1 presents the absolute performance gains (the reduced WER) comparing the results of the task-specific TORGO LMs in Table 5.4 and the results of the *TORGO LM* in Table 5.3 (the second and the third rows) by drawing the improvement lines for the isolated word and sentence recognition tasks. Not surprisingly, both task-specific TORGO LMs achieve better results than the *TORGO LM* evaluated on the corresponding utterance type subset. It is seen that as the dysarthria severity degree increases, the improvement made by the *TORGO unigram LM* on the isolated word task increases (3.66%, 17.32% and 19.23% respectively). In contrast, the opposite case occurs for the *TORGO trigram LM* recognising sentences. The consistent improvement across speakers on the words performance is caused by the constraint made by the unigram LM, which forces the ASR system to output a single word. It can also eliminate some of the insertion errors caused by the slow speaking rate characterised by the moderate and severe groups. The sentences performance of the mild speakers drops from 12.2% to a highly optimistic value (2.0%). This indicates that the constraint (e.g., reduction of training corpus) makes the trained *TORGO trigram LM* highly tuned to the test set by removing the isolated words from the training set. Therefore, the evaluation results are overly optimistic.

Task-specific TORGO LMs						
	Severe		Moderate		Mild	
LM	GMM	DNN	GMM	DNN	GMM	DNN
TORGO unigram LM	**61.5**	62.8	54.9	**48.2**	19.2	**15.9**
TORGO trigram LM	59.7	**41.8**	16.0	**12.8**	3.1	**2.0**

Table 5.4: WER using different AMs and the task-specific TORGO LMs for isolated words (*TORGO unigram LM)* and sentences (*TORGO trigram LM*) tasks, averaged for speakers with different dysarthria severity.

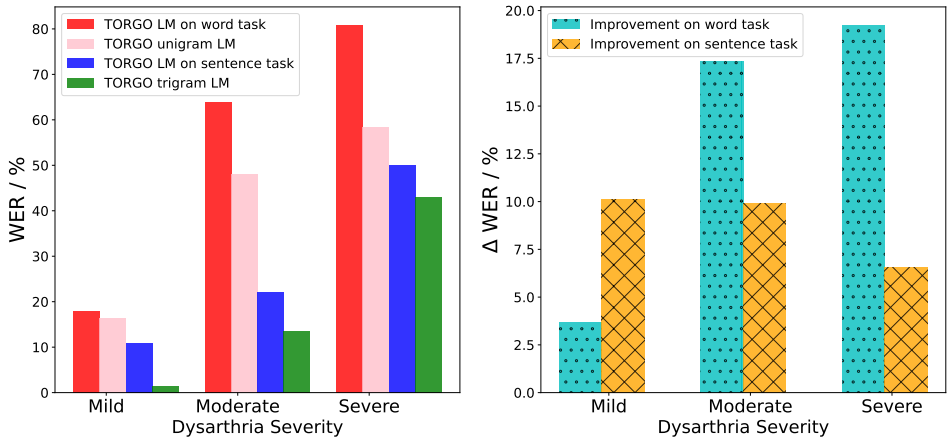


Figure 5.1: Comparison between the task-specific TORGO LMs and the full (both tasks) TORGO LM.

5.3.2 Out-of-domain LibriSpeech Language Models

To measure the impact of the biases introduced by the within-corpus TORGO LMs, the ASR performances to those obtained with LMs built from non-TORGO texts are compared. For this purpose, the LibriSpeech corpus [Panayotov et al., 2015] is introduced to as the OOD text corpus for LM training. It is a read speech dataset based on LibriVox's audiobooks, containing 1000 hours of speech sampled at 16 kHz. Around 803 million tokens from 14,500 public domain books and 900,000 unique words taken from Project Gutenberg books [Panayotov et al., 2015] are used for the LM training. The 200,000 most

frequent words are selected to be in the lexicon, and the Sequitur G2P toolkit [Bisani and Ney, 2008] is used to generate pronunciations for words not present in the CMU pronunciation dictionary [Kominek and Black, 2004]. The corpus has been made freely available for download, along with separately pre-built LMs which could be pruned into different LM sizes.

Two types of OOD LMs: *LibriSpeech unigram LMs* and *LibriSpeech trigram LMs* are used for the isolated word and sentence task, respectively. The ***LibriSpeech unigram LMs*** are built for isolated words over a range of vocabulary sizes: {2k, 5k, 10k, 15k, 20k, 25k, 30k, 35k, 40k, 45k, 50k, 100k, 150k, 200k}, by gradually extending the vocabulary originating from LibriSpeech in line with the decreasing order of word frequency. For instance, a 2k vocabulary size represents a vocabulary list containing the 2k most frequently occurring words in the LM. As vocabulary size increases, the LM complexity also increases. The optimal LM complexity for speakers is explored by using different vocabulary sizes.

The ***LibriSpeech trigram LMs*** are built for continuous dysarthric speech by pruning the pre-trained and pruned 3-gram LibriSpeech LM using the CHANGE-LM-VOCAB method in the *SRILM* toolkit. In particular, the CHANGE-LM-VOCAB method modifies the LM size by putting constraint on the vocabulary list. The method has three parameters: VOCAB (a list of vocabulary used in the LM), LM (a LM file, usually it is an arpa file fromat) and WRITE-LM (a new LM file). During the LM pruning process, the OOV words are converted to the <UNK> tag in the unigram while any N-grams containing OOV words are removed, and then the model is re-normalised. Compared with the isolated words, the sentences are more likely to cover the most frequent word such as pronouns. For this reason, the ***LibriSpeech trigram LMs*** are built over a range of vocabulary sizes starting from 0.1k for the sentence task. Compared with unigram LMs, additional smaller vocabulary sizes {0.1k, 0.2k, 0.5k, 1k, 1.5k} are introduced to ensure that the OOV rate is not too low at the beginning to mislead the result.

For the detailed neural network configurations, the reader is directed to the released Kaldi scripts[1].

[1]The Kaldi scripts for this work's experiments have been released at

5.3.3 Results and Discussion

In addition to WER, the OOV rate, Correct Rate (CorrR) and recognition confusion (Conf) are employed to measure the ASR systems. The CorrR refers to the proportion of the correctly recognised words in the reference utterance. The Conf is defined to measure how much confusion the system experiences when attempting to recognise words it is aware of, i.e., the in-vocabulary words. It could also be explained as the proportion of in-vocabulary words that have been wrongly recognised. The recognition confusion $Conf$ can be calculated as follows:

$$Conf = \frac{i - c}{i} = 1 - \frac{c}{i} \tag{5.2}$$

where c denotes the number of correctly recognised words, and i is the number of in-vocabulary words. It could also be written in the form that is related to CorrR and OOV rate:

$$Conf = 1 - \frac{\frac{c}{n}}{\frac{n-o}{n}} = 1 - \frac{CorrR}{1 - (OOV\,rate)} \tag{5.3}$$

where n denotes the number of words in the reference utterance, and o is the number of OOV words. It is notable in the uniform LM: $WER = 1 - CorrR$. This is because instead of three types of errors (i.e., insertion, deletion and substitution errors), only the substitution error is presented in the output. In this case, the recognition confusion $Conf$ can also be calculated as follows:

$$Conf = 1 - \frac{1 - WER}{1 - (OOV\,rate)} \tag{5.4}$$

In addition, to measure how well a LM predicts a token (a word or a sentence), the perplexity (ppl) [Katz, 1987] and ppl1 are applied. The ppl measures the geometric average of 1/(probability of each token), i.e., the perplexity. And ppl1 denotes the average perplexity per word excluding the </s> tokens [1]. The formulas of ppl and ppl1 are written

`https://github.com/zhengjunyue/CADSR-LM`

[1] </s> is an end-of-sentence (EOS) token which makes the n-gram grammar a true probability distribution [Jurafsky, 2000]. The </s> token prevents the probability of the whole language being infinite by limiting how long strings in a language can get. Only if the sentence ends in EOS, will the distribution over strings of any length P(EOS — ...) be high enough that the sentence is always guaranteed to stop after having generated a finite number of words.

as follows:

$$ppl = 10^{\frac{-logprob}{(words-OOVs+sentences)}} \tag{5.5}$$

$$ppl1 = 10^{\frac{-logprob}{(words-OOVs)}} \tag{5.6}$$

where the logprob refers to the log probability of a sentence, i.e., the log sum of probabilities for each n-gram in a sentence. The OOV words are silently ignored during the calculation. A low perplexity indicates the LM is good at predicting the sample.

The relationship between the impaired speech severity and the complexity of the LibriSpeech LMs is explored. The increasing vocabulary size indicates the increasing LM complexity, since more OOD words are introduced, resulting in more confusion in the search process in the ASR system. Figure 5.2a and 5.2b plot the WER, OOV rate and Conf for a range of vocabulary sizes for different severity groups (plotted in different colours). A base-10 log scale is used for the x-axis (vocabulary size). On each line, the lowest WER achieved by the LM with the specific vocabulary size is annotated with a coloured circle. The black line represents how the OOV rates vary by different vocabulary sizes. As the vocabulary size increases, more words are introduced to the LM, and the OOV rate gets smaller. For some of the speakers, the OOV rate will never become zero even though the LM reaches the largest size. This happens because some of the words in the TORGO prompts are OOV words of the CMU dictionary. The solid and dashed lines represent the DNN and GMM AMs, respectively.

Table 5.5 shows the results from these selected vocabulary sizes (indicated by 'optimal vocab size') for the LibriSpeech LM[1]. Comparing the results in Table 5.5 with the TORGO LM (Table 5.3) for the DNN AM on the isolated word task, the LibriSpeech unigram LM showed improvements across speakers with moderate and severe dysarthria. This might be because it reduces a large number of insertion errors resulting from the slow speaking rate, by constraining the output to be a single word. However, for mildly impaired speakers, since their speaking rate is similar to the typical speakers, although the LibriSpeech unigram LM constrains the output to make the task easier, it still degrades

[1]The lowest WER are used instead of the 'knee' of each WER line for results comparison.

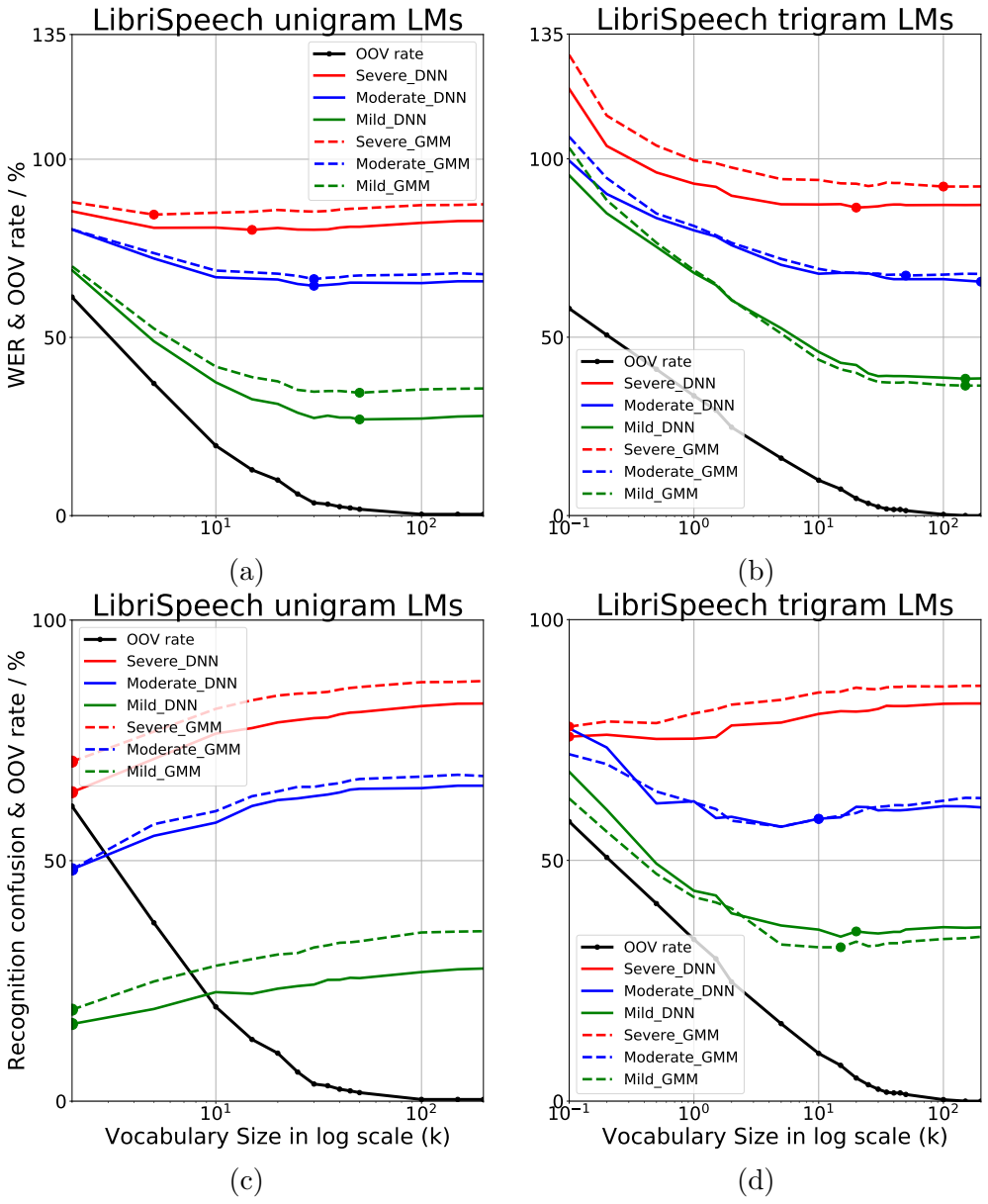


Figure 5.2: WER, recognition confusion and OOV rate for LibriSpeech LMs for speakers with different dysarthria severity levels.

the performance due to the reduced complexity. Comparing the sentence performances in Table 5.5 (86.4%, 65.6% and 38.4% WER) and those using TORGO LM (the last row of Table 5.3 (48.3%, 22.7% and 12.2% WER)), the WER obtained by the LibriSpeech trigram LMs are on average relatively 40.5% worse for moderate and severe speakers and even 26.2% for mild speakers. In contrast to the unrealistically small WERs of the TORGO LM, these results present a fairer evaluation.

LibriSpeech unigram LMs; isolated word task						
	Severe		Moderate		Mild	
	GMM	DNN	GMM	DNN	GMM	DNN
The lowest WER (%)	84.5	**80.2**	66.4	**64.5**	34.5	**27.0**
Optimal vocab size	5k	15k	30k	30k	50k	50k
LibriSpeech trigram LMs; sentences task						
	Severe		Moderate		Mild	
	GMM	DNN	GMM	DNN	GMM	DNN
The lowest WER (%)	92.3	**86.4**	67.3	**65.6**	**36.4**	38.4
Optimal vocab size	100k	20k	50k	200k	150k	150k

Table 5.5: WER using different AMs and the OOD LibriSpeech LMs for isolated words (*LibriSpeech unigram LM*) and sentences (*LibriSpeech trigram LM*) tasks, averaged for speakers with different dysarthria severity levels.

It is seen that, in general, speakers with dysarthria at different levels of severity require the LibriSpeech LMs with different vocabulary sizes: the higher the severity level, the smaller the optimal vocabulary size. To explain the possible reasons, the Conf rate is plotted across speakers with different degrees of dysarthria in Figure 5.2c and 5.2d. It is seen that there is more confusability in the speech as the severity level is higher. Therefore, reducing the vocabulary size reduces the chance of poorly pronounced common words being mistaken for low-frequency words that might be better acoustic matches. Typically, as for the word recognition task, as the vocabulary size increases, the confusion sees a monotonic increase across all the speakers. While in the sentence recognition task, the confusion rates reach the minimum point with 0.1k, 10k and 20k vocabulary sizes individually for speakers with severe, moderate and mild dysarthria. This might be because the continually reducing OOV rate and the increasing number of utterances

available, offset the extra confusions (i.e., some of the extended words are in a recognisable range to reduce some substitution errors caused by OOV words). The greater the severity of dysarthria, the less compensation is made by the decreasing OOV rate. Comparing different AMs, when further increasing the vocabulary size after the optimal vocabulary sizes required by the LibriSpeech LMs, the recognition confusion of the GMM systems will increase more than that of the DNN models.

Figure 5.3 further explores the effect of different AMs on the best LM size for each degree of dysarthria, where the best performance for each AM (GMM and DNN) of certain LM size is marked with a red circle and purple triangle, respectively, onto the OOV rate line as well. Considering the large inter-speaker variability characterised by dysarthric speech, three representative speakers at each level (*severe*, *moderate* and *mild*) are selected to investigate the effect of the LibriSpeech unigram LM and LibriSpeech trigram LMs. When looking at the plots in each column, besides the conclusion made above, it is seen that for the severe speaker F01, DNN requires a larger vocabulary size for both OOD LMs that performs best combined with the GMM AM, which results in dramatically decreased OOV rate to achieve the best performance, where the Conf line starts to be flat. Although, for speakers with less severe dysarthria, DNN seems to influence also the best LM size, the effect is small both in terms of OOV rate and Conf rate. This is presented in Table 5.5 where the best performance of the OOD LMs are denoted with "Min_WER", and also the vocabulary size for the LM that achieve the best WER, indicated by "Min_Size" in the table, is shown.

Especially for the LibriSpeech trigram LMs used for testing the phrase-based speech data, the CorrR lines are also plotted in Figure 5.3. As shown in Figure 5.4, as the vocabulary size increases, the perplexities of all speakers increase monotonically. However, the result is unlikely to be monotonously worse, and the correctness does not drop much after reaching an optimal value. Thus, the optimal LM complexity is driven by several factors, for instance, dysarthria severity which means highly speaker-dependent and AM quality. Although the result obtained by the LibriSpeech trigram LMs is worse than the TORGO LM, which is on average 37.23% for moderate and severe speakers and even 24.28% for mildly dysarthric speakers, unlike to get the unrealistic small perplexity and

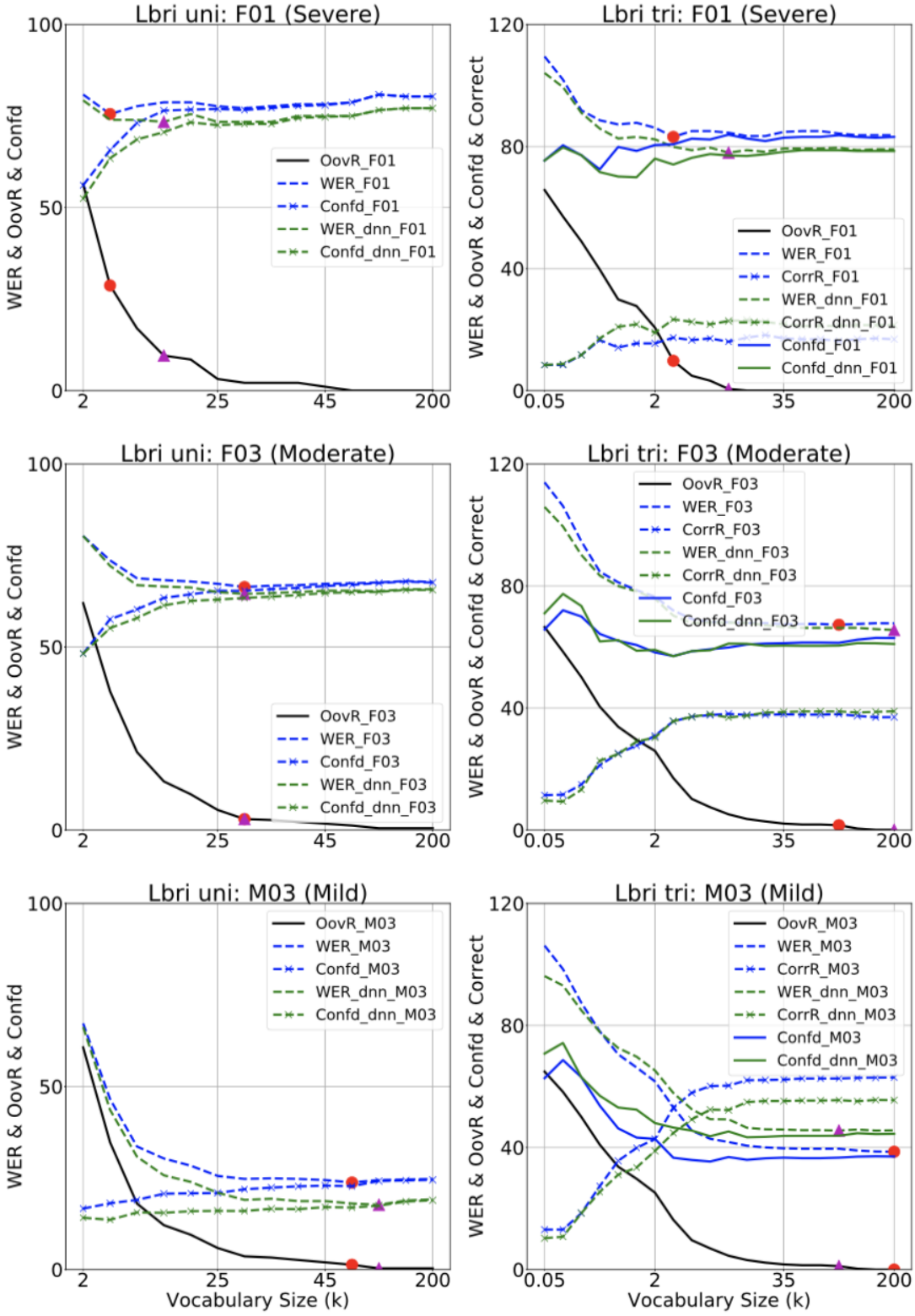


Figure 5.3: Results of LibriSpeech LMs; see text for further details.

WER, it guarantees better fairness and could be further explored for better performance in the future.

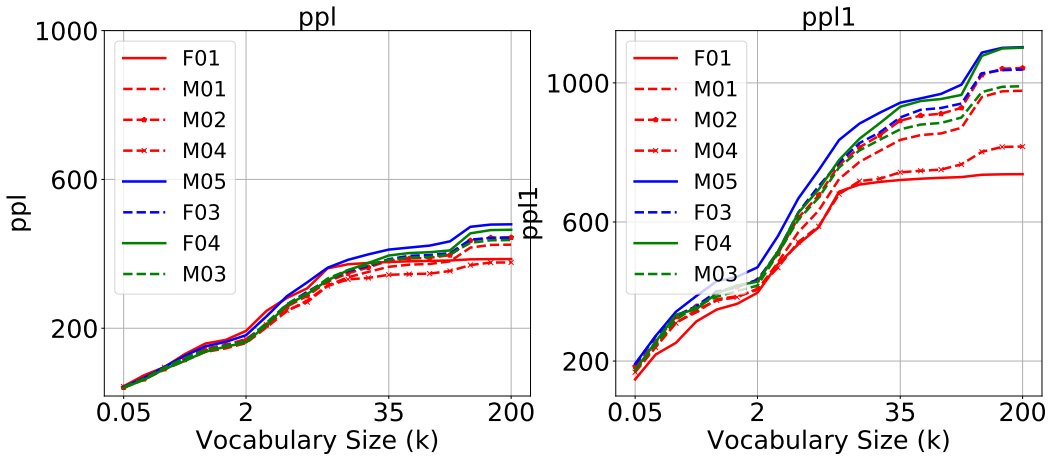


Figure 5.4: Perplexity vs. vocabulary size for LibriSpeech trigram LMs.

5.4 Conclusion

Starting with a pilot ASR study to build a baseline on TORGO, this chapter presented an in-depth analysis comparing LMs trained on TORGO text prompts with LMs trained on varying vocabulary-sized subsets of LibriSpeech. It was found that the TORGO LMs (used widely in literature) give a hugely overestimated performance of ADSR because of prompt overlap between training and test parts. In comparison, the LibriSpeech models offer a lower but fairer performance which will better allow for a more meaningful decoding space (in terms of WER). Exploring different vocabulary sizes for the LibriSpeech LMs, it was found that in general, the lowest WERs are achieved with the largest vocabulary size. The greater the severity, the less complex the LM is required to have for the best results. In real applications, speaker-specific LMs may be appropriate as, depending on the severity and *when not asked to read prompts*, speakers would choose to use different language constructs and words to counteract specific speech impairments. The baseline with OOD LibriSpeech LMs provides a solid and fair benchmark for continuous dysarthric speech recognition with an appropriate evaluation framework. Chapter 6 will build more robust ASR systems for continuous dysarthric speech on top of the baseline in this chapter,

exploring a novel speech representation learning framework and advanced acoustic models.

Chapter 6

A Novel Speech Representation Learning Framework

Contents

6.1 Introduction

In Chapter 5, a continuous dysarthric speech recognition baseline framework was built using the TORGO corpus which employed out-of-domain (OOD) language models. This work provides a fair evaluation approach for automatic continuous dysarthric speech recognition (ACDSR) task which will be used in the following experiments. As discussed in the previous chapters, data scarcity is a major issue in the automatic dysarthric speech recognition (ADSR) task. There are even less continuous dysarthric speech data available. As a result, the amount of dysarthric data is usually insufficient to train robust continuous speech systems using conventional approaches. Exploiting OOD data is a good way to address this issue [Christensen et al., 2013; Xiong et al., 2020; Yılmaz et al., 2019].

The feature extraction process is an essential part of an automatic speech recognition (ASR) system. The performance of the feature extraction stage underpins the performance of the ASR system. No matter how well-designed the acoustic and language models are, the ASR system cannot perform well if the features do not capture the useful information in the signal. Given that the acoustics of dysarthric speech are highly variable and with low intelligibility, it is difficult to capture the robust acoustic cues by using the hand-crafted features that carry less useful information. With the popularity of deep learning approaches, there is a growing interest in applying deep learning methods for speech representation learning. Due to the deep architecture, neural networks can learn richer representations than the hand-crafted features or learn complementary information to the hand-crafted features which may also be applicable to dysarthric speech representation learning. However, a large amount of data is required for neural network training, and the performance is constrained due to the lack of dysarthric data.

Previous studies have demonstrated the effectiveness of employing speech representations such as bottleneck (BN) features [Takashima et al., 2015; Yılmaz et al., 2019] to support acoustic features for improving acoustic modelling of dysarthric speech. The BN features are extracted from a neural network bottleneck layer trained with a supervised criterion such as phoneme prediction accuracy [Grezl and Fousek, 2008]. The BN features have been shown to capture complementary information for dysarthric speech that can

be beneficially fused with standard short-time spectral input features [Takashima et al., 2015; Yılmaz et al., 2019].

Recently, there has been growing interest in autoencoder-based bottleneck features (e.g., autoencoder bottleneck (AE-BN) features) [Chorowski et al., 2019; Sainath et al., 2012]. In contrast to the conventional BN features, AE-BN features are learnt by reconstructing the input features in an unsupervised manner [Sainath et al., 2012]. Unsupervised learning has the advantage that neither the transcription of the training data nor a linguistic pronunciation lexicon is required. It can be very useful when there are a large of unlabeled dysarthric speech available for training. Autoencoder is also able to exploit the local dependencies in the sequential data [Chorowski et al., 2019]. For this reason, it has the potential to improve continuous speech recognition performance since the power to capture local dependencies is important when modelling continuous speech. The autoencoder (AE)-based models have been used in the context of isolated-word ADSR. For instance, in Bhat et al. [2018]; Vachhani et al. [2017], the AEs were applied for dysarthric speech feature enhancement by learning non-linear mappings from the dysarthric speech to the typical speech. The enhanced dysarthric features tended to be like typical speech, and were used for recognition obtaining higher accuracy. This AE-based feature enhancement approach is limited to corpora with parallel recordings for both typical and dysarthric speech. To make the approach applicable to a wider range of datasets and tasks, the AE-BN features are proposed to be applied which are extracted using the reconstruction objective driven by the same input and output in this chapter.

Given the data scarcity issue, the OOD typical-data pretraining framework can increase the robustness of the dysarthric speech representation learning process when little in-domain training data is available. The pretraining framework which applied OOD typical speech data at the feature extraction stage was introduced in Christensen et al. [2013] for tandem features learning. Various BN feature extractors using both typical and dysarthric data were investigated in Yılmaz et al. [2019]. It was concluded that using a large amount of OOD typical speech data to train the BN feature extractor achieved the best recognition performance on dysarthric speech.

This chapter explores the second research question **RQ2**: What is a good way to lever-

age typical speech, which is OOD, to learn more robust representations for continuous dysarthric speech? Motivated by the above previous studies, the work proposes a novel speech representation learning framework for ACDSR: an AE-BN feature extractor making use of the OOD knowledge with multi-task optimisation techniques. In particular, the feature extractor is built using an AE-BN architecture pretrained on OOD typical speech data to increase the robustness of the dysarthric speech representation learning process. In addition, to accommodate for the possible drawback of the unsupervised AE-BN feature learning approach, the framework allows for jointly optimising the AE-BN feature extractor and the speech recogniser. This enables the speech recogniser to engage in and influence the feature extraction process. The extracted speech representations, therefore, can benefit the phoneme classification specifically. Monophone regularisation is applied as a multi-task learning strategy to provide further improvement.

6.2 System Overview

This section gives a system overview including the system architecture and the description of several main components in the system.

6.2.1 System Architecture

Figure 6.1 depicts the architecture of the proposed ACDSR system. The red box on the left shows the AE-BN feature extractor and the blue box on the right represents the speech recogniser. In particular, the AE-BN feature extractor is first trained on the 100-hour subset of LibriSpeech [Panayotov et al., 2015] corpus, which is a large typical read speech dataset. The pretrained feature extractor is then fine-tuned using dysarthric data in TORGO, and the extracted dysarthric AE-BN features are concatenated with the input acoustic features and fed into the speech recogniser.

6.2.2 Autoencoder Bottleneck Feature Extractor

Autoencoder is a type of neural network which learns an efficient data representation (encoding) in an unsupervised manner [Liou et al., 2014]. Dimensionality reduction is

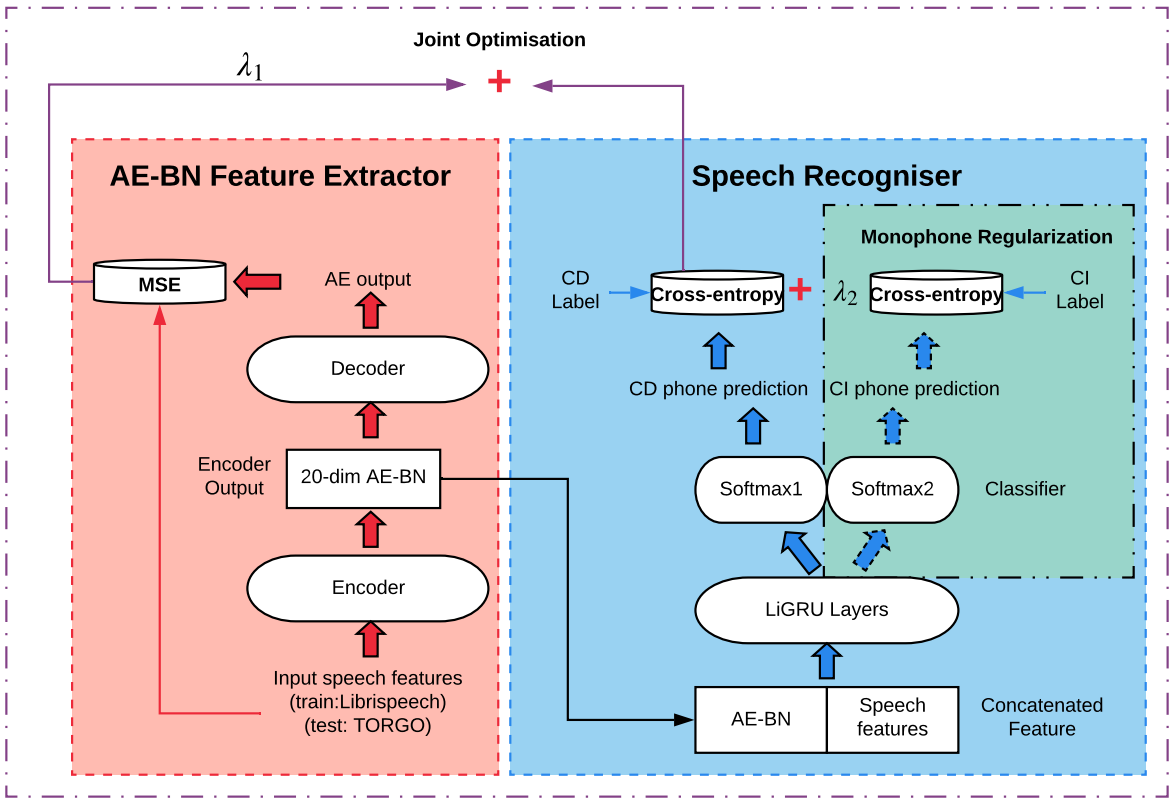


Figure 6.1: System architecture.

conducted by AE. It consists of two parts: an encoder and a decoder. The encoder encodes the high dimensional input feature vectors into lower-dimensional latent variables (in the following called AE-BN features). The decoder reconstructs the original input (as close as possible to its original input) from the generated latent variables. This two processes could be defined as transitions ϕ and ψ:

$$\begin{aligned} &\phi : X \to F \\ &\psi : F \to X \\ &L(X, (\phi \cdot \psi)X) = ||X - (\phi \cdot \psi)X||^2 \\ &\phi, \psi = \arg\min L(X, (\phi \cdot \psi)X) \end{aligned} \tag{6.1}$$

where X represents the input vectors and F represents the latent variables (the AE-BN features). AEs are trained to minimise reconstruction errors $L(X, (\phi \cdot \psi)X)$ between X and $(\phi \cdot \psi)X$. The reconstruction error is typically the mean square error calculated

between the reconstructed feature vector y_i and the true input feature vector x_i:

$$Loss_{AE} = (\frac{1}{n}) \sum_{i=1}^{n} (y_i - x_i)^2 \tag{6.2}$$

where n is the number of frames.

The AE-BN feature is driven by two opposing constraints: i) the reconstruction objective which forces the AE-BN feature to capture as much of the input data characteristics as possible, and ii) the bottleneck (i.e., the dimension reduction) which forces the network to discard the redundant information that is not needed for the inversion. AE-BN features are expected to capture complementary information for dysarthric speech to the acoustic features. However, redundant information might also be captured by the unsupervised feature learning approach, i.e., information that is needed for signal reconstruction but not important for phoneme classification (e.g., speaker variability, pitch). Designing effective regularisation (described in Section 6.2.4 and Section 6.2.5) techniques either for the feature extractor or the acoustic model (AM) has the potential to compensate for the deficiency.

6.2.3 Light Gated Recurrent Units Acoustic Model

The speech recogniser uses light gated recurrent unit (LiGRU) as the AM. The performance improvements on ADSR have been made by exploring various deep learning architectures such as convolutional neural network (CNN), time-delay neural network (TDNN) and long short-term memory (LSTM) [Espana-Bonet and Fonollosa, 2016; Hermann et al., 2020; Kim et al., 2018; Mengistu and Rudzicz, 2011] in the past few years. Recently, the LiGRU [Ravanelli et al., 2018] have been shown to outperform existing architecture on large typical speech datasets such as LibriSpeech and TIMIT [Zue et al., 1990]. Recurrent neural networks are effective tools to process sequential data such as speech [Li and Wu, 2015]. As an advanced recurrent neural network (RNN), the LiGRU model has the capability to exploit large time contexts and to capture long-term speech modulations. Compared with LSTMs Hochreiter and Schmidhuber [1997], LiGRUs have a simpler cell design that allows for faster training. The design also avoids the numerical issue of learn-

ing long-term dependencies and mitigates the vanishing gradient problem.

The LiGRU model has not been used for ACDSR. To demonstrate the effectiveness of using the LiGRU model, a test is done by only replacing the AMs with the LiGRU-based model while keeping the same experimental settings as in Espana-Bonet and Fonollosa [2016] and Chapter 5. Evaluating on the sentence subset of TORGO, it is found that the performance achieved by the LiGRU AM in is comparable to other AMs presented in previous papers. For instance, the TDNN model achieves 70.72% word error rate (WER) averaged across all speakers, while LiGRU achieves 71.08%. For speakers with severe dysarthria, the LiGRU model performs even better (83.90% vs 86.40%). These comparable results are achieved in Pytorch-Kaldi [Ravanelli et al., 2019] without the benefits of the (computationally expensive) lattice-free maximum mutual information training used in the systems using the Kaldi [Povey et al., 2011] toolkit. The LiGRU AM will be used in the remainder of this thesis.

6.2.4 Joint Optimisation

The feature extractor and speech recogniser are often designed independently in the previous studies. This means that the feature extractor is tuned according to criteria not directly related to ASR performance. Recently, deep neural network (DNN)s have made the integration of various components of a typical ASR system possible. In Ravanelli et al. [2016] a DNN-based integrated network for distant speech recognition was proposed that combined speech enhancement and speech recognition modules allowing for the joint updating of parameters. It was shown that the joint training achieves better results than training each part separately. It was also demonstrated that a pretraining strategy with a fine-tuning phase improves performance. The core idea of joint training is that the feature extractor should provide more discriminative representations for the ASR task as it is in part guided by the speech recognition cost function [Ravanelli et al., 2016]. In this case, the speech recognition gradient is also back-propagated through the feature extraction module.

In this chapter, in addition to training the AE-BN feature extractor and the speech recogniser separately, an integrated framework where these two parts are jointly optimised

is proposed. The feature extractor is pretrained on LibriSpeech data and fine-tuned using TORGO dysarthric data. The recently proposed PyTorch-Kaldi framework provides a platform to implement the joint optimisation which can be difficult to perform in Kaldi. The parameters are updated by back-propagating a weighted sum of the AE reconstruction loss and the cross-entropy loss,

$$Loss_{Joint} = \lambda_1 * Loss_{AE} + Loss_{ASR} \qquad \lambda_1 \in (0, 1) \tag{6.3}$$

where λ_1 controls the trade-off between the reconstruction quality of the feature extractor and the effectiveness of the speech recogniser.

6.2.5 Monophone Regularisation

To train a more robust AM, the multi-task learning (MTL) technique has been applied to hybrid DNN systems in Bell et al. [2016] by added a secondary task of predicting alternative ccontext-dependent (CD) (i.e., triphone) or context-independent (CI) (i.e., monophone) targets. Consistent improvements were achieved over the standard single target training approach on large-vocabulary typical speech recognition tasks. Importantly, this strategy does not require additional data. This makes it suitable for the low-resource data domain, for instance, dysarthric speech. This MTL scheme can be regarded as a technique to regularise the AM. The regularisation prevents the AM from over-fitting to a single senone target classification by learning additional CI or CD labels. This encourages a better presentation of the data to be learnt by the AM (and by extension, by the auto-encoder when joint optimisation is engaged).

In this work, multi-task regularisation is applied to the AM. Particularly, monophone classification is used as a secondary task by adding another softmax classifier to estimate the CI states. The joint optimisation cost function becomes the sum of the $Loss_{CD}$ and the cross-entropy loss $Loss_{CI}$ between the true CI labels and the predictions:

$$Loss_{ASR} = Loss_{CD} + \lambda_2 * Loss_{CI} \tag{6.4}$$

where λ_2 indicates the weighting between each task's loss.

6.3 Experiments and Results

6.3.1 Experimental Setup

The AE-BN feature extractor used in this work consists of a four-layer encoder and a two-layer decoder. The encoder contains two convolutional layers to learn rich local representations and two multi-layer perceptron (MLP) layers to flatten the feature vectors and encode the high dimensional features into a lower-dimensional representation. The decoder comprises two MLP layers fed by the learned AE-BN features and aims to produce an output matching the original input.

The LiGRU-based AM follows the design from [Ravanelli et al., 2018], containing five stacked bidirectional LSTM layers [Graves et al., 2013] and a final softmax classifier. Recurrent dropout (0.15) is used as a regularisation technique. The minibatch sizes are 128 and 16 for the AE-BN feature extractor and the AM, respectively. Stochastic gradient descent optimisation is used in the feature extractor and RMSProp in the LiGRU model. Learning-rate annealing is applied with a factor of 0.5. The 200k vocabulary size trigram language model originating from the OOD LibriSpeech data is used for evaluation as in Chapter 5.

6.3.2 The Training Setup for the TORGO Corpus

TORGO does not come with a pre-defined training and test partition. An *N-fold cross-training* (N=5) setup is applied, with the total dataset (including all speakers) being divided into five folds (i.e., one fifth of each speaker in every fold)[1]. According to Section 5.2.3, TORGO features a lot of repeated prompts across speakers. The *N-fold cross-training* maximises the available training and test data while maintaining the need for disjoint training and test sets. Table 6.1 summarises the duration of the recordings in each fold (after excluding the recordings that are shorter than 25 ms and any wrongly annotated audio). The ratio of the duration of the two utterance type subsets (isolated word vs. sentence) is about 1.5:1.1.

[1]The pre-defined training and test partition set is available at `https://github.com/zhengjunyue/bntg`.

Table 6.1: Duration (hours) of the training and test data in each fold using the 5-fold cross-training setup

subset	fold 1	fold 2	fold 3	fold 4	fold 5
train_all	10.71	10.69	10.71	10.83	10.57
train_sentence	4.63	4.54	4.60	4.71	4.59
train_word	6.10	6.15	6.11	6.12	6.16
test_all	2.71	2.73	2.72	2.59	2.67
test_sentence	1.14	1.22	1.17	1.06	1.18
test_word	1.57	1.51	1.55	1.53	1.49

Most of the previous TORGO-based work used the leave-one-speaker-out (LOSO) approach to train speaker-independent (SI) models [Espana-Bonet and Fonollosa, 2016; Hermann et al., 2020; Mengistu and Rudzicz, 2011]. Most of them reported results averaged for speakers at different dysarthria severity levels. However, when looking at the results for individual speakers, it is found the performance for each speaker varies a lot, even for those at the same severity level of dysarthria. The LOSO approach trains different AMs for each speaker with a different amount of data. This makes the trained models not comparable. The unbalanced data yields greatly varying recognition results even for speakers within the same dysarthria severity. In addition, with only eight speakers, there are insufficient speakers in TORGO to capture the wide inter-speaker variability observed in dysarthria. In a LOSO SI setting, speaker performances will be more determined by the chance degree of matched-ness of the target speaker to the few others in the training set, i.e., rather than to any intrinsic difficulty of the speech itself.

The previously published work using UASpeech [Kim et al., 2008a] which employed 2:1 disjoint training and test partition scheme for isolated word recognition provides a good inspiration to the 5-fold cross-training setting. The UASpeech-based work split the data of each speaker into three disjoint blocks, and each block consists of non-repeated 245 words. Blocks 1 and 3 are used for training and block 2 for the test. TORGO does not have enough unique utterances as in UASpeech, so the dataset is split manually into five folds. The *N-fold cross-training* approach ensures a good trade-off between having

a reasonably large training set while providing some matched speaker training data to allow for a more meaningful comparison of recognition performance across speakers.

6.3.3 Results

Table 6.2: WER using different speech representations and AMs for per (F)emale or (M)ale speaker with dysarthria at different severity levels, and the averaged result of all speakers 'M/S': moderate to severe level of dysarthria.

	Severe				M/S	Moderate	Mild		
Models	F01	M01	M02	M04	M05	F03	F04	M03	**Average**
MFCC	77.93	77.91	76.17	91.66	85.46	51.47	22.27	22.04	59.22
fMLLR	73.86	76.36	73.12	88.66	83.74	49.18	21.71	21.69	57.33
fMLLR+BN20	69.84	71.55	72.26	85.97	78.9	47.06	19.75	19.86	54.70
fMLLR+BN20 + mono	71.47	69.3	70.88	**79.91**	**77.18**	**44.21**	**18.26**	**18.23**	**52.37**
fMLLR+BN20 + joint	**69.29**	70.54	71.65	83.37	80.4	47.74	19.5	19.65	54.05
fMLLR+BN20 + mono + joint	70.65	**69.07**	**70.81**	81.82	78.4	45.18	18.42	19.15	52.99
fMLLR+BN20 (TORGO)	75.22	76.91	73.65	89.03	83.55	48.54	20.96	20.03	57.83

Baseline result: Results are shown in Table 6.2. The 1^{st} row displays the baseline system using the LiGRU AM trained on 39-D Mel-frequency cepstrum (MFCC) feature and without using the AE-BN feature extractor. Compared with the results in Yue et al. [2020b], the baseline achieves consistent improvement on ASR task for speakers with all dysarthria severity levels. The MFCC features are then substituted with the 40-D feature-space MLLR (fMLLR) features (the 2^{nd} row). It is seen that fMLLR features outperform the baseline MFCCs, reducing WER by 3% for speakers with moderate and severe dysarthria. Therefore, fMLLR features are used as the input in the following experiments.

AE-BN feature result: The AE-BN feature extractor is pretrained on the 100-hour OOD LibriSpeech data for a more generalised model than trained on the task-specific TORGO dysarthric data. The latter case has bad results since the TORGO dysarthric dataset is too small to train a suitable feature extractor. When introducing the AE-BN feature extractor, since the recognition loss depends on the width of the bottleneck, the optimal dimensionality of the AE-BN features is explored. It is found that 20 is the best

dimensionality for this task, with results reported in the 3^{rd} row in Table 6.2. Introducing the AE-BN features reduced WER by a further 1.77% to 4.84% absolute.

Monophone regularisation result: Further improvements are made by applying multi-task optimisation techniques. λ_2 is set to be 1 as the previous work did. Comparing the 3^{rd} and 4^{th} rows in Table 6.2, the AM regulariser successfully reduces WER by an absolute 2.33% across speakers. For speakers with severe dysarthria, the WERs are reduced by from 1.83% to 6.06% with the exception of speaker F01 (where the WER is even higher). For speakers with moderate dysarthria, there is a 2.85% recognition performance improvement. This indicates that a single set of triphone targets is not optimal for the discriminative clustering process (i.e., phoneme classification). The additional CI label learning step strengthens the dysarthric AM.

Joint optimisation of AE and ASR result: Since users can employ their own features with PyTorch-Kaldi, it is possible to train a cascade between a speech representation extractor and a speech recogniser. The extractor aims to generate BN features which are then concatenated with the original inputs to be fed into the ASR model to predict CD phone states. In this case, the LibriSpeech AE parameters are retrained using TORGO dysarthric data, and the AE-BN feature extractor and the speech recogniser are jointly trained by back-propagating the sum of the AE loss (i.e., mean square error) and the cross-entropy ASR prediction loss. When tuning the jointly optimised model, different values of λ_1 (Eq. 6.3) ranging from 0.1 to 1 are tested with 0.2 producing the best ASR performance. The λ_1 tuning results are presented in Table 6.3 averaged for all speakers with dysarthria. Comparing the 3^{rd} and the 5^{th} rows in Table 6.2, the joint optimisation technique achieves a WER reduction of 0.65% absolute compared to the model that trains the feature extractor and AM separately.

Table 6.3: The averaged WER for speakers with dysarthria when using different λ_1s.

λ_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
WER (%)	56.61	54.05	55.73	57.67	59.13	60.20	65.33	68.59	74.56	89.76

The "BN20+fMLLR + mono + joint" system in the last row in Table 6.2 applies

the joint optimisation technique to the AM with monophone regularisation. Comparing the last three rows shows that the monophone regularisation technique provides a further improvement on the joint optimisation model and vice versa except for some speakers with severe dysarthria. Almost all the benefits seen in the last row are coming from monophone regularisation, therefore it appears that the joint optimisation provides no significant benefit when coupled with a sufficiently strong AM. The possible reason is that the joint training is actually performed as a fine-tuning procedure, and the hyperparameters such as learning rate need to be selected properly to take advantage of the pretraining. Although the joint optimisation does not provide the benefits expected, it remains an under-explored research direction deserving of further investigation. The overall best result (52.37% WER) is obtained when employing monophone regularisation alone.

6.3.4 Discussion

Effect of Utterance Type: The results show that achieving an acceptable performance for a continuous dysarthric speech recogniser remains challenging. This is exacerbated by the fact that some speakers with dysarthria produce many repetitions and false starts when having to speak in full sentences. Figure 6.2 illustrates WERs for not just the TORGO sentence task, but also for the isolated word task and the full, combined test set across all speakers. In general, and as expected, the sentence task is harder for all speakers; and for some speakers (e.g., M04 and M05), the sentence performances are much worse. Inspection of the audio confirmed that the ASR transcription had many insertions caused by disfluencies typical for speakers with dysarthria.

Effect of Microphone Type: The acoustic data in TORGO is simultaneously recorded by a head-mounted and a single directional microphone (in the following called array microphone). It is interesting to explore whether the microphone type affects the performance. It is observed that the amount of data recorded by different microphone types (head vs array) per speaker per session is different as shown in Table 6.4. Some data was removed from the original dataset because of the severe Gaussian acoustic noise caused by the electric field when the electromagnetic midsagittal articulography (EMA) interfered with the microphones. To explore the effect of the channel type, the results for

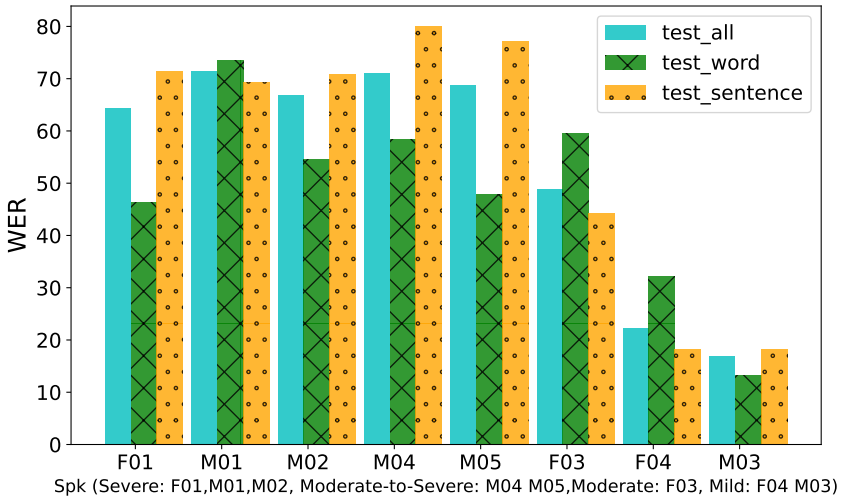


Figure 6.2: WER for different utterance subsets using the proposed "fMLLR+BN20 + mono" model (modified).

different channel subsets are presented in Figure 6.3.

Table 6.4: Number of utterances recorded by array and head microphones per dysarthric speaker per Session. 's': Session.

		Severe				**M/S**	**Moderate**	**Mild**	
Microphones		F01	M01	M02	M04	M05	F03	F04	M03
array	total	134	386	400	520	-	848	448	416
	s1	134	100	240	126	-	204	199	-
	s2	-	286	160	294	-	435	249	416
	s3	-	-	-	-	-	209	-	-
head	total	132	386	409	298	523	577	250	421
	s1	132	100	240	-	130	204	-	-
	s2	-	286	169	298	393	159	250	421
	s3	-	-	-	-	-	214	-	-

Figure 6.3 indicates that the performance on the word subset of the two channels varies a lot[1]. Notably, for speakers M01 and F03, the performance on utterances recorded by the array microphone is approximately 40% worse than the head microphones. This

[1]Array microphone recordings of speaker M05 are removed from the dataset since they are distorted.

might be because the head microphone is more sensitive to the electric field generated by the electromagnetic articulograph. For the sentence subset, there is no significant difference between the performance of the two channels. However, it is notable that for speaker F03, the array channel result is better than the head channel on word utterances while the opposite case happens on sentence utterances. The potential reason might be the different amount of utterances recorded by these two channels for different utterance types.

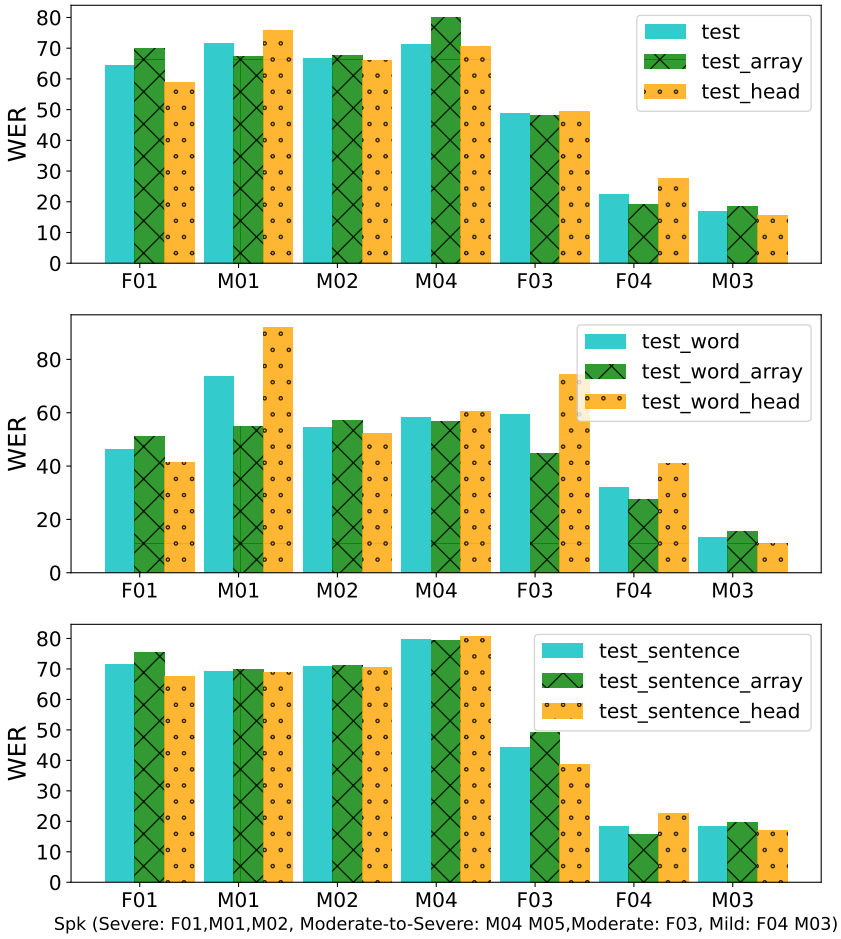


Figure 6.3: WER for different utterance subsets using the "fMLLR+BN20 + mono" system.

6.4 Conclusion

This chapter has explored a proposed novel speech representation learning framework for ACDSR, including a pretrained and fine-tuned AE-BN feature extractor and multi-task optimisation techniques. One of the current advanced LiGRU architecture is used for the AM, and the system is evaluated with the fairly designed OOD language model proposed in Chapter 5. The results demonstrated the effectiveness of augmenting conventional acoustic features with the extracted AE-BN features, reducing WERs by 2.63% absolute on average compared with the MFCC baseline. More WER reduction was achieved on higher dysarthria severity levels. The multi-task optimisation techniques: monophone regularisation and joint optimisation made further recognition performance improvements by reducing 2.33% and 0.65% absolute WER. However, no consistent additional benefit was found by using the joint optimisation technique when applied in conjunction with monophone regularisation. Overall, the proposed framework learns useful speech representation for the phoneme classification task on continuous dysarthric speech and demonstrates a way of exploiting OOD data for speech representation learning. Besides acoustic information, the additional information source of speech production may carry complementary information for dysarthric speech clues. The following two chapters Chapter 7 and Chapter 8 will systematically explore the articulatory data using TORGO and exploit it to build robust ACDSR systems.

Chapter 7

Incorporating Articulatory Information

Contents

7.1 Introduction

The previous chapters present the work using acoustic information for automatic continuous dysarthric speech recognition (ACDSR). Given that the acoustics of dysarthric speech are highly variable, a typical phonetic token can be pronounced differently. There are often no robust acoustic cues for a specific phoneme. The single acoustic modality might not be a good solution for acoustic modelling trained on a limited amount of data. The multimodal automatic speech recognition (ASR) utilises the data from other modalities to facilitate the task when it is insufficient using the single data modality. Attempts have been made to harness alternative or additional sources of knowledge captured during speech production. One such additional source of information is the articulatory information, which captures the movements of speakers' articulators (e.g., lips and tongue). The positioning of the articulators plays an important role in human speech production. Compared with acoustic representations, the articulatory information directly models the signal in the speech production domain. It has been shown to be more noise-robust [Wrench and Richmond, 2000], less speaker-variant [Fujimura, 1986] and more suitable to model the coarticulation variability [Frankel and King, 2001; Kirchhoff et al., 2002]. Articulatory information may therefore hold complementary information that could be exploited by automatic dysarthric speech recognition (ADSR) and ACDSR systems.

The articulatory data can be acquired from electromagnetic midsagittal articulography (EMA) [Schönle et al., 1987], laryngography [Gilbert et al., 1984] and electropalatog raphy (EPG) [Hardcastle and Gibbon, 1997] equipments. Recently, the most commonly used system for articulatory data collection is EMA. An EMA system collects the 2-D or 3-D recordings of articulatory movements inside and outside the vocal tract by attaching sensor coils to articulators of the participants. The participants wear a special helmet that produces an alternating magnetic field. The movement of the sensors results in changes in the magnetic field, which generate electrical currents. The electrical currents are then converted to the movement angle and location of the articulators, which can simulate the shape of the vocal tract.

However, dysarthric articulatory data is difficult to collect and currently, very few

dysarthric datasets contain aligned acoustic and articulatory data. Due to the limited amount of dysarthric articulatory data, synthetic (often referred to as pseudo) articulatory data obtained via learnt acoustic-to-articulatory mappings has been employed to support acoustic features for improving acoustic modelling of dysarthric speech [Xiong et al., 2018; Yılmaz et al., 2018]. By learning the mapping from the acoustic to the articulatory space, the synthesiser estimates the articulatory data from the acoustic representations. *Gnuspeech* [Hill et al., 2017] and TADA [Nam et al., 2004] are the two popular articulatory data synthesisers. The potential drawback of synthetic articulatory features is that they might not represent the real dysarthric articulatory space effectively. That is, the synthesisers are normally trained on typical speech and then applied to generate dysarthric speech while assuming the articulatory-acoustic mapping remains invariant between typical and dysarthric speech. In contrast, the real recorded dysarthric articulatory data can better reflect the dysarthric articulatory space.

Given the substantial differences between the typical and dysarthric speech signals, there can be a significant mismatch between the acoustic-articulatory mappings for typical and dysarthric speech. This leads to uncertainty as to whether the synthesised articulatory data conformed to actual dysarthric speech properties. Therefore, using acoustic-articulatory mappings learned for typical speech is suboptimal and accompanied by a significant error. In this chapter, the recorded dysarthric EMA data in the TORGO dysarthric speech dataset will be exploited. Does using the real dysarthric articulatory data minimise such errors, and is it a more reliable speech representative? Detailed analysis of the articulatory space of dysarthric speech is essential to answer these questions. It can provide evidence for the key differences in the production of the dysarthric and typical speech regarding the articulatory side.

The rest of the chapter is organised as follows: The 2-D articulator movement trajectory and the 3-D point cloud of some samples are visualised in Section 7.2, which provides intuitive observation of the difference between dysarthric and typical speech. Section 7.3 presents a systematical comparison of the dysarthric and typical articulator movement patterns by analysing the statistical articulatory space distribution of the articulatory data using the maximum articulator motion range (MAMR) indicator. Section 7.4 gives

a summary of this chapter.

7.2 TORGO Articulatory Data Visualisation

In this section, the 2-D articulator movement trajectory and the 3-D point cloud of some articulatory data samples in TORGO are visualised. The visualisation provides intuitive observation of the articulatory difference between dysarthric and typical speech. First, the details of the recorded articulatory data in TORGO are presented.

7.2.1 TORGO EMA data

The articulatory data in TORGO were collected through a 3-D AG500 EMA system (also known as EMA data). The EMA data samples are measured by 12 sensors capturing articulatory movements in the 3-D space, each returning sensor positions in Cartesian coordinates (x, y, z) along with the six spatial orientation angles. The sensors are attached to the tongue back (TB), tongue middle (TM), tongue tip (TT), forehead, bridge of the nose (BN), upper lip (UL), lower lip (LL), lower incisor (LI), left mouth (LM), right mouth (RM), left ear (LE) and right ear (RE). The sensors attached behind each ear are used for reference purposes and to record the head motion. Figure 7.1 illustrates the placement of some of the sensor coils attached to the articulators during data collection in TORGO.

After removing the data samples that were not well recorded, TORGO consists of 7177 EMA data samples. Since the acoustic data (16,363 audio data samples) is recorded by a head-mounted and array microphone simultaneously, one set of EMA data usually is associated with two sets of acoustic data. Therefore, TORGO contains aligned acoustic and articulatory recordings for most of the utterances (13,127/16,363), which are used in the following (Chapter 8) experiments. Table 7.1 presents the number of EMA recordings in each session of the individual speaker in TORGO. Not all speakers in the dataset have articulatory recordings, and there is some missing EMA data in some sessions of the speakers. The missing EMA data is usually removed due to the dropped sensors or the disturbing magnetic field during the recording session.

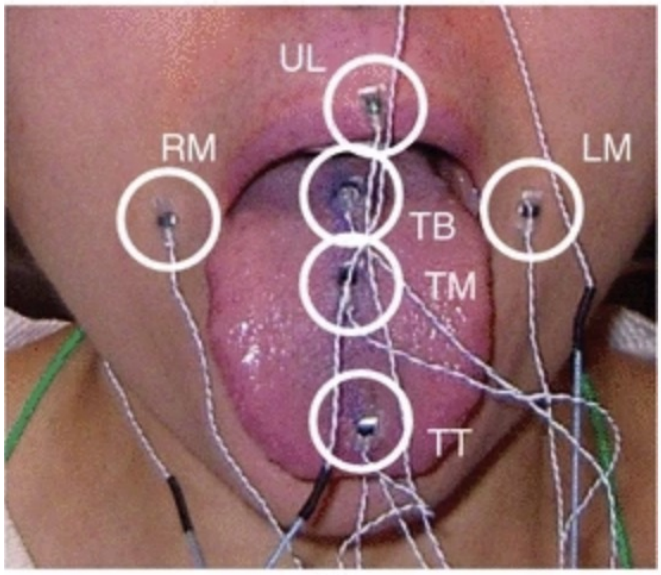


Figure 7.1: The placement coils on the RM, LM, UL, TT, TM and TB in the AG500 EMA system. The figure is adapted from the original article [Rudzicz et al., 2012b].

Visartico [Ouni et al., 2012] is a useful articulatory data visualisation tool. By simulating the sensor attachment, it is first used to check whether the name of the sensors used corresponds to the correct sensor and whether there are any problematic sensors. The sensor configuration in Visartico is illustrated in Figure 7.2

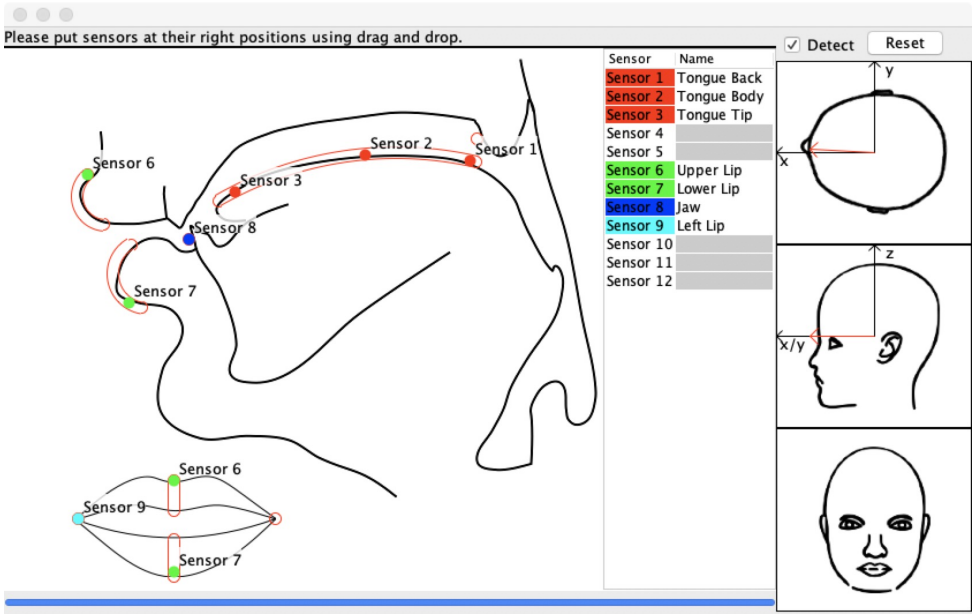


Figure 7.2: Sensor configuration.

Table 7.1: The number of EMA recordings of each speaker in TORGO. '-' indicates the missing recordings, 's' represents 'Session'.

	Severe				**M/S**	**Moderate**	**Mild**	
Session	F01	M01	M02	M04	M05	F03	F04	M03
total	-	100	169	298	523	661	250	421
s1	-	100	-	-	130	-	-	-
s2	-	-	169	298	393	447	250	421
s3	-	-	-	-	-	214	-	-

	Typical speakers						
Session	FC01	FC02	FC03	MC01	MC02	MC03	MC04
total	225	999	795	754	696	600	656
s1	225	-	405	333	388	600	656
s2	-	-	390	-	308	-	-
s3	-	999	-	421	-	-	-

In the following, some observations are presented.

1. The order of the sensor data saved for typical speakers and speakers with dysarthria is different. Table 7.2 presents the channel sequence for the dysarthric and typical groups. It is seen that the 4^{th}, 5^{th} and 10^{th} sensors are attached differently (marked with the red colour). Care should be taken when using the EMA data of all sensors from all speakers. The sensor sequences of the tongue and lip articulators (i.e., the 1^{st}, 2^{nd}, 3^{rd}, 5^{th} and 10^{th} sensors) are the same for the two groups. So using the lip or tongue information will not be affected by the mismatching sensor order.

2. Some problematic sensors are observed for some speakers. One case is that some attached sensors move to other places accidentally during the recording. This happens for speakers M04 and MC01, where the TB sensor moves out of their mouths. This might be because the electromagnetic signal of this sensor is interfered with by the external signal or the sensor drops off during the recording. Another observation is that some sensors do not work for speaker M05, where the TM and TB sensors have no data recorded, i.e., zero values for TM and TB.

3. The UL and LL sensors have no obvious attached problem across speakers. Therefore, the lip articulatory information might be a better choice to use as articulatory features

compared with the tongue articulators. Other sensors attached to the forehead, BN, LE and RE, are used for reference purposes and to record head motion. The following experiments will use the data from the lip region (UL and LL) and the tongue region (TT, TB and TM) as articulatory features.

Table 7.2: The EMA data channel sequence attached for the typical and the dysarthric group

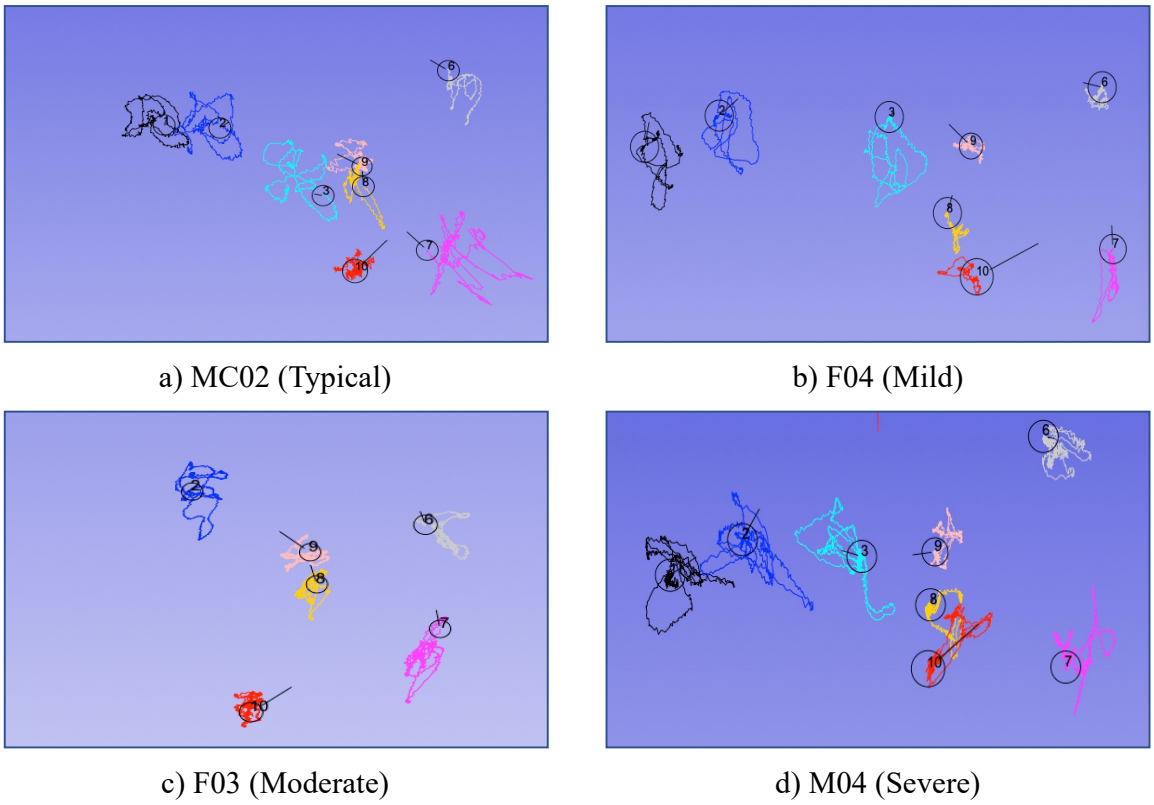
Channel No.	Typical	Dysarthric
1	Tongue back (TB)	Tongue back (TB)
2	Tongue middle (TM)	Tongue middle (TM)
3	Tongue tip (TT)	Tongue tip (TT)
4	Right mouth (RM)	Forehead
5	Forehead	Bridge of the nose (BN)
6	Upper lip (UL)	Upper lip (UL)
7	Lower lip (LL)	Lower lip (LL)
8	Lower incisor (LI)	Lower incisor (LI)
9	Left mouth (LM)	Left mouth (LM)
10	Bridge of the nose (BN)	Right mouth (RM)
11	Left ear (LE)	Left ear (LE)
12	Right ear (RE)	Right ear (RE)

7.2.2 2-D Articulator Movement Trajectory

Figure 7.3 depicts the 2-D articulator movement trajectory for the utterances spoken by four speakers with different severity levels: MC02 (typical), F04 (mild), F03 (moderate) and M04 (severe) with the same prompt *"The pair of shoes was new"*. The sensors corresponding to the numbers labelled in Figure 7.3 can be found in Figure 7.2.

It is seen that the typical speaker has the clearest tongue articulator movement trajectories. The tongue articulators' (i.e., sensors 1, 2 and 3) movement ranges of the typical speakers and the speakers with mild dysarthria are smaller than the speakers with moderate and severe dysarthria. For instance, for sensor 2, the movement ranges (mm) along the X-axis and Y-axis are (5.38, 5.32) for typical speech, (5.56, 6.43) for mild speech, (5.78, 7.53) for moderate speech and (8.03, 10.86) for severe speech. It suggests that the tongues of speakers with moderate and severe dysarthria are less flexible to move around.

The lip sensors of the mild speaker have smaller movement ranges than the severe speaker, where the movement area (mm^2) for sensor 6 and sensor 7 is (2.13, 11.55) for speaker F04 and (15.96, 27.88) for speaker M04). This indicates that the typical speaker does not need to move (open or close) his or her mouth that much to make a pronunciation. In contrast, the speakers with dysarthria have struggled to open and close their mouths during speaking in an uncontrolled shape.



a) MC02 (Typical) b) F04 (Mild)

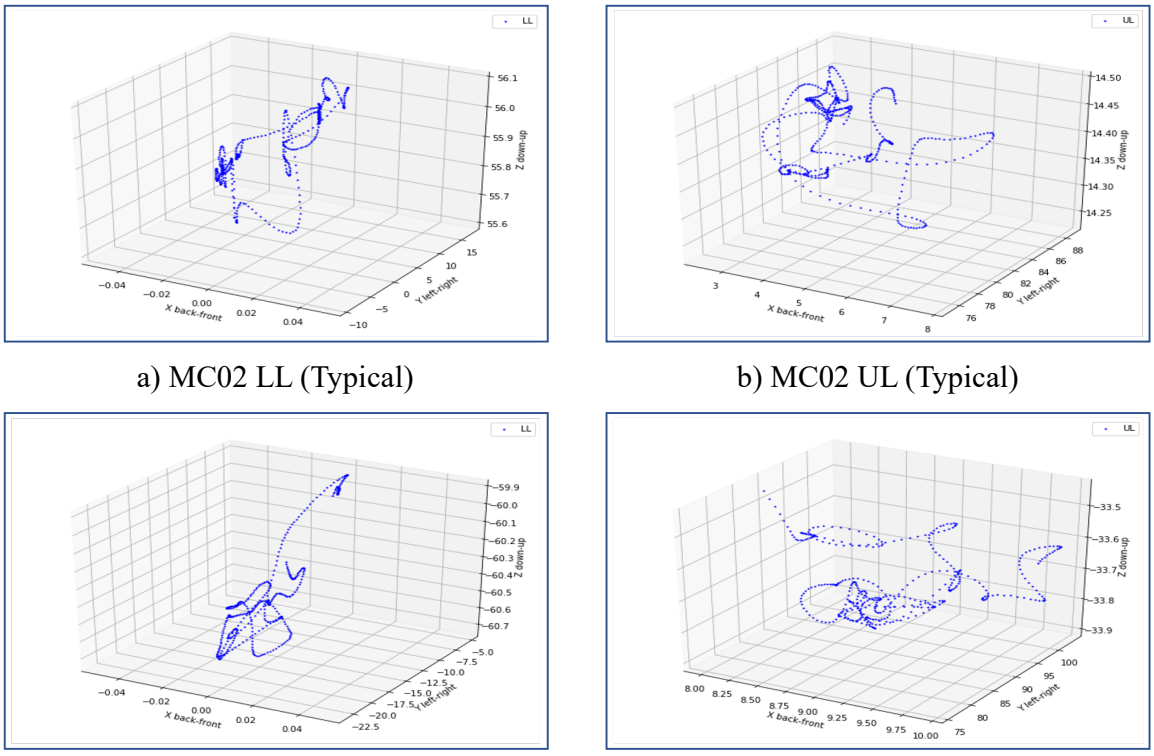
c) F03 (Moderate) d) M04 (Severe)

Figure 7.3: 2-D articulator movement trajectory for the utterance "*The pair of shoes was new*" for speakers with different dysarthria severity levels.

7.2.3 3-D Point Cloud Plots

The 3-D point cloud of the UL and LL articulators for speaker F03 with moderate dysarthria and the typical speaker MC03 are plotted in Figure 7.4. It is seen that the typical speaker has a clearer lip articulator movement trajectory in the 3-D space. The lip articulators' movement ranges of the typical speech are lower than the dysarthric speech along the left-right (UL: 15.1mm Typ 34.2mm Dys, LL: 6.50mm Typ 19.33mm Dys) and

up-down (UL: 0.28mm Typ 0.71mm Dys, LL: 0.61mm Typ 0.93mm Dys) directions. The difference between the two groups along the front-back direction is small. The observation is consistent with the findings of the 2-D articulator movement trajectory in Section 7.2.2.



a) MC02 LL (Typical)

b) MC02 UL (Typical)

c) F03 LL (Moderate)

d) F03 UL (Moderate)

Figure 7.4: 3-D point cloud of the UL and LL for the utterance "*The pair of shoes was new*" for speakers MC02 (typical) and F03 (moderate dysarthria).

This section is an intuition of the articulatory difference between the dysarthric and typical speech by visualising the EMA data. However, the 2-D articulator movement trajectory and 3-D point cloud plots of several samples are not representative of the motion patterns of all dysarthric and typical utterances. To this end, a thorough quantitative analysis on articulator motion patterns will be presented in Section 7.3 based on the Cartesian coordinates position such as the MAMR [Duan et al., 2020].

7.3 Statistical Articulatory Space Distribution

In this section, quantitative analysis on the statistical articulatory space distribution of the lip and tongue regions is carried out using a part of the TORGO dataset. The analysis is made over the dysarthric and typical speech. The MAMR is used as an indicator of the articulatory space distribution, which measures the difference between the maximum and the minimum articulator position values within a single utterance.

7.3.1 Data Selection and Preparation

In order to make better comparison, the utterances where the prompts are overlapping between speakers are used. Speaker M04 and M05 are left out due to dropped sensors being observed. Therefore, the recordings from four speakers with dysarthria (M01, F03, F04 and M03) and four typical speakers (MC01, MC02, FC02 and FC03) are eventually used for the analysis. Speaker MC02 is selected as the base speaker to determine the utterances used to compare with other speakers, since he has the largest number of utterances with the prompts overlapping with other speakers. Table 7.3 presents the number of utterances where the prompts are overlapping between speaker MC02 and other speakers. For instance, 80 means that speaker M01 and MC02 have 80 utterances where the prompts are overlapping.

Table 7.3: The number of utterances where the prompts are overlapping between speaker MC02 and other speakers.

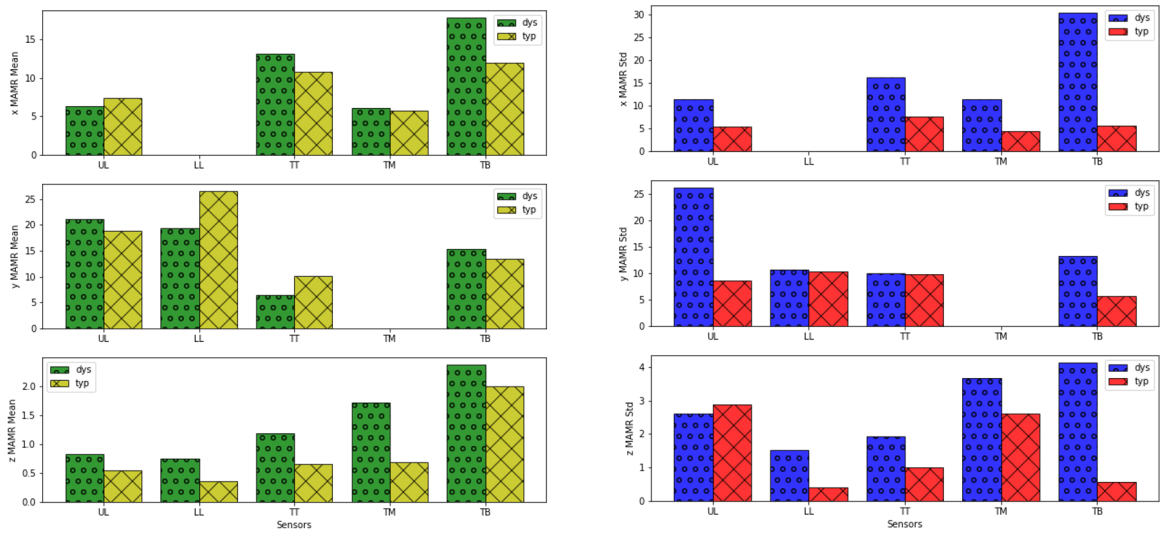
Speaker	M01	F03	F04	M03	FC02	FC03	MC01
MC02	80	274	209	339	756	579	555

7.3.2 Maximum Articulator Motion Range

Figure 7.5a and Figure 7.5b compares the MAMR mean (μ) and standard deviation (σ) of different articulators (TT, TM, TB, UL and LL[1]) along three directions: X (front-back),

[1]There are initially 12 sensor coils attached (shown in Table 7.2), and five of them are chosen to be used (UL, LL, TT, TM, TB) as articulatory features while others are used as reference sensors.

Y (left-right) and Z (up-down). The values are averaged for the typical speech (speaker MC02) and the dysarthric speech (speaker M01, F03, F04 and M03).



(a) MAMR mean (μ). (b) MAMR standard deviation (σ).

Figure 7.5: Statistics of MAMR between dysarthric and typical speech.

Table 7.4 presents the MAMR μ and σ of the two groups. In general, the σ of the typical speech is lower than the dysarthric speech (except for the UL articulator along the Z direction). It indicates that the dysarthric speech exhibits more fluctuation than the typical speech. TB has much higher σ along the X direction ($\sigma = 30.52$) for the dysarthric speech than the typical speech ($\sigma = 5.62$), suggesting that the speakers with dysarthria have less control over the muscle of the root of the tongue when producing speech. Looking at μ, all sensors exhibit higher MAMR for the dysarthric speech than the typical speech along the Z direction. It suggests that people with dysarthria tend to move their articulators up and down in a broader range than the typical speakers. However, along the X and Y directions, the dysarthric speech does not always have higher MAMR than the typical speech.

Large speaker variability exists for dysarthric speech. Given that speakers M01, F03, F04 and M03 have different severity levels, the MAMR distribution statistics are analysed for speakers with dysarthria individually in Table 7.5. The values in Table 7.6 examine whether the large MAMR distribution statistical difference exists among different typical speakers (MC01, MC02 and FC02). The difference between the μs of all five articula-

tors among different typical speakers varies from 0.68 mm to 2.47 mm, from 0.74 mm to 8.41 mm and from 0.83 mm to 3.08 mm along the X, Y and Z direction, respectively. In contrast, the difference between the μs is much bigger among different dysarthric speakers.

The μ difference between the moderate and mild speakers is much smaller than the μ difference between the severe and moderate, the severe and mild speakers. It indicates that the articulatory movement patterns are different at various severity levels. The moderate and mild speakers are more similar and are in turn more similar to the typical speakers, while speakers with severe dysarthria have much higher MAMR μ and σ values. This indicates that the speakers with severe dysarthria have less control over their articulators and exhibit much more variability in their speech.

Note that there is an exception along the Y direction for UL and LL articulators in Table 7.5. Although the σs of the severe speaker are higher than the moderate and mild speakers, the μs of the severe speaker (5.97 mm and 9.89 mm) are lower than the moderate (27.76 mm and 20.48 mm) and mild (26.95 mm 23.36 mm) speakers. This might be because that the severe speaker moves her mouth with poorer flexibility when moving her mouth along the left-right direction. It is also seen that the motion ranges of all five articulators are narrower in the Z direction compared with the other two directions. Take TT as an example, the μs are 1.18 mm (13.10 mm and 6.42 mm along the X and Y directions) and 0.66 mm (10.72 mm and 10.05 mm along the X and Y directions) across speakers. Besides, the σs display less fluctuation along the Z direction, for both speakers with dysarthria and typical speakers.

Table 7.4: The MAMR statistics (μ and σ) for different articulators averaged for dysarthric and typical speech.

Articulator	Direction	Dysarthric speech μ /mm	Dysarthric speech σ	Typical speech μ /mm	Typical speech σ
TT	X	13.1	16.18	10.72	7.58
	Y	6.42	10.04	10.05	9.8
	Z	1.18	1.94	0.66	1.0
TM	X	6.06	11.45	5.68	4.47
	Y	-	-	-	-
	Z	1.71	3.67	0.68	2.62
TB	X	17.8	30.52	11.93	5.62
	Y	15.26	13.21	13.42	5.73
	Z	2.38	4.15	2.0	0.57
UL	X	6.27	11.46	7.43	5.5
	Y	21.07	26.26	18.68	8.68
	Z	0.83	2.62	0.54	2.88
LL	X	-	-	-	-
	Y	19.26	10.61	26.55	10.31
	Z	0.74	1.51	0.36	0.39

Table 7.5: The MAMR statistics (μ and σ) for different articulators of different speakers with dysarthria.

Articulator	Direction	M01 (Severe) μ /mm	M01 (Severe) σ	F03 (Moderate) μ /mm	F03 (Moderate) σ	F04 (Mild) μ /mm	F04 (Mild) σ
TT	X	21.74	30.11	10.4	5.21	9.85	7.44
	Y	11.76	11.12	5.72	4.75	5.73	6.71
	Z	3.26	1.42	1.11	1.37	0.81	0.93
TM	X	23.35	23.03	3.25	2.64	4.53	4.59
	Y	-	-	-	-	-	-
	Z	4.64	6.17	1.79	2.67	0.91	1.47
TB	X	48.23	74.17	9.78	6.2	15.05	6.66
	Y	23.08	29.87	14.19	13.52	12.03	6.8
	Z	3.35	1.51	2.83	1.47	2.42	1.41
UL	X	5.72	10.03	4.14	3.28	5.5	17.24
	Y	5.97	7.72	27.76	44.01	26.95	11.34
	Z	2.0	5.21	0.83	1.69	0.60	0.72
LL	X	-	-	-	-	-	-
	Y	9.89	19.73	20.48	10.62	23.36	9.94
	Z	1.58	2.01	0.59	1.15	0.61	0.47

Table 7.6: The MAMR statistics (μ and σ) for different articulators of typical speakers.

Articulator	Direction	MC01		MC02		FC02	
		μ /mm	σ	μ /mm	σ	μ /mm	σ
TT	X	12.67	6.31	12.23	11.55	12.91	10.88
	Y	6.76	5.63	7.34	8.87	5.67	3.82
	Z	11.40	9.47	10.93	1.19	9.79	3.58
TM	X	4.76	5.60	6.50	7.08	6.66	11.27
	Y	-	-	-	-	-	-
	Z	3.48	2.59	3.79	1.08	6.15	2.17
TB	X	11.89	9.48	11.66	6.97	14.13	5.85
	Y	16.91	6.68	13.20	5.71	15.61	14.9
	Z	5.68	7.55	1.94	1.60	5.02	2.49
UL	X	7.94	8.55	8.06	6.72	6.95	13.04
	Y	27.65	11.96	19.24	9.15	25.32	16.57
	Z	0.07	0.03	0.51	0.57	1.31	1.41
LL	X	-	-	-	-	-	-
	Y	23.52	8.88	23.55	9.53	22.78	12.13
	Z	0.13	0.05	0.42	0.45	0.96	0.96

7.4 Summary

This chapter illustrated the articulatory mismatch between dysarthric and typical speech by visualising and systematically analysing the real articulatory data in the TORGO dataset. Specifically, the 2-D and 3-D point cloud visualisation of several articulatory data samples provided intuitive observation of the mismatch. Then the statistical space distribution regarding MAMR was compared between dysarthric and typical speech, demonstrating the key differences of the production of the dysarthric speech regarding the articulatory side. This exploration provided evidence that instead of using estimated articulation parameters from acoustic signals using knowledge about typical speech, the

real articulatory data can only conform to actual dysarthric speech properties with less uncertainty. In the next chapter, the recorded articulatory data will be applied to build robust acoustic-articulatory speech systems for dysarthric speech.

Chapter 8

Multimodal Acoustic-articulatory Speech Recognition Systems

Contents

8.1 Introduction

In Chapter 7 the statistical distribution of the articulatory movement has been analysed. It demonstrates the articulatory restrictions that speakers with dysarthria have, as well as the motion mismatch between the dysarthric and typical speech in the articulatory space. This indicates that the articulatory information can reflect the essence of the dysarthria and can capture the variability of dysarthric speech. In addition, the acoustic and articulatory mismatch between dysarthric and typical speech leads to the mismatch between the acoustic-articulatory mappings for typical and dysarthric speech. Using the synthetic dysarthric articulatory information obtained via acoustic-to-articulatory mappings learnt for typical speech can be accompanied by a significant error. In contrast, the real articulatory information is a more reliable speech representative that can better reflect the actual dysarthric speech properties for automatic dysarthric speech recognition (ADSR).

Multimodal speech recognition has received increased attention in recent years. This is because deep learning provides effective frameworks for fusing different data modalities. Previous studies have demonstrated the benefit of incorporating articulatory features by building acoustic-articulatory automatic speech recognition (ASR) systems for typical speech [Badino et al., 2016; Mitra et al., 2017]. As mentioned in the previous chapter, compared with acoustic representations, the articulatory information is more noise-robust [Wrench and Richmond, 2000], less speaker-variant [Fujimura, 1986] and more suitable to model the coarticulation variability [Frankel and King, 2001; Kirchhoff et al., 2002]. It is expected that the additional articulatory information integrated with conventional acoustic features can be more representative to help improve the performance of dysarthric speech recognition. However, most of the research on dysarthric speech recognition has focused on using the acoustic feature representations. Incorporating real articulatory data with acoustic features has not been widely explored in the dysarthric speech community. Limited research on multimodal ADSR has been carried out due to the lack of parallel multimodal data in the dysarthric domain. Synthetic articulatory data have been jointly used with the acoustic features in Xiong et al. [2018]; Yılmaz et al. [2018], and made

recognition improvement for ADSR. However, the synthetic dysarthric articulatory data was modelled from the acoustic-articulatory mappings learned for typical speech, which is accompanied by a significant error owing to the dysarthric and typical speech mismatch. The real dysarthric articulatory data needs to be exploited to minimise such error.

TORGO is a widely used dysarthric speech dataset [Rudzicz et al., 2012b], consisting of aligned acoustic and articulatory data for both dysarthric and typical speech. There have been several TORGO-based studies incorporating its articulatory data [Rudzicz, 2009, 2010b,c; Rudzicz et al., 2012a], and some have demonstrated improved recognition performance after incorporating articulatory information with acoustic features on speaker-dependent (SD) models [Rudzicz, 2009]. It is noticed that most of these TORGO-based studies employed the Gaussian mixture model (GMM)-hidden Markov model (HMM) or simple deep neural network (DNN) acoustic models. Whether the articulatory information is still beneficial when combined with state-of-the-art acoustic models proposed recently needs to be verified.

In addition, the better fusion scheme for fusing articulatory and acoustic features is not clear. The concatenation of the acoustic and articulatory features in the input level, although simple, is in fact suboptimal [Loweimi et al., 2020]. This is because the acoustic and articulatory representations encode different information in various formats and with different importance to the task. Consequently, the optimal set of filters to process each stream will differ. Pre-processing each stream individually, and then fusing the processed streams at a higher level is necessary. Direct feature concatenation at the input level does not allow such per stream pre-processing separately.

This chapter demonstrates the effectiveness of multimodal acoustic modelling for dysarthric speech recognition using conventional acoustic features along with articulatory information. It extends the previous acoustic-articulatory ADSR studies with more recent acoustic modelling architectures. It also discusses how to better apply the articulatory information and what is a better information fusion scheme in an acoustic-articulatory speech recognition system. Section 8.2 presents data processing for electromagnetic mid-sagittal articulography (EMA) data in TORGO. Various multimodal acoustic-articulatory ADSR system are explored in Section 8.3. Section 8.4 concludes this chapter.

8.2 Data Processing

The raw EMA data in TORGO is stored as a 1-D array in binary files. The length of the array is (N_samples * N_variables * N_sensors), where N_samples refers to the number of samples, and the number of sensors (N_sensors) is 12. The number of variables recorded for each sample channel is 7, including 3 Cartesian sensor positions in (x, y, z) directions, 3 spatial orientation angles and an extra 0. In order to make better use of the EMA data, it needs to be processed properly. First, the data is reshaped to a 3-D array (N_samples, 7, 12). Like most previous studies, the sensor positions in the (x, y, z) directions are used to measure the articulatory movements in this work. The 3-D array is then reshaped to three 12-D vectors: *data_x*, *data_y* and *data_z*. Figure 8.1 depicts the 12^th channel of *data_x* of a sample as an example.

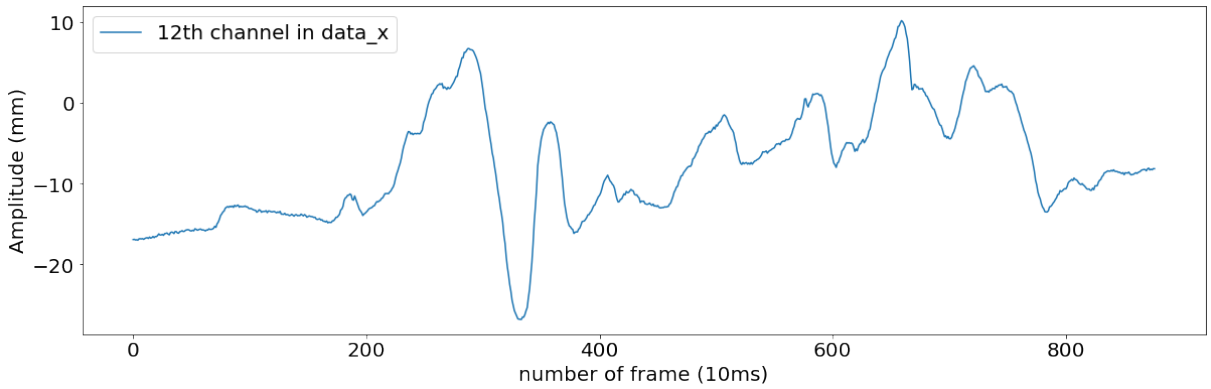


Figure 8.1: An example of a clean channel of an EMA data sample.

Before being employed as a feature, the raw EMA data needs to be appropriately processed. It is pre-processed by three steps: low-pass filtering, downsampling and channel selection, as depicted in Figure 8.2.

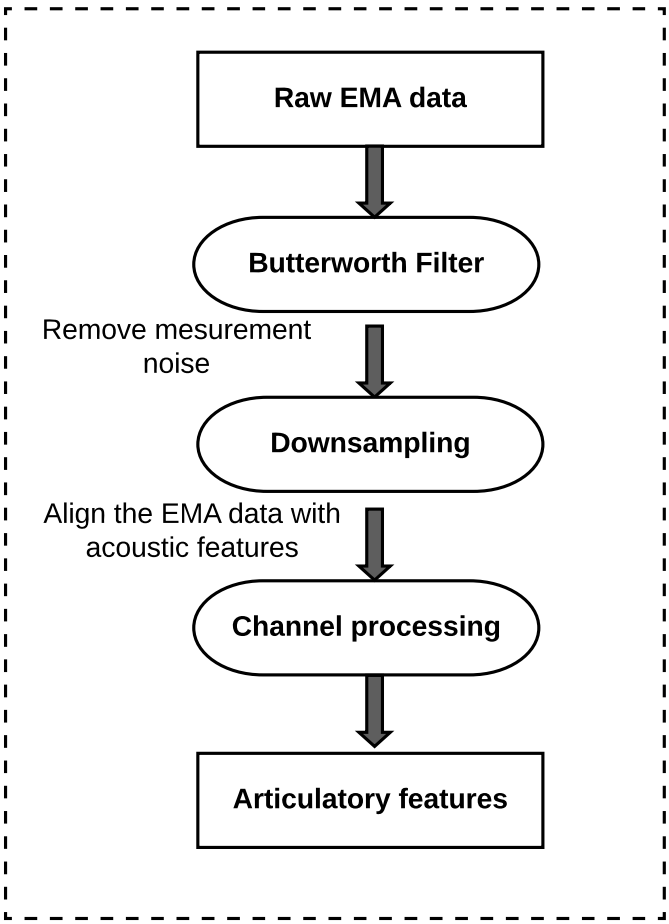


Figure 8.2: EMA data pre-processing.

A Butterworth low-pass filter is first applied to reduce the high-frequency measurement noise existing in most of the EMA data channels. The high cutoff frequency of the filter is set to 10 Hz, and the order is 5. The acoustic (e.g., Mel-frequency cepstrum (MFCC)) and the articulatory data have different frame rates, namely 100 Hz and 200 Hz, respectively. Therefore, the EMA data needs to be downsampled to align with the acoustic features (100 Hz). Figure 8.3 compares the two downsampling approaches: the anti-aliasing filtering and the Fourier resampling. It is seen that the downsampled data using the Fourier resampling fits better to the raw EMA data. A Python script was created to read and pre-process the raw EMA data in TORGO, which is available at `https://github.com/zhengjunyue/art_tg`. The EMA data consists of 12 channels, and each channel corresponds to a sensor which attached to an articulator. Since different articulators work differently, it might be the case that not all EMA channels can benefit the

recognition task for dysarthric speech. To avoid redundant information, the task-beneficial EMA channels which can complement the acoustic features towards achieving the highest recognition result are selected as the articulatory features. This process is called channel selection, which is done along with the acoustic-articulatory speech recognition task by comparing the performance for each system in terms of word error rate (WER).

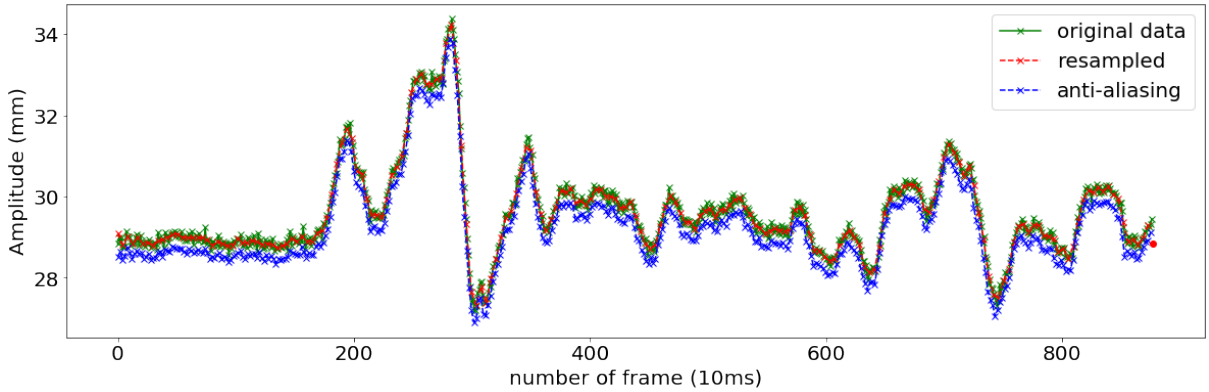


Figure 8.3: EMA data downsampling.

8.3 Acoustic-articulatory Dysarthric Speech Recognition Systems

In this section, various acoustic-articulatory ADSR systems are explored by comparing various articulatory feature settings, training configurations and feature fusion level. The impact of transfer learning on the multimodal framework is also investigated.

8.3.1 Experimental Setup

Figure 8.6 depicts the structure of the proposed multimodal acoustic-articulatory speech recogniser, where each stream is first pre-processed by Network-1 and Network-2, and then the fused streams are post-processed by Network-3 before reaching the output layer. The 39-D MFCCs and 3-D EMA features are used as inputs, with splicing of ± 5 contextual frames. The training data is augmented using speed perturbation (using factors 0.9, 1.0 and 1.1). Including the feature scheme, the proposed acoustic models are cascades of convolutional neural network (CNN), fully-connected multi-layer perceptron (MLP) and

light gated recurrent unit (LiGRU) [Ravanelli et al., 2018]. The CNNs are a cascade of three 1-D convolutional layers as used in Loweimi et al. [2020]. The subsequent structure, including the LiGRU and MLP layers, is the same as proposed in Chapter 6. The 5-fold cross-training TORGO setup used in Chapter 6 is applied. The independent 200k vocabulary size Librispeech trigram language model proposed in Chapter 5 is employed for decoding.

8.3.2 Exploration of Appropriate Measures for Articulatory Features

There are different ways to employ the EMA data. Three measures of the processed EMA data are considered as articulatory features:

1. The Cartesian coordinates (x,y,z) positions.
2. The Euclidean distance between the articulatory sensors and the origin (0,0,0), so-called origin Euclidean distance.
3. The pair-wise Euclidean distance between sensors.

Figure 8.4 illustrates the three articulatory measures in the 3-D space.

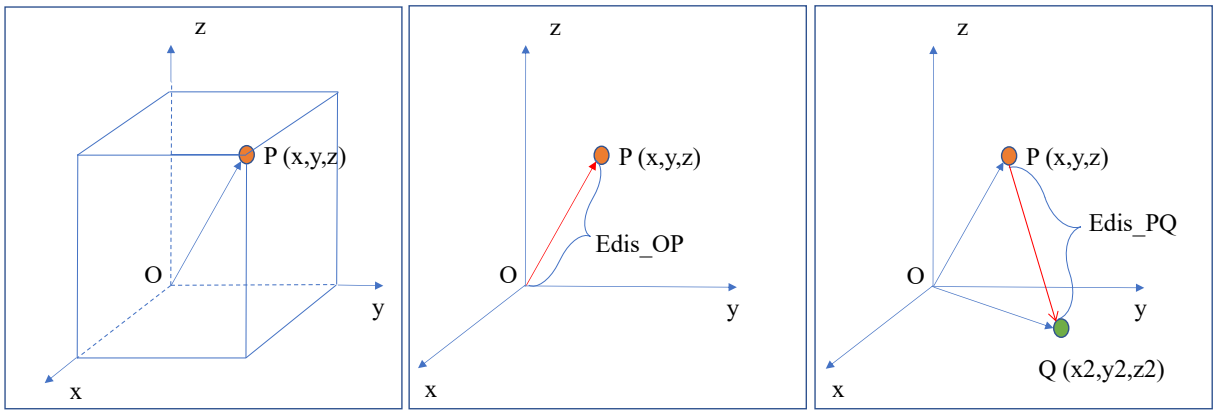


Figure 8.4: Three measures of EMA data. E_dis: Euclidean distance.

The Cartesian coordinates (either (x,y) or (x,y,z)) of the articulators are the most commonly used articulatory features in previous ADSR studies [Rudzicz, 2009; Rudzicz et al.,

2012a]. Instead of measuring where the articulators are, the Euclidean distance measures the distance between the articulators. The Euclidean distance between the articulator and the origin (0,0,0) and the pair-wise Euclidean distance between two articulators (e.g., in the lip region[1]) are exploited as the articulatory features.

Figure 8.5 plots the maximum articulator motion range (MAMR) distribution maps of the three measures of the lip sensors along the front-back direction for the dysarthric and typical speech separately.The MAMR in the top and bottom figures obeys the log-normal distribution, and the middle figure follows the bimodal distribution as it has two peaks. The local maximums for the typical speech are 26 and 64, while the local maximums for the dysarthric group are 50 and 120. Comparing the envelopes of the MAMR distribution, the pair-wise Euclidean distance appears to distinguish between the dysarthric and typical speech the best. Although the MAMR of the origin Euclidean distance is the most dispersed among the three measures, the overall envelope is more overlapped than the pair-wise Euclidean distance. It is also observed from the origin Euclidean distance figure that the MAMR distribution converges more slowly for the dysarthric speech than the typical speech, and there tend to be more utterances with high MAMR (e.g., higher than 100 ms). This is owing to more abnormal MAMR in the dysarthric speech.

Table 8.1 compares the MAMR mean (μ) and standard deviation (σ) of the three articulatory measures of the lip sensors for dysarthric and typical speech. The standard deviation of the dysarthric speech is higher than the typical speech among the three measures. With higher standard deviation, speakers with dysarthria exhibit more fluctuation than typical speakers. The MAMR mean for the dysarthric speech is bigger than the typical speech of the first two measures (i.e., the Cartesian coordinates and the origin Euclidean distance). In contrary, the MAMR mean for the dysarthric speech is smaller than the typical speech of the third measure (i.e., the pair-wise Euclidean distance). The MAMRs of Cartesian coordinates and the origin Euclidean distance measure the absolute displacement of the articulators while the pair-wise Euclidean distance measures the relative displacement. This indicates that the speakers with dysarthria tend to move or shake

[1]The Euclidean distance between the UL (UL_x,UL_y,UL_z) and LL (LL_x,LL_y,LL_z).

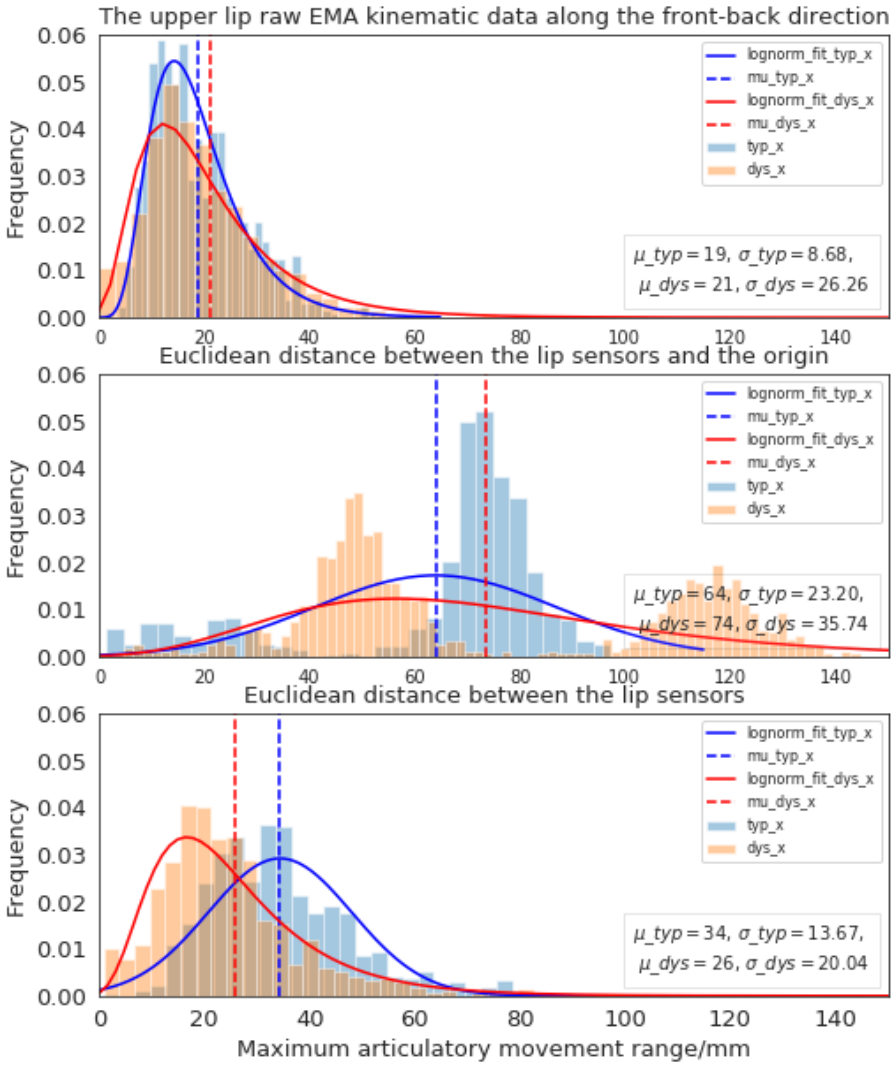


Figure 8.5: MAMR distribution map of three articulatory measures of the lip sensors for dysarthric and typical speech.

their bodies or heads while speaking. They work harder to move their articulators (i.e., resulting in smaller MAMR means). This demonstrates the pair-wise Euclidean distance's ability to remove the body/head movement influence. The articulatory features therefore are implicitly normalised.

Table 8.1: The MAMR statistics (μ and σ) of the three articulatory measures of the lip sensors for dysarthric and typical speech.

	Dysarthric speech		**Typical speech**	
Articulator	μ /mm	σ	μ /mm	σ
Lip	21	26.26	19	8.68
Lip_ud_origin	74	35.74	64	23.20
Lip_ud	26	20.04	34	13.67

Table 8.2 compares the WER of systems trained on different input features using different articulatory measures. The training data is the combination of the dysarthric and typical speech. The results are averaged for the dysarthric and typical speech. The 1^{st} row displays the baseline results using only MFCC acoustic features. The results of the systems concatenating the lip and tongue Cartesian coordinates with the MFCC are reported in the 2^{nd} and the 3^{rd} rows. It is observed that both *MFCC+Tongue* and *MFCC+Lip* systems outperform the baseline *MFCC* system reducing WER by 0.12% and 1.07% absolute on average for the dysarthric speech. 0.41% and 0.43% performance gains are also achieved for the typical speech by integrating the lip and tongue information. The lip information adds more benefit to the ADSR model than the tongue information. As a result, the lip information will be employed in the following experiments.

Although the overall improvement is obtained by employing the Cartesian coordinates of the lip articulators, it is found that it does not provide consistent improvement across all speakers. The 4^{th} and 5^{th} rows in Table 8.2 present the results of using the Euclidean distance-based articulatory features. It shows that the *MFCC+Lip_ud* system outperforms any other articulatory measures, reducing WER by 1.91% and 0.53% for the dysarthric and typical speech, respectively, compared with the baseline MFCC sys-

tem. The most task-beneficial articulatory measure appears to be the pair-wise Euclidean distance system using the lip information[1] (Lip_ud), which will be used in the following experiments.

Table 8.2: WER for systems trained on different input features and different articulatory measures.

Input Features	Dysarthric	Typical
MFCC	47.80	16.38
MFCC+Tongue	47.66	15.97
MFCC+Lip	46.73	15.95
MFCC+Lip_ud_origin	46.33	16.01
MFCC+Lip_ud	**45.89**	**15.85**

8.3.3 Exploring the Effect of Different Training Sets

There have been comparative studies on the speaker-independent (SI), SD and speaker adaptation ADSR systems previously [Christensen et al., 2013, 2012b; Raghavendra et al., 2001] using acoustic information. In this section, whether SI and SD is better, and whether using out-of-domain (OOD) typical speech helps the recognition performance for the acoustic-articulatory ADSR are investigated. The following systems listed in Table 8.3 are explored.

Table 8.3: Systems trained on different training sets.

SD	Trained on the data of the target speaker
SI	Trained on the data of all other speakers
Train_Dys	Trained on the dysarthric speech only
Train_Typ	Trained on the typical speech only
Train_Both	Trained on both the dysarthric and typical speech

[1]The Euclidean distance between the UL (UL_x,UL_y,UL_z) and LL (LL_x,LL_y,LL_z).

Table 8.4: WER for systems trained on different training sets.

System	Dysarthric	Typical
Train_Both (SI)	**45.89**	15.85
Train_Dys (SI)	50.79	17.60
Train_Typ (SI)	86.61	**15.58**
SD	79.85	16.62

The results of systems listed in Table 8.3 are reported in Table 8.4. The 1^{st} to the 3^{rd} rows are the results for the SI systems trained on both dysarthric and typical speech (*Train_Both*), the dysarthric speech (*Train_Dys*) and the typical speech (*Train_Typ*). The last row presents the results of the *SD* system. It is observed that the *Train_Both* system achieves the best performance for both dysarthric and typical speech, followed by the *Train_Dys*, *SD* and the *Train_Typ* system. Comparing the results of the three SI systems, both *Train_Dys* and *Train_Typ* systems benefit from adding training data from the other domain when recognising the dysarthric speech. This indicates that adding typical speech data during training can help in recognising dysarthric speech. The best result for the typical speech is obtained by the *Train_Typ* system. The limited amount of training data is not compensated for by adding the dysarthric speech data. The variability and uncertainty of the trained model have been increased instead. This suggests that the additional dysarthric speech data confuses the system recognising the typical speech. This also indicates the significant mismatch between the dysarthric and the typical speech. It is notable that the performance of the *SD* system is much worse than the *Train_Dys* system for dysarthric speech. This is different from what was found in Rudzicz [2009]; Salama et al. [2014]. The reason might be that the amount of training data is too small to train recent robust acoustic models. With the best performance, the *Train_Both* system using the MFCC+Lip_ud features as input will be applied in the following experiments.

8.3.4 Acoustic and Articulatory Feature Fusion

The early (feature-based) integration and late (model-based) integration are two widely used methods to fuse different data modalities. The former concatenates two different

features in a single feature vector frame-by-frame before feeding them into the acoustic model. The latter implements the integration at a later stage of the speech recognition process, i.e., by combining the results of the classifiers for different feature modalities as the final phoneme classification result. Based on previous knowledge, the model trained on the articulatory data alone cannot capture acoustic information and therefore performs poorly in recognition. The late integration is not suitable in the acoustic-articulatory ADSR task. The following experiments apply the feature-based integration approach as illustrated in the yellow box in Figure 8.6.

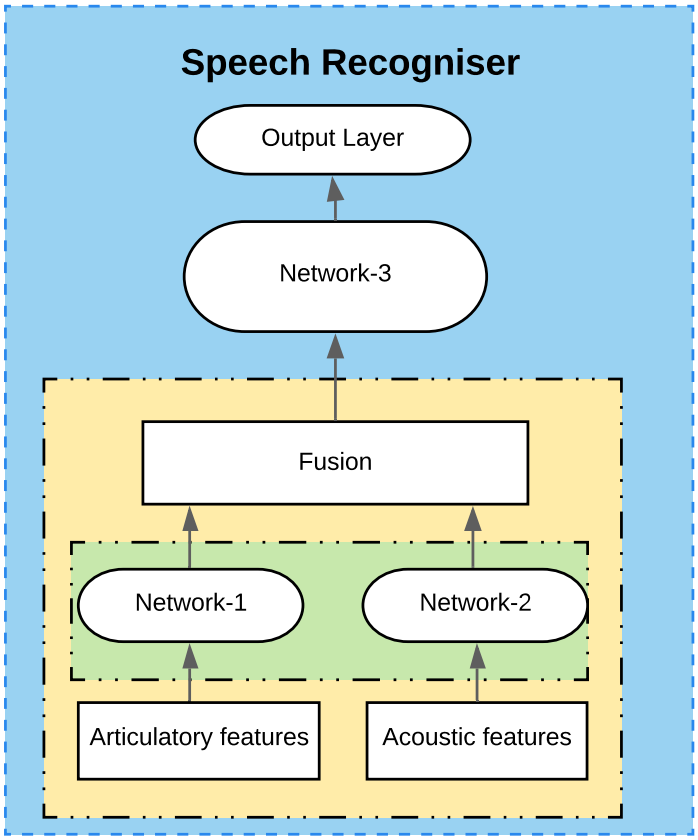


Figure 8.6: Proposed speech recogniser.

The simplest way of integrating the articulatory features is concatenating them with acoustic features in a single vector on a frame-by-frame basis and at the network's input level. However, a direct concatenation of these features might be suboptimal. This is owing to the fact that these two information streams carry different types of information, encoded in different forms with different importance to the given task. This necessitates

applying a bespoke per-stream pre-processing before fusion at an optimal level of abstraction. The Network-1 and Network-2 components in the green box in Figure 8.6 perform multi-stream per-processing while Network-3 carries out post-processing after fusion and before the output layer. In Loweimi et al. [2020], multi-stream acoustic modelling and fusion at low, medium and high levels were studied. Inspired by such a framework, the optimal fusion level is explored by concatenating individually processed streams at three stages in this section. The following fusion levels are studied: at the input level (**concat-1**), at the medium level after the last convolutional layer (**concat-2**), and at the high level after the last recurrent layer and before the output layer (**concat-3**). **concat-0** represents the direct concatenation and does not include convolutional layers, where the baseline acoustic model consists of recurrent (LiGRU) and fully-connected MLP layers. Figure 8.7 illustrates various concatenation levels and fusion schemes.

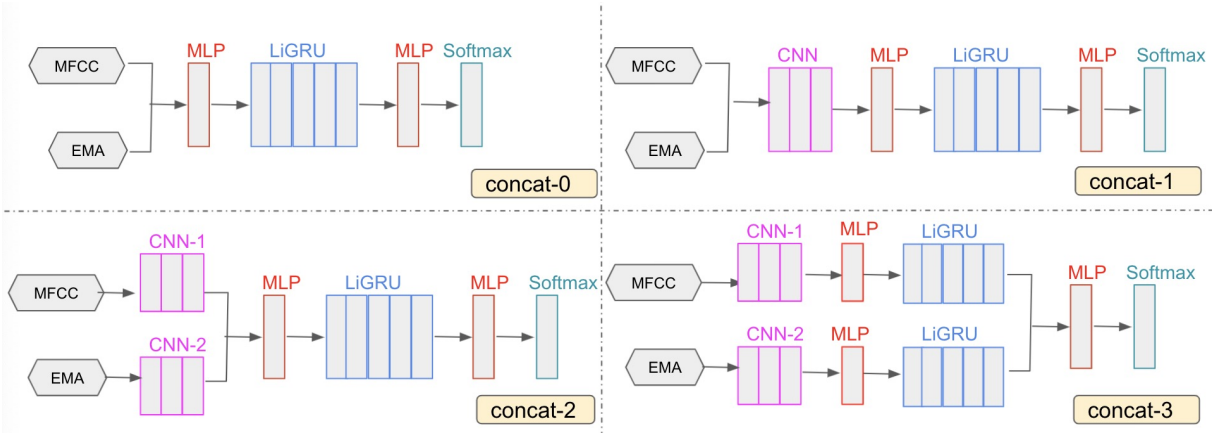


Figure 8.7: The proposed architectures fusing the acoustic and articulatory features at different levels.

The results of different feature fusion systems are reported in Table 8.5. The 1^{st} and 2^{nd} rows display the results of the MFCC and MFCC+Lip_ud systems (also referred to *concat-0*). It is seen that the direct feature fusion at the input level outperforms the baseline MFCC system. On average, it reduces WER by 1.91% and 0.53% (absolute) for dysarthric and typical speech, respectively. The 3^{rd} to 5^{th} rows in Table 8.5 show the results of introducing the multi-stream CNN feature fusion schemes. Comparing the *concat-1* and *concat-2* with *concat-0* (i.e., (MFCC+Lip_ud)) shows 0.49% and 2.68% absolute WER reductions for dysarthric speech, respectively. The *concat-2* system appears to be the

best fusion scheme while *concat-3* leads to the poorest performance.

Table 8.5: WER for different feature fusion systems averaged for dysarthric and typical speech.

System	Dysarthric	Typical
MFCC	47.80	16.38
concat-0	45.89	15.85
concat-1	45.40	15.02
concat-2	**43.21**	**12.88**
concat-3	60.42	35.15

The number of trainable parameters (*#params*) is counted for each model in Table 8.6. As seen, fusion at higher levels greatly increases *#params*. For example, *#params* of the *concat-3* system is 1.5 times of the *concat-1* and *concat-2* systems. This makes the model more liable to overfitting, especially in low-resource data scenarios. Furthermore, concatenating the streams close to the output layer could give rise to insufficient post-processing (after fusion). Our experimental results for dysarthric speech verify the conclusion in Loweimi et al. [2020] regarding the optimal fusion level: it should be high enough to effectively pre-process each information stream for the given task and low enough to leave sufficient capacity after fusion for post-processing the fused streams.

Table 8.6: Number of parameters (in millions) for different fusion schemes.

	MFCC	concat-0	concat-1	concat-2	concat-3
#params	11.1	11.3	15.1	15.0	24.9

Figure 8.8 compares the evolution of the cross-entropy (CE) loss of various proposed fusion schemes during training. It illustrates that the *concat-3* system converges faster than other models. The *concat-2* system tends to have the lowest training and validation loss which is due to the fact that it provides the best trade-off in terms of pre- and

post-processing.

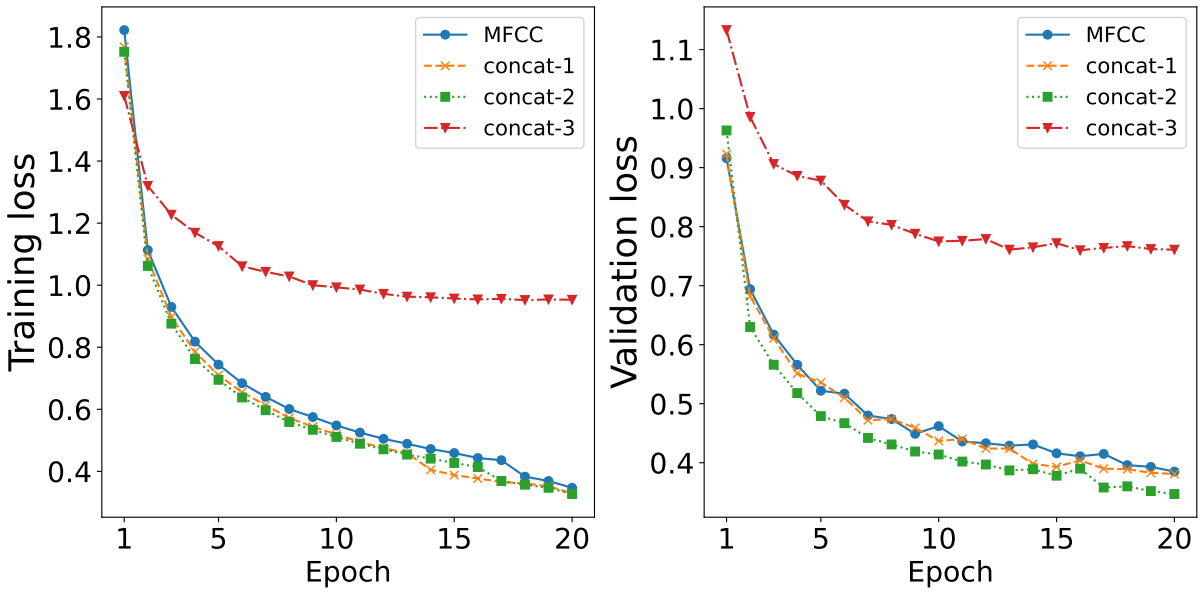


Figure 8.8: CE loss for different fusion schemes.

The performance evolution of the *concat-2* system across different epochs was then explored. The results for speakers with dysarthria and typical speakers are plotted in Figure 8.9. As seen, the WER improvement for speakers with severe dysarthria is notably limited and does not continuously improve during training, contrary to the typical or mild conditions. Moreover, the performance reaches a plateau after 10 epochs for dysarthric speech while for the typical speech, the performance keeps significantly improving up to 15 epochs.

8.3.5 Exploring the Effect of Transfer Learning

To utilise the existing data more efficiently, the transfer learning strategy is applied to adapt the learnt model to each target speaker. In particular, the acoustic model is pre-trained on both dysarthric and typical data. Then some layers are re-trained on the test speaker's training data. After several explorations, it was found that freezing the learnt parameters in the first six layers, then re-training the last fully connected layer together with the classifiers is the best configuration for the transfer learning system. Typically, the first several layers in the neural networks learn to capture the global information of

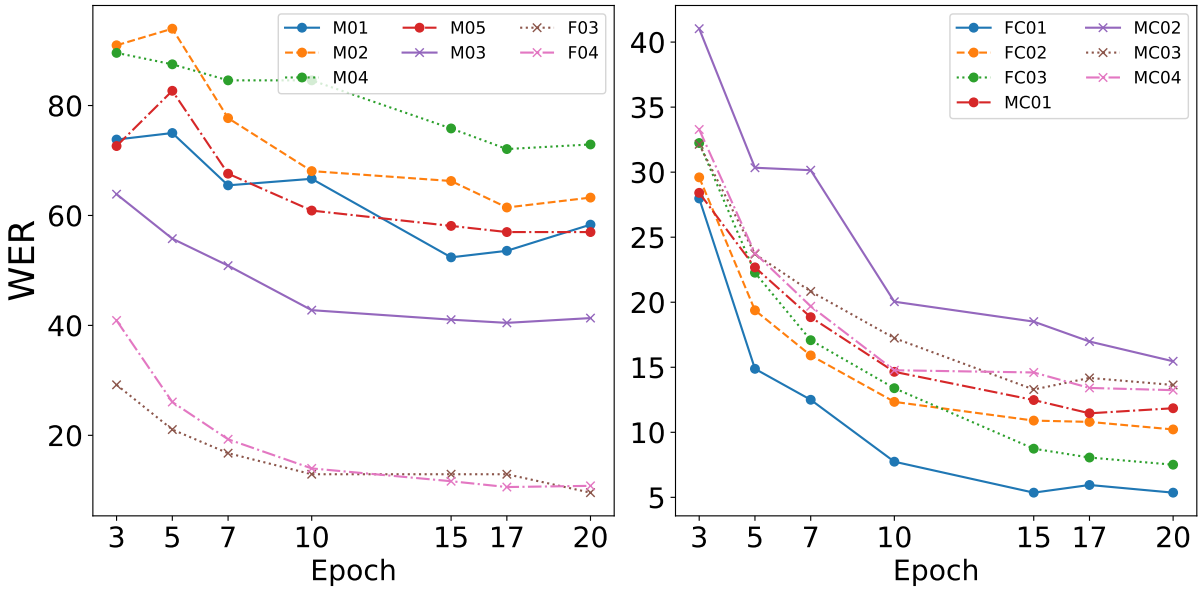


Figure 8.9: WER at different epochs in the *concat-2* system for speakers with dysarthria (left) and typical speakers (right).

the representations, and the last few layers are more likely to learn the fine features.

Table 8.7 reports WERs along with the relative gain obtained by applying transfer learning for each system. The transfer learning strategy successfully reduces WER for all three systems (*MFCC*, *MFCC+Lip_ud* and *concat-2*). Applying transfer learning, the *concat-2* system achieves the best performance with 3.62% and 8.39% relative gains for the dysarthric and typical speech. It is also observed that both *MFCC+Lip_ud* and *concat-2* systems obtain more relative performance gains than the *MFCC* system when applying the transfer learning strategy. This indicates that the additional articulatory information helps learn better information from the target speaker during the transfer process. That is, the transfer learning strategy is of greater benefit for the fused acoustic and articulatory features than for the acoustic-only features.

8.3.6 Results for the Separate Sentence and Word Tasks

In order to investigate how the systems perform on the continuous speech, the results for the separated word and sentence task are reported in Table 8.8. It shows that the additional lip articulatory features effectively improve the recognition performance on the sentence task for both dysarthric and typical speech. The improvement made by the lip

Table 8.7: WER for systems applying transfer learning (TF).

System	Dysarthric	Typical
MFCC	47.80	16.38
MFCC TF	46.91 (1.86%)	15.55 (5.07%)
MFCC+Lip_ud	45.89	15.85
MFCC+Lip_ud TF	43.77 (**4.62%**)	14.26 (**10.03%**)
concat-2	43.21	12.88
concat-2 TF	**41.70** (3.62%)	**11.08** (8.39%)

information on the sentence subset is greater than that made on the full (sentence+word) testset for the dysarthric speech. In contrast, the lip articulatory features benefit more on the word task for the typical speech. The above results suggest that the additional lip articulatory information is effective on the sentence recognition task for dysarthric speech. It helps severely dysarthric continuous speech particularly.

Table 8.8: WER for systems applying transfer learning (TF) for different utterance types.

System	Dysarthric		Typical	
Subsets	Sentence	Word	Sentence	Word
MFCC	53.63	41.23	18.10	10.07
MFCC+Lip_ud	50.11	39.88	17.85	9.35

Figure 8.10 compares the WER reduction employing the transfer learning strategy on the *MFCC* and *MFCC+Lip_ud* systems for the word, the sentence and the word+sentence tasks across speakers with dysarthria. It is found that, in general, the transfer learning strategy is more beneficial when incorporating articulatory information for the continuous, severely dysarthric speech.

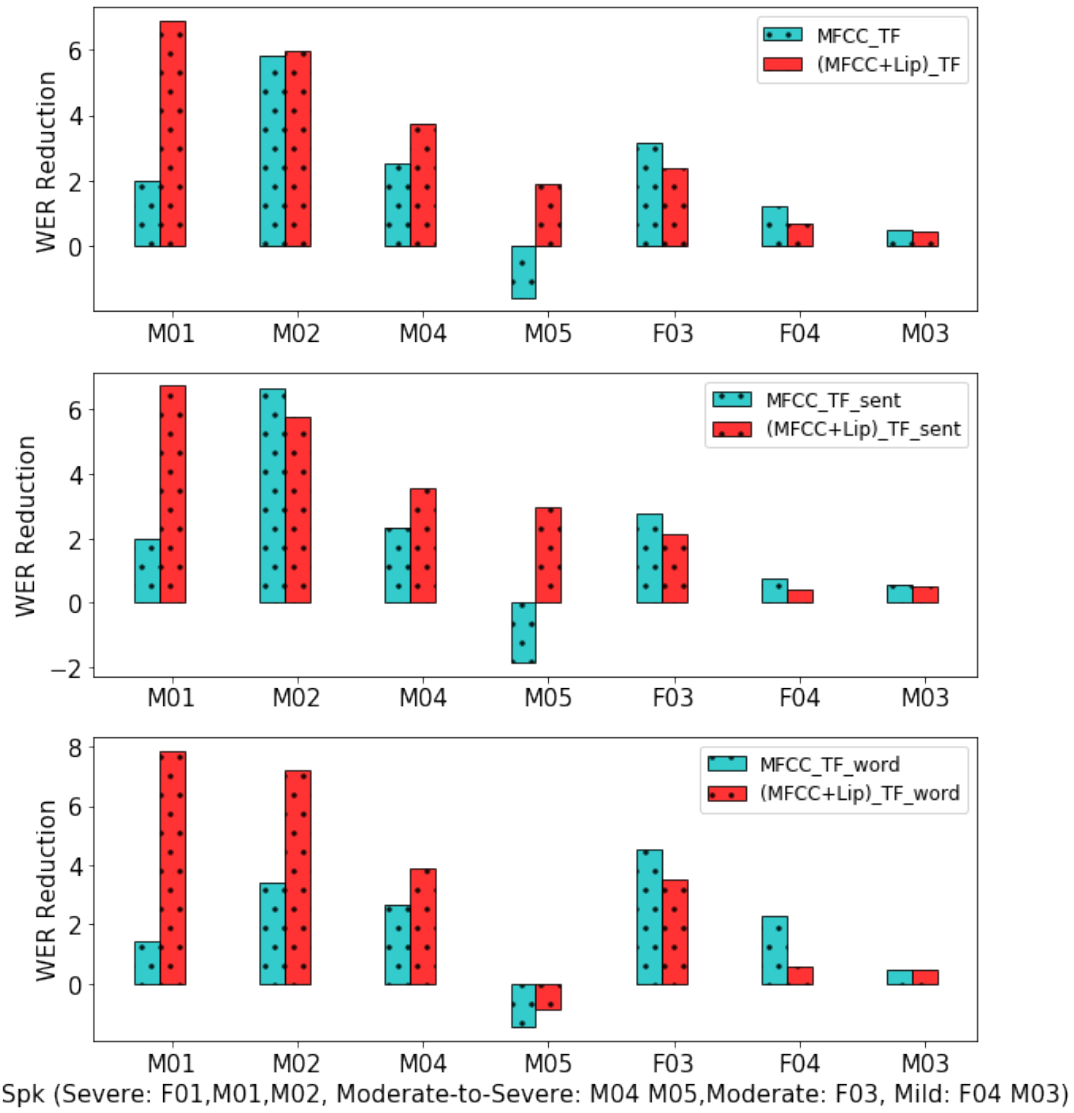


Figure 8.10: The WER reduction employing transfer learning (TF) on the MFCC and the (MFCC+Lip_ud) systems.

8.4 Conclusion

This chapter demonstrated the effectiveness of incorporating the real articulatory information along with acoustic features constructing multimodal acoustic-articulatory speech recognition systems with recent advanced acoustic models on the TORGO dataset. The pair-wise Euclidean distance of the articulators in the lip region was shown to be the most appropriate articulatory feature for the ADSR task. The proposed multi-stream acoustic models consist of convolutional, recurrent and fully-connected layers allowing the multi-

modal features to be fused via various schemes and at different levels of abstraction. The best performance was achieved by fusing the acoustic and articulatory information at the medium level trained on both dysarthric and typical data, which reduced the absolute WER by 4.6% compared with the MFCC baseline. Further improvement was achieved by exploiting the transfer learning strategy resulted in a 1.5% (absolute) WER reduction. It is also notable that transfer learning is more effective on fused acoustic and articulatory features than on acoustic-only information. The work systematically addressed the research question **RQ4** with several explorations. It built a benchmark for multimodal acoustic-articulatory ADSR system which also achieves performance gain on continuous speech. The next chapter will summarise the work in this thesis and present future work that is promising to be explored.

Chapter 9

Conclusion and Future Work

Contents

9.1 Conclusions

After introducing the topic of this thesis in **Chapter 1**, the thesis began by providing a background overview of dysarthria and dysarthric speech in **Chapter 2**. Dysarthria is a common speech disorder stemming from disruption in the neuro-motor interface [Gowers, 2001]. People with dysarthria often produce atypical speech due to the poorer motor control of their speech articulators. As a result, the intelligibility of dysarthric speech to listeners and the ability of machines to recognise dysarthric speech is affected. This then causes increasing social exclusion for people with dysarthria. In addition, people with dysarthria often have physical disabilities, meaning that simple tasks in daily life, such as turning on the light, can be affected. This demonstrates a significant need for automation and voice-enabled interfaces to help this group of people communicate better and live independently. However, the atypical speech makes it hard to interact with the devices using their voice.

There is, therefore, an urgent need for a reliable automatic speech recognition (ASR) system capable of recognising dysarthric speech with high accuracy, especially for speech assessed at the severe level of dysarthria. However, due to data scarcity, significant dysarthric and typical speech mismatch and high speaker variability, the automatic dysarthric speech recognition (ADSR) task is still challenging, and lags far behind the mainstream ASR systems for typical speech in terms of performance. ASR on continuous dysarthric speech is under-explored. This thesis was based on deploying state-of-the-art technologies in speech and deep learning to build benchmark systems for automatic continuous dysarthric speech recognition (ACDSR).

To improve ACDSR systems, this thesis attempted to explore and find the answers to the fundamental research questions put forward in Section 1.2. Each of the research questions will be restated now along with a summary of how this thesis addressed it.

(RQ1): What is an appropriate evaluation framework for continuous dysarthric speech recognition, given current data limitations?

The pilot study in Chapter 5 showed that most of the existing dysarthric datasets (including TORGO) contain many overlapped prompts across speakers. By evaluating

the word and sentence recognition tasks separately, the examination of the existing TORGO-based ADSR literature highlights the problems with how these systems are evaluated. The training and test sets for the in-domain TORGO language model is non-disjoint, leading to overly optimistic results for recognising sentences. Two task-specific TORGO language models were introduced to further support the fact. To choose appropriate language models in this low-resource task, language models trained on varying vocabulary-sized subsets of the external LibriSpeech dataset were proposed. The results demonstrated that the out-of-domain (OOD) language models provided lower but fairer performance. They helped the model better generalise to truly unseen utterances and allowed for a much larger decoding space. In addition, it was found that the vocabulary size of the OOD language models affected the WERs. In general, the lowest WERs are achieved with the largest vocabulary size. A reproducible benchmark for the current state-of-the-art ACDSR system with a fairly designed OOD language model was established. It is the first study to systematically explore and improve the evaluation methods on continuous dysarthric speech. The OOD language model was demonstrated to provide a more appropriate evaluation framework than the in-domain language models.

(RQ2): What is a good way to leverage typical speech, which is out-of-domain, to learn more robust representations for dysarthric speech?

Typical speech can be leveraged for dysarthric speech recognition in various ways. Inspired by Christensen et al. [2013]; Takashima et al. [2015]; Yılmaz et al. [2019]'s studies, an improved ACDSR system with a pretrained autoencoder bottleneck (AE-BN) feature extractor and applying multi-task optimisation techniques was established in Chapter 6. The results demonstrated the effectiveness of employing AE-BN features extracted from a feature extractor pretrained on OOD LibriSpeech data for the ACDSR task by reducing WERs by 2.63% absolute on average. Joint optimisation of the AE-BN feature extractor and the speech recogniser resulted in better AE-BN features achieving 0.65% absolute recognition improvements. The acoustic model was strengthened by applying monophone regularisation as an auxiliary task achieving 2.33% absolute WER reduction. However, the joint optimisation technique provided no consistent additional benefit when applied

together with monophone regularisation. The robust representations for dysarthric speech can be learnt by fine-tuning a pretrained OOD AE-BN feature extractor with joint optimisation.

(RQ3): How can articulatory information characterise continuous dysarthric speech, and what are the advantages of incorporating articulatory information?

Instead of using synthetic speech data, Chapter 7 focused on analysing the real articulatory data recorded in TORGO. Visualising the 2-D articulator movement trajectory and the 3-D point cloud of several sentence samples provided evidence that the articulators of speakers with moderate and severe dysarthria were less flexible. The statistical articulatory space distribution of the articulatory data was quantitatively analysed using the maximum articulator motion range (MAMR) indicator, which measures articulator motion patterns. This demonstrated the mismatch between dysarthric and typical speech in the articulatory space. It also showed that articulatory information was capable of capturing the speaker variability. The additional articulatory data stream integrated with conventional acoustic features could be more representative for dysarthric speech.

(RQ4): How can articulatory information be incorporated effectively to build multimodal ACDSR systems using recent acoustic models?

This research question was explored in Chapter 8. First, exploring the effectiveness of different components of articulators and various measures of the articulatory data, the results suggested that employing the information of the lip articulators achieved better results on dysarthric speech than others. The pair-wise Euclidean distance between the lip articulators was demonstrated to be the best articulatory measure for the ADSR task. Then, various training configurations were investigated and combining dysarthric and typical speech data for training was shown to be the best training configuration for the acoustic-articulatory system. Multi-stream architectures consist of convolutional, recurrent and fully connected layers allowing for per-stream pre-processing were established to investigate the optimal fusion level of the acoustic and articulatory features. It was found that the optimal fusion level should be high enough to effectively pre-process each information stream for the given task and low enough to leave sufficient capacity after

fusion for post-processing the fused streams. The optimal fusion scheme achieved a notable performance gain on dysarthric speech. Finally, the transfer learning strategy was applied to further improve the recognition performance. The proposed model leads to significant performance gains for ACDSR.

In conclusion, this thesis made promising progress in improving continuous dysarthric speech recognition. The frameworks are available at `https://github.com/zhengjunyue/CADSR-LM` and the experimental results are fully reproducible.

9.2 Future Work

The previous section described the main contributions towards improving the performance on continuous dysarthric speech. However, some challenges still need to be addressed to improve the overall performance of the proposed systems. Future directions for this work are discussed below.

9.2.1 More Data

The main issue of working with medical datasets is having limited access to useful data. The lack of data is mainly caused by the limited number of recordings and the ethical issues of sharing data. However, the deep learning based techniques largely depend on the quantity and quality of the data. For instance, collecting data from more speakers with dysarthria would allow for a better understanding of the speaker variability, which might further allow us to train acoustic models that are more generalised to different speakers. The performance of the proposed systems might be further improved with more data by training better language and acoustic models using deep learning techniques. At the time of writing this thesis, more dysarthric datasets have been collected, which contains richer recordings (e.g., [Turrisi et al., 2021]). These data can be utilised to train more robust systems for ACDSR.

Data augmentation is an alternative way to access more data. A speed perturbation method has been used in this thesis. Other data augmentation approaches could be employed in future work, such as voice conversion [Harvill et al., 2021; Jin et al., 2021] and SpecAugment [Park et al., 2019].

9.2.2 Employing Speech Representations from Other Components

In this thesis, the speech representations are based on handcrafted spectral features such as MFCCs. There are many other types of features that could be explored. Recent ASR research has achieved improved performance by learning speech information directly from the waveform. The raw waveform avoids the limitation of the spectral representations where useful information would have been lost during the feature computation, which might help learn better speaker and speech information in the dysarthric domain. The raw waveform has been successfully exploited in the dysarthric detection task [Millet and Zeghidour, 2019; Zeghidour, 2019]. End-to-end systems enable a learnable front-end where the speech representations are learnt from raw waveform jointly with the phoneme classifier [Zeghidour et al., 2018], which will be optimal for the given task. Learning low-level representation of speech has the potential to push state-of-the-art for ACDSR even further with more dysarthric speech data.

Another widely applied approach is source-filter modelling [Chiba and Kajiyama, 1958], which is among the fundamental techniques in speech processing. Based on the properties of the human speech production system, this model characterises the speech signal as a temporal convolution of some random or quasiperiodic excitation (Exc) signal (source) passing through a linear filter representing the vocal tract (VT). The filter component is primarily associated with the linguistic content of the speech signal while the source component is associated with the speaker attributes. Loweimi et al. [2021] demonstrated the efficacy of building acoustic models using the raw magnitude spectra of the source and filter components. Source-filter modelling has the potential to offer a special advantage in the context of ACDSR. When the model takes two inputs characterising the lingual content (vocal tract) and speaker attributes (excitation), among others, it learns to normalise the speaker-associated properties reflected in the source component while extracting the lingual content of the speech from the filter component. Such implicit speaker normalisation is highly desirable in recognising dysarthric speech with high speaker variability.

9.2.3 Extension on the Multimodal Acoustic-articulatory Speech Recognition Framework

Although the effectiveness of incorporating articulatory data with acoustic data have been demonstrated in Chapter 8, the paired acoustic and articulatory data is required in both training and testing. However, it is impractical to record articulator movements in real-life speech recognition scenarios if more paired data is required for training. This constraint requires ASR systems to utilise articulatory data only for training, i.e., to recognise without the articulatory data.

Articulatory-to-acoustic inversion [Atal et al., 1978] is one potential approach to solving this problem, where the articulatory features are generated from the acoustic signal through a learnt mapping. This, however, is challenging since the mapping between acoustic and articulatory data spaces is non-linear and not unique [Richmond, 2002]. Moreover, the limited paired dysarthric acoustic and articulatory data hinder the effectiveness of learning the mapping. Using OOD typical data to train the articulatory-to-acoustic inversion mapping system might help with this problem. Still, care needs to be taken considering the dysarthric and typical speech mismatch in the articulatory aspect. Another way is to embed the articulatory data in the model to adjust the parameters and optionally the structure of the acoustic model. In this case, the articulatory data is not required during testing and can be deployed during training. Recently, the distillation training [Yu et al., 2016] approach has been widely used to incorporate knowledge into the network. The knowledge can be transferred through soft targets from a "teacher" model trained with additional features to a "student" model with no access to those features. In this study, the "teacher" model could be trained with both articulatory and acoustic data, and then the outputs of the "teacher" model are used as additional targets during training the "student" model with only acoustic data.

9.2.4 Concluding Remarks

This thesis has built novel ASR systems to improve human-human communication and human-machine interaction for people with dysarthria. We explored techniques and methodologies in deep learning and speech technology. The evaluation of the ASR sys-

tems demonstrated promising results which confirms that there is a potentially bright future for assistive speech-driven devices for people with dysarthria.

References

Abraham, B., Umesh, S., and Joy, N. (2017). Joint estimation of articulatory features and acoustic models for low-resource languages. In *Interspeech*, pages 2153–2157. 49

Afouras, T., Chung, J., Senior, A., Vinyals, O., and Zisserman, A. (2018). Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*. 49

Atal, B., Chang, J., Mathews, M., and Tukey, J. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, 63(5):1535–1555. 139

Baayen, R., Piepenbrock, R., and Gulikers, L. (1996). The celex lexical database (cd-rom). 39

Badino, L., Canevari, C., Fadiga, L., and Metta, G. (2016). Integrating articulatory data in deep neural network-based acoustic modelling. *Computer Speech & Language*, 36:173–195. 49, 114

Bahl, L., Brown, P., de Souza, P., and Mercer, R. (1989). A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):1001–1008. 41

Bell, P., Swietojanski, P., and Renals, S. (2016). Multitask learning of context-dependent targets in deep neural network acoustic models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(2):238–247. 88

Beukelman, D. and Yorkston, K. (1979). The relationship between information transfer

and speech intelligibility of dysarthric speakers. *Journal of communication disorders*, 12(3):189–196. 21

Bhat, C., Das, B., Vachhani, B., and Kopparapu, S. (2018). Dysarthric speech recognition using time-delay neural network based denoising autoencoder. In *Interspeech*, pages 451–455. 45, 83

Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451. 72

Borrie, S., McAuliffe, M., and Liss, J. (2012). Perceptual learning of dysarthric speech: A review of experimental studies. In *ASHA*. 18

Brown, J. and Aronson, A. (1970). Ataxic dysarthria. *International Journal of Neurology*, 43(5):302–318. 23

Caballero Morales, S. and Cox, S. (2009). Modelling errors in automatic speech recognition for dysarthric speakers. *EURASIP Journal on Advances in Signal Processing*, 2009:1–14. 55

Chen, S. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394. 70

Chiba, T. and Kajiyama, M. (1958). The vowel: Its nature and structure. In *Phonetic Society of Japan, Tokyo*. 138

Choi, D., Kim, B., Kim, Y., Lee, Y., Um, Y., and Chung, M. (2012). Dysarthric speech database for development of QoLT software technology. In *LREC*, pages 3378–3381. Citeseer. 58

Chorowski, J., Weiss, R., Bengio, S., and van den Oord, A. (2019). Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12):2041–2053. 45, 83

Christensen, H., Aniol, M., Bell, P., Green, P., Hain, T., King, S., and Swietojanski, P. (2013). Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech. In *Interspeech*, pages 3642–3645. 6, 44, 47, 82, 83, 123, 135

Christensen, H., Cunningham, S., Fox, C., Green, P., and Hain, T. (2012a). A comparative study of adaptive, automatic recognition of disordered speech. In *Interspeech*, pages 1776–1779. 57

Christensen, H., Cunningham, S., Fox, C., Green, P., and Hain, T. (2012b). A comparative study of adaptive, automatic recognition of disordered speech. In *Thirteenth Annual Conference of the International Speech Communication Association*. 25, 46, 69, 123

clinic, C. (2021). Dysarthria. *Cleveland Clinic*. 16

Crowley, T. and Bowern, C. (2010). *An introduction to historical linguistics*. Oxford University Press. 24

Darley, F., Aronson, A., and Brown, J. (1969). Clusters of deviant speech dimensions in the dysarthrias. *Journal of speech and hearing research*, 12(3):462–496. 2

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366. 36

Deller Jr, J., Liu, M., Ferrier, L., and Robichaud, P. (1993). The Whitaker database of dysarthric (cerebral palsy) speech. *The Journal of the Acoustical Society of America*, 93(6):3516–3518. 54, 58

Deng, J., Gutierrez, F., Hu, S., Geng, M., Xie, X., Ye, Z., Liu, S., Yu, J., Liu, X., and Meng, H. (2021). Bayesian parametric and architectural domain adaptation of LF-MMI trained TDNNs for elderly and dysarthric speech recognition. *Interspeech*, pages 4818–4822. 47

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 42

Doyle, P., Leeper, H.and Kotler, A., et al. (1997). Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of rehabilitation research and development*, 34:309–316. 26, 48

Duan, S., Zhang, X., Yan, M., and Zhang, J. (2020). Statistical distribution exploration of tongue movement for pathological articulation on word/sentence level. *IEEE Access*, 8:91057–91069. 105

Duffy, J. (2013). *Motor speech disorders-e-book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences. 16

Enderby, P. (1980). Frenchay dysarthria assessment. *British Journal of Disorders of Communication*, 15(3):165–173. 22

Espana-Bonet, C. and Fonollosa, J. (2016). Automatic speech recognition with deep neural networks for impaired speech. In *International Conference on Advances in Speech and Language Technologies for Iberian Languages*, pages 97–107. Springer. 12, 45, 58, 63, 65, 67, 86, 87, 90

Estellers, V. and Thiran, J. (2012). Multi-pose lipreading and audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*. 49

Fant, G. (1970). *Acoustic theory of speech production*. Walter de Gruyter. 38

Ferrier, L., Shane, H., Ballard, H., Carpenter, T., and Benoit, A. (1995). Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, 11(3):165–175. 22, 25

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188. 37

Frankel, J. and King, S. (2001). ASR-articulatory speech recognition. In *Seventh European Conference on Speech Communication and Technology*. 98, 114

Fujimura, O. (1986). Relative invariance of articulatory movements, in Invariance and variability in speech processes. *Lawrence Erlbaum*, pages 226–242. 49, 98, 114

Gemello, R., Mana, F., Scanzio, S., Laface, P., and De Mori, R. (2007). Linear hidden transformations for adaptation of hybrid ANN/HMM models. *Speech Communication*, 49(10-11):827–835. 64

Geng, M., Xie, X., Liu, S., Yu, J., Hu, S., Liu, X., and Meng, H. (2020). Investigation of data augmentation techniques for disordered speech recognition. In *Interspeech*, pages 696–700. 48

Gilbert, H., Potter, C., and Hoodin, R. (1984). Laryngograph as a measure of vocal fold contact area. *Journal of Speech, Language, and Hearing Research*, 27(2):178–182. 98

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264. 42

Gowers, W. (2001). Clinical speech syndromes of the motor systems. *Neurology for the Speech-Language Pathologist. Fifth edition. Philadelphia: Butter worth_ Heinemenn*, pages 196–203. 19, 134

Graves, A., Jaitly, N., and Mohamed, A. (2013). Hybrid speech recognition with deep bidirectional LSTM. In *ASRU*, pages 273–278. IEEE. 89

Grezl, F. and Fousek, P. (2008). Optimising bottle-neck features for LVCSR. In *ICASSP*, pages 4729–4732. IEEE. 44, 82

Han, K., Hahm, S., Kim, B., Kim, J., and Lane, I. (2017). Deep learning-based telephony speech recognition in the wild. In *Interspeech*, pages 1323–1327. 46

Hardcastle, W. and Gibbon, F. (1997). Electropalatography and its clinical applications. *Instrumental clinical phonetics*, pages 149–193. 98

Harvill, J., Issa, D., Hasegawa-Johnson, M., and Yoo, C. (2021). Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary. In *ICASSP*, pages 6428–6432. IEEE. 137

Hasegawa-Johnson, M., Gunderson, J., Perlman, A., and Huang, T. (2006). HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria. In *ICASSP*, volume 3, pages III–III. IEEE. 45

Hermann, E. et al. (2020). Dysarthric speech recognition with lattice-free MMI. In *ICASSP*. IEEE. 86, 90

Hermansky, H., Ellis, D., and Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. In *ICASSP*, volume 3, pages 1635–1638. IEEE. 44

Hill, D., Taube-Schock, C., and Manzara, L. (2017). Low-level articulatory synthesis: A working text-to-speech solution and a linguistic tool. *The Canadian Journal of Linguistics/La revue canadienne de linguistique*, 62(3):371–410. 99

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780. 86

Holmes, K., Judge, S., and Murray, J. (2010). Communication matters-research matters: An AAC evidence base. *Commuication Matters*. 2

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556. 41

Jiao, Y., Tu, M., Berisha, V., and Liss, J. (2018). Simulating dysarthric speech for training data augmentation in clinical speech applications. In *ICASSP*, pages 6009–6013. IEEE. 49

Jin, Z., Geng, M., Xie, X., Yu, J., Liu, S., Liu, X., and Meng, H. (2021). Adversarial data augmentation for disordered speech recognition. *arXiv preprint arXiv:2108.00899*. 49, 137

Johnson, W., Darley, F., and Spriesterbach, D. (1963). Diagnostic methods in speech pathology. In *Harper & Row*. 54

Joy, N. and Umesh, S. (2018). Improving acoustic models in TORGO dysarthric speech database. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(3):637–645. 45, 58, 63, 65, 67

Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India. 37, 73

Jurafsky, D. and Martin, J. (2009). *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 39

Kaneko, T., Kameoka, H., Hojo, N., Ijima, Y., Hiramatsu, K., and Kashino, K. (2017a). Generative adversarial network-based postfilter for statistical parametric speech synthesis. In *ICASSP*, pages 4910–4914. IEEE. 49

Kaneko, T., Takaki, S., Kameoka, H., and Yamagishi, J. (2017b). Generative adversarial network-based pstfilter for STFT spectrograms. In *Interspeech*, pages 3389–3393. 49

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401. 42, 73

Kent, R., Netsell, R., and Abbs, J. (1979). Acoustic characteristics of dysarthria associated with cerebellar disease. *Journal of Speech, Language, and Hearing Research*, 22(3):627–648. 23, 26

Kent, R., Rosen, K., and Maassen, B. (2004). Motor control perspectives on motor speech disorders. *Speech motor control in normal and disordered speech*, pages 285–311. 24

Kent, R., Weismer, ., Kent, J., and Rosenbek, J. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54(4):482–499. 21, 58

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., and Frame, S. (2008a). Dysarthric speech database for universal access research. In *Interspeech*. 3, 45, 48, 54, 90

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., and Frame, S. (2008b). Dysarthric speech database for universal access research. In *Interspeech*, pages 1741–1744. 56

Kim, M., Cao, B., An, K., and Wang, J. (2018). Dysarthric speech recognition using convolutional LSTM neural network. *Interspeech*, pages 2948–2952. 3, 46, 57, 86

Kingsbury, P., Strassel, S., McLemore, C., and MacIntyre, R. (1997). Callhome american english lexicon (pronlex). *Linguistic Data Consortium, Philadelphia*. 39

Kirchhoff, K., Fink, G. A., and Sagerer, G. (2002). Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, 37(3-4):303–319. 98, 114

Klasner, E. and Yorkston, K. (2005). Speech intelligibility in ALS and HD dysarthria: The everyday listener's perspective. *Journal of medical speech-language pathology*, 13(2):127–140. 22

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*. 48

Kominek, J. and Black, A. (2004). The CMU Arctic speech databases. In *Fifth ISCA workshop on speech synthesis*. 72

Kroos, C. (2008). Measurement accuracy in 3D electromagnetic articulography (Carstens AG500). In *Proceedings of the 8th international seminar on speech production*, pages 61–64. 58

Kwong, S. and He, Q. (2001). The use of adaptive frame for speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2001(2):1–7. 37

Lally, M., Kim, H., and Moon, L. (2019). Dysarthric speech perception: Comparison of training effects on human listeners versus automatic speech recognition tools. *The Journal of the Acoustical Society of America*, 145(3):1795–1795. 19

Lamel, L., Kassel, R., and Seneff, S. (1989). Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Speech Input/Output Assessment and Speech Databases*. 58

Landow, G. (1993). *The digital word: Text-based computing in the humanities*. MIT Press. 58

Levelt, W. (1999). Models of word production. *Trends in cognitive sciences*, 3(6):223–232. 17

Li, X. and Wu, X. (2015). Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *ICASSP*, pages 4520–4524. IEEE. 86

Lillvik, M., Allemark, E., Karlström, P., and Hartelius, L. (1999). Intelligibility of dysarthric speech in words and sentences: development of a computerised assessment procedure in Swedish. *Logopedics Phoniatrics Vocology*, 24(3):107–119. 22

Liou, C., Cheng, W., Liou, J., and Liou, D. (2014). Autoencoder for words. *Neurocomputing*, 139:84–96. 84

Liss, J., Spitzer, S., Caviness, J., Adler, C., and Edwards, B. (1998). Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech. *The journal of the acoustical society of America*, 104(4):2457–2466. 24

Loweimi, E., Bell, P., and Renals, S. (2020). Raw sign and magnitude spectra for multi-head acoustic modelling. In *Interspeech*, pages 1644–1648. 115, 119, 126, 127

Loweimi, E., Cvetkovic, Z., Bell, P., and Renals, S. (2021). Speech acoustic modelling using raw source and filter components. In *Interspeech*, pages 716–720. 138

Maciuszek, J. (2018). Lexical access in the processing of word boundary ambiguity. *Social Psychological Bulletin*, 13(4):1–11. 24

Makhoul, J. and Schwartz, R. (1995). State of the art in continuous speech recognition. *Proceedings of the National Academy of Sciences*, 92(22):9956–9963. 4, 25

Mathew, J., Jacob, J., Sajeev, K., Joy, J., and Rajan, R. (2018). Significance of feature selection for acoustic modeling in dysarthric speech recognition. In *2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 1–4. IEEE. 43

McClelland, J. and Elman, J. (1986). The TRACE model of speech perception. *Cognitive psychology*, 18(1):1–86. 18

Menendez-Pidal, X., Polikoff, J., Peters, S., Leonzio, J., and Bunnell, H. (1996). The Nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1962–1965. IEEE. 48, 54

Mengistu, K. and Rudzicz, F. (2011). Adapting acoustic and lexical models to dysarthric speech. In *ICASSP*, pages 4924–4927. IEEE. 25, 26, 46, 65, 86, 90

Meunier, C., Fougeron, C., Fredouille, C., Bigi, B., et al. (2016). The TYPALOC corpus: A collection of various dysarthric speech recordings in read and spontaneous styles. In *Language Resources and Evaluation Conference (LREC)*, pages p–4658. 58

Millet, J. and Zeghidour, N. (2019). Learning to detect dysarthria from raw speech. In *ICASSP*, pages 5831–5835. IEEE. 138

Mitra, V., Sivaraman, G., et al. (2017). Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks. In *ICASSP*, pages 5205–5209. IEEE. 49, 114

Morris, M., Meier, S., Griffin, J., Branda, M., and Phelan, S. (2016). Prevalence and etiologies of adult communication disabilities in the United States: Results from the 2012 national health interview survey. *Disability and health journal*, 9(1):140–144. 2

Naik, D. (1995). Pole-filtered cepstral mean subtraction. In *ICASSP*, volume 1, pages 157–160. IEEE. 37

Nakashika, T., Yoshioka, T., Takiguchi, T., Ariki, Y., Duffner, S., and Garcia, C. (2014). Dysarthric speech recognition using a convolutive bottleneck network. In *12th International Conference on Signal Processing (ICSP)*, pages 505–509. IEEE. 44

Nam, H., Goldstein, L., Saltzman, E., and Byrd, D. (2004). Tada: An enhanced, portable task dynamics model in MATLAB. *The Journal of the Acoustical Society of America*, 115(5):2430–2430. 99

Nicolao, M., Christensen, H., Cunningham, S., Green, P., and Hain, T. (2016). A framework for collecting realistic recordings of dysarthric speech - the homeService corpus. In *Proceedings of LREC*. European Language Resources Association. 48, 54, 56

Ouni, S., Mangeonjean, L., and Steiner, I. (2012). VisArtico: a visualisation tool for articulatory data. In *Thirteenth Annual Conference of the International Speech Communication Association*. 101

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *ICASSP*, pages 5206–5210. IEEE. 7, 48, 71, 84

Parascandolo, G., Huttunen, H., and Virtanen, T. (2016). Recurrent neural networks for polyphonic sound event detection in real life recordings. In *ICASSP*, pages 6440–6444. IEEE. 48

Park, D., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E., and Le, Q. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*. 137

Paul, D. and Baker, J. (1992). The design for the wall street journal-based csr corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. 7

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (1802). Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*. 42

Piczak, K. (2015). Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE. 48

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., and Khudanpur, S. (2018). SemiOrthogonal low-rank matrix factorisation for deep neural networks. In *Interspeech*, pages 3743–3747. 64

Povey, D., Ghoshal, A., Boulianne, G., et al. (2011). The kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society. 63, 87

Purohit, M., Parmar, M., Patel, M., Malaviya, H., and Patii, H. (2021). Weak speech supervision: A case study of dysarthria severity classification. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 101–105. IEEE. 56

Raghavendra, P., Rosengren, E., and Hunnicutt, S. (2001). An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *Augmentative and Alternative Communication*, 17(4):265–275. 23, 123

Rao, K. and Vuppala, A. (2014). *Speech processing in mobile environments*. Springer. 37

Ravanelli, M., Brakel, P., Omologo, M., and Bengio, Y. (2016). Batch-normalized joint training for DNN-based distant speech recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 28–34. IEEE. 87

Ravanelli, M., Brakel, P., Omologo, M., and Bengio, Y. (2018). Light gated recurrent units for speech recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):92–102. 86, 89, 119

Ravanelli, M., Parcollet, T., and Bengio, Y. (2019). The Pytorch-Kaldi speech recognition toolkit. In *ICASSP*, pages 6465–6469. IEEE. 87

Richmond, K. (2002). Estimating articulatory parameters from the acoustic speech signal. In *The University of Edinburgh*. 139

Rudzicz, F. (2007). Comparing speaker-dependent and speaker-adaptive acoustic models for recognising dysarthric speech. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 255–256. 55

Rudzicz, F. (2009). Applying discretized articulatory knowledge to dysarthric speech. In *ICASSP*, pages 4501–4504. IEEE. 50, 115, 119, 124

Rudzicz, F. (2010a). Articulatory knowledge in the recognition of dysarthric speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):947–960. 50

Rudzicz, F. (2010b). Correcting errors in speech recognition with articulatory dynamics. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 60–68. 115

Rudzicz, F. (2010c). Learning mixed acoustic/articulatory models for disabled speech. In *NIPS*, pages 70–78. Citeseer. 45, 50, 115

Rudzicz, F. (2011). Articulatory knowledge in the recognition of dysarthric speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):947–960. 50, 58

Rudzicz, F., Hirst, G., and van Lieshout, P. (2012a). Vocal tract representation in the recognition of cerebral palsied speech. *Journal of Speech, Language, and Hearing Research*. 50, 115, 119

Rudzicz, F., Namasivayam, A., and Wolff, T. (2012b). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46(4):523–541. xiii, 7, 12, 26, 39, 48, 54, 101, 115

Sainath, T., Kingsbury, B., and Ramabhadran, B. (2012). Auto-encoder bottleneck features using deep belief networks. In *ICASSP*, pages 4153–4156. IEEE. 45, 83

Sainath, T., Vinyals, O., Senior, A., and Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. In *ICASSP*, pages 4580–4584. IEEE. 46

Salama, E., El-Khoribi, R. A., and Shoman, M. (2014). Audio-visual speech recognition for people with speech disorders. *International Journal of Computer Applications*, 96(2). 49, 124

Salamon, J. and Bello, J. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283. 48

Samuel, A. and Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, 71(6):1207–1218. 18

Schönle, P., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., and Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31(1):26–35. 98

Selouani, S., Dahmani, H., Amami, R., and Hamam, H. (2012). Using speech rhythm knowledge to improve dysarthric speech recognition. *International Journal of Speech Technology*, 15(1):57–64. 44

Shahamiri, S. (2021). Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pages 852–861. 45

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*. 63

Sy, B. and Horowitz, D. (1993). A statistical causal model for the assessment of dysarthric speech and the utility of computer-based speech recognition. *IEEE Transactions on Biomedical Engineering*, 40(12):1282–1298. 22

Takashima, R., Takiguchi, T., and Ariki, Y. (2020). Two-step acoustic model adaptation for dysarthric speech recognition. In *ICASSP*, pages 6104–6108. IEEE. 47

Takashima, Y., Nakashika, T., Takiguchi, T., and Ariki, Y. (2015). Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1411–1415. IEEE. 44, 82, 83, 135

Tatham, M. and Morton, K. (2006). *Speech production and perception*. Springer. 18

Therapists, S. . L. (2006). *Communicating quality 3: RCSLT's guidance on best practice in service organisation and provision*. Royal College of Speech and Language Therapists. 2

Thomas-Stonell, N., Kotler, A., Leeper, H., and Doyle, P. (1998). Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition accuracy. *Augmentative and Alternative Communication*, 14(1):51–56. 22

Tsujii, J. (2011). Computational linguistics and natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 52–67. Springer. 68

Tu, M., Berisha, V., and Liss, J. (2017). Interpretable objective assessment of dysarthric speech based on deep neural networks. In *Interspeech*, pages 1849–1853. 46

Turner, G., Tjaden, K., and Weismer, G. (1995). The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, 38(5):1001–1013. 23

Turrisi, R., Braccia, A., Emanuele, M., Giulietti, S., Pugliatti, M., Sensi, M., Fadiga, L., and Badino, L. (2021). EasyCall corpus: a dysarthric speech dataset. *arXiv preprint arXiv:2104.02542*. 58, 137

Vachhani, B., Bhat, C., Das, B., and Kopparapu, S. (2017). Deep autoencoder based speech features for improved dysarthric speech recognition. In *Interspeech*, pages 1854–1858. 83

Vachhani, B., Bhat, C., and Kopparapu, S. (2018). Data augmentation using healthy speech for dysarthric speech recognition. *Interspeech*, pages 471–475. 48

Wachter, M., Demuynck, K., Compernolle, D., and Wambacq, P. (2003). Data driven example based continuous speech recognition. In *Eighth European Conference on Speech Communication and Technology*. 4

Weide, R. et al. (1998). The carnegie mellon pronouncing dictionary. *release 0.6, www. cs. cmu. edu*. 39

Whurr, R. (1988). Clinical management of dysarthric speakers. *Journal of neurology, neurosurgery, and psychiatry*, 51(11):1467. 21

Wilson, B. and Blaney, J. (2000). Acoustic variability in dysarthria and computer speech recognition. *Clinical Linguistics & Phonetics*, 14(4):307–327. 22, 24, 25, 26

Witten, I., T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *Ieee transactions on information theory*, 37(4):1085–1094. 42

Wong, K., Yeung, Y., Chan, E., Wong, P., Levow, G., and Meng, H. (2015). Development of a Cantonese dysarthric speech corpus. In *Sixteenth Annual Conference of the International Speech Communication Association*. 58

Wrench, A. and Richmond, K. (2000). Continuous speech recognition using articulatory data. In *International Speech Communication Association*. 49, 98, 114

Xiong, F., Barker, J., and Christensen, H. (2018). Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. In *Speech Communication; 13th ITG-Symposium*. VDE. 3, 50, 57, 99, 114

Xiong, F., Barker, J., and Christensen, H. (2019). Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In *ICASSP*, pages 5836–5840. IEEE. 48, 57

Xiong, F., Barker, J., Yue, Z., and Christensen, H. (2020). Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In *ICASSP*. IEEE. 4, 6, 46, 57, 64, 82

Yılmaz, E., Mitra, V., Bartels, C., and Franco, H. (2018). Articulatory features for ASR of pathological speech. *arXiv preprint arXiv:1807.10948*. 50, 99, 114

Yılmaz, E., Mitra, V., Sivaraman, G., and Franco, H. (2019). Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech. *Computer Speech & Language*, 58:319–334. 6, 44, 47, 82, 83, 135

Yorkston, K., Beukelman, D., Minifie, F., and Sapir, S. (1984a). Assessment of stress patterning. *The dysarthria: Physiology, acoustics, perception, management*, pages 131–162. 58

Yorkston, K., Beukelman, D., Strand, E., and Hakel, M. (1999). *Management of motor speech disorders in children and adults*, volume 404. Pro-ed Austin, TX. 21

Yorkston, K., Beukelman, D., and Traynor, C. (1984b). *Computerized assessment of intelligibility of dysarthric speech*. CC Publications. 22

Young, S. (1996). Large vocabulary continuous speech recognition: A review. *IEEE signal processing magazine*, 13(5):45–57. 4

Yu, J., Markov, K., and Matsui, T. (2016). Articulatory and spectrum features integration using generalized distillation framework. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE. 139

Yu, J., Xie, X., Liu, S., Hu, S., Lam, M., Wu, X., and Ho, K. (2018). Development of the CUHK dysarthric speech recognition system for the UASpeech corpus. *Interspeech*, pages 2938–2942. 4, 57

Yue, Z., Christensen, H., and Barker, J. (2020a). Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition. In *Interspeech*. 8

Yue, Z., Loweimi, E., Cvetkovic, Z., Christensen, H., and Barker, J. (2022). Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition. In *ICASSP*. IEEE. 10

Yue, Z., Xiong, F., Christensen, H., and Barker, J. (2020b). Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition. In *ICASSP*. IEEE. 7, 91

Zeghidour, N. (2019). *Learning representations of speech from the raw waveform*. PhD thesis, Paris Sciences et Lettres (ComUE). 138

Zeghidour, N., Usunier, N., Synnaeve, G., Collobert, R., and Dupoux, E. (2018). End-to-end speech recognition from the raw waveform. *arXiv preprint arXiv:1806.07098*. 138

Zue, V., Seneff, S., and Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech communication*, 9(4):351–356. 86