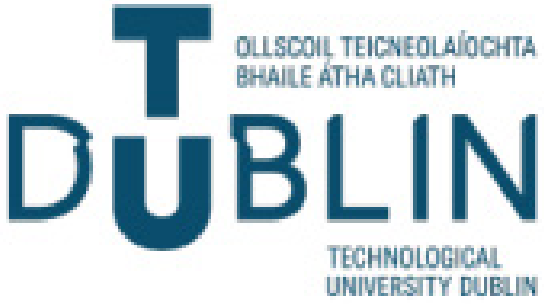
2022

Improving Dysarthric Speech Recognition by Enriching Training Datasets

Sophie Cullen
Technological University Dublin, Ireland

Improving Dysarthric Speech Recognition by enriching training datasets



Sophie Cullen

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computing (Data Analytics)

6 March 2022.

DECLARATION

I certify that this dissertation, which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed : Sophie Cullen

Date: 06/03/2022

ABSTRACT

Dysarthria is a motor speech disorder that results from disruptions in the neuro-motor interface and is characterised by poor articulation of phonemes and hyper-nasality and is characteristically different from normal speech. Many modern automatic speech recognition systems focus on a narrow range of speech diversity therefore as a consequence of this they exclude a groups of speakers who deviate in aspects of gender, race, age and speech impairment when building training datasets. This study attempts to develop an automatic speech recognition system that deals with dysarthric speech with limited dysarthric speech data. Speech utterances collected from the TORGO database are used to conduct experiments on a wav2vec2.0 model only trained on the Librispeech 960h dataset to obtain a baseline performance of the word error rate (WER) when recognising dysarthric speech. A version of the Librispeech model fine-tuned on multi-language datasets was tested to see if it would improve accuracy and achieved a top reduction of 24.15% in the WER for one of the male dysarthric speakers in the dataset. Transfer learning with speech recognition models and preprocessing dysarthric speech to improve its intelligibility by using general adversarial networks were limited in their potential due to a lack of dysarthric speech dataset of adequate size to use these technologies. The main conclusion drawn from this study is that a large diverse dysarthric speech dataset comparable to the size of datasets used to train machine learning ASR systems like Librispeech,with different types of speech, scripted and unscripted, is required to improve performance.

Keywords

Bias, disability and AI, speech augmentation, dysarthria, speech intelligibility, dataset bias, CycleGAN, wav2vec2.0, speech processing,

ACKNOWLEDGMENTS

I would like to thank my supervisor John Gilligan for all his support over the past six months and for his patience in helping me write this dissertation.

Additionally, I would like to thank my family, friends and work colleagues for supporting me throughout and for listening to my many rants.

Finally, I would like to thank the writer Jillian Weise who wrote the essay ‘Common Cyborg’ for showing me the impact technology can have on disabled people and the potential it has for creating real life cyborgs.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

1.1 - Introduction

This dissertation focuses on improving the recognition of dysarthric speech with speech recognition systems by enriching datasets with dysarthric speech and preprocessing dysarthric speech to make it more intelligible.

The World Economic Forum (2018) published the 'White Paper: How to Prevent Discriminatory Outcomes in Machine Learning'. In this paper, the issue of bias in machine learning datasets is addressed along with an analysis of the cause and ways to mitigate it in machine learning projects in the future. Bias in machine learning is caused by a number of factors including unrepresentative datasets, wrong model choice for the project, discriminatory model features, lack of human oversight and involvement, unpredictable and inscrutable systems and unchecked and intentional discrimination(World Economic Forum, 2018). Bias has the potential to violate human rights and it is the duty of businesses and institutions to respect these rights(Executive Office of the President of US, 2016)

Automatic speech recognition (ASR) technology is becoming increasingly more important in modern society as it is being integrated into services such as emergency response centers, domestic voice assistants and search engines(Feng et al., 2021). As ASR has become apart of machine learning practices, it also is susceptible to bias. Recent studies have shown that state of the art ASR systems struggle with the large variation of speech, that includes accents, age, gender and speech impairments(Feng et al., 2021). The main source of this bias is attributed to unrepresentative datasets that are mainly composed of native speakers, within a selective age range who do not have any speech impairments(Feng et al., 2021).

People with disabilities are also affected by bias in machine learning and are affected differently by bias in comparison to other protected attributes like gender, race or age (Trewin, 2018). This can be attributed to the wide diversity of the way disability can present itself and how people adapt(Trewin, 2018). Additionally, individuals can be hesitant to share sensitive information about their disability for the fear of discrimination.

The diversity of disability can be troubling for machine learning as it is designed to establish patterns and to form groups where in a dataset there may not be enough examples of a specific disability for it to be recognised(Trewin, 2018).

Dysarthria is defined as a motor speech disorder that results from disruptions in the neuro-motor interface and is characterised by poor articulation of phonemes and hyper-nasality(Rudzicz, 2010). Dysarthric speech has been shown to produce high word error rates when using state of the art ASR systems, for example, De Russis and Corno (2019) showed that Google Cloud Speech has a WER rate of 4.9% for normal speech but a WER of 59.81% for dysarthric speech. While there are many large datasets available for normal speech for training ASR models e.g; Librispeech (Panayotov et al.,2015), there are only a small amount of dysarthric speech datasets of a significant size, for example TORGO or the neumours database (Jiao et al., 2018). It can be difficult to collect dysarthric data due to factors such as variable recording conditions, unbalanced samples between speakers and uncontrolled body movements(Jiao et al., 2018).

1.2 - Research Background

As discussed in the introduction, individuals with disabilities face different a different kind of bias in machine learning due to the diversity of disability (Trewin,2018). As those who have dysarthric speech do so as a result of diseases such as ALS, Cerebral Palsy or Parkinson's disease which may limit their mobility, ASR systems can be of huge benefit in completing daily tasks unaided such as using a personal assistant like Apple's Siri to control appliances in the home, bringing independence to their lives(Cave & Bloch, 2021). Although other speech aids have been created to assist communication for people with dysarthria such as talking keyboard, eye tracking device , and communication board (Chen et al., 2020), most individuals desire to use their own voice for communication (Cave & Bloch, 2021).

Recent studies in machine learning have shown that the utilisation of large datasets in training can produce high performance by mining data-driven features directly from the data(Jiao et al., 2018). In the context of large vocabulary continuous speech recognition, this would require thousands of hours of speech consisting of varied accents, age and

environment to build a robust deep neural network acoustic model(Jiao et al., 2018; Yu & Li, 2017).

There are very few public datasets available for dysarthric speech that would be relatively large enough for a DNN (Jiao et al., 2018) and as a result research has pivoted to trying to solve the lack of data issue.

Dysarthia is characterized by poor articulation of phonemes. According to the TORGO database, for example, the deletion of affricatives and plosive phonemes in word final positions which do not occur in non-dysarthric speech(Rudzicz., 2010). Moreover, another common feature of dysarthric speech is its speed. All vowels produced by dysarthric speakers are significantly slower than those produced by healthy speakers at a 95% confidence interval and have been shown to be up to twice as long on average (Rudzicz., 2010).

Research has proposed speech synthesis as a solution to increasing data for dysarthric speech (Jiao et al., 2018; Hue et al., 2021, Shahamiri, 2021). Though moderately successful, it is computationally intensive and requires a large amount of time to produce large amounts of data. Additionally, it could be considered insensitive under the idea of active inclusion as it does not include dysarthric individuals as a part of the solution(World Economic Forum, 2018). Bragg et al. (2018) discussed the issue of using animation or avatars as a stand in for sign language data. This generated sign language data may fail to fully represent the complexity of the original data and this can be considered an issue when augmenting dysarthric speech data.

Furthermore, the use of GANs (General Adversarial Networks) and other models such as CNNs has become popular for speech enhancement and can be used to augment dysarthric speech to make it more intelligible(Phan et al., 2020 ;Chen et al., 2020; Yang & Chung; 2020). (to enrich datasets to look at preprocessing speech before its input)

Finally, Takashima et al. (2019) showed the potential for using dysarthric speech in two different languages to try increase training data for a model. It could be investigated if multi-language speech recognition models have a greater accuracy when recognising

dysarthric speech due to the greater potential to be exposed to a wider amount of speakers and phoneme variation.

1.3 - Research Problem

There is not enough adequate training data for modern ASR models to improve recognition accuracy when recognising dysarthric speech(Jiao et al, 2018). This can be considered an issue of bias in machine learning ASR datasets. This study asks " Can bias against dysarthric speech in an ASR model be mitigated by enriching training datasets with dysarthric speech and preprocessing dysarthric speech samples with general adversarial networks to improve its intelligibility?"

1.4 - Research Objectives

The research objectives are as follows:

- Obtain a dataset that represents dysarthric speech and review its suitability
- Review current literature to establish main issues that cause bias against ASR for dysarthric speech
- Review current literature to gather current techniques available to help mitigate bias against dysarthric speech in speech datasets
- Obtain a suitable ASR model for experimentation
- Set up a series of experiments to test techniques established in literature
- Review and evaluate experiments for suitability to research problem

1.5 - Research Methodologies

A review is conducted on previous literature based on the subject of improving the recognition accuracy of speech recognition systems with dysarthric speech. A suitable dataset is selected that represents dysarthric speech, the TORGO dataset (Rudzicz et. This work is therefore secondary research as the data being analysed and used for experimentation has been gathered by a third party.

A suitable and accessible speech recognition system is chosen to be utilised for experimentation, wav2vec2.0 (Baevski et al., 2020). The hypotheses, that a speech recognition systems accuracy when recognising dysarthric speech can be improved by enriching training datasets with dysarthric speech and preprocessing dysarthric speech with GANs can improve its intelligibility when being recognised by a speech recognition system, were derived from the literature reviewed in chapter two. In chapter three and four, experiments were designed to test these hypotheses and their results were compared, evaluated in line with the literature reviewed. This work is empirical research as conclusions are drawn from experimentation and observation . Quantitative methods are employed to determine and compare accuracy in the form of the Word Error Rate (WER). The WER was the metric of accuracy for all experiments and allows their results to be compared fairly.

Finally, this research can be considered deductive as it begins with a theory and its conclusion is based on the facts, literature, case studies and results presented.

1.6 - Scope and Limitations

The scope of the dissertation is to establish whether a fully representative dataset can be achieved for dysarthric speech by testing model adaptation and the augmenting of speech data.

This study faces a number of significant challenges. As speech recognition is a very computationally heavy task, the hardware used during experimentation limits the models chosen and their parameters along with the outcomes of the results.

Additionally, there are very few datasets available for dysarthric speech and those datasets are very small and only contain a small sample of people(Jiao et al., 2018).Therefore, only the specific type of dysarthria captured in the dataset used

Moreover, this study only addresses the improvement of bias within one specific ASR model, wav2vec2.0 and will not address the improvement of bias in other models.

1.7 - Document Outline

In chapter two, the literature review will describe the various different approaches to speech recognition and will focus on its machine learning implementation. The characteristics of dysarthria will be considered along with its relationship with speech recognition technology. Bias and fairness in machine learning practices will be detailed, followed by an overview of speech datasets for machine learning. The architecture of general adversarial networks will be discussed along with its applications for speech recognition. Finally, a review of all current literature on improving speech recognitions accuracy when recognising dysarthric speech will be explored along with a concluding overview of the research problem and its requirements for improvement.

In chapter three, the requirements defined by the literature review will be discussed and the CRISP-DM approach will be utilised to explain the design and implementation of the experiments. Experiments to evaluate the wav2vec2.0 Librispeech 960h models accuracy (measured in WER) for normal speech and dysarthric speech will be expanded upon. Further experiments to evaluate the wav2vec2.0 model fine-tuned on multi-language datasets accuracy for normal speech and dysarthric speech will also be expanded upon. An experiment to employ transfer learning with the wav2vec2.0 base model by training it on dysarthric speech will be detailed. Finally, preprocessing of dysarthric speech with GANs to try and improve its intelligibility before it is passed to the wav2vec2.0 multi-language model.

In chapter four, the results of the experiments testing the WER for normal speech and dysarthric speech when they are recognised by a wav2vec2.0 model trained on English speech alone versus a wav2vec2.0 model that has been trained on multiple languages is compared and the results are discussed in the context of the literature review from chapter two. The performance of the wav2vec2.0 model that had be retrained with dysarthric speech is detailed along with an analysis of why it rejected the hypothesis. Finally, the performance of the selected CycleGAN-VC-2 model to convert the dysarthric speech samples to the speech style of their normal speech control pairs is evaluated and an analysis of why it also rejected the hypothesis is detailed. There is a thorough discussion on the original aims of the study along with the outcome of the experimentation. The

approach of the experimentation is critiqued and suggestions are made to improve future work.

In chapter five, the entirety of the study is summarised and put in the context of the original aims of the study. The main conclusion of the study, that more dysarthric speech data is required for the improvement of ASR systems in recognising dysarthric speech is expanded upon. The contributions and impact of the results are detailed along with suggestions for future work.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This study is concerned with the improvement of ASR systems when recognising dysarthric speech. The proposed approach is to try enrich training datasets with dysarthric speech and use GANs for preprocessing dysarthric speech samples to improve their intelligibility. In this chapter we will look at the challenges of machine learning implementations of speech recognition. The architecture of machine learning ASR systems will be detailed and the problem of dysarthria in relation to this technology will be discussed. The issue of bias in machine learning along with speech recognition datasets will be highlighted to display lack of representation in certain ASR training datasets. Speech pre-processing and GANs in the context of ASR will be discussed to show their potential in being applied to the research problem. The conclusion will give an overview of the discussed methods viability for solving the research problem.

2.1.1 - An overview of Speech Recognition Approaches

There are three approaches to ASR: The acoustic phonetic approach, which is based on finding a finite set of distinctive phonetic units and applying labels to these sounds, the pattern recognition approach, which involves the pattern training and pattern matching of speech representations and the artificial intelligence approach, which is considered a hybrid of the previous two approaches(Anusuya & Katti, 2009). All three approaches are discussed in more detail below:

1.The acoustic phonetic approach - is the least adopted of the three, entails distinguishing between phonetic units within spoken language which are de scribed by a set of acoustic properties. These acoustic properties are highly variable and it is hypothesised that the rules that govern this variability between sounds are straightforward and can be easily learnt by a machine.

2.The pattern recognition approach - extracts patterns based on certain criteria

and to separate samples into one class or another. It is a model based approach and takes the form of a template method or a stochastic method. The tem plate method has become outdated but involved attempting to match a speech sample against a set of preexisting templates. Meanwhile, the highly successful and most widely adopting method until recently, the stochastic mod el uses probabilistic models to determine the word most likely spoken from a speech sample. The most popular model, the hidden markov modeling (HMM), is a finite state Markov model with a set of output distributions. Its transition parameters are temporal variability and the output distribution model parame ters are spectral variability. These variabilities are the centre of speech recognition.

3. The artificial intelligence approach - is a compromise between the previous two approaches. Some research involved using acoustic information to develop a model which develops classification rules for speech data. While other devel opments use the template based approach.

2.2 - Artificial Neural Network Speech Recognition Architecture

A typical speech recognition system usually consists of an acoustic front-end, which includes an input device e.g; microphone along with signal processing and feature extraction of the input, acoustic model, which represents the acoustic features for phonetic units to be recognised, language model, which is a body of constraints on the possible sequence of words that is acceptable in a given language and decoder, which finds the most likely word sequence based on the observation sequence and the acoustic-phonetic-language model(Karpagavalli & Chandra, 2016).

2.2.1 - Acoustic Front End

The acoustic front-end involves two processes of signal processing and feature extraction(Karpagavalli & Chandra, 2016). Feature extraction is considered one of the most important steps in the ASR pipeline as it distinguishes different speech samples

from one another. Each different type of speech has characteristics individual to it in its utterances. These characteristics are extracted through a wide array of different techniques each which work optimally with different ASR methodologies(Vimala & Radha, 2012). There is typically three stages to feature extraction, firstly, a version of spectra temporal analysis of the speech signal is performed which generates raw features which detail the envelope of the power spectrum of those speech signals, secondly, an extended feature vector is compiled with static and dynamic features and finally, in the final stage which only occurs in certain circumstances, these representations are transformed to be more compact and robust vectors which are then passed on to the recogniser (Karpagavalli & Chandra, 2016). There is no set criteria for what constitutes a suitable feature set for applications but they should fulfill the following; allow the system to differentiate between similar sounding speech sounds, allow the system to create an acoustic model for these sounds without the requirement for large amounts of training data and should create statistics which are mostly invariant no matter what speaker or speaking style(Karpagavalli & Chandra, 2016).

Two widely used feature extraction techniques include Linear Predictive Coding (LPC) and . Mel Frequency Cepstral Coefficients (MFCC).

LPC allows for an accurate estimate of speech parameter and is efficient speech model computation wise.LPC approximates a speech sample as a linear combination of past speech samples. It minimises the sum of squared differences, over a finite interval, from the actual speech samples to the predicted values, it determines an individual set of parameters or predictor coefficients. This analysis allows for the capability for calculating the linear prediction model of speech over time. Through the transformation of the predictor coefficients, a more robust group of parameters called cepstral coefficients are created(Vimala & Radha, 2012).

One of the most popular and widely used feature extraction technique is MFCC due to the frequency bands being positioned logarithmically, it approximates human auditory reception more imilarly than any other technique(Vimala & Radha,2012). The technique has many steps which are detailed below:

1.Pre-emphasis - The speech sample's high-frequencies are amplified to allow these regions to be more recognisable during training of a Hidden Markov model training and recogniser.

2.Windowing - The speech sample is sliced into discrete time segments using a window where N is the number of milliseconds wide and at offsets of M mi li-seconds long. Hamming windows are often employed as they stop the sharp edges that can occur with a rectangular window.

3.Discrete Fourier Transform - A DFT is an algorithm which transforms signals in the time domain to its representation in the frequency domain. The DFT is exercised on the windowed signal and represents the magnitude and phase information of the signal.

4.Mel Filter Bank - This step tailors the output to be relevant to human hearing by altering it with the logarithmic Mel scale. Humans hearing is less strong at frequencies above 1000 Hz. A bank of triangular filters are applied to the DFT spectrum which said filters spaced equally below 1000 Hz and spaced logarithmically above 1000 Hz. The output is known as the Mel spectrum.

5.Log - The Mel spectrum co-efficients are created by taking the loga rithm of he Mel spectrum.

6.DCT - The final phase, a discrete cosine transform is performed on the Mel spectrum co-effcients and the Mel-cepstral coeffcients of 13th order is created.

7.Delta MFCC Features - To capture changes in the speech samples from frame to frame, the first and second derivative of the MFCC co-effcients are calculat ed(Karpagavalli & Chandra, 2016).

2.2.2 - Acoustic Models

Choice of acoustic model makes a large impact on the output of an ASR system. The

acoustic modelling of speech refers to the statistical representation of the speech feature vector sequences gathered from the previosu feature extraction stage. A Hidden Markov Model (HMM) is a common statistical model used to create acoustic models.There is a wide range of choices available including segmental models, supersegmental models (including hidden dynamic models), neural networks, maximum entropy models, (hidden) conditional random fields and end to end models (Karpagavalli & Chandra, 2016). For the context of this literature review the context of an acoustic model will be explored for a HMM and an end to end model.

Hidden Markov Model

For decades, the Hidden Markov Model has been the most widely adopted choice for a mainstream large vocabulary speech recognition system and boasted the best results. The acoustic model for a HMM maps speech input to a feature sequence which is usually a phoneme or sub phoneme sequence(Wang et al., 2019). The acoustic model is built by introducing a large speech database, known as a speech corpus, and applying training algorithms to it to create a bank of statistical representations for each phoneme. Each phoneme has its own individual HMM. The decoder then tries to find a match for input speech in the acoustic model. For each spoken word, W, in an input sequence, it is broken down into a series of basic sounds known as base phones. The acoustic model produces the probability of a specific observation occuring given a base phone(Karpagavalli & Chandra, 2016). There are three current approaches to calculating this probability, HMM-GMM (Gaussian Mixed Model), HMM-ANN (Artificial Neural Networks) and HMM-DNN(Deep Neural Networks).

GMM is considered a generative learning. It was considered state of the art before the introduction of DNNS. With the introduction of the expectation-maximization (EM) algorithm in the eighties, GMM's took over as a tool to represent the relationship between HMM states and the acoustic input.

The use of GMMs was advantageous as with a sufficient number of components they can model probabality distrubutions to any level of accuracy needed and are easy to fit with the use of the EM algorithm. They are so successful that it was hard to produce a

method that beat their performance for acoustic modelling(Hinton et al., 2012).

Despite this, they have one major drawback. GMMs are not very statistically efficient for modelling data that lies on or near a nonlinear manifold in the dataspace. Speech is created by modulating a relatively small amount of parameters of a dynamic system which implies its structure underneath is far less complex than what is initially apparent in a window that holds hundreds of coefficients(Hinton et al., 2012).

ANN is considered a discriminative approach which involves either using a discriminative model or applying it to a generative model. Neural networks are successful at enabling discriminative training to occur in a natural and efficient fashion. Despite their usefulness in the classification of short time units like isolated phones and words, they do not perfom well for continuous recognition tasks as they have poor ability in modelling temporal dependencies. As a result, they have found success as a pre-processing approach for feature transformation and dimension reduction for HMM based recogntion (Karpagavalli & Chandra, 2016). For example, the accuracy of a HMM-GMM can be improved if the input features are altered with bottleneck features gener
ated by neural networks(Hinton et al., 2012).

DNN is considered as representation learning or unsupervised feature learning. When they were first introduced, they were trained discriminatively. In the past decade, research showed a significant increase in performance could be achieved by adding an initial stage of generative pretraining that disregards the end goal of the system. This pretraining is far more helpful in deep neural nets than in shallower ones, particularily if labelled training data is scarce. This reduces overfitting and the time required for discriminative fine-tuning with backpropagation, which was the main reason neural networks were chosen over DNNs to replace GMMs in the nineties. They are particularily adpet at exploiting information in neighboring frames and from modeling tied context-dependent states. Pretraining helps reducing overfitting and the time required for fine tuning. It has overtaken GMM as they have no issue modeling multiple simultaneous events within one frame or window as it has the ability to use different subsets of its hidden units to model different events and have out performed GMMs in many tasks such as large vocabulary speech recognition. Unfortunately, they have one drawback,

they struggle to make good use of large cluster machines to train on massive data sets in comparison to GMM. This is negated by the fact they not require as much data to achieve the same performance (Hinton et al., 2012).

End to End

The end-to-end model is a system that directly maps input audio sequences to a sequence of words or other phonemes. Most systems consist of the following parts, an encoder, which maps an input speech sequence to a feature sequence, an aligner, which recognises the alignment between the feature sequence and language and finally, a decoder, which decodes the final identified result. As an end to end model is a complete structure, these divisions may not exist for every model and at times it can be difficult to tell which part of the system does which subtask. This is wholly different from the previously discussed HMM which is composed of multiple modules, its acoustic model as a seperate enity, where in this architecture it is fully intergrated. Additionally, End to End replaces multiple modules with a deep network, recognising the direct mapping of acoustic signals into label sequences without a carefully-designed intermediate states. Therefore, there is no need to excercise backend processing on the output. This enables the model to joint train and persue globally optimal results where a HMM can't. As a result, the end to end model simplifies the building of speech recogntion systems(Wang et al., 2019). End to end models consist of Connectionist temporal classification (CTC) , RNN-Transducer and attention based models.

Connectionist temporal classification (CTC)

The CTC model was developed as a way to avoid the requirement of frame-level alignment of target labels for training utterances(). Before CTC, when researchers were modelling time-domain features with RNN or CNN they faced a data alignment problem as both models loss function are defined at each point in the sequence, therefore in order to facilitate training, the aligment between the RNN output sequence and the target sequence must be known(Wang et al., 2019). By solving this issue, it allowed for a greater employment of DNN in speech recogntion. Additionally, it simplified model architectures by negating the need for further processing on outputs to retrieve a final transcrip-

tion where traditional models in the past would output phonemes(Wang et al., 2019).

The model operates by editing the target labels with an additional blank symbol. For each input sequence of a certain length, the encoder encodes it into a feature sequence, the vectors of which have a dimension one greater than that of the length of the vocabulary. Through a softmax operation, the feature sequence is converted to a probability distribution sequence. Here the probabilities for whether a label or blank label occurs at the output at each time step in the sequence(Wang et al., 2019).

Furthermore, using these calculated probabilities the input sequence is then mapped to a certain label at each input frame. The mapping of the input sequence to a path is thought of as a hard-aligning process. This process assumes that each element in the output sequence is independent of each other, therefore this encoder is entirely an acoustic model and does not have the ability to model language (Wang et al., 2019). Finally, if the output path length and input speech length are not equal, usually the transcription is much shorter than the corresponding speech sequence, then multiple paths are aggregated into a shorter label sequence. This includes the deletion of blank and the merging of consecutive identical labels (Wang et al., 2019).

RNN-Transducer

The RNN-Transducer model exploits the nature of the CTC model being similar to an acoustic model in an ASR system and augments the encoder of the CTC model with a separate recurrent prediction network over the output symbols(Wang et al., 2019).Meanwhile the prediction network can be thought of as similar to a language model.

The RNN Transducer model makes up for some of the short-comings of the CTC model. Where CTC can only map input sequences to output sequences that are shorter than it, the RNN-Transducer can map an input to any finite, discrete output sequence. Additionally, where CTC is unable to model interdependencies for the output sequence as it assumes indepencence between output elements, the RNN-Transducer can model these dependencies(Wang et al., 2019).

Although similar to CTC, the RNN-Transducer model takes a different approach to their path generation processes and path probability calculation diverge. The model consists of three subnetworks, a transcription, prediction and joint network. The transcription network acts as an acoustic model as an encoder, the prediction network is a decoder that acts like a language model modelling interdependcies within the output label sequences and finally the joint network aligns the input and output sequences(Wang et al., 2019).

The design of the RNN-Transducer architecture through the joint network enables the interdependence between the input sequence and output sequence by using both the language model and acoustic model to calculate probability, allowing them to be training together(Wang et al., 2019).

Despite its advantages over CTC, the RNN-Transducer has its own drawbacks.Research has found that this type of model is difficult to train from scratch, pre-training its various parts has been proposed as a solution. Additionally, due to the models flexibility, its calculation process produces many unreasonable paths, which is also a problem for any model that begins enumerating all possible paths and then aggregates them(Wang et al, 2019).

Attention Based Models

An Attention Based model, for example a Listen-Attend-Spell (LAS) model consists of an encoder network, like the previously discussed RNN Transducer model but differs by using a single decoder with an attention method at each output step to assign different weights to each vector in a specific sequence. The following output step is then determined by a historical output sequence and a weighted summation of the encoding result sequence. This allieviated the problem of requiring all the text information being encoded to a fixed length vector, allowing even long sentences to have a good encoding effect(Wang et al., 2019).

Additionally, the encoder-decoder along with the attention method doesn’t need the pre-segment aligment of data. It learns the soft alignment between input and output sequences with attention, which is highly beneficial for ASR systems(Wang et al., 2019).

2.2.3 - Language Model

A language model is a group of constraints that denote what is an acceptable sequence of words in a language. These may be represented by rules of generative grammar or statistics on each word pairing estimated in a training corpus. Humans are naturally adept at distinguishing words with similarly sounding phones as they will understand the context of the sentence and have general knowledge regarding what words and phrases can occur in that context. A language model provides this context to a ASR system. Usually, language models are trained using n-gram probabilities that are estimated by observing sequences of words in large corpora of text that contain millions of word tokens. This perplexity is then reduced on training data. Despite this, reduced perplexity does not necessarily mean more accurate results. it then follows, that algorithms that have the ability to improve language models through their interaction with ASR are very appealing. Common language models are bigram and trigram models. These models hold a particular number of words in a sequence, respectively. Tools for language modeling include the CMU Statistical Language Modeling (SLM) Toolkit and the Stanford Research Institute Language Modeling Toolkit(Karpagavalli & Chandra, 2016).

2.2.4 - Decoder

The decoder aims to establish the most likely word sequence W, given an observation sequence O along with the acoustic-phonetic-language model. This can be solved using dynamic programming algorithms. Instead of evaluating all possible model paths that generate O, they focus on a single path through the network which is the best match for O. The Viterbi algorithm is frequently used for this purpose. In the context of large vocabulary tasks, it would be unfeasible to consider every possible word during the recursive part of the Viterbi algorithm, therefore a beam search is employed to only consider paths above a certain probability threshold when extending paths to the next time step. This may speed up the searching process but decrease decoding accuracy. An issue of the algorithm is that it considers the best paths at time t to be an extensio of the best paths ending at time t-1, which is not always correct. The path that seems least probable at the start may be the best as a whole. An extended Viterbi and forward-backward algorithm has been implemented to try mitigate this problem(Karpagavalli & Chandra,

2.3 - Charcteristics of Speech Recognition Systems

Speech recognition systems come in a number of different varieties for different purposes. They can be differentiaed by the different characteristics discussed in this section. The speech type, utterance approach and style that a ASR system is trained to recognise can impact how it performs with other types of speech it is then asked to recognise.

Speaker type

Speaker dependent -where the system is tailored to recogise one speaker but its accuracy decreases for other speakers

Speaker independent-where the system is tailored to recognise a wide array of speakers but has a trade off in accuracy as a result

Speaker adaptive- a system which a compromise between the two previous sys tems, they are tailored to be able to learn new speaking patterns but this func tionality can affect processing times and accuracy(Malik et al. 2020).

Utterance approach

Isolated words - where the user is required to take pauses after each word and the system processes them one word at a time

Connected words - where users speak multiples words at a time and the system processes multiple words at a time(Malik et al., 2020).

Utterance style

Continous speech - where the user is able to speak naturally to a point, and the system processses the speech sequence as a whole.

Spontaneous speech- where the user speaks completely naturally with the inclu

sion of non speech utterances such as coughing. These systems are harder to produce due to the need for a larger vocabulary and the ability to differentiate between speech and non-speech(Malik et al., 2020)

2.4 - Metrics to Evaluate Speech Recogniton Systems

There have been many different metrics to evaluate ASR systems performance. Different papers use different evaluation metrics and this can make it hard to compare performance. The two popular accuracy metrics are disscussed in this section. WER (Word Error Rate) is the chosen metric for this study as it is the most used metric in the literature regarding speech recognition and its perfomance with dysarthric speech. Additionally, the model chosen in chapter three, wav2vec2.0 and its training dataset Librispeech use WER in both of their related papers to measure accuracy (Baevski et al., 2020; Panayotov et al., 2015).

Two popular methods to measure the accuracy of ASR include:

1.Word error rate - Word error rate or WER solves the issue of measuring the accuracy of ASR when the output produced is not the same length as the ground truth. It calculates the error on word rate rather than on phoneme level. WER is calculated by dividing the number of substitutions, deletions and inser tions in the prediction divided by the number of total words in the ground truth(Malik et al.,2021) .

2. Word recognition rate - Word Recognition Rate (WRR) is considered a variation of WER . The WRR is calculated by taking the total number of words predicted completely (number of words in the ground truth minus the substitu tions and deletions) minus the number of insertions in the prediction divided by the total number of words in the ground truth(Malik et al., 2021).

2.5 - Recent Advancements in Open Source Automatic Speech Recognition Systems and Methodologies

Before the introduction of machine learning to the task of ASR, with the increasing capabilities of the hardware, the existing techniques for ASR, which were based primarily on the hidden Markov model (HMM) with Gaussian mixture output distributions had begun to plateau in performance(Deng & Li, 2013). The improvement of computational resources coupled with the increase of data available for training driven by the internet, the application of machine learning could be applied to speech recognition to improve it and push its accuracy further than it had before(Deng & Li, 2013). A brief overview of recent state-of-the-art open-source machine learning frameworks will be discussed below and their suitability for experimentation for the research problem will be discussed later in this chapter.

2.5.1 - wav2vec2.0

Wav2vec2.0 is a framework for the self-supervised learning of speech representations. Representations (the 'vec' vectors) are learnt from speech audio (the 'wav') alone and then labelled speech is used for fine-tuning. Developed by Facebook AI, many state of the art models trained on the Librispeech dataset and other multi-language datasets are available to use and deploy from Fairseq which is a sequence modeling toolkit also made available by Facebook(Baevski et al., 2020; Ott et al., 2019).

2.5.2 - DeepSpeech

DeepSpeech is an open-source speech-to-text engine provided by Mozilla which uses model trained on techniques based on the 2014 paper by Hannun et al. It is a end to end speech recognition system and as a result does not require hand-designed componenent to model speaker variation or background noise but learns it directly. This technique allows the model to be more robust to these effects than models previous and other state of the art systems. Phoneme dictionaries are not required. The model is an RNN that utilises multiple GPU's and data synthesis techniques to gather a significant amount of

2.5.3 - seq2seq

seq2seq is a machine learning framework that is concerned with the training models to convert a sequence from one domain to another. This has been applied to speech recognition. They have become popular as they are able to fold acoustic, pronunciation and language models into a single network in comparison to traditional ASR systems that are composed of seperate modules and as a result make training easier. Examples of seq2seq models include Recurrent Neural Network Transducer (RNNT) ,Listen, Attend and Spell (LAS) , Neural Transducer, Monotonic Alignments and Recurrent Neural Aligner (RNA). Initally, their performance was not comparable to state-of-the-art speech recogniton systems but recently have began to outperform them(Chiu et al., 2018).

Additionally, researchers have developed open-source toolkits to aid in the development and training of these models, two prominent toolkits, openseq2seq and Kaldi are discussed below:

1.Openseq2seq - In 2018 Kuchaiev et al. presented Openseq2seq for Nvidia. It is a TensorFlow based toolkit for training sequence to sequence models. It can handle a wide range of tasks such as neural machine translation, automatic speech recognition and speech synthesis. It has been shown to have bench marks for speech recognition tasks give state of the art performance with less training time in comparison to other models(Kuchalev et al., 2018). Additionally, offers a library of widely used encoders and decoders and con tains a large group of models for speech recognition including DeepSpeech2 and Wav2Letter+ (Kuchalev et al., 2018).

2.Kaldi ASR - Kaldi is an open-source toolkit developed by Povey et al. in 2011. It provides a speech recognition system that is based on finite-state trans ducers coupled with scripts and documentation for building a complete speech recognition system. Its core library includes supports for the modelling of arbitrary phonetic-context sizes, acoustic modeling with subspace Gaussian mixture models (SGMM) , standard Gaussian mixture models, along with all commonly used linear and affine transforms(Povey et al., 2011).

2.6 - Dysarthria and Speech Recognition

Dysarthria is a motor speech disorder which is classed by the underlying neuropathology that causes it(Enderby et al.,2013), for example, ALS (Amyotrophic lateral sclerosis), Multiple Sclerosis, brain injury or tumor, stroke among many others. It is related to symptoms of laryngeal function, airflow direction, and articulation which cause difficulties of speech quality and intelligibility. Dysarthria can be split into six relevant groups: flaccid dysarthria related to lower motor neuron impairment, spastic dysarthria related to damaged upper motor neurons which are linked to the motor areas of the cerebral cortex, ataxic dysarthria which is mostly caused by cerebellar dysfunction, and hyperkinetic dysarthria and hypokinetic dysarthria, which is asscociated with a disorder of the extrapyramidal system. The final group is named as a mixed dysarthria and is caused by damage in more than one areas previously mentioned, resulting in an overlap of speech characteristics of at least two of the other groups(Enderby et al.,2013).

Dysarthric speech is distinctly different to normal or healthy speech and can be a barrier to communication. The severity of an individuals Dysarthria can be measured by using standardised assessments such as the Frenchay Dysarthria Assement which analyses the motor functions of its subject and Assessment of Intelligibility of Dysarthric Speech which analyses the intelligibility of words and sentences of its subjects along with their speaking rate(Rudzicz et al., 2010).

During the creation of the TORGO database for Dysarthric speech by Rudzicz et al. in 2010, they outlined the differences they found between the Dysarthric speech and normal speech collected in their research. They found dysarthric speech contained more mispronouced phonemes, in particular plosives phonemes (/p,b,t,d,k,g/ in the English Language), than in healthy speech. In addition,dysarhric speech contains a high prevalence of deleted affictives and plosives in final word positons which was not mirrored in the healthy speech data and took twice as long on average pronouncing vowels in comparison to healthy speakers. Furthermore, the greatest pronunciation difference was found between nasal consonants and the dysarthric speech was discovered to have a measurable amount of non speech noise such as glottal noise, repitition or noisy swallowing problems.

Collecting a fully representative Dysarthric speech dataset is difficult due to the many different types of Dysarthria and as the severity of the Dysarthira greatens, the length of time a person is able to speak for without fatigue or weakening in speech quality decreases. For this reason, many datasets of Dysarhric speech are very small, containing a small amount of participants, a small amount of source material that has been read from and will usually only focus on one type of dysarthria. Due to a lack of data and only a few publicly available datasets such as the TORGO Database (Rudzicz et al., 2012) majority of experiments have been performed on small local datasets built by universities.

With the rapid development of machine learning techniques over the decades discussed, the excitement of progress left little room to ensure artificial intelligence would be built with all users in mind. As most of these emerging technologies are built upon datasets,the contents of those datasets are of the upmost importance and these algorithms cannot function without them. Therefore, if there are underlying biases within the datasets, they will transfer over and affect these algorithms.

As speech recognition technology improved, researchers were highly interested in its application in assisting disabled users in their day to day life. In the 1980s, various researchers investigated the usefulness of speech recognition technology for disabled persons with normal speech for activities such as controlling a wheelchair. Into the 1990s, the use of this technology was explored for people with dysarthric speech. At the time researchers and clinicians identitfied two main reasons speech recognition technology could be beneficial for people with dysarthria; coupled with the use of a speech synthesiser, an ASR system could improve the intelligibility of dysarthric speech and make communication easier for those individuals. In addition, it could be used as a tool to improve the articulation accuracy of the dysarthric speaker, like speech therapy, where the system would be set up to give feedback on the quality of articulation(Patel, 2000).

At the time researchers and clinicians identified two main reasons speech recognition technology could be beneficial for people with dysarthria; coupled with the use of a speech synthesiser, an ASR system could improve the intelligibility of dysarthric speech and make communication easier for those individuals. In addition, it could be used as a

tool to improve the articulation accuracy of the dysarthric speaker, like speech therapy, where the system would be set up to give feedback on the quality of articulation(Patel, 2000).

Furthermore, Researchers realised that people with dysarthric speech may have additional physical disabilities that would prohibit or complicate their use of a keyboard. Without the use of a keyboard, the independence of these individuals is severely limited as they are unable to express or articulate themselves in writing, which may become a barrier to certain services or parts of society e.g.; academia. Before the availability of commercial ASR technologies, assisted communication for dysarthric speakers included the use of objects, pictures, alphabet boards, sign language among others as alternatives to speech. Other devices such as head pointer devices, sip and puff switches and scanning devices were also employed. These methods were found to be too slow and in some cases too onerous on the user, who by the nature of their disabilities could become fatigued easily. Moreover, these forms of communication were still too limiting, as many dysarthric speakers would still use speech to communicate emotion or to gain attention(Patel,2000).

Researchers found success with Dragon Dictate systems in the ninites as a writing tool for users with moderate Dysarthric speech, reaching accuracy rates of 80% with the Dragon Dictate 1.01A. Though it was also found the better the intelligiblity of the speaker, the training times would be less and the accuracy would be higher. This posed a problem for users with more severe forms of Dysarthria. Accuracy rates were variably and the required training time needed was much higher, which could put the technology out of reach from certain users who would fatigue quickly due to their condition. Despite this users with severe Dysarthria still wanted to communicate with speech recogntion with its limitations. A study was conducted by Treiranus et al. (1991) to compare to access techniques, the traditional scanning method or speech recognition along with the scanning method. Out of the eight participants, seven prefferred speech recognition along with the scanning method due to an increased rate of communication along with a decrease in fatigue when using speech as an input(Patel,2000).

With the demand for a usable speech recognition system, researchers began developing techniques to try and improve accuracy for dysarthric speech. They recognised that as dysarthric speech had different characteristics than normal speech different approaches would have to be used.Research focused on speaker dependent models, which would capture the large variablity between different speakers with different severities of dysarthria. Chen and Kostov (1997) built an ASR system that reached 90% accuracy on a very small vocabulary of 10 digits (zero to nine) with fifteen repititions of each. HMM techniques were used to model the speech and MFCC (Mel frequency cepstrum coefficients) were employed to convert the speech samples into a time varying representation(Patel, 2000). Despite some progress, systems designed for users with severe dysarthria lagged behind those for users with normal speech. Hux et al. (2000) conducted a survey of the performance of three state-of-the-art speech recognition systems when dictating speech. While the control speakers with non-dysarthric speech achieved high accuracy rates, see figure 2.6.1, the accuracy rates for the dysarthric speakers , see figure 2.6-2, was significantly lower. The three systems reviewed were speaker dependent and required the participants to complete training excercises provided by the systems. For each system, five training sessions were performed (Probes), and in most cases the accuracy improved with each session. The researchers attributed the gap in performance to issues that arise with the intelligibility of dysarthric speech, such as slower rate of speech and the way the systems were set up to capture data, very long training times to achieve high accuracy which may be too strenuous for certain individuals with dysarthria(Hux et al., 2000). This paper also highlighted the need to address the needs of the individual when selecting an ASR system as dysarthria can be caused by a number of conditions which will dictate what each individual is capable of, for example discrete speech may be more suitable for certain individuals who take a long time to form sentences(Hux et al., 2000).

Figure 2.6-1. *Results of three ASR systems when dictating normal speech. Reprinted from "Accuracy of three speech recognition systems: Case study of dysarthric speech", by Hux et al., 2000, Augmentative and Alternative Communication, 16(3), 192. Copyright 2000 by ISAAC.*

	Microsoft Dictation		Dragon NaturallySpeaking®		VoicePad Platinum®	
	Preselected	Novel	Preselected	Novel	Preselected	Novel
Probe 1	83.50	78.31	88.35	85.57	77.67	75.53
Probe 2	89.32	75.28	91.26	84.38	80.58	83.51
Probe 3	93.20	77.67	92.23	93.33	94.17	86.92
Probe 4	85.44	76.84	96.12	85.71	89.32	85.47
Probe 5	88.35	86.14	92.23	89.61	87.38	80.37

Figure 2.6-2. *Results of three ASR systems when dictating dysarthric speech. Reprinted from "Accuracy of three speech recognition systems: Case study of dysarthric speech", by Hux et al., 2000, Augmentative and Alternative Communication, 16(3), 191. Copyright 2000 by ISAAC.*

	Microsoft Dictation		Dragon NaturallySpeaking®		VoicePad Platinum®	
	Preselected	Novel	Preselected	Novel	Preselected	Novel
Probe 1	45.63	45.83	68.93	54.17	35.92	52.86
Probe 2	47.57	51.14	66.02	55.84	45.63	57.89
Probe 3	54.37	58.57	67.96	68.67	59.22	64.86
Probe 4	49.51	44.44	69.99	66.28	55.34	57.35
Probe 5	64.08	64.71	64.08	64.86	38.83	50.79

Pushing forward to modern times, as mobile devices have become a main staple in most peoples pockets and the intergration of technology in general requires most members of society to have a basic level of computer literacy, accessible technology for users with disabilities is imperative so they can be fully involved with society. Unfortunately, with the pivot towards cloud based platforms for ASR, their usability for users with dysarthria has been mixed.

For a modern selection of ASR systems which are speaker independent, De Russis and Corno (2019) tested the usability of ASR cloud platforms for users with Dysarthria. Google Cloud Speech had the best results with an average WER of 59.81%, a WER of 16.11% for those with mild dysarthria, a WER of 78.21% for speakers with severe dysarthria yet for users with normal speech the WER was 3.95%. All systems tested very poorly with users with severe dysarthria with the number of sentences being transcribing

correctly almost numbering to zero.

The cause of machine learning based ASR systems poor performance in recognising dysarthric speech can be attributed to the lack of dysarthric speech data available(Jiao et al., 2018). There are very few datasets available of an adquate enough size for training such as the TORGO database, Nemours database and the UA speech dataset(Jiao et al,. 2018 ; Kim et al., 2008).

Fortunately companies and researchers are taking action to improve these accuracy rates. In 2019, Google announced they would be recruiting people with speech disabilities to donate speech samples. They have partnered with two non profit organisations, ALS Therapy Development Institute and ALS Residence Initiative, to obtain speech samples from individuals with ALS. They aim to use these samples to help build a model that can understand 'impaired speech'. They cite lack of data for why an algorithm like this hasn't been built before(The ASHA Leader, 2019).

2.7 - Bias in Machine Learning and Active Inclusion

The World Economic Fourm (2018) defined Active Inclusion as ' The development and design of ML applications must actively seek a diversity of input, especially of the norms and values of specific populations affected by the output of AI systems'.

There is discrimination and lack of fairness and in the world and this has sometimes transferred over into algorithms(World Economic Forum, 2018). Due to the data driven nature of AI systems, if the dataset used to train a system contains biases, it will learn these biases and include them in its predictions. In some circumstances, depending on the algorithm, these biases can become amplified. In addition, algorithms by design can display biased behaviour even if the data does not contain any biases. The outcomes of these biases in the real world can perpeptuate them further, help inform users descisions and cause biases to appear in data for training algorithms in the future(Mehrabi et al., 2021).

There are many kinds of bias. For the context of this literature review, the bias experienced by disabled users will be focused upon.

Disabled people are considered a minority group in society, therefore even when datasets are created from a representative pool of people, miniority groups only make up a small amount of the dataset. Therefore the data its its nature becomes imbalaced. This imbalance becomes an issue in the context of machine learning as these minorities can be ignored by models and then lead to poor performance with the classification of these minorities(Trewin, 2018).

Additionally, disabled people can be cautious to release their senstive information due to fear of discrimination(Trewin, 2018). Therefore, datasets can become unrepresentative due to an expectation of discrimination. Additionally, in Europe, new GDPR laws that enforce annomity of data can potentially cause bias against disabled people as the explicit information about disabililty that can be passed on to the algorithms to apply fairness tests and corrections is not accessible(Trewin, 2018). There fore, the handling of the data of disabled people must be executed differently to those who are not disabled. In some cases, the inclusion of too many categories in a training dataset can affect performance and it must be assessed whether disabled persons require a systen of their own(Trewin, 2018).

Under the definition of Active Inclusion, it is integral to include disabled persons in the designs of machine learning systems. It must also be discussed with developers and disabled users whether a pure machine learning approach is appropriate to solve issues of bias. The datasets must be evaluated for possible data patterns that may not interact well with a machine learning algorithm and establish the proper fairness technique to deal with outliers in the dataset(Trewin, 2018).

There have been some developments in developing toolkits to mitigate bias in machine learning. In 2002, the SMOTE algorithm was proposed to tackle this issue(Fernandez et al., 2018). This algorithm used a oversampling approach to rebalance datasets which employs synthetic examples as its main tool. The new data is created by interpolating between several minority class instances that are within a defined neighborhood(Fer-

nandez et al.,2018). In this way, the algorithm focuses on the values of the features and their relationship rather than just the data samples as a whole(Fernandez et al., 2018).

Furthermore, Bellamy et al. (2018) produced the AI Fairness 360 toolkit for IBM. This toolkit contains a multitude of algortihms designed to mitigate various types of bias and fairness metrics to evaluate bias in models and datasets.

2.8 - Datasets for Speech Recognition

The topic of datasets for speech recogntion can be a controversial one. With the curation of a dataset, there is always the potential for exclusion. In the context of speech recognition, who is included and how the samples were captured have a huge impact on the functionality of the system.

In the early years of speech recognition, systems could only handle isolated words or phonemes. In 1952, Bell Labs created the first complete speech recogniser, which could recognise 10 digits from a single speaker, therefore the requirement for vast and detailed datasets was a long way off. As the technology improved, the size of the datasets grew to push progress further(Anusuya & Katti, 2009).

Initially, speech datasets where very small and select but as the capability of speech recognition systems increased, there becamer a demand for a large standardised speech dataset to further research and for comparison between systems. Therefore, MIT, TI, and SRI jointly created the TIMIT corpus and was published in 1988. The dataset contained a total of 6,300 sentences from 639 speakers, representing over 5 hours of speech material. Before the takeover of the internet by big tech giants such as google, gathering datasets of this nature was a very time consuming task. The creators of the database designed TIMIT to balance “utility and manageability” and held small samples of speech from American speakers of different sexes in eight different dialects recorded in a number of different enviroments. For this reason it is widely used in a vast amount of speech recognition research(Garofolo et al., 1992).

With the advent of big data, and the superabu bundance of information and me-

dia available from the internet and the corresponding internet of things, the ability to capture large amounts of speech data was capitalised.

Panayotov et al. (2015) published the Libirispeech corpus, which was derrived from the Libirivox project, a group of worldwide volunteers who read and record texts in the public domain and publish them as audibooks for free to download from their website and digital libraries hosted on the internet. The dataset contains over 1000 hours of speech sampled at 16 kHz and is in MP3 compressed format. DNNs trained on this corpus achieved promising WER rates.

2.8.1 - Defining Datasets

Parameters of Interest for Speech Datasets:

Intra-speaker variability consists of variable speaking rate, changing emotions or other mental variables and enviromental noise.

Inter-speaker variability consists of differences that occur due to the individual variability in vocal systems involving source excitation, vocal tract articulation, lips and/or nostril radiation.

The acoustic enviroment the samples are recorded in are also vital in categorising a dataset for purpose. They are either recorded in noise free enviroments or with enviromental noise from a home or office enviroment for example.

Inter-speaker variability is particularily relevant and important to the context of this literature review. What does and does not get included in a dataset intended to train a speech recognition system informs what people it will be able to interact with. There are many variables to be considered, for example, how old are the people sampled in the dataset. Usually, the vocal chords of children are shorter and lighter than the vocal chords of adults causing the fundemental and resonant frequencies to be higher and therefore, causing a greater range of spectral variability(Yu et al., 2021). Futhermore, the accent and its corresponding dialect and vocabulary can also affect ASR performance

Futhermore, the accent and its corresponding dialect and vocabulary can also affect ASR performance depending how much it differs from the training data (Koenecke et al., 2020).

In the past, when datasets were being curated, not many of these factors were taken into account. With the rise of awarness around fair A.I. and inclusivity, more representative datasets are being created. The IEEE Spoken Language Technology Workshop held a challenge in 2021 for the improvement of children's ASR by releasing a large dataset for benchmarking. In addition, in 2020, a similiar challenge was enacted for accented english by Interspeech. By providing more inclusive datasets to researchers, it gives great potential to improve the robustness of ASR systems. To allow speech data, in all its forms to be open source is problematic. A corpus like Librispeech is easy to legitimise logistically, the participants willingly donated their data to the project and the content that they are reading from is within the public domain. Trying to capture unscripted speech, for example, is far trickier as for it to be truly unscripted it has to be of a sensitive and private nature to the speaker. With the introduction of GDPR by the EU, there are much more constraints on the collection and dissemination of data by individuals and institutions. Therefore, it becomes harder to collect and allow for speech datasets to be made open source(Trewin, 2018).

Dysarthric speech falls under the problem of lack of representation within speech datasets. There are very few datasets with dysarthric speech available to the public, and those datasets are relatively small with only a small range of dysarthric severity and a smaller range of utterance style and approach(Jiao et al., 2018).

2.9- Preprocessing for Speech Recognition

The preprocessing of speech has been successful in improving the robustness and accuracy of hearing aids, mobile communication and speech recognition systems. Preprocessing of speech for ASR systems is known as Speech Enhancement. In general, speech enhancement deals with the improvement of degraded speech signals by reducing noise. In this context, it is assumed the noise is additive and the noise characteristics change very slowly in comparison to the signal(Chaudhari & Dhonde, 2015).

Speech enhancement techniques can be classified by the number of microphones employed, either single channel, dual and multi-channel. Multi-channel speech enhancement provides the best performance of the three yet due to the convience of single channel, a significant amount of research is pursued in its direction (Chaudhari & Dhonde, 2015).

The types of noise that can affect speech signal include periodic noise, wide band noise, interfering speech and impulsive noise. Each has different algorithms and methods available to try lessen their effect. Traditional techniques include:

Periodic Noise Removal

Techniques to remove periodic noise include stationary filters, adaptive filters and or transform domain filters. Stationary filters involves a group of notch filters, like T-filters, whch act as a comb filter to remove the noise. Adaptive filters use a forward prediction error as an inverse filter which will attempt to remove the noise. Finally, the transform domain filter technique exploits the ability to be able to be observe and manipulate periodic noise in the transform domain(Chaudhari & Dhonde, 2015).

Wide Band Noise

Wide band noise can be mitigated by Spectral Subtraction (SS) and adaptive cancellation. Spectral subtraction estimates the noise spectrum and subtracts it from the noisy spectrum. This is achieved by obtaining the magitude spectrum, which is the square root of the original signal spectrum, and combining it with the phase of the noisy signal. The signal is then gathered using the Inverse Discrete Fourier Transform (IDFT).

Adaptive cancellation removes noise by using an impulse response which must cause the filtered channel noise to match the signal noise and therefore can be tuned to remove it. The coefficients are continuously updated until the signal outputs noise is at a minimum(Chaudhari & Dhonde, 2015).

Interrferring Speech

Speech enhancement is not optimal in the case of two speech signals interrferring. Pitch separation is sometimes employed to try isolate voices, this is achieved using tracked voiced segements. Addtionally, a comb filter can be used to recover the voice of interests harmonics if the pitch value is known, the same is true for the transform domain technique. With the assumption of pitch values for each speaker, a discrete fourier transform (DFT) can be employed to trace the harmonics of the fundemental frequencies of each speaker. If it is possible to isolate the DFT outputs, then an IDFT is taken to recover the individual voice of interest(Chaudhari & Dhonde, 2015).

General Adverserial Networks

Furthermore, in recent years, GANS, general adversarial networks, have been employed for speech enhancement to improve the robustness of speech recognition.

GANS are an unsupervised generative models which learn to create convincing samples from a dataset from low-dimensional, random latent vectors(Donahue et al., 2018). They are composed of two models, a generator and a discriminator, which are played off each other in an adversarial framework. The generator maps latent vectors taken from a selection of samples known prior which the discriminator decides is real or fakes(Donahue et al., 2018). In 2018, Donahue et al. showed GANS were useful in removing reverberant noise. They achieved this by using a refinement of the SEGAN method, FSEGAN. The noisy speech samples were converted from the time domain into time frequency spectral representations which were then fed into the fully convolutional generator in second long windows which enhances the noisy spectra. The discriminator then decides if they original or enhanced samples are real or fake. Through this, the model is then trained to output clean speech. Their results were postitive but did not exceed results achieved by using multi-style training, which is training with multiple speech styles and quality of recording. With the implementation of FSEGAN and multi-style training the WER increased slightly, still an improvement overall, but the authors hypothesise that this may be due to artifacts left by the enhancement process.

2.10 - General Adversarial Networks and Speech Recognition

General adversarial networks have been shown to be useful in the domains of speech conversion and speech synthesis(Jiao et al., 2018: Donahue et al., 2018). There applications and architecture are discussed here for the context of improving the intelligibility of dysarthric speech.

GANS or General Adversarial Networks originate from the zero-sum game in game theory. This is defined as a game where the gains of one opponent is bound to bring loss to the other opponent, where the gains and losses add up to zero(Cheng et al., 2020). In the context of a GAN, a discriminator judges samples created by a generator and tries to distinguish whether it is authentic or false. As the images created by the generator become more realistic, it becomes more difficult for the discriminator to distinguish between the two. Furthermore, early on in training, it is easier for the discriminator to identify poor quality samples as false. Eventually, an equilibrium is reached in the game known as the Nash equilibrium, which is a strategy for both opponents to maximise their own interests within the game. This process is a main component of GANS(Cheng et al.,2020). Additionally, the generator does not have direct access to the authentic images and can only learn through interacting with the discriminator, only the discriminator has access to both the synthetic and real samples.

2.10.1 - Training General Adversarial Networks

When training GANS, parameters that maximise the classification accuracy of the discriminator and parameters that maximise the confusion to the discriminator are found. In figure 2.9-1, the training cost function is depicted.

$$\max_{\mathcal{D}} \min_{\mathcal{G}} V(\mathcal{G}, \mathcal{D}),$$

where

$$V(\mathcal{G}, \mathcal{D}) = \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \log \mathcal{D}(\mathbf{x}) + \mathbb{E}_{p_z(\mathbf{x})} \log (1 - \mathcal{D}(\mathbf{x})).$$

Figure 2.9-1. *Training cost function of GAN. Reprinted from "Generative Adversarial Networks: An Overview", by Creswell et al., 2018,IEEE Signal Processing Magazine, 35(1), 6.*

While training the parameters of one model are updated while the parameters of the other are fixed. It follows that the generator, G, is at is optimum when the discriminator predicts 0.5 for all samples drawn from x. Preferably, the discriminator will be trained until its optimum point in tandem with the current generator, which after this point is updated. Though in some circumstances it may be trained for a small number of iterations and then updated alongside the discriminator. In this instance a non saturating training criterion would be used for the generator, which replaces minGLog(1-D (G(z))) with maxGLogD(G(z)) (Creswell et al., 2018) .

While training the parameters of one model are updated while the parameters of the other are fixed. It follows that the generator, G, is at is optimum when the discriminator predicts 0.5 for all samples drawn from x. Preferably, the discriminator will be trained until its optimum point in tandem with the current generator, which after this point is updated. Though in some circumstances it may be trained fora small number of iterations and then updated alongside the discriminator. In this instance a non saturating training criterion would be used for the generator, which replaces minGLog(1-D (G(z))) with maxGLogD(G(z)) (Creswell et al., 2018).

Unfortunately, training GANS can be challenging. It can be difficult to get the pair of models to converge, the generative model may collapse due to having similar samples as inputs and the discriminator model loss may converge too quickly to zero which affects the gradient updates of the generator (Creswell et al., 2018).

2.10.2 - General Adversarial Network Architectures

Fully Connected General Adversarial Networks

A GAN architecture that uses a fully connected neural network for both generator and discriminator models. This type of GAN was useful for small, uncomplicated

datasets(Creswell et al., 2018).

Convolutional General Adversarial Networks

In traditional neural network architecture, previously mentioned, each layer is mapped to the next by an activation function. The disadvantage of fully connected layers is that the model can become affected by a large number of parameters such as weights, slow training convergence and poor generalisation effect. One solution to this is Convolutional neural networks or CNN, which is a feed forward neural network with convolutional calculation. It uses the back-propagation algorithm to train its weights and in turn obtain classification results. Like the usual NN structure, CNNs have a convolutional calculation. It uses the back-propagation algorithm to train its weights and in turn obtain classification results. Like the usual NN structure, CNNs have a input layer, hidden layer and output layer. The hidden layer consists of a convolutional layer, pooling layer and full connection layer. Weights are shared from the convolutional layer and this helps to mitigate the issue of poor training efficiency due to overbearing amounts of parameters in NNs (Cheng et al., 2020).

CNNs were successful in supervised learning but were under utilised in unsupervised contexts. This changed with the introduction of DCGAN by Radford et al. in 2015 and allowed for unsupervised feature extraction from images. In this particular model, the fully connected hidden layer was removed to allow training to be more efficient. Additionally, the activation functions for the generator, ReLU, and the discriminator, Leaking ReLU are altered. Furthermore, the tangent h function is used for the output of the generator to better fit the pixel range (Cheng et al., 2020).

Conditional General Adversarial Networks

Conditional GANs have a generator and discriminator which are class conditional. With this they have the ability to provide better representations for multi modal data generation (Creswell et al., 2018).

General Adversarial Networks with Inference Models

In their initial iteration, GANs did not have the ability to map a given observation, x, to a vector in latent space, this is known as an inference mechanism. Solutions like adversarially learned inference (ALI) and bidirectional GANs have been proposed. They allow for a simple yet useful extension which provides an inference network which a discriminator can analyse in joint pairs. In this context, the generator is composed of two networks, the inference network or encoder and decoder. Therefore, the discriminator receives pairs of vectors to distinguish whether they are an authentic tuple consisting of genuine image samples and their encoding or fake samples and their corresponding latent-space input to the generator. This model is thought of as a reconstruction. Usually, the constancy of the reconstructed data synthesised using this framework is low. Their constancy is hypothesised to be improved upon with an additional adversarial cost on the samples and their reconstructions (Creswell et al., 2018).

Adversarial Autoencoders

Autoencoders are networks which are composed of an encoder and decoder. They learn to map data to internal latent representations and back out again. They learn a mapping from both the encoder and decoder, these two mappings are then composed to reconstruct an image as close as possible to the original image. They are thought to be similar to perfect-reconstruction filter banks that are common in image and signal processing(Creswell et al., 2018).

2.10.3 - GANS vs traditional signal processing

GANS differ from traditional signal processing techniques such as PCA, ICA, Fourier, and wavelet representations due to the complexity their models can reach while mapping vectors from latent space to image space. This complexity is possible due to the non-linearities of the generator networks of which can almost be of any random length(Creswell et al., 2018).

2.10.4 - Latent Space

Like neural network models such as word2vec and VAE (Variational Autoencoders)

GANs model their own representations of data they are presented with and produce structured geometric vector spaces for different domains. Usually, the data is mapped to a vector space which has smaller dimensions than the data itself, therefore it must find novel structures in the data and represent it efficiently. The representations of this data at the inception end of the generator can be highly structured and could support high level semantic operations. Additionally, certain GANs have an additional encoder which supports inverse mapping. This gives GANs the ability to develop and explore concept vectors within its latent space. For example, when analysing images, this could be a facial expression. These vectors can then be applied to scaled offsets within the latent space to direct the actions of the generator(Creswell et al., 2018).

2.10.5 - Applications of General Adversarial Networks

The potential for the application of GANs is vast. They are particularly adept when dealing with image data. DCGAN has been used for image classification, LAPGAN has been used for image synthesis, CycleGAN and pix2pix models have been used for image to image translation and finally SRGAN has been used for heightening the resolution of low resolution images (Creswell et al., 2018).

In the context of speech processing, the application of GANs is highly promising but has a lot of groundwork still to be covered in comparison to the amount of research conducted on images. Research has shown success when using raw audio and its visual representations to GANs for synthesis. Donahue et al. (2019) investigated the validity of using waveforms and spectrograms for generating one second slices of audio with GANs. The spectrogram iteration, called SpecGAN, adopted an representation that allowed for approximate inversion and bootstrapped the two dimensional deep convolutional GAN framework. Meanwhile the waveform iteration, WaveGAN, flattens the DCGAN architecture into one dimension which resulted in a model with the same amount of parameters and numerical operations as its two dimensional analog counterpart. This in turn has provided a way for other image generation methods to be applied to waveforms.

In adapting GANs for audio, Donahue et al. use PCA to illustrate the intrinsic differences between images and audio. The principal components of images consist of inten-

sity, gradient and edge characteristics meanwhile audios form a periodic basis which degrade the audio into constituent frequency bands. It follows that they are more likely to show periodicity than image and as a result correlations span large windows. This informs the size of the filters with receptive fields required to process audio. Both these techniques could also extend to classification and speech preprocessing.

2.11- Recent Advancements in Automatic Speech Recognition for Dysarthric Speech

With the advent of awareness about fairness and active inclusion within the development of artificial intelligence, there has been a wealth of research over the last decade in improving ASR systems for speakers with dysarthria. There have been many avenues perused, from the preprocessing of data to the use of General Adversarial Networks.

There has been a rich bank of research gathered to discover and improve pre-processing techniques to improve the intelligibility of Dysarthric speech. Tolba and El-Torgoman(2009) displayed improvement in recognition accuracy by modifying the first and second formants, which are considered more fundamental than the others, of words spoken by dysarthric speakers. Rudzicz (2011) proposed further techniques to alter dysarthric speech including splicing to correct dropped and inserted phoneme errors, morphing the speech in time to bring in to the rate of normal speech, altering the frequency of vowel formants to tackle the hyper-nasality of dysarthric vowel pronunciation and finally by adapting a voice transformation system by Kain et al. (2007). Although the voice transformation gave poor performance, the splicing technique gave the highest improvement in accuracy and highlighted the importance of lexically correct phoneme sequences.

Moreover, Chen et al. (2018) developed a Gated Convolutional-based Voice Conversion System to improve the intelligibility of dysarthric speech. This CNN model was trained with parallel dysarthric speech and target healthy speech and then deployed to alter the dysarthric speech samples which averaged an improvement in recognition accuracy from 17.1% to 81.0%.

Furthermore, Hetsugi et al. (2020) proposed a method for "phonological control of vowels using a standardized space to control vowels in the normalized articulation space, normalized for speaker individuality". This was achieved by analysing the formant frequencies of the inputed dysarthric speech, these formant frequencies are then mapped to a normalised articulation space that the individual differences of vocal tract length for each speaker and finally the initial speech samples are synthesised with the newly made formant frequencies from the normalisation phase.

Additionally, there has been a wealth of research in recent times in using CNNs samples to train with, as the greater the amount of samples, the greater the chance of improving accuracy (Jiao et al., 2018). Vachani et al. (2018) proposed a solution to the lack of available datasets for dysarthric speech by altering healthy speech to hold the characteristics of dysarthric speech, specifically the speed and tempo. An overall improvement of 4.24% and 2% was achieved in comparison to training with healthy speech data alone.

Additionally that year, Jiao et al. (2018) attempted to mitigate this issue for clinical applications by simulating dysarthric speech. This was orchestrated through adversarial training, where a GAN used both generative and discriminative models to transform healthy speech samples to have similar spectral features to dysarthric speech. Also, the pitch and speed of the normal speech was altered to mimic dysarthric speech. The generated speech was found to be acoustically and spectrally similar to actual dysarthric speech but has yet to be deployed on a large scale machine learning ASR system.

Furthermore Hu et al. (2021) also attempted to bolster the amount of Dysarthric speech data available by generating synthetic samples. This was achieved by altering various deep learning based text to speech generators configurations and adapting them using transfer learning. The models were pre-trained with normal speech data and then experimentation was carried out on freezing the layers to preserve the skills learnt in order to then retrain with dysarthric speech. By using this simulated speech data in training, a improvement of 5.6% in word recognition accuracy was achieved although no difference in accuracy occurred when recognising mild dysarthric speech.

It is difficult to compare the results of the three different attempts of synthesis-

ing dysarthric speech as they all utilise different datasets with different vocabularies of various sizes. In addition, as the datasets differ, the metric for measuring accuracy changes (WER (word error rate) vs WRA(word recognition accuracy)making the comparison of perfomance difficult. According to Hu et al. (2021) their configuration has an advantage over that proposed by Jiao et al. as their use of GANs which can be more troublesome due to issues with non-convergance. Additionally, Jiao et al. (2018) method requires healthy speech samples to be manually added to be converted while Hu et al. generate their own samples. Meanwhile, Hu et al. also claim that the pre-processing effects applied by Vachani et al. (2018) requires more testing on a larger vocabulary of diverse words and phonemes to prove the alterations can capture the wide range of characteristics of dysarthric speech.

To date, the most successful implementations of ASR systems for users with dysarthric speech have been isolated word ASR models and conventional ASR algorithms, such as artificial neural networks (Shamiriri & Binti Salim, 2014). Despite recent research being successful in recognising small vocabularies of dysarthric speech, there has been no large-vocabulary dysarthric speech recognition system available yet (Zaidi et al., 2021).

Most dysarthric speech recognition systems in use today are built on statisitical approaches such as HMM which models the sequential structure of speech signals. These HMMs are usually based on GMMs (Gaussian Mixed Models) which are regarded as the optimum statistical representation of the spectral distributions of speech waveforms. This probability based modelling remains effective when it is joined with flexible time dimension representation of uncertainity and may be useful in the context of dysarthric speech synthesis. Unfortunately, it is not suitable in the deployment of dysarthric speech recognition as it requires a large amount of speech data to become robust which the pool of dysarthric speech sets available is very small(Zaidi et al., 2021).

Based on these challenges, research has pivotted towards Deep Neural Networks as a possible solution, of which convolutional neural networks (CNNs) and long short-term memory (LSTM) networks have been shown to be very promising. In a performance comparison of DNN architectures to date for dysarthric speech recogntion by Zaidi et al. in 2021, the performance of various auditiory based input features, models (HMM-

GMM, CNN, LSTM) and activation functions were all analysed and weighed. The CNN architecture was shown to be the most successful, and was robust enough even with very severe cases of dysarthric speech. It demonstrated its capability in the context of a speaker-dependent application, in the capture of timing artifacts and is suitable as a robust recogniser of dysarthric speech.

Moreover, the best audio based input parameter was found to be Perceptual lin
ear prediction (PLP) over Mel-frequency Cepstral coefficients (MFCCs) and Mel-frequency spectral coefficients (MFSCs). The optimum activation function for the CNN model was shown to be the Poly1ReLU which as a polynomial activation function, learns nonlinearity and approximates continuous real values of input data in order to provide the best discriminative model.

In 2021, Shahamiri et al. developed an end to end deep learning based dysarthric speech ASR system called Speech Vision. Improving on previous attempts to date, Speech Vision tries to recognise the shape of a word spoken from dysarthric speakers by converting them into a visual feature representation, voicegrams, instead of trying to recognise phonemes. This ensures the model is more robust to the nature of variable phonemes associated with dysarthric speech and the difficulty of labeling them, The shape of the words was learnt by the model using transfer learning. First the model is training on normal speech representations, neuron freezing is then employed and the model is retrained on dysarthric speech. They also created synthetic voice data using text-to-speech generators to boost the amount of training samples available. Speech Vision achieved a WRA 61.11% just using dysarthric speech and 64.71% with using dysarthric speech along with synthetic speech samples. The study was limited by the production of the synthetic data as they could only produce one additional sample per word for each speaker.

2.12- Conclusion

There is still much groundwork to cover in the development of a robust LVCSR(Large Vocabulary Continuous Speech Recognition) system for individuals with dysarthria.

There is an argument to be made that a continuous speech recognition system maybe unsuitable for certain individuals with dysarthria who may find it easier communicating with single words due to the fatigue they may suffer from speaking at length(Hux et al., 2000). It could be investigated if ASR systems should be tailored in this instance to single word communication. Additionally, ASR systems developed for dysarthric speech tend to have higher word recognition accuracy when focusing on single words. It would be imperative that there would be survey conducted to assess what these individuals would prefer and what would suit their particular needs best and for what context.

Furthermore, a recurrent theme among the majority of papers reviewed within this literature review is a lack of dialogue between researchers and Dysarthric individuals. For a task that is attempting to stop these individuals from being ignored by advancements in technology it is unusual that they are not included in these conversations. There is little to no information on what users with Dysarthric speech would be looking for from an ASR system(Cave & Bloch, 2021), which could impede a fully functioning systems design and deployment in the future(World Economic Forum, 2018).

Moreover, lack of data is continuously cited as a roadblock in the creation of an ASR system for dysarthric speakers. To date there is still no publicly available substantially large enough dataset to allow for the progression of the technology and importantly, to be able to benchmark and compare approaches(Jiao et al., 2018).

Finally, experimentation with GANs to create dysarthric speech and to augment it is still only beginning. As most experimentation has adapted normal speech to dysarthric speech, these synthetic samples still haven't completely reached the authenticity of the samples they are mimicking. GANs require large amounts of data and could be potentially held back by the requirement for data in pairs for style transfer tasks like voice conversion. New emerging techniques such as 'CycleGAN' could useful, to allow for more data to be used by disregarding paired utterances.

CHAPTER THREE: EXPERIMENT DESIGN AND METHODOLOGY

3.1 - Introduction

The experimentation in this chapter looks to evaluate if enriching training datasets with dysarthric speech for a wav2vec2.0 model improves its performance with dysarthric speech and if using GANs for style transfer can improve the intelligibility of dysarthric speech by converting the speech style to normal speech. Additionally the experimentation will investigate if using models trained with a single language or multiple languages has an effect on performance.

The rationale of all the experiments is based of the main issue identified in the literature in chapter two, the lack of available dysarthric speech data. Experiment five, transfer learning with the wav2vec2.0 model attempts to try enrich the training dataset with dysarthric data. While experiment six, addresses the issue of the lack of data by trying to improve the intelligibility of the dysarthric data to try mitigate the issue.

3.1.1 - Requirements from Literature Review

From the findings of the literature review, in order to experiment to improve the performance of an ASR system with dysarthric speech, a model must be selected. The main issue that affect accuracy was identified as lack of representation of dysarthric speech within machine learning speech training datasets. In order to try increase the representation, more data samples are required. Without access to more dysarthric data due to a lack of availability, machine learning techniques such as enriching datasets and pre-processing to try and increase representation and reduce bias must be employed.

3.1.2 - Chapter Overview

This chapter details all experiments involved with improving the accuracy of an automatic speech recognition system in recognising dysarthric speech. Detailed descriptions of the datasets, model frameworks are discussed as well. This is then followed

by rationale for the two different approaches in experimentation to try and improve accuracy.

The first set of experiments involve comparing a wav2vec2.0 model trained on the English language alone against a wav2vec2.0 model trained on multiple languages. Both the dysarthric speech samples and their matching normal speech control speakers were tested to obtain word error rates for comparison.

The fifth experiment attempts to use transfer learning to retrain a wav2vec2.0 model with dysarthric speech samples to try and improve the WER obtained in the previous four experiments.

The sixth and final experiment uses CycleGAN-VC-2 (Kaneko et al., 2019) to convert dysarthric speech samples to the speech style of normal speech to try and improve the dysarthric speech's intelligibility. The speech will be tested on the wav2vec2.0 multi-language model to compare WER results.

3.1.3- CRISP-DM

This chapter follows the CRISP-DM methodology, which was released in 2000 and has become an industry standard for data science projects. It consists of a six step process model from business understanding to deployment (Schröer et al., 2020) . A brief overview is detailed below:

1. Business Understanding - The project is assessed to gather an overview of all available and required resources. The main goal of the project is identified along with the model type and success criteria. A project plan is also created (Schröer et al., 2020).

2. Data Understanding - Data is collected for the project from data sources and is assessed and explore to ensure it is of adequate quality (Schröer et al., 2020).

3. Data Preparation - Data is then selected by a inclusion and exclusion crite

ria. Data cleaning is applied to data of bad quality. Depending on the project, attributes may have to be derived from the data (Schröer et al., 2020).

4. Modelling -Informed by the data and project problem, modeling techniques and parameters for the model are chosen. The assessment of the model is chosen by evaluating all evaluation criteria and selecting the most suitable (Schröer et al., 2020).

5. Evaluation- The results are then evaluated in respect to the project problem and objectives. Based on this evaluation, further actions will be decided. At this stage, the entire process is reviewed in general (Schröer et al., 2020).

6. Deployment - When the project is considered complete, its deployment is planned along with its maintenance and a plan for monitoring its perfor mance(Schröer et al., 2020).

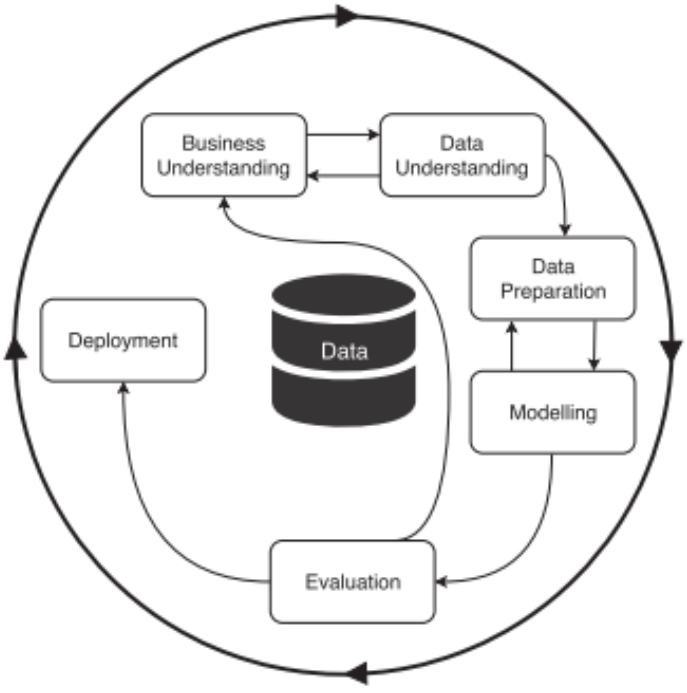


Figure 3.1.3-1. *CRISP-DM Lifecycle. Reprinted from "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories", by Martínez-Plumed et al. ,2021, IEEE Transactions on Knowledge and Data Engineering, 33(8), 3049. Copyright 2019 by IEEE.*

3.2 - Project approach and design

The improvement of dysarthric speech with ASR systems is a multi-faceted problem not only in a technical sense but in an ethical sense also. For decades representation in speech datasets for training ASR systems has been very poor. In addition to lack of representation for people with dysarthria, children's speech and speech of people with a non-native accent has also been neglected(Yu et al., 2021; Shi et al., 2021). As a result research has pivoted to two approaches to tackle the issue of lack of data. Firstly, Preprocessing techniques to improve the quality and intelligibility of dysarthric speech to increase accuracy with ASR systems pre-trained on normal speech(Rudzicz, 2011; Rudzicz, 2013; Chen et al., 2020) . Secondly, artificially trying to boost the amount of data available by creating synthetic data samples using General Adversarial Networks(-Jiao et al.,2018; Hue et al., 2021;Shahamiri, 2021; Harvill et al., 2021).

Wav2vec2.0 was decided as model framework for these range of experiments over other current popular frameworks such as DeepSpeech and Seq2Seq due to its high performance by training on large amounts of unlabeled raw audio and fine-tuning on small amounts of labeled data(Baevski et al., 2021). The model achieved a 7.9% WER ("other" dataset which includes inaccuracies) training on only 10 minutes of labelled data(Baevski et al., 2021). This is advantageous for training with dysarthric speech as there isn't a large amount of labeled speech data available. It was a considered decision to choose a pre-existing model rather than training from scratch. Since the landscape of ASR has pivoted towards very large datasets and deep learning(Malik et al., 2021), dysarthric speech can be left behind due to the aforementioned lack of data. There is a limit to how well a DNN model can be trained to perform with these small representations of dysarthria. Therefore, to integrate dysarthric speech recognition with the most up to date and successful ASR models, the use of pre-existing models is required.

Additionally the chosen frameworks for this set of experiments is dictated by the quality of hardware available.

The purpose of these experiments is to determine whether a preprocessing approach

or model adaptation via transfer learning or both is optimal in improving the wav-2vec2.0 model in recognising dysarthric speech(the reduction of the WER for dysarthric speech).

3.2.1 - Research Questions

Does employing transfer learning by retraining a Wav2Vec 2.0 model with dysarthric speech samples improve its WER when recognising Dysarthric speech?

Does a model trained on multiple languages rather than a single language model improve the WER when recognising Dysarthric speech?

Can using GANS for speech style transfer improve the intelligibility of Dysarthric Speech Samples when been recognised by a Wav2Vec2.0 model trained on normal speech?

3.3 - Dataset description and Understanding

All experiments conducted make use of either or both the Librispeech ASR corpus and the TORGO Database(Acoustic and articulatory speech from speakers with dysarthria). The Librispeech dataset is composed of speakers reading from audiobooks from the public domain from Project Gutenberg. Specifically, the recordings originate from the Librivox project, where individuals volunteer to read texts for free. It boasts 1000 hours of speech audio sampled at 16kHz. The dataset has been carefully constructed ensuring there is an equal balance between speaker gender and that no speaker overlap occurred between training, development and test sets, ensuring that each sample is attributable to a single speaker. The audio in the Librispeech dataset passes through two alignment stages. The first involves using the Smith-Waterman alignment algorithm to determine the best alignment between the audio and its corresponding text. Then the audio is split into 35 seconds or less using a dynamic programming algorithm and is only split if the silence exceeds 0.5 seconds to ensure confidence in each piece of split audio. The second alignment stage entails removing segments carried over from the first stage which may be inaccurate. Inaccuracies include mistakes in Project Gutenberg texts, reader-introduced insertions deletions, substitutions and transpositions, and involuntary disfluencies.

Two versions of the dataset exist, one clean dataset with all inaccuracies removed and a second “other” dataset which retains all inaccuracies. In the context of research dealing with the dysarthric speech, the “other” dataset is used for all experiments as speech data with inaccuracies is more likely to have similar acoustic properties to dysarthric speech, than clean normal speech. Furthermore, the speech samples are then split further with two different approaches. The training data is split if there is silence interval exceeding 0.3 seconds and for the test data the samples were only split at sentences breaks, which is more intuitive for language modeling(Panayotov et al., 2015).

It is important to note the content within the Librispeech dataset which primarily contains texts from the 19th and 20th centuries which may be semantically different from modern speech and may influence the performance of models trained on it(Panayotov et al., 2015).

The TORGO dataset is a collection of speech samples and 2D and 3D articulatory features from dysarthric speakers who have either cerebral palsy or amyotrophic lateral sclerosis. These particular types of dysarthria were selected for this database as they are the two most common causes of speech disability (Rudzicz et al., 2010).

Therefore, the experiments discussed in this chapter will only deal with the recognition improvement of this specific type of dysarthria. The dataset consists of samples from seven individuals with dysarthria, four male and three female, all with cerebral palsy bar one participant with ALS. All the acoustic data in this dataset was collected through two microphones, the first being an Acoustic Magic Voice Tracker array microphone which uses amplitude information to locate the speaker within a range and mitigate noise by spatial filtering and typical amplitude filtering. This audio was recorded at 44.1 kHz. Meanwhile the second microphone was a head mounted and recorded audio at 16 kHz. Although the audio recordings from the first microphone are of better quality, it is necessary to use the audio from the head microphone as it has the same sampling rate as the samples from the Librispeech dataset in order to be able to draw comparison during the experiments. For certain participants, there were multiple recording sessions as people with dysarthria may fatigue quicker due to their medical conditions which may affect the quality of the recordings. Though it may be argued that it could be useful to capture fatigued dysarthric

speech to have a more representative dataset, only speech from well rested participants is included(Rudzicz et al., 2010).

The prompts used for the participants consist of non-words, short words, restricted sentences and unrestricted sentences. The non-words consist of plosive consonants in the presence of high and low vowels and high and low pitch vowels, which hold specific dysarthric speech characteristics such as the hyper-nasality of vowel sounds. The short words consist of various selections such as 50 words from the word intelligibility section of the Frenchay Dysarthria Assessment, 360 words from the word intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech, the 10 most common words in the British National Corpus, phonetically contrasting pairs that affect intelligibility and repetitions of English digits such as 'yes' or 'no'. or 'no'. The restricted sentences include 162 sentences from the sentence intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech, 460 TIMIT-derived sentences used as prompts in the MOCHA-TIMIT database, "The Grandfather passage" and curated phoneme dense sentences such as "The quick brown fox jumps over the lazy dog". Finally, the unrestricted sentences were developed by prompting the participants with 30 images form the Webber Photo Cards: Story Starters collection. This speech is the only unscripted speech in the database and as a result will contain the most inaccuracies and non speech noise.

It is important to note this dataset is dated in its purpose for training an ASR system. With the introduction of deep learning to ASR research, a dataset this size is no longer optimal. The publicly available version of this database is only 18GB in size, which, half of the samples are normal speech control samples (Rudzicz et al., 2010).

3.4 - Dataset Cleaning and Preprocessing

The TORGO dataset required a significant amount of cleaning. Every sample in the dysarthric speech dataset and the control normal speech dataset had to be played to check its annotation text matched the contents of the file. A small proportion of samples were mislabeled, for example, a sample would be labelled as a word but when the file was played it contained no speech sounds. Additionally, certain samples were over two

minutes long, unsegmented recordings of the data collection sessions and were removed as they were too long to be processed by the model and segmented versions of these recordings already exist as samples within the dataset. All of the corresponding annotation texts were stored in separate text files. In order to prepare the dataset to be instantiated as a pytorch dataset, a excel sheet had to be collated where each wav file had its location and text contents detailed.

After all the text annotations were inputed into the excel sheet, all unscripted speech samples were removed as the Librispeech 960h model was only trained on scripted samples, in order for the WER to be fairly calculated. Additionally, all punctuation was removed from the corresponding text labels as the Librispeech models vocabulary does not recognise punctuation, apart from blank spaces and apostrophes which were not removed. Also, all corresponding labels were converted to lower case in line with the Librispeech model.

3.5 - Machine Configuration

All experiments were ran on jupyter notebooks on WSL 2 on Windows 11. The machines GPU was a Nvidia GeForce 2080 Ti along with a AMD Ryzen 7 3700X 8-Core processor and 16GB of RAM.

3.6 -Wav2Vec2.0 Architecture

The Wav2Vec 2.0 framework entails the self-supervised learning of representations from raw audio data. The model is composed of a multi-layer convolutional feature encoder which accepts raw audio as input and gives latent speech representations as output for a number of time steps. This encoder is composed of a group of blocks which have a temporal convolution, layer normalisation and a GELU activation function. The input is normalised to a zero mean and unit variance. The encoders total stride decides the number of time-steps that are then fed to the transformer. When the representations are transferred to the transformer, it builds representations that capture information from the entire sequence. The output from the feature encoder is transferred to a context network

which adheres to the transformer architecture. A convolutional layer is then used to act as a relative positional embedding, the output of which is added to the inputs along with a GELU and is followed by the application of layer normalisation. Then, the outputs from the feature encoder are discretised to a finite set of speech representations by product quantisation. This is achieved by deciding quantised representations from a group of codebooks and concatenating them. The Gumbel softmax is then employed to choose discrete codebook entries in a way that is fully differentiable (Baevski et al., 2020).

The model is pre-trained by masking a particular proportion of times steps in the latent feature encoder space. The feature encoder outputs are masked by transferring them to the context network and replacing them with a trained feature vector that is shared between all the masked time steps. Additionally, the inputs passed to the quantisation model are not masked. The outputs are masked by randomly sampling a selection of time steps to be starting indices and masking the subsequent consecutive time steps from every sampled index, of which spans may overlap. The model attempts to identify the correct quantised latent audio representation in a group of distractors for each masked timestep and its final form is fine-tuned on labeled data. The contrastive loss from this activity is derived by calculating the cosine similarity between the context representations and the quantised latent speech representations. While the contrastive task relies on the codebook for positive and negative examples, the diversity loss is configured to increase the use of the quantised codebook representations.This is implemented by enabling the equal use of entries in every codebook by maximising the entropy of averaged softmax distribution over the codebook entries for each codebook across a batch of utterances (Baevski et al., 2020).

The model is then fine-tuned by the addition of a random initialised linear projection on top of the context network which are placed into classes representing the vocabulary of the task. For the Librispeech 960h model, there are 29 tokens for character targets and additionally a word boundary token. The model is optimised by minimising CTC loss and a modified version of SpecAugment is applied also by masking to time-steps and channels during training which improves the final error rates significantly and delays overfitting (Baevski et al., 2020).

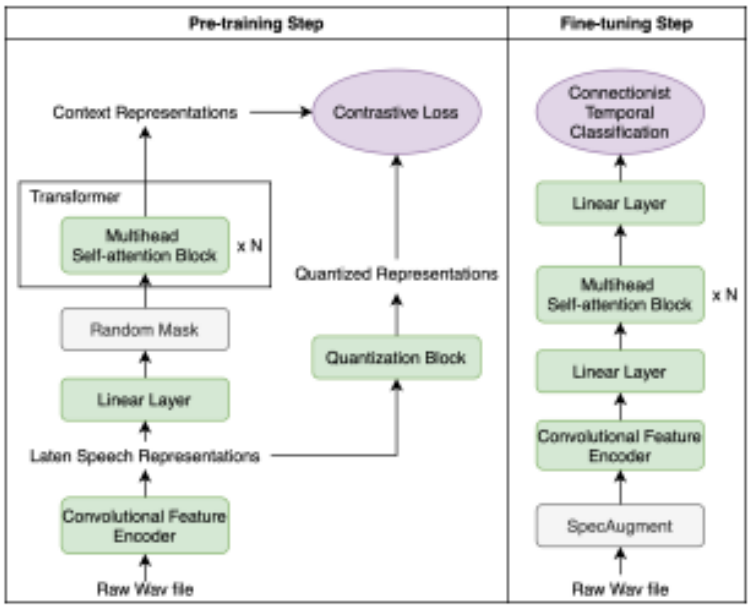


Figure 3.6-1. *wav2vec2.0 framework. Reprinted from "Interpreting A Pre-trained Model Is A Key For Model Architecture Optimization: A Case Study On Wav2Vec 2.0", by Chen & Asgari, 2021, ArXiv:2104.02851 [Cs], 2.*

Furthermore, the model was implemented in Fairseq, an open-source sequence modeling toolkit(Ott et al., 2019). 49% of all time steps were masked with a mean span length of 14.7 seconds. Its feature encoder has seven blocks and the temporal convolutions in each block have 512 channels with strides (5,2,2,2,2,2,2) and kernel widths (10,3,3,3,3,2,2). Its output frequency is 49hz with an approximate stride of 20ms between each sample and a receptive field of 400 input samples or 25ms of audio. The convolutional layer modeling the relative positional embeddings has a kernel size of 128 and 16 groups((Baevski et al., 2020).

The transformer holds 24 blocks with a model dimension of 1,024 , inner dimension of 3,072 and 16 attention heads. The model was trained on 128 V100 GPUs over 2.3 days where 320,000 samples were cropped with a limit of 1.2 samples per GPU. The total batch size was 2.7hrs. The dropout rate for the transformer, at the output of the feature encoder and the input to the quantisation module was 0.1, while layers are dropped at a rate of 0.2((Baevski et al., 2020).

The model is optimised with Adam for the first 8% of updates to a peak of 3 x 10-4 and then linearly decays from this point. It trained for 250,000 updates. The weight $\alpha = 0.1$ was used for the diversity loss, two codebooks were used along with 320 entries of size 384, resulting in a theoretical total of 102,400 codewords. The Gumbel softmax temperature is annealed from 2 to a minimum of 0.1 for a factor of 0.999995 at every update while the temperature for the contrastive loss is set to 0.1.for fine tuning, the model is optimised using Adam along with a tri-state rate schedule where the learning rate increased for the first 10% of updates, held at a constant for the next 40% and then linearly decayed for the remainder. The model batched 1.28 million samples on each GPU while fine tuning on 24 GPUs, resulting in a batch size of 1,920 seconds. The ouput classifier is only trained for the first 10,000 updates after which the Transformer is also updated. Additionally, the feature encoder included to be trained during fine-tuning((Baevski et al., 2020).

The language model employed is a Transformer trained on the Librispeech LM corpus and contains 20 blocks, has a model dimension of 1,280, an inner dimension of 6,144 and 16 attention heads. The weights of the language model were tuned with an interval [0,5] and a word insertion penalty [-5,5] with Bayesian optimisation. 128 trials were ran with beam 50 to choose the best set of weights according to the performance on the "dev-other" version of the Librispeech dataset. The test performance was measured with a beam of 500 for the language model. The beam search utilised is from Wav2letter++: A fast open-source speech recognition system(Baevski et al., 2020).

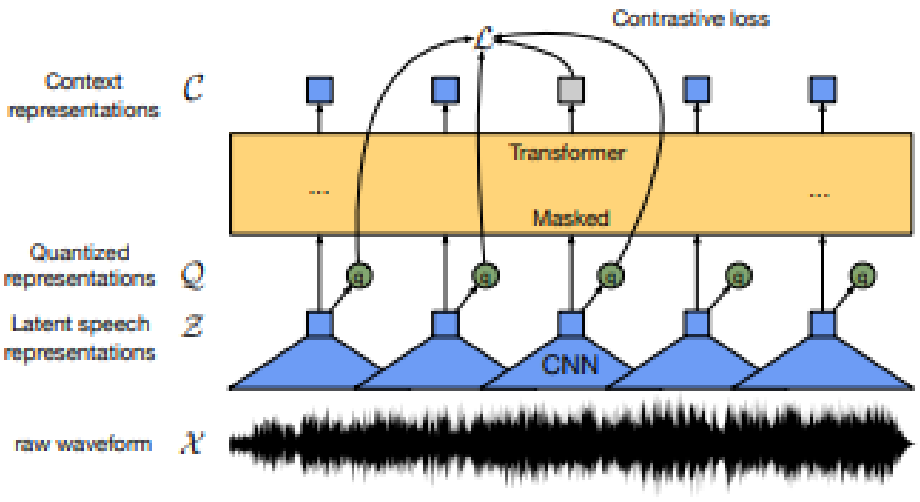


Figure 3.6-2. *wav2vec2.0 architecture. Reprinted from "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", by Baevski et al., 2020, ArXiv:2006.11477 [Cs, Eess], 2.*

3.7- Experiment One

This first experiments purpose is to obtain a control WER of normal speech on a Wav2Vec 2.0 model trained on normal speech to compare against the models performance with dysarthric speech. The model chosen was the Wav2Vec 2.0 960h model trained on the Librispeech dataset from the first experiment. The model was ran on WSL 2 (Windows Subsystem for Linux) on Windows 11. WSL 2 was selected as to implement the Wav2Vec 2.0 framework, the open source machine learning library 'Flashlight' was utilised which was developed entirely in C++ and required a Linux based operating system. The TORGO dataset contains 5, 036 usable samples for testing. Each speaker is tested separately, to test to see if any other attributes of the speech could be attributed to a higher WER.

3.8- Experiment Two

This second experiments purpose is to determine the WER of dysarthric speech on a Wav2Vec 2.0 model trained on normal speech. The same model is implemented from the first experiment to compare to results. The TORGO dataset contains 2815 usable samples of dysarthric speech. The WER is calculated for each speaker rather than as a whole, as the severity of dysarthria is different between speakers sampled in the dataset(Rudzicz et al., 2010).

3.9-Experiment Three

The third experiments purpose is to determine whether a model fine-tuned on multiple languages performs differently with a model fine-tuned on the English language alone. Takshima et al. (2019) conducted experiments that showed reduction in the character error rate (CER) for a Japanese speaker with an articulation disorder could be improved by training with a dataset consisting of articulation orders in a different language, specifically the TORGO dataset, for a Listen Attend Spell (LAS) model. Furthermore, this study showed that training with normal speech and the speech with articulation disorder

The model was acquired from pytorchs Fairseq library(Ott et al., 2019). The model is a version of the Librispeech 960h Wav2Vec2.0 model which has been additionally fine-tuned with the Common Voice and Babel multi-language datasets(Zu et al., 2021). In figure 3.9-1 there is a table from the paper by Zu et al. detailing all languages utilised for the fine-tuning.

For comparison between the single language fine-tuned model, WER is the metric for accuracy. Additionally, the exactly the same data is used from the first experiment.

3.10-Experiment Four

The fourth experiments purpose is to determine whether a model fine-tuned on multiple languages performs differently than a model fine-tuned on the English language alone when dealing with dysarthric speech. The same model is used from the third experiments and the same dataset is used from the second dataset in order for comparison. WER is used for accuracy in line with the previous three experiments.

Split	Languages
CommonVoice (CV)	
train	Esperanto(eo), Lithuanian(lt), Welsh(cy), Tamil(ta), Swedish(sv-SE), German(de), English(en), Oriya(or), Hindi(hi), Persian(fa), Japanese(ja), Assamese(as), Indonesian(id), Catalan(ca), Spanish(es), French(fr), Portuguese(pt), Arabic(ar), Chinese(zh-CN), Chinese(zh-TW), Turkish(tr), Estonian(et), Hungarian(hu), Russian(ru), Czech(cs)
dev	Italian(it)
test	Basque(eu), Interlingua(ia), Latvian(lv), Georgian(ka), Irish(ga-IE), Dutch(nl), Greek(el), Punjabi(pa-IN), Romanian(ro), Maltese(mt), Chinese(zh-HK), Tatar(tt), Finnish(fi), Slovenian(sl), Polish(pl), Kirghiz(ky)
BABEL (BB)	
train	**Amharic(am)**, **Bengali(bn)**, Cebuano(ceb), Igbo(ig), Haitian(ht), **Javanese(jv)**, Mongolian(mn), Swahili(sw), Tamil(ta), **Vietnamese(vi)**, Assamese(as), Dholuo(luo), Guarani(gn), Kazakh(kk), Pashto(ps), **Georgian(ka)**, Tagalog(tl), Telugu(te), Turkish(tr), **Zulu(zu)**
dev	CV-Italian(it)
test	Cantonese(yue), Lao(lo)

Figure 3.9-1. *All languages used in the Common Voice and Babel datasets. Reprinted from "Simple and Effective Zero-shot Cross-lingual Phoneme Recognition", by Xu et al., 2020, ArXiv:2109.11680 [Cs], 2.*

3.11 - Experiment Five - Transfer Learning wav2vec2.0

Based on the results of the first two experiments, in order to try improve results,transfer learning for the Wav2Vec 2.0 Base model was implemented. As the model has only been trained on normal scripted speech, it requires retraining in order to be able to recognise dysarthric speech. The proposed method is to fine-tune the model with the TORGO dysarthric speech labeled data. It isn't feasible to retrain the entire model from scratch with the dysarthric speech data as there simply isn't enough data available to reach a significant accuracy. Even the smallest of the Librispeech models were pre-trained on at least 960h of speech, whereas in total, for the context of these experiments, the TORGO dataset has approximately 3000 usable speech samples between 1 and 20 seconds long. Despite this, the 960h model trained on the "other" dataset has been shown to reach a respectable accuracy, roughly under 11% WER, with fine-tuning on only 10 minutes of labeled data(Baevski et al.,2020) .

The dysarthric fine-tuning was implemented on the Librispeech base model which has not been fine-tuned. Pre-trained versions of the models along with the necessary model checkpoints were downloaded from the Transformer Hub, which is a library of pre-trained models that perform tasks within the different fields of text, vision, and audio. The dysarthric speech samples don't require any data cleaning as that has already been detailed and accomplished in the previous two experiments.

The models are fine-tuned using Connectionist Temporal Classification (CTC) as it suitable for speech recognition tasks where the alignment between the input and the output are not known(Wang et al., 2019). In order for the models to transcribe the speech samples to text, they require a feature vector and tokeniser. They are implemented using Wav2Vec2CTCTokenizer and Wav2Vec2FeatureExtractor from the Transformers library. A vocabulary is built on the contents of the dysarthric speech samples as an enumerated dictionary along with tokens for the special characters ' and " " and a unknown token along with a blank token, which is required for the alignment part of the CTC algorithm. This dictionary is then saved as a json file which is then passed to instantiate an object of the Wav2Vec2CTCTokenizer class.

When the pre-trained model is loaded, its feature extractor is frozen as the first component of the model is concerned with extracting features from raw speech samples which has already been sufficiently pre-trained and does not require fine-tuning. Like all previous experiments, in order to fairly compare results, WER is used as the metric for performance. The WER for normal speech is also tested against these models for comparison of performance.

Model Parameters	
Number of epochs	30
Batch Size	8
Training test data split	80:20
Learning Rate	1e -4
Weight decay	0.005

Table 3.1 - *wav2vec2.0 model parameters*

Feature Extractor Parameters	
Feature Size	1
sampling_rate	16000
padding_value	0.0
do_normalize	True
return_attention_mask	False

Table 3.2 - *wav2vec2.0 feature extractor parameters*

3.12 - Experiment Six - General Adversarial Networks

Jiao et al. (2018) showed promising results converting normal speech samples to dysarthric speech samples, specifically the spectral features, using a DCGAN. Previous research has shown promising results preprocessing dysarthric speech samples to improve their intelligibility before they are passed to an ASR system to improve accuracy (Rudzicz et al., 2013).

A disadvantage to using GANs for voice conversion is that in many models, the source and target speech must be paired, the word utterances must be the same for each pair. For example, Rudzicz (2013) showed promise by synthesizing dysarthric speech samples with Kaneko et al. (2019) produced 'CycleGAN-VC2', a non parallel voice con-

suming and difficult to gather matching target data of the same utterance for all source data. In the publicly available version of the TORGO dataset, the utterances between the control speakers and the dysarthric speakers are not wholly the same, therefore using other voice conversion methods would cause a significant amount of data to go unused. By using CycleGAN, this issue is mitigated and allows all data to be utilised which is integral as it is already very compact.

The aim of using CycleGAN to alter the dysarthric speech samples is to alter the Mel-cepstral coefficients to change the qualities of the speech to be more similar to normal speech. Mel-cepstral coefficients represent the hearing of the human ear which has high resolution at low frequencies and (Fukadat et al., 1992). They represent a small set of features which describe the overall shape of a spectral envelope, which is one point in time. It represents the timbre of the sound, what distinguishes one voice from another. By altering the Mel-cepstral coefficients it changes the timbre of the sound but retains the pitch, loudness and duration. Therefore, this technique may not address all the differences between normal speech and dysarthric speech, such as the hyper-nasality. Despite this it is important to try and retain as much of the original speech information as possible or else it could be considered that the data samples are being replaced by normal speech.

The experiment was set up by selecting one dysarthric speaker and its matching control normal speaker from the TORGO dataset. All samples are preprocessed by being padded to the same length, while 34 Mel-cepstral coefficients , the logarithmic fundamental frequency and a periodicities are extracted every 5ms from each sample
using the WORLD analyser(Kaneko et al., 2019; Morise et al., 2016).

The MCEPs are then converted from the source speech data, dysarthric speech,to the target speech data, the normal speech. The fundamental frequency is converted using a logarithm Gaussian normalized transformation and the a periodicities are used without being modified(Kaneko et al., 2019). The WORLD vocoder is then used to synthesize the speech.

When the speech samples have been converted, they will be tested using the ASR mod-

el from experiment two, to compare their WER results with the unconverted samples.

Model Parameters	
Optimiser	Adam
Batch Size	1
Epochs	1000
Generator Learning Rate	0.0002
Discriminator Learning Rate	0.0001

Table 3.3 - *CycleGAN-VC-2 model parameters*

CHAPTER FOUR: RESULTS & EVALUATION

The purpose of this study is to try improve the performance of an ASR model with recognising dysarthric speech. This was attempted by enriching training datasets with dysarthric speech and using GANs for speech style transfer (from dysarthric to normal).

In chapter two, the main issue for the poor performance of dysarthric speech with machine learning based ASR models was identified as a lack of dysarthric speech data.

In chapter three, based on the problem outline in chapter two, six experiments were conducted, experiments one to four were to establish the performance of a wav2vec2.0 models performance with recognising dysarthric speech when it had been trained on the English language alone versus multiple languages. Two further experiments were performed, an implementation of transfer learning on the wav2vec2.0 base model, to see how the model would perform, with respect to WER, if it had been trained with dysarthric speech. The second experiment attempted to tackle the intelligibility of dysarthric speech, as there is not much dysarthric training data available. A GAN voice converter system CycleGAN-VC-2 was selected to try transfer the speech style of the control normal speakers to the paired dysarthric speakers.

This chapter will review the results from the six experiments, evaluate their performance and design and will suggest possible amendments and work for the future.

4.1 -Experiment One to Four Results Comparison

Table 4.1 illustrates the results of experiments one to four. The speakers WER have been separated for each speaker due to the differences in severity in dysarthria between them.

	Single Language WER	Multi Language WER
Female 01	77.87%	73.93%
Female 03	78.21%	74.64%
Female 04	77.5%	74.29%
Male 01	103.95%	79.80%
Male 02	91.68%	72.42%
Male 03	11.83%	10.45%
Male 04	108.99%	97.15%
Male 05	84.47%	61.86%
Female Control 01	17.94%	15%
Female Control 02	13.36%	10.05%
Femaile Control 03	15.72%	12.17%
Male Control 01	16.99%	12.14%
Male Control 02	24.57%	19.40%
Male Control 03	15.73%	12.63%
Male Control 04	11.80%	11.54%

Table 4.1 - *Experiment One to Four Results*

From Table 4.1 it is clear there is a significant improvement in the results between the single language (English) wav2vec 2.0 and the multi-language Wav2Vec2.0 model. There is a small decrease in WER for the female dysarthric speakers who have milder dysarthria in comparison to most of the male speakers in the dataset (Rudzicz et al., 2010). It is significant that the worse the WER rate performance from the single language model, the greater the improvement with the Multi Language model.

These results follow on from Takashima et al. (2019) when they found training with combined datasets of speech articulation disorders in two different languages could improve accuracy. These results show that the a multi-language fine-tuned model does improve WER for speakers with dysarthric speech along with speakers with normal speech.

4.2 - Experiment Five Results

The results of fine-tuning the Wav2Vec2.0 base model were unsuccessful. Throughout the duration of the models training, the WER never reduced from 100%. When validating the model, the prediction text for all speech samples were repetitions of the tokenziers 'BOS' token which denotes the beginning of a sentence. An example prediction from the model looked like " <s> <s> <s> <s> <s> <s> <s> <s>.....". Therefore, the fine-tuning of the model failed to learn the speech representations from the dysarthric speech.

4.3 - Experiment Six Results

The results of converting the dysarthric speech sample to health speech samples was unsuccessful using CycleGAN within the limits of this experiment. With the machine configuration available, the amount of time required to train the GAN (at least 1000 epochs) to convert approximately 100 dysarthric samples to healthy speech would be approximately 14 days. This length of training time and its required resources were not feasible within the confines of this experiment. After approximately 48hrs of training, the model had reached 30,000 iterations out of 269000 total iterations to train the model efficiently. It was decided training would be ceased at this point as it was no longer feasible.

To show the potential of this technology for future work, example converted GAN samples were taken from the researchers of CycleGAN-VC-2 model and tested against the wav2vec2.0 multi-language model. For the twelve samples provided by Kaneko et al. (2019), which were male to female and female to male normal speech style conversions, they achieved an overall WER of 3.54%. The quality of the conversion is shown to be very high from the WER and shows the potential it could have when converting dysarthric speech. It should be noted, since this GAN only adapts the MFCC features, it would not address issues with dysarthric speech such as the issue of the deletion of phonemes. This would have to be addressed with a different approach.

4.4 - Discussion

The aim of this research is to try improve an ASR models performance when recognising dysarthric speech by enriching datasets with dysarthric speech and using GANs to improve dysarthric speech intelligibility.

The main takeaway from the set of results presented in this chapter is that the recognition of dysarthric speech becomes more successful when the dataset used for training is larger and more diverse. The multi-language model outperformed the single language model and this can be attributed to the models greater exposure to a may diverse set of phonemes. In Takashima et al. (2019) hypothesised that their multi-lingual dysarthric speech dataset when training the listen portion of their Listen Attend Spell module would benefit and create a better high-level representation over a listen portion only trained on a single language dysarthric dataset. For one of the speakers studied the top CER (character error rate) was 72.7% on a model trained with no dysarthric data and reduced to a top CER of 23.2% when tested on the multi language dysarthric model.

It can then be determined that the Wav2Vec2.0 model when fine-tuned on the Babel and Common Voice datasets had a better exposure to a wider range of voices and pronunciation of phonemes. Additionally, the Common Voice dataset was gathered differently to the Librispeech dataset. The speech samples were crowd sourced using the a browser and mobile app and was validated by contributers and users of the app (Ardila et al., 2020). The prompts were taken from Wikipedia articles which contrast against the texts used in the Gutenberg project which are much older and have a different semantic style to modern speech(Ardila et al., 2020; Panayotov et al., 2015). Additionally, due to the recordings not being recorded in a controlled environment, being recorded on a mobile phone, this would expose the model to noisy and more imperfect speech. These differences can account for the difference in WER rate between the Librispeech model and the multi-language model. This denotes that the more diverse a dataset, the lower the WER will be, not only for dysarthric speech, but for normal speech also.

The poor performance of the wav2vec2.0 model pre-trained on the Librispeech dataset to successfully develop speech representations can be attributed to lack of data and the

diverse difference between the Librispeech dataset and the TORGO dataset. Additionally, the wav2vec2.0 base model, which was chosen due to hardware constraints, was trained on the clean version of the Librispeech dataset(Baevski et al., 2020). This could also be a reason for the poor performance of the model as the model had not been previously exposed to speech abnormalities or impairments in comparison to the 960h large model.

It is important to evaluate the process of the experimentation and the results under the principles of the CRISP-DM life cycle outlined in chapter three.

Did the outcomes of the experimentation fit the project problem and are they feasible to be deployed in a real world setting?

Experiment three and four were not practical in solving the issue of poor performance of speech recognition systems when recognising dysarthric speech. Experiment three performed poorly due to a lack of available dysarthric data and is not plausible to mitigate the issue. Experiment four, when performed under more advantageous conditions has been shown in other research to be successful in lowering the WER in speech recognition models(Donahue et al., 2018). Despite this, the use of GANs to modify dysarthric speech samples is very computationally expensive and intensive, if it were to be deployed in a real world setting, it would not be very feasible. Additionally, if viewed against the World Economic Forums (WEF) paper ‘How to Prevent Discriminatory Outcomes in Machine Learning’ (2018) this approach does not meet the guidelines for fair machine learning. This approach in particular violates the guidelines of fairness and active inclusion. People with dysarthria have been excluded from speech datasets and continue to be, of all the large speech datasets mentioned in this study, of which have been gathered relatively recently, none include dysarthric speech. The practice of trying to use GANs to augment voice samples or generate synthetic samples can be viewed just as exclusionary as the practice of excluding this speech type from datasets. The paper by the WEF specifies to aim for a ‘diversity of input’, these practices do the opposite. By augmenting the speech samples, they are not being fully represented in the data and in a convoluted manner are being removed. Bragg et al. (2018) discuss the issue of using stand in data to replace sign language data and highlight that the unnatural blending of

features would not represent the language properly. This critique could be applied to the practice of generating speech samples with GANs. Without fully representational datasets for dysarthric speech, it makes it difficult for a GAN to generate fully representational dysarthric speech. This technique could be beneficial in the future if more dysarthric samples, with diverse characteristics captured, were available to boost the amount of training data if necessary. This issue can not be fully solved with a pure machine learning approach and requires changes in the real world first before algorithms can be introduced.

It must be assessed whether machine learning techniques are able to solve the issue of a lack of representative data. There is a lack of communication with dysarthric individuals on this issue. There is little literature published on the opinions and impact this issue has on the dysarthric community and their feelings towards the current practices employed to try solve the problem(Cave & Bloch, 2021). There is a requirement for a study to be conducted similar to the paper by Bragg et al. (2018) where the opinions and needs of the dysarthric community can be communicated to researchers. There is a presumption in the literature in this field that people with dysarthric speech will use ASR technology in the same manner as people with normal speech. Furthermore, dysarthric speech datasets (Rudzicz et al., 2010; Kim et al., 2008) focus primarily on capturing the acoustic qualities of dysarthric speech but do not consider if the data collected represents how a dysarthric speaker would use an ASR system. These datasets are over a decade old and contain prompts from the likes of intelligibility assessments but do contain more up to date common phrases that would be used to interact with a personal assistant such as Apple's 'Siri' or Amazon's 'Alexa'. The UA speech dataset contains a selection of uncommon words as prompts taken from the Gutenberg project(Kim et al., 2008), although they display the variety of sounds in dysarthric speech, they do not display how dysarthric speakers talk day to day. None of the datasets include a significant amount of samples that represent unscripted speech. The collection of data in the TORGO dataset did not include gathering samples when the participants became fatigued. This data collection style could potentially affect the robustness of a ASR system in recognising dysarthric speech in the future, as the users will not be able to output ideal, unfatigued speech all of the time.

The amount of resources required to perform these highly complex computational tasks

needs to be addressed. As natural language processing systems and algorithms improve alongside the hardware they are used on, their carbon footprint increases(Strubell et al., 2019). The benefit of running this computationally intensive tasks must be weighed against the cost and its impact on the environment. The collection of more dysarthric speech samples maybe logistically more difficult and time consuming due to factors such as obtaining permission to collect the data and dealing with the constrained data collection sessions due to the conditions of individuals with dysarthria. In the long run, this practice is the most ethical choice and the most cost effective overall.

CHAPTER FIVE: CONCLUSION

5.1 -Introduction

In this chapter, the original aims of the project, to improve an ASR systems performance in recognising dysarthric speech by enriching datasets and using GANs to increase its intelligibility, will be reviewed against the results obtained in chapter four. The results in chapter four, suggest that more dysarthric data needs to be gathered in order to progress the performance of speech recognition systems when recognising dysarthric speech. If this was possible, experiment five, when the wav2vec2.0 model was fine-tuned with dysarthric speech may have performed better. Additionally, under better experimental conditions, experiment six that used CycleGAN-VC-2 has the potential to improve the intelligibility of dysarthric speech has the ability to help solve the lack of data in the short term. Despite this, dialogue still needs to be opened with the community of people with dysarthria to determine what they need from an ASR system and how they would like their data to be represented.

5.2 - Summary

The aim of this research was to improve an ASR models accuracy performance by enriching speech datasets for training and to use GANs to improve the intelligibility of dysarthric speech. Dysarthric speech has performed poorly with a lot of modern ASR systems and the main proponent that causes this is a lack of representation of dysarthric speech in ASR training datasets.

A literature review was conducted in chapter two to understand the landscape of speech recognition and the problem of dysarthric speech. An overview of speech recognition technologies and the architecture of machine learning ASR systems was discussed along with an overview of the characteristics of dysarthria and its history of performance with ASR systems to date. Bias in machine learning and its effect on the disabled community was detailed to understand the problem in a real world context. The practice of building speech datasets was explored to understand their requirements ,how they are affected by bias and how they can be improved. Speech preprocessing

and GANs were reviewed to understand how they can be applied to the issue. Finally, an overview of all current literature that concerns improving ASR systems performance with dysarthric speech was detailed. The requirements and main aims of the experimentation were then concluded.

The aim of all experiments in chapter three derived from the research gathered in chapter two was to try enrich training datasets with dysarthric speech and to pre-pro cess the speech to be more intelligible. The word error rates for normal speech and dysarthric speech were tested against a wav2vec2.0 model trained on English only and the same model trained on multiple languages to compare performance. Transfer learning was employed on an un-fine-tuned wav2vec2.0 model trained on English speech to observe if the WER rate from the previous experiments could be improve upon. Finally CycleGAN -VC-2 was used to try and convert dysarthric speech samples to have the same speech style as normal speech to try improve intelligibility.

In chapter four, the results showed that using a model that had been fine-tuned on a larger variety of speech data (multi-language) performed better than a model only trained on normal English speech. The fine-tuning of the Wav2Vec2.0 model was unsuccessful due to the insufficient size of the dataset for training and may have suffered due to the model choices base training set. An attempt to convert dysarthric speech samples to have the speech style of normal speech using CycleGAN-VC-2 was unsuccessful due to hardware limitations but shows potential for further research with better conditions.

5.3 - Conclusions: Experimentation, Evaluation & Results

The main conclusion that can be drawn from the experimentation in this study is that there is a greater diversity needed in speech datasets for training ASR systems, specifically more representation of dysarthric speech. From the results of experiments one to four, even the inclusion of more speech data from different languages, that isn't dysarthric data, improves the WER significantly.

Although the transfer learning of the wav2vec2.0 model performed poorly and the speech transfer style with CycleGAN requires further experimentation to be brought to full

potential, they still confirm the need for more diverse and available dysarthric speech data. Without it, further techniques to fine-tune the performance of ASR systems when recognising dysarthric speech cannot fully progress to their full potential. These experiments try to mitigate the issue of the lack of data, without directly solving the issue. If more data became available in future studies they could be used to improve the accuracy of ASR systems further but may not be viable as a complete solution.

Speech recognition is a very computationally heavy task. The hardware used during the experimentation of was not of adequate strength to handle some of the models. Additionally, speech recognition has a steep learning curve and requires a lot of time to become well versed enough to be able to perform experimentations. It requires more research than what was possible in the time frame of this study.

5.4 - Contributions and Impact

Throughout this project the issue of the lack of representation of dysarthric speech within datasets has been flagged and discussed. Many of the experiments in this study have cited lack of data as the reason for unsuccessful results. This research problem requires the inclusion of the dysarthric community to help solve it. It cannot be solved with machine learning techniques alone. More diverse speech datasets are required to improve the robustness of ASR systems in general. People with dysarthria are not the only community affected by lack of representation in datasets, peoples gender, accent and age can affect how well their voice will be recognised by an ASR system. Additionally, the way datasets are gathered may need to be addressed, allowing for longer data gathering windows, considering how suitable certain texts are as prompts (Librispeech dataset and UA dataset contain texts which are over 100 years old (Panayotov et al., 2015; Kim et al., 2008)) and aiming to capture as many different speech styles as possible. The Librispeech dataset contains a cleaned version and an ‘other’ version which contains disfluencies(Panayotov et al., 2015), is perfect speech with no errors the right kind of speech to train a robust ASR with? Additionally, the viability of computationally intensive processes to mitigate issues due of lack of data may need to be addressed.

Many studies on the improvement of dysarthric speech recognition, do not address the ethical and environmental impacts of using synthetic data and the augmentation of speech. The findings of the research of this study suggest a more practical approach to solving this issue, collecting more data. From a review of all current literature there is little interaction from researchers and individuals with dysarthria (Cave & Bloch, 2021). With clearer outlines from the dysarthric community on how they use and would like to use ASR technologies would streamline research and help improve ASR systems.

5.5 - Future Work and Recommendations

Research in this field would benefit greatly from a larger dysarthric speech dataset from a large pool of speakers of different varying severities of dysarthria. Additionally, there are very few examples of unscripted conversational dysarthric speech and dysarthric speech recorded in noisy environments. A publicly available dataset that matches this criteria could have a massive impact on the improvement of the issue. Although Google is running a project to gather more dysarthric speech samples (The ASHA Leader, 2019), it has not been announced whether this dataset will be made available to the public for research. It would also be beneficial to study how dysarthric speech changes with fatigue and try and capture this type of speech to ensure for more robust models which meet a dysarthric individuals needs.

There is no literature detailing how dysarthric users use and would like to use an ASR system or their expectations of this technology (Cave & Bloch, 2021). Under the principles of Active Inclusion (World Economic Forum, 2018), researchers could include individuals with dysarthria more the principles of Active Inclusion (World Economic Forum, 2018), researchers could include individuals with dysarthria more readily in speech recognition studies to ensure the outcomes of these studies are beneficial and fair to the users they are intended for.

The use of GANS for converting dysarthric speech samples could be explored further. Although not brought to its full potential in this study, with hardware that can handle the lengthy training time of CycleGAN-VC-2, it could be beneficial for improving the intelligibility of dysarthric speech. Due to its non-requirement for paired data

unlike other GANS, it has the potential to use more data and makes gathering data samples easier by allowing different datasets with different prompt texts to be used.

BIBLIOGRAPHY

Anusuya, M. A., & Katti, S. K. (2009). Speech Recognition by Machine: A Review. 6(3), 25.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. ArXiv:1912.06670 [Cs]. http://arxiv.org/abs/1912.06670

Arora, S., & Singh, R. (2012). Automatic Speech Recognition: A Review. International Journal of Computer Applications, 60, 34–44. https://doi.org/10.5120/9722-4190

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. ArXiv:2006.11477 [Cs, Eess]. http://arxiv.org/abs/2006.11477

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018a). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. ArXiv:1810.01943 [Cs]. http://arxiv.org/abs/1810.01943

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018b). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. ArXiv:1810.01943 [Cs]. http://arxiv.org/abs/1810.01943

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. Speech Communication, 49(10), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Brown, M. K., McGee, M. A., Rabiner, L. R., & Wilpon, J. G. (1991). Training set design for connected speech recognition. IEEE Transactions on Signal Processing, 39(6), 1268–1281. https://doi.org/10.1109/78.136533

Butzberger, J., Murveit, H., Shriberg, E., & Price, P. (1992). Spontaneous Speech Effects In Large Vocabulary Speech Recognition Applications. Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992. HLT 1992. https://aclanthology.org/H92-1068

Cave, R., & Bloch, S. (2021). The use of speech recognition technology by people living with amyotrophic lateral sclerosis: A scoping review. Disability and Rehabilitation: Assistive Technology, 0(0), 1–13. https://doi.org/10.1080/17483107.2021.1974961

Chaudhari, A., & Dhonde, S. B. (2015). A review on speech enhancement techniques. 2015 International Conference on Pervasive Computing (ICPC), 1–3. https://doi.org/10.1109/PERVASIVE.2015.7087096

Chen, C.-Y., Zheng, W.-Z., Wang, S.-S., Tsao, Y., Li, P.-C., & Lai, Y.-H. (2020a). Enhancing Intelligibility of Dysarthric Speech Using Gated Convolutional-based Voice Conversion System. Proc. Interspeech 2020, 4686–4690. https://doi.org/10.21437/Interspeech.2020-1367.

Chen, F., & Kostov, A. (1997). Optimization of dysarthric speech recognition. Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. “Magnificent Milestones and Emerging Opportunities in Medical Engineering” (Cat. No.97CH36136), 4, 1436–1439 vol.4. https://doi.org/10.1109/IEMBS.1997.756975

Chen, L., & Asgari, M. (2021). Interpreting A Pre-trained Model Is A Key For Model Architecture Optimization: A Case Study On Wav2Vec 2.0. ArXiv:2104.02851 [Cs]. https://doi.org/10.48550/arXiv.2104.02851

Chen, X., Liu, X., Wang, Y., Ragni, A., Wong, J. H. M., & Gales, M. J. F. (2019). Exploiting Future Word Contexts in Neural Network Language Models for Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(9), 1444–1454. https://doi.org/10.1109/TASLP.2019.2922048

Cheng, J., Yang, Y., Tang, X., Xiong, N., Zhang, Y., & Lei, F. (2020). Generative Adversarial Networks: A Literature Review. KSII Transactions on Internet and Information Systems (TIIS), 14(12), 4625–4647. https://doi.org/10.3837/tiis.2020.12.001

Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., & Bacchiani, M. (2018a). State-of-the-art Speech Recognition With Sequence-to-Sequence Models. ArXiv:1712.01769 [Cs, Eess, Stat]. http://arxiv.org/abs/1712.01769

Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., & Bacchiani, M. (2018b). State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4774–4778. https://doi.org/10.1109/ICASSP.2018.8462105

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine, 35(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

De Russis, L., & Corno, F. (2019). On the impact of dysarthric speech on contemporary ASR cloud platforms. Journal of Reliable Intelligent Environments, 5(3), 163–172. https://doi.org/10.1007/s40860-019-00085-y

De Wachter, M., Matton, M., Demuynck, K., Wambacq, P., Cools, R., & Van Compernolle, D. (2007). Template-Based Continuous Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, 15(4), 1377–1390. https://doi.org/10.1109/TASL.2007.894524

Deng, L., & Li, X. (2013). Machine Learning Paradigms for Speech Recognition: An Overview. IEEE Transactions on Audio, Speech, and Language Processing, 21(5), 1060–1089. https://doi.org/10.1109/TASL.2013.2244083

Donahue, C., Li, B., & Prabhavalkar, R. (2018). Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5024–5028. https://doi.org/10.1109/ICASSP.2018.8462581

Donahue, C., McAuley, J., & Puckette, M. (2019). Adversarial Audio Synthesis. ArXiv:1802.04208 [Cs]. http://arxiv.org/abs/1802.04208

Enderby, P. (2013). Chapter 22 - Disorders of communication: Dysarthria. In M. P. Barnes & D. C. Good (Eds.), Handbook of Clinical Neurology (Vol. 110, pp. 273–281). Elsevier. https://doi.org/10.1016/B978-0-444-52901-5.00022-8

Executive Office of the President. (2016). Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021a). Quantifying Bias in Automatic Speech Recognition. ArXiv:2103.15122 [Cs, Eess]. http://arxiv.org/abs/2103.15122

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018a). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research, 61, 863–905. https://doi.org/10.1613/jair.1.11192

Ferrer, L., Lei, Y., McLaren, M., & Scheffer, N. (2014). Spoken Language Recognition Based on Senone Posteriors. 5. https://doi.org/10.21437/Interspeech.2014

Fukadat, T., Tokudatt, K., Kobayashzjtt, T., & Imaittt, S. (1992). An Adaptive Algorithm for Mel-Cepstral Analysis of Speech.

Furui, S., Ichiba, T., Shinozaki, T., Whittaker, E., & Koji, I. (2005). Cluster-based modeling for ubiquitous speech recognition. 2865–2868. https://doi.org/10.21437/Interspeech.2005-838

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., & Zue, V. (1992). TIMIT Acoustic-phonetic Continuous Speech Corpus. Linguistic Data Consortium.

Georgescu, A.-L., & Cucu, H. (2018). Automatic Annotation of Speech Corpora Using Complementary GMM and DNN Acoustic Models. 2018 41st International Conference on Telecommunications and Signal Processing (TSP), 1–4. https://doi.org/10.1109/TSP.2018.8441374

Ghai, W., Singh, N., College, M. G., & Punjab, F. S. (n.d.). Business Studies, Mohali, Punjab.
Google Seeks Voice Samples From People With Dysarthria (world). (2019, January 19). [News in Brief]. The ASHA Leader; American Speech-Language-Hearing Association. https://doi.org/10.1044/leader.NIB4.24072019.16

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. ArXiv:1412.5567 [Cs]. http://arxiv.org/abs/1412.5567

Harvill, J., Issa, D., Hasegawa-Johnson, M., & Yoo, C. (2021). Synthesis of New Words for Improved Dysarthric Speech Recognition on an Expanded Vocabulary. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6428–6432. https://doi.org/10.1109/ICASSP39728.2021.9414869

Hetsugi, Y., Sakata, T., & Ueda, Y. (2020). A Phonological Control Method on A Speech Compensation System for Dysarthria Using A Standardized Space. 2020 5th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 158–162. https://doi.org/10.1109/ICIIBMS50712.2020.9336404

Hines, A., & Harte, N. (2012). Speech intelligibility prediction using a Neurogram Similarity Index Measure. Speech Communication, 54(2), 306–320. https://doi.org/10.1016/j.specom.2011.09.004

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine, 29(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hu, A., Phadnis, D., & Shahamiri, S. R. (2021). Generating synthetic dysarthric speech to overcome dysarthria acoustic data scarcity. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-021-03542-w

Hux, K., Rankin-Erickson, J., Manasse, N., & Lauritzen, E. (2000). Accuracy of three speech recognition systems: Case study of dysarthric speech. Augmentative and Alternative Communication, 16(3), 186–196. https://doi.org/10.1080/07434610012331279044

Ibrahim, Y. A., Odiketa, J. C., & Ibiyemi, T. S. (2017). PREPROCESSING TECHNIQUE IN AUTOMATIC SPEECH RECOGNITION FOR HUMAN COMPUTER INTERACTION: AN OVERVIEW. 6.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2018). Image-to-Image Translation with Conditional Adversarial Networks. ArXiv:1611.07004 [Cs]. http://arxiv.org/abs/1611.07004

Jassim, W. A., Skoglund, J., Chinen, M., & Hines, A. (2021). WARP-Q: Quality Prediction For Generative Neural Speech Codecs. ArXiv:2102.10449 [Eess]. http://arxiv.org/abs/2102.10449

Jiao, Y., Tu, M., Berisha, V., & Liss, J. (2018). Simulating Dysarthric Speech for Train-

ing Data Augmentation in Clinical Speech Applications. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6009–6013. https://doi.org/10.1109/ICASSP.2018.8462290

Kain, A. B., Hosom, J.-P., Niu, X., van Santen, J. P. H., Fried-Oken, M., & Staehely, J. (2007). Improving the intelligibility of dysarthric speech. Speech Communication, 49(9), 743–759. https://doi.org/10.1016/j.specom.2007.05.001

Kaneko, T., Kameoka, H., Tanaka, K., & Hojo, N. (2019). CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion. ArXiv:1904.04631 [Cs, Eess, Stat]. http://arxiv.org/abs/1904.04631

Karatzoglou, A., Jablonski, A., & Beigl, M. (2018). A Seq2Seq learning approach for modeling semantic trajectories and predicting the next location. Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 528–531. https://doi.org/10.1145/3274895.3274983

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Watkin, K., & Frame, S. (2008). Dysarthric speech database for universal access research. 1741–1744.

Kim, M., Yoo, J., & Kim, H. (2013). Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (p. 3626).

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Kuchaiev, O., Ginsburg, B., Gitman, I., Lavrukhin, V., Li, J., Nguyen, H., Case, C., & Micikevicius, P. (2018). Mixed-Precision Training for NLP and Speech Recognition with OpenSeq2Seq. ArXiv:1805.10387 [Cs]. http://arxiv.org/abs/1805.10387

Kuligowska, K., Kisielewicz, P., & Włodarz, A. (2018). Speech synthesis systems: Disadvantages and limitations. International Journal of Engineering & Technology, 7(2.28), 234. https://doi.org/10.14419/ijet.v7i2.28.12933

Kumar, S. A., & Kumar, C. S. (2016). Improving the intelligibility of dysarthric speech towards enhancing the effectiveness of speech therapy. 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 1000–1005. https://doi.org/10.1109/ICACCI.2016.7732175

Kumar, Y., & Singh, N. (2019). A Comprehensive View of Automatic Speech Recognition System—A Systematic Literature Review. 2019 International Conference on Automation, Computational and Technology Management (ICACTM), 168–173. https://doi.org/10.1109/ICACTM.2019.8776714

Li, B., Sainath, T. N., Sim, K. C., Bacchiani, M., Weinstein, E., Nguyen, P., Chen, Z., Wu, Y., & Rao, K. (2018). Multi-Dialect Speech Recognition with a Single Sequence-to-Sequence Model. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4749–4753. https://doi.org/10.1109/ICASSP.2018.8461886

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: A survey. Multimedia Tools and Applications, 80(6), 9411–9457. https://doi.org/10.1007/s11042-020-10073-7

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 115:1-115:35. https://doi.org/10.1145/3457607

Mengistu, K. T., & Rudzicz, F. (2011). Adapting acoustic and lexical models to dysarthric speech. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4924–4927. https://doi.org/10.1109/ICASSP.2011.5947460

Molla, K. I., & Hirose, K. (2004). On the effectiveness of MFCCs and their statistical distribution properties in speaker identification. 2004 IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2004. (VCIMS)., 136–141. https://doi.org/10.1109/VECIMS.2004.1397204

Morgan, N. (2012). Deep and Wide: Multiple Layers in Automatic Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1), 7–13. https://doi.org/10.1109/TASL.2011.2116010

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Transactions on Information and Systems, E99D(7), 1877–1884. https://doi.org/10.1587/transinf.2015EDP7457

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. ArXiv:1609.03499 [Cs]. http://arxiv.org/abs/1609.03499

O'Shaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. Pattern Recognition, 41(10), 2965–2979. https://doi.org/10.1016/j.patcog.2008.05.008

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. ArXiv:1904.01038 [Cs]. http://arxiv.org/abs/1904.01038

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Interspeech 2019, 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Park, J. H., Seong, W. K., & Kim, H. K. (2011). Preprocessing of Dysarthric Speech in Noise Based on CV–Dependent Wiener Filtering. In R. L.-C. Delgado & T. Kobayashi (Eds.), Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop (pp. 41–47). Springer. https://doi.org/10.1007/978-1-4614-1335-6_6

Parker, M., Cunningham, S., Enderby, P., Hawley, M., & Green, P. (2006). Automatic speech recognition and training for severely dysarthric users of assistive technology: The STARDUST project. Clinical Linguistics & Phonetics, 20(2–3), 149–156. https://doi.org/10.1080/02699200400026884

Patel, R. (2000). IDENTIFYING INFORMATION-BEARING PROSODIC PARAMETERS IN SEVERELY DYSARTHRIC VOCALIZATIONS. University of Toronto, 116.

Permanasari, Y., Harahap, E. H., & Ali, E. P. (2019a). Speech recognition using Dynamic Time Warping (DTW). Journal of Physics: Conference Series, 1366(1), 012091. https://doi.org/10.1088/1742-6596/1366/1/012091

Phan, H., McLoughlin, I. V., Pham, L., Chén, O. Y., Koch, P., De Vos, M., & Mertins, A. (2020). Improving GANs for Speech Enhancement. IEEE Signal Processing Letters, 27, 1700–1704. https://doi.org/10.1109/LSP.2020.3025020

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., & Vesel, K. (2011). The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

Purohit, M., Patel, M., Malaviya, H., Patil, A., Parmar, M., Shah, N., Doshi, S., & Patil, H. A. (2020). Intelligibility Improvement of Dysarthric Speech using MMSE DiscoGAN. 2020 International Conference on Signal Processing and Communications (SPCOM), 1–5. https://doi.org/10.1109/SPCOM50965.2020.9179511

Rabiner, L., & Juang, B. H. (2004). Automatic Speech Recognition—A Brief History of the Technology Development. Undefined. https://www.semanticscholar.org/paper/Automatic-Speech-Recognition-A-Brief-History-of-the-Rabiner/1d199099a2f4f-8749c7e10480b29f5adaecad4a1

Ramirez, J. M., Montalvo, A., & Calvo, J. R. (2019). A Survey of the Effects of Data Augmentation for Automatic Speech Recognition Systems. In I. Nyström, Y. Hernández Heredia, & V. Milián Núñez (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (pp. 669–678). Springer International Publishing. https://doi.org/10.1007/978-3-030-33904-3_63

Ravanelli, M., Parcollet, T., & Bengio, Y. (2019). The PyTorch-Kaldi Speech Recognition Toolkit. ArXiv:1811.07453 [Cs, Eess]. http://arxiv.org/abs/1811.07453

Rudzicz, F. (2007). Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech. Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility, 255–256. https://doi.org/10.1145/1296843.1296899

Rudzicz, F. (2011). Acoustic transformations to improve the intelligibility of dysarthric speech. Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, 11–21. https://aclanthology.org/W11-2302

Rudzicz, F. (2013). Adjusting dysarthric speech signals to be more intelligible. Computer Speech & Language, 27(6), 1163–1177. https://doi.org/10.1016/j.csl.2012.11.001

Rudzicz, F., Namasivayam, A., & Wolff, T. (2010). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evalu-

ation, 46, 1–19. https://doi.org/10.1007/s10579-011-9145-0

S, K., & Chandra, E. (2016). A Review on Automatic Speech Recognition Architecture and Approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition, 9, 393–404. https://doi.org/10.14257/ijsip.2016.9.4.34

Sarı, L., Hasegawa-Johnson, M., & Yoo, C. D. (2021). Counterfactually Fair Automatic Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3515–3525. https://doi.org/10.1109/TASLP.2021.3126949

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Selouani, S.-A., Dahmani, H., Amami, R., & Hamam, H. (2012). Using speech rhythm knowledge to improve dysarthric speech recognition. International Journal of Speech Technology, 15(1), 57–64. https://doi.org/10.1007/s10772-011-9104-6

Shahamiri, S. R. (2021). Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29, 852–861. https://doi.org/10.1109/TNSRE.2021.3076778

Shahamiri, S. R., & Binti Salim, S. S. (2014). Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. Advanced Engineering Informatics, 28(1), 102–110. https://doi.org/10.1016/j.aei.2014.01.001

Shao, Y., Wang, Y., Povey, D., & Khudanpur, S. (2020). PyChain: A Fully Parallelized PyTorch Implementation of LF-MMI for End-to-End ASR. ArXiv:2005.09824 [Cs, Eess]. http://arxiv.org/abs/2005.09824

Sharma, H. V., & Hasegawa-Johnson, M. (2013). Acoustic model adaptation using

in-domain background models for dysarthric speech recognition. Computer Speech & Language, 27(6), 1147–1162. https://doi.org/10.1016/j.csl.2012.10.002

Shi, X., Yu, F., Lu, Y., Liang, Y., Feng, Q., Wang, D., Qian, Y., & Xie, L. (2021). The Accented English Speech Recognition Challenge 2020: Open Datasets, Tracks, Baselines, Results and Methods. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6918–6922. https://doi.org/10.1109/ICASSP39728.2021.9413386

Sidi Yakoub, M., Selouani, S., Zaidi, B.-F., & Bouchair, A. (2020). Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. EURASIP Journal on Audio, Speech, and Music Processing, 2020(1), 1. https://doi.org/10.1186/s13636-019-0169-5

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. ArXiv:1906.02243 [Cs]. http://arxiv.org/abs/1906.02243

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. ArXiv:1409.3215 [Cs]. http://arxiv.org/abs/1409.3215

Swamy, S., & Ramakrishnan, K. V. (2013). Evolution of Speech Recognition – A Brief History of Technology Development. 5.

Tan, K., & Wang, D. (2018). A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement. Interspeech 2018, 3229–3233. https://doi.org/10.21437/Interspeech.2018-1405

Tolba, H., & El_Torgoman, A. (2009). Towards the Improvement of Automatic Recognition of Dysarthric Speech. Computer Science and Information Technology, International Conference On, 0, 277–281. https://doi.org/10.1109/ICCSIT.2009.5234947

Trewin, S. (2018). AI Fairness for People with Disabilities: Point of View. ArXiv:1811.10670 [Cs]. https://doi.org/10.48550/arXiv.1811.10670

Trewin, S., Basson, S., Muller, M., Branham, S., Treviranus, J., Gruen, D., Hebert, D., Lyckowski, N., & Manser, E. (2019a). Considerations for AI fairness for people with disabilities. AI Matters, 5(3), 40–63. https://doi.org/10.1145/3362077.3362086

Tzimas, D., & Demetriadis, S. (2021). Ethical issues in learning analytics: A review of the field. Educational Technology Research and Development, 69(2), 1101–1133. https://doi.org/10.1007/s11423-021-09977-4

Vachhani, B., Bhat, C., & Kopparapu, S. K. (2018). Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. Interspeech 2018, 471–475. https://doi.org/10.21437/Interspeech.2018-1751

Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. International Journal of Intelligent Systems, 36(4), 1619–1643. https://doi.org/10.1002/int.22354

van Nood, R., & Yeomans, C. (2021). Fairness as Equal Concession: Critical Remarks on Fair AI. Science and Engineering Ethics, 27(6), 73. https://doi.org/10.1007/s11948-021-00348-z

Walshe, M., & Miller, N. (2011). Living with acquired dysarthria: The speaker's perspective. Disability and Rehabilitation, 33(3), 195–203. https://doi.org/10.3109/09638288.2010.511685

Wang, D., Wang, X., & Lv, S. (2019). An Overview of End-to-End Automatic Speech Recognition. Symmetry, 11(8), 1018. https://doi.org/10.3390/sym11081018

Wang, D., & Zheng, T. F. (2015). Transfer learning for speech and language processing. 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 1225–1237. https://doi.org/10.1109/APSIPA.2015.7415532

World Economic Forum. (2018). World Economic Fourm White Paper Report 2018: How to Prevent Discriminatory Outcomes in Machine Learning. World Economic

Fourm White Paper Report 2018. Global Future Council on Human Rights 2016-2018.

Xu, Q., Baevski, A., & Auli, M. (2021). Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. ArXiv:2109.11680 [Cs]. http://arxiv.org/abs/2109.11680

Yang, S., & Chung, M. (2020a). Improving Dysarthric Speech Intelligibility using Cycle-consistent Adversarial Training: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies, 308–313. https://doi.org/10.5220/0009163003080313

Yu, F., Yao, Z., Wang, X., An, K., Xie, L., Ou, Z., Liu, B., Li, X., & Miao, G. (2021). The SLT 2021 Children Speech Recognition Challenge: Open Datasets, Rules and Baselines. 2021 IEEE Spoken Language Technology Workshop (SLT), 1117–1123. https://doi.org/10.1109/SLT48900.2021.9383608

Yuan, J., & Bao, C. (2020). Multi-channel Speech Enhancement with Multiple-target GANs. 2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 1–5. https://doi.org/10.1109/ICSPCC50002.2020.9259454

Zaidi, B. F., Selouani, S. A., Boudraa, M., & Sidi Yakoub, M. (2021). Deep neural network architectures for dysarthric speech analysis and recognition. Neural Computing and Applications. https://doi.org/10.1007/s00521-020-05672-2

Zhao, Y., Kuruvilla-Dugdale, M., & Song, M. (2020). Voice Conversion for Persons with Amyotrophic Lateral Sclerosis. IEEE Journal of Biomedical and Health Informatics, 24(10), 2942–2949. https://doi.org/10.1109/JBHI.2019.2961844

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2020). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. ArXiv:1703.10593 [Cs]. http://arxiv.org/abs/1703.10593

Zue, V. W., & Seneff, S. (1996). - Transcription and Alignment of the TIMIT Database. In H. Fujisaki (Ed.), Recent Research Towards Advanced Man-Machine In-

terface Through Spoken Language (pp. 515–525). Elsevier Science B.V. https://doi.org/10.1016/B978-044481607-8/50088-8