# Transfer Learning Using L2 Speech to Improve Automatic Speech Recognition of Dysarthric Speech

Hillel Steinmetz

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2023

Reading Committee:
Gina-Anne Levow, Chair
Michael Tjalve

Program Authorized to Offer Degree:
Computational Linguistics

University of Washington

## Abstract

Transfer Learning Using L2 Speech to Improve Automatic Speech Recognition of Dysarthric Speech

Hillel Steinmetz

Chair of the Supervisory Committee:
Associate Professor Gina-Anne Levow
Department of Linguistics

Dysarthria is a class of speech disorders associated with impairments to a person's motor system. Dysarthric speech is diverse but is broadly characterized by reduced prosodic, phonation, and articulatory precision (Rowe et al., 2022). Non-native English speech, or L2 English speech, shares acoustic and phonetic features with the speech of several dysarthria subtypes, such as slower and more variable speech rate compared to native, non-dysarthric English speech (Baese-Berk and Bradlow, 2021; Hertrich et al., 2021). L2 English speech also has different phonetic correlates than native-English speech, with phonetic variation more closely resembling a speaker's first language (Flege, 1981). Since L2 speech both shares acoustic features with dysarthric speech and has more diverse phonetic correlates of phonological segments, it should facilitate knowledge transfer when training an ASR model on dysarthric recognition tasks. This study finetunes Wav2vec2 models on two English dysarthric speech datasets, UA-Speech and TORGO, and one English L2 speech dataset, L2-Arctic, using standard finetuning and multitask learning paradigms. It examines whether including L2 speech in the training data improves dysarthric speech recognition in speaker-dependent, speaker-independent, and zero-shot settings. Our results suggest that including L2 speech in the training data improves dysarthric speech recognition in speaker-dependent and speaker-

independent settings, with models trained using multitask learning performing better than those trained using standard finetuning.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

ALS: Amyotrophic lateral sclerosis. A neurodegenerative disorder that causes dysarthria.

ASR: Automatic speech recognition.

CER: Character error rate. A metric of a model's performance on speech classification tasks. It uses minimum edit distance to compare two character sequences.

CNN: Convolutional neural network.

CP: Cerebral Palsy. A motor disorder associated with dysarthria.

CTC: Connectionist temporal classification. A loss function used to train ASR models.

DANN: Domain adversarial neural network. A neural network architecture that maximizes the loss of an "adversarial" classifier to adapt a model from one domain to another.

DAT: Domain adversarial training. See DANN.

E2E: End-to-end (model). A term describing an ASR model learning paradigm that trains a model on speech inputs directly on the labeled speech, without the need to train intermediate representations separately.

HMM: Hidden Markov Model. A statistical model whose emissions (sequence of outputs) are determined by a sequence of hidden states.

L1 SPEECH: Speech spoken in a person's native language.

L2 SPEECH: Speech spoken in a person's non-native language (also referred to as accented speech or non-native speech).

MAPSSWE: Matched-Pair Sentence Segment Word Error. A statistical measure that compares WER across models.

VOT: Voice onset time. The duration between stop releases and the onset of voicing. Languages use differences in VOTs to mark phonemic contrasts.

WER: Word error rate. A metric of a model's performance on speech classification tasks. It uses minimum edit distance to compare two word sequences.

SSL: Self-supervised learning. A training method used by ASR models like wav2vec2.0 to pretrain models.

# ACKNOWLEDGMENTS

I am incredibly grateful to the many people who provided me with support, feedback, and encouragement as I wrote this thesis. First, I would like to thank my advisor, Dr. Gina-Anne Levow, whose guidance helped me navigate an area of research to which I had no prior exposure and whose advice and wisdom kept me focused. I would also like to thank Dr. Michael Tjalve whose comments and feedback provided me with new perspectives on this thesis's experiments, methods, and results. Finally, I would like to thank my partner, Emily Knopf, and my friends and family, for their relentless words of kindness.

# DEDICATION

To my parents, Chaim and Lisa, who have never stopped encouraging my curiosity.

# Chapter 1

# INTRODUCTION

Dysarthria is a motor speech disorder that impedes a person's ability to accurately coordinate the articulator movement necessary for speech (Hertrich et al., 2021). There are four subtypes of dysarthria, each associated with a different set of acoustic and phonetic characteristics. Some dysarthria subtypes are associated with specific neuromuscular or movement disorders, such as Parkinson's disease and hypokinetic dysarthria; however, many motor disorders cannot be linked to a particular subtype, such as Cerebral Palsy (CP) or multiple sclerosis (MS) (Hertrich et al., 2021; Rowe et al., 2022). Broadly, many dysarthrias are acoustically characterized by slower and more variable speech rates, reduced variations in pitch, increased nasality, and imprecise consonant articulation. Still, there can be considerable variation in the acoustic qualities within and between subtypes (Rowe et al., 2022).

Since people with dysarthria often have difficulty performing tasks requiring motor coordination, they stand to benefit from ASR systems, like virtual assistants. Many individuals may find speech gestures easier to perform than the fine-motor movements required by keyboards and touchscreens. However, off-the-shelf ASR models often perform poorly on dysarthric speech. For instance, Gutz et al. (2022) found that Google Cloud ASR's word recognition rates for dysarthric speech decreased sharply for speakers with lower intelligibility ratings. The study found a nonlinear relationship between the system's word recognition rates and human recognition rates: the system's performance degrades quadratically in relation to linear decreases in human word recognition rates. For instance, when human word recognition rates decreased from 94% to 83% between mild and moderate dysarthria groups, the system's word recognition rates decreased from 77% to 50%. Speaker-adaptive models are good alternatives to off-the-shelf models, performing quite well on dysarthric speech

recognition. However, speaker-adaptive models can take some time to train and require the speaker to provide it with data, which can be effortful for the speaker (Rowe et al., 2022).

Because of the issues with existing dysarthric speech training sets[1] and the difficulty associated with collecting dysarthric speech data, it is worthwhile to consider how speech data from other domains can be used to develop ASR models for dysarthric speech recognition. If knowledge from other speech domains can be successfully transferred to dysarthric speech recognition, it could lessen the need for data collection. While these methods cannot substitute high-quality dysarthric speech data, they can improve models in the interim or help models converge faster and with fewer resources.

Transfer learning is a common approach to improve models when there is insufficient data to train a model on a particular task. Transfer learning broadly describes techniques that train a model using data from similar but different domains. It theoretically teaches the model representations of features shared across the two domains, improving its performance on the target task (Jurafsky and Martin, 2022). This study is interested in transfer learning because it can teach models to represent certain acoustic or phonetic features of dysarthric speech using non-dysarthric speech data. It can also teach models to learn features that are robust to acoustic variations found in dysarthric speech by training on data from multiple similar domains.

Identifying appropriate source domains for transfer learning is of particular interest to this thesis. Hernandez et al. (2022) found that multilingual speech models serve as better feature extractors than monolingual models for models trained to recognize monolingual dysarthric speech. The authors attribute the improved results to the multilingual model learning shared representations of similar phonemes across languages. Because these phonemes differ phonetically across languages, the model learns shared representations that span a more extensive set of phonetic correlates. Following this hypothesis, leveraging data from a domain containing similarly diverse phonemic representations to multilingual speech while more closely re-

[1] See Rowe et al. (2022) for more on the lack of diversity within dysarthric speech datasets.

sembling English dysarthric speech would lead to further improvements in English dysarthric speech recognition.

L2 speech (also known as non-native speech or accented speech) often has acoustic and phonetic patterns similar to those observed in a speaker's native language (Flege, 1981). L2 speech also shares acoustic characteristics with dysarthric speech, such as reduced speech rates and variability in voice onset times (VOTs). In other words, L2 speech contains more varied phonetic correlates, like multilingual speech data and acoustic features specific to dysarthric speech. For these reasons, L2 speech may be a good source domain for transferring knowledge to dysarthric speech recognition.

L2 speech might also address potential drawbacks associated with using multilingual speech as a source domain for dysarthric speech recognition. For a multilingual model to acquire representations robust to the acoustic alterations seen in dysarthric speech, it needs to generalize across languages. However, different languages have different phonotactic constraints and phonetic alternations. Consequently, a multilingual model may be sensitive to contextual information in latent speech representations. For instance, a multilingual model might incorrectly classify a speech segment with phonetic correlates resembling phoneme $p$ in language $l$ if $p$ is found in a context that violates $l$'s phonotactic constraints. The model may instead categorize the speech segment as another phoneme that better satisfies the phonotactic constraints in $l$. In other words, a language's phonotactic constraints may bias a model. For dysarthric speech recognition, this bias can lead to incorrect predictions since phones may appear in previously unseen contexts. L2 speech might diminish the influence of a phone's neighborhood on a model's classification decision. L2 speech data, especially multi-accent speech data, explicitly trains a model to transcribe more phonetically-varied data on the same set of phoneme or grapheme sequence labels. Training a model on speech labeled by the same grapheme sequences might reduce context bias. Additionally, the phonetic features of L2 speech tend to occupy an intermediate space between L1 and L2 phonetic targets while still varying in ways similar to the speaker's L1 (Flege, 1981; Vaughn et al., 2019). The higher proximity to L1 targets might make L2 speech more effective at training

a model to learn shared representations for phonemes: L2 speech phones might overlap with L1 speech phones along phonetic features at a higher frequency, making it more challenging for a model to learn accent-specific decision boundaries.

### *1.1 Research questions*

This thesis examines whether transfer learning techniques that use L2 speech as a source domain improve a model's performance on dysarthric speech recognition tasks. Specifically, it investigates whether finetuning Wav2vec2 on dysarthric and L2 speech datasets improves dysarthric speech recognition compared to models trained solely on dysarthric speech datasets. Secondary research questions of the study include identifying acoustic characteristics of L2 speech and dysarthric speech that degrade performance and comparing two training paradigms. The present study's research questions can be summarized as follows:

1. Can L2 speech data improve dysarthric speech recognition?

    (a) Can L2 speech improve dysarthric speech recognition without including dysarthric speech data at training time? In other words, are the features of L2 speech similar enough to dysarthric speech features to facilitate knowledge transfer in a zero-shot setting, or do the models need to learn these shared representations explicitly?

    (b) If L2 speech data does improve dysarthric speech recognition, does a multitask learning architecture lead to further improvements?

2. What acoustic characteristics of L2 and dysarthric speech have the greatest effect on model performance?

### *1.2 Study contributions*

This study addresses the research questions above by comparing Wav2vec2 models finetuned on L2 and dysarthric speech datasets to models trained solely on dysarthric speech

datasets. We train models under three experimental settings: speaker-dependent, speaker-independent, and zero-shot, to examine the extent to which models can successfully generalize the knowledge transferred from L2 speech. The speaker-dependent and speaker-independent experiments suggest that adding L2 speech to the training data improves word error rates (WERs) and character error rates (CER). In speaker-dependent experiments, a multitask architecture further improved model performance. However, in speaker-independent experiments, only the multitask model consistently outperformed the model finetuned solely on dysarthric speech. In zero-shot experiments, L2 speech is only marginally beneficial and primarily relegated to a specific subset of the data, requiring further study.

Taken together, the results indicate L2 speech can improve the performance of dysarthric speech recognition. The results of this study contribute to research identifying appropriate source domains to facilitate transfer learning for dysarthric speech. They also underscore the importance of using diverse speech data when developing off-the-shelf ASR technologies. The study demonstrates that the Curb-Cut Effect also applies to the development of ASR technologies: designing technologies accessible to one demographic of users often benefit a larger population than initially planned.

### *1.3 Outline of thesis*

Chapter 2 discusses the linguistic similarities between L2 and dysarthric speech and recent advancements in speech recognition. The Chapter also provides an overview of previous approaches to improving dysarthric and L2 speech recognition. Chapter 3 discusses the datasets, methods, and model architectures used by the study to finetune Wav2vec2 for dysarthric speech detection. Chapter 4 details the preprocessing steps, experiments, and model training procedures. Chapter 5 presents the outcomes of the experiments, and Chapter 6 discusses these results. Chapter 7 summarizes the study and its conclusions.

# Chapter 2

# LITERATURE SURVEY

Dysarthria is a speech disorder associated with motor system dysfunction. These impairments to the motor system impact a person's ability to coordinate the muscle movements involved in phonation, including the production of an airstream, voicing, and articulator movements (Darley et al., 1975). Dysarthria occurs in people with neuromuscular disorders like Parkinson's Disease or Amyotrophic Lateral Sclerosis (ALS), movement disorders like Cerebral Palsy (CP), or as a result of traumatic brain injuries (Hertrich et al., 2021).

Impairments in motor control can make it difficult for a person with dysarthria to coordinate articulator movements and reach articulatory targets (Darley et al., 1975). Various acoustic features are associated with dysarthric speech, and several dysarthria subtypes are used to classify different auditory patterns perceived by a listener. However, the typology can be imprecise since it relies on listeners' perceptions (Rowe et al., 2022). Broadly, several dysarthria subtypes can be characterized by mono-loudness (and reduced loudness), monotonous pitch, effortful or "breathy" speech, imprecise consonant and vowel articulation, and nontypical speech rate (including a slow speech rate or more variable speech rate) (Hertrich et al., 2021; Darley et al., 1975). The severity of these speech patterns can change over time, depending on the nature or progression of its associated disorder. It can also vary over the course of a day due to fatigue or medication (Rowe et al., 2022).

## *2.1 Comparing L2 Speech and Dysarthric speech*

L2 speech and dysarthric speech share several acoustic features. L2 speech and dysarthric speech are generally spoken at a slower rate. Within-speaker speech rates are also often more variable for L2 and dysarthric speech (Rowe et al., 2022; Baese-Berk and Bradlow, 2021).

At the segmental level, stop consonants have more variable voice onset times (VOTs) (Rowe et al., 2022; Chodroff et al., 2022). Table 2.1 summarizes several acoustic features observed in both L2 and dysarthric speech.

Table 2.1: Comparison between acoustic features of L2 speech and Dysarthric speech.

| Acoustic Feature | | Dysarthric Speech | L2 Speech |
|---|---|---|---|
| Prosody | Speech rate | Slower speech rate (Rowe et al., 2022) | Slower speech rate (Baese-Berk and Bradlow, 2021) |
| | Speech rate variability | More variable speech rate (for some subtypes) (Rowe et al., 2022) | More variable speech rate (Baese-Berk and Bradlow, 2021) |
| | Syllable duration | Reduced durational variability (Liss et al., 2009) | Durational variability depends on L1 (Ordin and Polyanskaya, 2015) |
| Vowel acoustics | Vowel dispersion | Reduced vowel dispersion (not true of all speakers) (Lansford and Liss, 2014) | Vowels disperse along different formants (Xie and Jaeger, 2020) |
| Consonants | Voice onset time | Greater variability in VOTs (Rowe et al., 2022) | Greater variability in VOTs but consistent covariation (Chodroff et al., 2022) |
| | Articulation | Imprecise articulation of consonants (Rowe et al., 2022) | Phonetic variation similar to the variation seen in L1 (Flege, 1981) |

Of particular interest to this study is the fact that both L2 speech and dysarthric speech have slower and more varied speech rates and syllable durations than native English speech.[1] Previous ASR studies using data augmentation techniques have found that uniformly altering speech speed and pitch improves both dysarthric speech recognition (Bhat et al., 2022) and L2 speech recognition (Fukuda et al., 2018). Slower L2 speech rates have been observed in

[1] The observation that L2 speech rates were slower than L1 English speech rates was accurate for all L1s found in the L2-Arctic corpus used in this study, except for Hindi. See Appendix A for more details.

French and German, including among speakers more proficient in their L2 (Trouvain and Möbius, 2014).

Outside of the similarities between L2 and dysarthric speech, this study is interested in L2 speech because it offers opportunities for an ASR model to learn a more diverse set of phonetic correlates. Mapping a more diverse set of phonetic correlates to phonemes could make the model less sensitive to the phonetic variations in dysarthric speech. For instance, people with dysarthria tend to undershoot articulatory targets, resulting in smaller vowel spaces (Hertrich et al., 2021; Lansford and Liss, 2014). This articulatory difficulty results in reliably lower measurements of mean F1 × F2 vowel dispersion, especially front vowel dispersion, to the extent that it can serve as a good diagnostic metric (Lansford and Liss, 2014). L2 speech, on the other hand, differs from L1 speech in vowel dispersion patterns. For instance, English vowels of native Mandarin speakers vary differently along F1 and F2 formants, with more variation along one formant and less variation along another (Xie and Jaeger, 2020). Training an ASR model on speech data with more varied vowel dispersion patterns might result in it learning larger vowel spaces. A wider range of phonetic correlates along all formants can increase the likelihood of dysarthric speech falling into those ranges. In other words, because vowels in dysarthric speech have smaller ranges of corollary formant values, wider ranges along more dimensions are needed to categorize these vowels correctly—these larger vowel spaces can be learned from L2 speech data. This hypothesis implies that speakers whose L1 has many vowels or vowels with dispersion patterns substantially from English would be wise choices when selecting a source domain.

### *2.2 Recent Advances in Automatic Speech Recognition*

Neural networks became a popular choice for the development of ASR systems in the 2010s, replacing traditional "hybrid" systems that used Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs). These new, neural-network-based end-to-end (E2E) systems have achieved high benchmark scores on ASR tasks (Li, 2022). The development of the Connectionist Temporal Classification (CTC) algorithm facilitated the increased use of

neural network-based models. The algorithm adapts the beam search algorithm to calculate grapheme probabilities directly from a sequence of inputs corresponding to audio frames (Graves et al., 2006). Calculating losses directly from a sequence of probabilities does not require intermediate representations (such as GMMs and mel-frequency spectrograms) or domain-specific knowledge (such as phonetic annotations of text data) to train a model (Jurafsky and Martin, 2022; Bahdanau et al., 2016; Li, 2022).

### *2.2.1 Wav2vec and self-supervised learning*

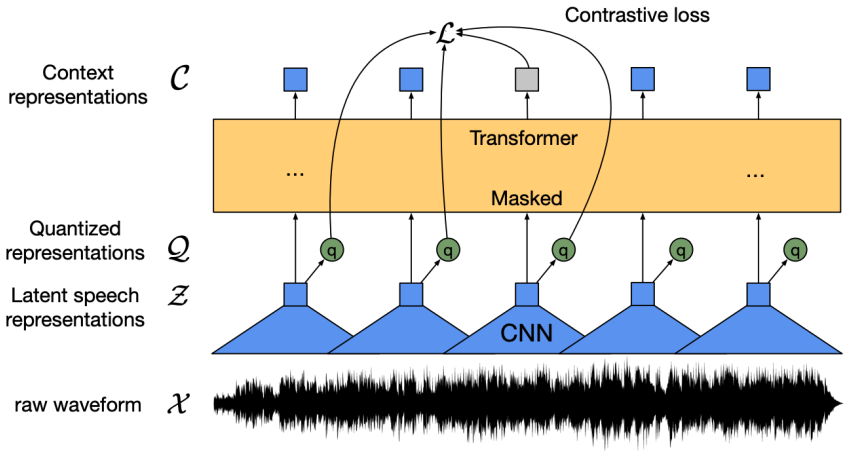



Figure 2.1: System overview of wav2vec2.0. The model jointly learns both to generate quantized speech representations and to interpret a sequence of quantized representations in context. The CNN is pretrained using the noise contrastive loss before the network is finetuned using the CTC loss (Baevski et al., 2020).

A more recent development in ASR is the use of unsupervised and self-supervised learning (SSL) techniques to train E2E neural-network-based ASR models. Wav2vec (Schneider et al., 2019) and Wav2vec2 (Baevski et al., 2020) pioneered SSL techniques using noise contrastive estimation (NCE) to distinguish intermediate representations obtained from a convolutional neural network (CNN) feature extractor (Mnih and Teh, 2012). The NCE loss function trains the model to distinguish one intermediate representation (or quantized unit)

in the audio sequence from the remaining representations (Schneider et al., 2019). The pre-trained model is then finetuned on labeled data with the CTC objective function. Wav2vec2 improved upon the original Wav2vec model by using a transformer to decode the sequences of quantized units as graphemes instead of a convolutional neural network like the one used in the original wav2vec model (Baevski et al., 2020). The authors found that the transformer layer substantially improved word error rates and reduced the data needed to achieve strong performance.

## *2.3 Previous Approaches to Improve Dysarthric and L2 Speech Recognition*

The common approaches to improve dysarthric and L2 speech recognition models can be categorized into three groups: data augmentation, domain adversarial training, and parameter-sharing methods.

### *2.3.1 Data augmentation*

Data augmentation involves either synthesizing new dysarthric speech data or transforming dysarthric speech data to resemble non-dysarthric speech. Vachhani et al. (2018) took the former approach. They performed acoustic transformations to make non-dysarthric speech data more closely resemble dysarthric speech. Specifically, they modified speech duration and tempo, the latter modification applying to the most sonorant portions of a syllable. The authors found that these augmentation methods reduced word error rates (WERs) for DNN-HMM models, with the largest improvements being observed with the recognition of more severe dysarthric speech (Vachhani et al., 2018). In addition to perturbing duration, tempo, and volume, Bhat et al. (2022) reduced WER scores in an autoencoder model by randomly applying acoustic transformations to simulate hypernasality, breathiness, and stuttering. They noticed significant improvement by simply altering speech speed and tempo.

Modifying speech rate can also improve L2 speech recognition. Fukuda et al. (2018) found that simply modifying the speed of L1 utterances by a multiplicative factor between 0.9 and 1.1 (and, in effect, simultaneously altering pitch) improved L2 speech WER of CNN-based

models for native speakers of Latin American Spanish or East Asian languages. Zhang et al. (2022) found that data augmentation improves the performance of HMM-DNN systems on recognizing non-native Dutch. However, unlike the previous study, they found alterations to pitch to be more effective than speed alterations at lowering word error rates. Zhang et al. (2022) also scaled loudness by multiplicative factors between 0.9 and 1.1 but did not notice significant improvements in WER from alternations to loudness.

### *2.3.2 Parameter-sharing methods*

This study defines parameter-sharing methods as a broad class of methods that remedy data shortages by sharing model parameters learned from one task to improve performance on another. These methods include pretraining on one task and finetuning them on another or multitask learning (Jurafsky and Martin, 2022; Chen et al., 2021).

#### *Finetuning*

Finetuning involves training an existing model on a new but similar task. Several studies have found that finetuning can improve dysarthric speech recognition. Shor et al. (2019) finetuned a pretrained model to recognize L2 and dysarthric speech. They found that finetuning a model pretrained on unaccented speech can effectively train speaker-specific dysarthric and L2 speech models. They also found that finetuning specific parameters resulted in more improvements than finetuning the whole model. Vásquez-Correa et al. (2021) found that pretraining and finetuning CNNs on source and target languages improved dysarthric speech detection compared to models trained with randomized parameters, but only when the models trained on the source data were sufficiently accurate.

#### *Pretrained E2E models*

Several studies have shown that E2E models can be successfully finetuned on new domains and possibly learn from similar features in different domains. Xu et al. (2022) found that

pretraining large acoustic models on multiple languages improves phoneme recognition compared to monolingual pretraining, indicating that models might benefit from diverse phonemic correlates. Shibano et al. (2021) found that finetuning Wav2vec2 is an effective method for improving the recognition of L2 speech. Finetuning Wav2vec2 on many accents enabled the model to correctly recognize speech from unseen accents.

Few studies have investigated using pretrained SSL E2E models to improve dysarthric speech models, but a few have shown some promise. Pretrained, SSL E2E models have shown some promise in detecting Alzheimer's disease in speech (Gauder et al., 2021). The speech of people with Alzheimer's tends to involve prosodic alterations and an increase in pauses, indicating that Wav2vec2 can potentially detect similar prosodic alterations in dysarthric speech. Hernandez et al. (2022) found that speech embeddings obtained from pretrained SSL models outperform filter banks. Additionally, they found that models trained on embeddings obtained from Wav2vec2-XLSR, a multilingual version of Wav2vec2, outperformed monolingual models.

Other studies have found that pretrained SSL models might not successfully finetune to recognize dysarthric speech. Baskar et al. (2022) found that Wav2vec2 would not converge when finetuned on specific speakers without providing additional acoustic features as inputs. Similarly, Violeta et al. (2022) found that finetuning Wav2vec2 had lower performance than models trained on mel-scale spectrograms and models pretrained using SSL. However, these studies only trained models on the UA-Speech dataset, leaving open the question of whether training on data from multiple domains can help Wav2vec2 converge during finetuning.

*Multitask learning*

Multitask learning (MTL) is a parameter-sharing method that facilitates transfer learning across domains by simultaneously training a model on multiple distinct tasks (Zhang and Yang, 2017). MTL trains a base model on multiple similar tasks to improve performance on one or more tasks (Chen et al., 2021). There are many configurations for training on multiple tasks, but in its most straightforward configuration for neural networks, a model

branches off into separate sequences of layers where the loss is calculated separately for each task and propagated back to their shared parameters (Chen et al., 2021).

Multitask learning architectures have improved the recognition of dysarthric and L2 speech. Takashima et al. (2019) found that training a model to jointly recognize English and Japanese dysarthric speech data improved Japanese dysarthric speech recognition. Vásquez-Correa et al. (2018) improved a CNN classifier of dysarthric and non-dysarthric speech by training on an auxiliary task of predicting the speaker's scores on the Frenchay Dysarthria assessment. Ding et al. (2021) and Korzekwa et al. (2019) both trained models with an auxiliary task that aimed to use latent representations to reconstruct mel-spectrograms that more closely resemble healthy speech. Korzekwa et al. (2019) 's neural network-based model improved the classification of dysarthric speech while Ding et al. (2021) improved the word error rates of its neural network-based model.

Other studies found MTL to be an effective training paradigm for improving L2 speech recognition. One study, Jain et al. (2018), trained a neural network-based model to jointly classify accents and recognize accented speech, improving WER over baselines. Speech recognition of different accents can also be considered independent tasks; Yang et al. (2018) trained a BiLSTM model that branched into separate nodes to classify American-accented and British-accented speech and found that WER decreased for both accents.

### *2.3.3 Domain adversarial training*

Domain adversarial training (DAT) is a parameter-sharing technique that transfers knowledge from a source domain to a target domain by maximizing the similarity between features across both domains. DAT involves training a domain adversarial neural network (DANN) consisting of a target task and a domain classifier that predicts whether the output belongs to a domain. The technique trains the network by minimizing the loss of a source domain task while maximizing the loss of a domain classifier that predicts whether the output belongs to a domain. It accomplishes this task by defining a gradient reversal layer (GRL) so that both the base model and domain classifier can update their parameters in the opposite

direction of the gradients, but with the domain classifier's losses later reversed when updating the parameters shared by the domain classifier and source task (Ganin et al., 2016). The advantage of DAT is that it can be done in an unsupervised manner; models can be trained without knowing the labels in the target domain, forcing models to learn domain-invariant features (Ganin et al., 2016).

DAT is useful for classifying dysarthric speech when labeled dysarthric speech data is unavailable or inaccurate. Woszczyk et al. (2020), used DAT to train a CNN-based model alongside a domain classifier tasked with classifying speech as dysarthric or not. The authors found that DAT performs better than baseline models when trained on unlabeled dysarthric speech data (Woszczyk et al., 2020). DAT can also be used when there are differences between two dysarthric speech datasets. Wang et al. (2021) found that treating one dysarthric speech dataset (UA-Speech or TORGO) as a source domain and the other as a target domain improves model performance when the target domain is unlabeled.

DAT has also been shown to be effective at recognizing the speech of unseen accents. Das et al. (2021) pretrained a QuartzNet model that included a classifier whose objective was to classify accents. They found that including an adversarial classifier improves the model's ability to recognize non-US accents, even when accented speech was unlabeled.

## *2.4 Chapter Summary*

This chapter discusses acoustic and phonetic research to identify similarities between dysarthric and L2 speech. Some features, like speech rate, are shown to substantially improve dysarthric and L2 speech recognition in speech processing literature. This chapter also provides a theoretical foundation for employing transfer learning techniques to improve dysarthric speech recognition models using L2 speech data as a source domain. It discusses several machine learning techniques that can improve L2 and dysarthric speech, noting previous research that used data augmentation, finetuning, multitask learning, and domain adversarial training to improve speech recognition performance. It also discusses a few studies that finetuned pretrained SSL models for dysarthric speech detection.

# Chapter 3

# METHODS

To investigate whether training a model on L2 speech can improve its ability to recognize dysarthric speech, this study finetunes Wav2vec2 (Baevski et al., 2020) on different datasets that either include or excludes L2 speech data. This section of the thesis describes the tools, datasets, preprocessing pipelines, and evaluation methods used to conduct the study's experiments. The precise experimental setup is described in Chapter 4.

## *3.1 Datasets*

The study uses two dysarthric speech datasets, TORGO and UA-Speech, and L2-Arctic, an L2 speech dataset. It uses two dysarthric speech datasets for several reasons: (1) to better balance the number of dataset samples containing L1, dysarthric, and L2 speech, (2) to enable the model to learn more robust representations of dysarthric speech, and (3) to evaluate how well different models can generalize the information they learn from both datasets. Table 3.1 shows a breakdown of samples found in each dataset.

Table 3.1: Number of audio samples within each dataset

| Dataset | L2Arctic | UA-Speech | TORGO |
|---|---|---|---|
| Dysarthric | - | 11,437 | 30,94 |
| Control | - | 9,945 | 5,900 |
| Total | 26,877 | 21,382 | 8994 |

The Universal Access, or UA-Speech, dataset consists of English speech from 15 people with cerebral palsy (4 female and 11 male) and 13 people without dysarthria (4 female and 11 male) (Kim et al., 2008). 11 participants had a diagnosis of spastic dysarthria. The remaining 5 were diagnosed with athenoid dysarthria or a mix of dysarthria subtypes. One speaker, M06, is not included in the dataset used in this study because he did not consent to his data being redistributed. The dataset consists of wav files of single-word utterances from a microphone array of 8 microphones, sampled at 16kHz. The `noisereduce` algorithm (Sainburg, 2019) was used to remove noise from the recordings.[1] The recordings are divided into three blocks. Participants produced utterances of the same 155 words for each block. They also produced speech for 100 words that differed across blocks, for a total of 765 utterances and 455 unique words (Kim et al., 2008). In this study, recordings of the same prompts were included in the dataset and are treated as separate data points.

The TORGO dataset consists of English speech from 7 people (4 male, 3 female) with dysarthria and 7 people (4 male, 3 female) control participants (Rudzicz et al., 2011). All participants with dysarthria had CP, and one had both CP and ALS. The dataset consists of wav files of nonce utterances and single-word or multiple-word utterances taken from the TIMIT corpus. The recordings are saved as wav files and sampled at 16kHz. Recordings were collected across three sessions. The speakers were tasked with reading as many prompts as they could within each time-limited session. So, some speakers have multiple recordings of the same prompts, while others did not complete every prompt provided.

The TORGO and UA-Speech datasets categorize the intelligibility of dysarthric speech differently. For the UA-Speech data, five listeners with no background in language disorders or phonetic transcription provided transcriptions for each speech recording. Speakers were placed into very low, low, medium, and high intelligibility categories based on the accuracy scores of human raters. The TORGO dataset provides Frenchay assessment scores for each speaker with dysarthria, obtained by a speech-language pathologist (Rudzicz et al., 2011).

---

[1] The database was updated to include recordings with this preprocessing step in 2020.

The Frenchay assessment assesses people's ability to move their articulators by asking them to perform tasks like talking or swallowing water. The assessment includes an intelligibility category consisting of three tests for speech interpretability (Enderby, 1983 in Rudzicz et al., 2011).

The L2-Arctic dataset includes recordings of spoken English by 24 non-native English speakers, distributed evenly by gender and L1 (Zhao et al., 2018). The speakers' L1s were Arabic, Mandarin, Spanish, Vietnamese, Korean, and Hindi. Speech recordings for each language were obtained from 2 male and 2 female speakers for each L1. The dataset consists of wav files of short, prompted sentences sampled at 44.1kHz. For this study, the L2-Arctic data was converted to a sample rate of 16Hz.

### *3.2 Finetuning Wav2vec2*

The models for this study were created by finetuning the Wav2vec2 base model on the UA-Speech, TORGO, and L2-Arctic datasets. The Wav2vec2 model is trained on 960 hours of the Librispeech dataset Baevski et al., 2020, which consists of English recordings of audiobooks (Panayotov et al., 2015). Most of the dataset contains L1 English speech. Examining the dataset's metadata, about 2-3% of the Librispeech data is drawn from audio samples spoken by non-native English speakers.[2]

The Wav2vec2 model was downloaded and modified using the `transformers` package (Wolf et al., 2020) for Python. The experiments use the 960-hour checkpoint of the model, which finetuned the unsupervised model on 960h hours of audio from the Librispeech dataset. The study uses single linear layers as decoders for all experiments and paradigms. These decoders map the outputs of the Wav2vec2 model to a set of English characters.

Pasad et al. (2021) found that Wav2vec2 encodes less linguistic content in the last few layers and that these layers change the most during finetuning, leading them to suggest

[2] This figure may be inaccurate since the creators of the dataset state that its annotations are unreliable. Regardless, the percentage of L2 speech in the dataset is likely small.

reinitializing these layers before finetuning the model. So, before training, the weights of the last 2 layers of the transformer decoder are reinitialized.[3]

While conducting the study, several models using different configurations were generated for hyperparameter tuning and evaluating various model configurations. These models were assessed using the validation set. Information on hyperparameter selection and early exploratory experiments can be found in Appendix B, one of which serves as a sensitivity analysis for the study.

### *3.2.1 CTC Loss*

Wav2vec2 models are finetuned by calculating Connectionist Temporal Classification (CTC) loss. CTC loss maximizes the probability that the model input corresponds to its label. Contextual representations are obtained from the final hidden layer of the Wav2vec2 transformer. These representations can be transformed into grapheme probabilities using the softmax function. The CTC loss algorithm compares these fixed-length sequences of grapheme probabilities to labeled text of variable length. The output sequence is the same length as the contextual representation generated by the model (the quantized units in Wav2vec2); a longer output sequence is compared to shorter labels by reducing sequences of identical graphemes into a single grapheme and using a special blank grapheme label to mark a sequence of repeated graphemes. The alignment reduction leads to a many-to-one mapping between alignments and labels.

Let $B$ denote the function that reduces alignments, and its inverse $B^{-1}$ denote a mapping between a label $Y$ and a set of corresponding alignments. Assuming conditional independence, the total probability that a sequence of grapheme probabilities, $X$, corresponds to a label, $Y$, is:

$$P(Y|X) = \sum_{A \in B^{-1}(Y)} \prod_{t=0}^{N} P(a_t|X) \tag{3.1}$$

[3] The weights are reinitialized by sampling from a uniform distribution per Huggingface's implementation of Wav2vec2 on GitHub.

where $A = [a_1, ..., a_n]$ is a sequence in the set of sequences that maps to the label $Y$. The value $P(Y|X)$ can be efficiently calculated using a modified version of the beam search algorithm (Jurafsky and Martin, 2022; Hannun, 2017).

The model then minimizes the negative log-likelihood of the input mapping to the correct label. For a set of labels $L$ and a set of inputs $I$, the loss $\mathcal{L}$ is calculated as follows:

$$\mathcal{L} = -\sum_{Y \in L,\ X \in I} \log P(Y|X) \tag{3.2}$$

Because the dysarthric datasets largely contain single-word utterances, while the L2-Arctic dataset contains multi-word sentences, the CTC loss for samples taken from the L2-Arctic dataset tends to be larger than for samples from the dysarthric datasets. To avoid ascribing higher losses to samples from the L2-Arctic dataset, all CTC losses were divided by the number of characters in the true labels, effectively normalizing the loss values across datasets. The loss for a batch is calculated by averaging the losses for all batch inputs. So, for a batch of inputs $B = \{x_1, ..., x_k\}$ with labels $\{y_1, ..., y_k\}$ of lengths $\{T_1, ..., T_k\}$ the loss is calculated as follows:

$$\mathcal{L} = -\frac{1}{|B|} \sum_{i=1}^{k} \frac{\log P(y_k|x_k)}{T_k} \tag{3.3}$$

where $P(y_k|x_k)$ is calculated using equation 3.2.1 and $|B|$ is the number of items in the batch. Examining CTC losses of a subsample of the data during the first epoch of finetuning found that dividing the CTC losses by sequence lengths $T_k$ resulted in dysarthric speech receiving higher loss values (L2 mean loss: 2.18, dysarthric mean loss: 5.52). Without normalizing, we observed the opposite pattern (L2 mean loss: 44.99, dysarthric mean loss: 28.58), which was not desirable given the lower baseline WERs of the L2 data.

## *3.3 Evaluation*

The current study evaluates model performance using word error rate (WER) and character error rate (CER). WER is based on the minimum edit distance (or Levenshtein distance) algorithm, but it calculates substitutions, deletions, and insertions at the word level instead of the character level (Jurafsky and Martin, 2022). After computing the minimum edit distance between the predicted sentences and their corresponding labels, WER is computed as follows:

$$WER = \frac{S + I + D}{N = H + S + D} \tag{3.4}$$

where $S$, $I$, and $D$ are the count of substituted, inserted, and deleted words respectively, $N$ is the total number of words in the label, and $H$ is the number of correct words (Morris et al., 2004).

Because many of the utterances in the dysarthric speech datasets are single words, it is helpful to consider character error rate (CER). CER is calculated using equation (3.3), but the counts for $I$, $S$, $D$, and $H$ are obtained at the character level. However, because English orthography has little correspondence with the phonetic realizations of its represented words, CER may not be a good measure of model performance or of partial correctness.

### *3.3.1 Issues with WER and CER*

Although WER and CER are standard metrics in ASR studies, including dysarthric speech recognition studies, there are significant drawbacks to their use in this context. As noted earlier, there is little correspondence between English orthography and its phonetic realization. Training a model to generate English orthography implicitly trains it on English spelling conventions, obscuring our understanding of how the model processes phonetic information. The metrics provide little insight into how a model internally represents phone segments since the metrics don't account for phonetic features associated with each segment. For

example, the phones [p] and [b] in speech only differ in whether they are voiced (or [+voice] in distinctive feature theory), but CER and WER consider this substitution to be equivalent to replacing [p] with [i], even though the latter differs in voicing place of articulation, and manner of articulation among other differences. It also fails to account for homophony. For instance, two homophonous words (such as "male" and "mail") would be given a WER score of 1 despite being phonemically identical.

Despite their limitations, this study uses WER and CER because no pretrained monolingual Wav2vec2 models are trained on phonemic transcription. Creating a Wav2vec2 model trained on phonemic transcription was not considered since it would require resources not commensurate with the study's scope. While multilingual Wav2vec2 models are trained on phonemic transcription, including other languages would present a confounder for this study.

### *3.3.2 Matched-Pair Sentence Segment Word Error*

The Matched-Pair Sentence Segment Word Error (MAPSSWE) is a parametric statistical test that evaluates whether model outputs are significantly different (Jurafsky and Martin, 2022; Gillick and Cox, 1989). The test divides sequences of words into segments composed of one or more words and calculates a score, $W$, which is the difference in the number of errors each model makes within a segment. The advantage of segmenting the data is that it allows us to assume that errors across segments are independent if the data is segmented at a natural stopping point such as a pause (Gillick and Cox, 1989). If systems have similar WER, the mean difference in errors would be 0, or $\hat{\mu}_z = \frac{1}{n}\sum_{i=0}^{n} Z_i = 0$. In other words, the null hypothesis, $H_0$, is that the two WERs are not significantly different.

With a sufficiently large $n$, the distribution of errors $W$, should be approximately normal, allowing us to calculate:

$$W = \frac{\hat{\mu}_z}{\sigma_z/\sqrt{n}} \qquad \text{where, } \sigma_z^2 = \frac{1}{n-1}\sum(Z_i - \hat{\mu}_z)^2 \tag{3.5}$$

We can then calculate, $P(Z > |w|)$ where $w$ is the realized value of $W$ and $Z \sim \mathcal{N}(0, 1)$. If $P(Z > |w|) \geq \alpha$, for a significance level $\alpha$ (set to 0.05 in this study), we reject the null hypothesis (Gillick and Cox, 1989). We use the two-tailed version of the test to compare models trained with L2 and dysarthric speech to those trained on solely dysarthric speech since we are also interested in whether L2 speech significantly worsens model performance. We used the National Institute of Standards and Technology's SCTK software to segment text and calculate MAPSSWE scores (*SCTK* 2021).[4] MAPSSWE test were only performed on the dysarthric speech data—L2 and control data was removed before conducting the tests.

It is important to note that the MAPSSWE test assumes that the errors are normally distributed (Gillick and Cox, 1989). However, because most of the dataset consists of single-word utterances, most of the errors fall take on values of -1, 0, or 1. The limited range of errors and use of discrete values make it difficult to evaluate whether $W$ will be approximately normal. For that reason, the test's outcomes should not be regarded as definitive evidence of performance differences across models. Nevertheless, we report the test's results since it is a helpful tool for comparing model performances, especially because it can be challenging to determine whether WER and CER values are substantially different.

### *3.4 Training paradigms*

Two training paradigms are compared in this study: finetuning and multitask learning.

#### *3.4.1 Finetuning*

The finetuning paradigm involves training a pretrained model on additional data to transfer knowledge from its original domain to new domains or downstream tasks (Jurafsky and Martin, 2022). In this study, the finetuning paradigm is implemented with two datasets: one containing both the dysarthric speech datasets and L2-Arctic and another containing only the dysarthric speech datasets. The latter serves as a control group for the experiments.

[4] The software can be downloaded from Github: https://github.com/usnistgov/SCTK.

Wav2vec2 is then finetuned on these two dataset configurations. The weights of the last two layers of the transformer are reinitialized before training, as suggested by Pasad et al. (2021).

### *3.4.2 Multitask*

Multitask learning also aims to facilitate knowledge transfer, but it accomplishes this goal by training a model on two tasks simultaneously (Zhang and Yang, 2017). This study's multitask training procedure adapts the finetuning paradigm with a different model architecture. It branches the last two layers of Wav2vec2's transformer-based encoder into two separate branches. In other words, inputs for each task are passed into separate copies of the final 2 layers. The motivation for separating the final layers for each task was to allow the model to encode contextual information specific to each domain. Here too, the weights of these last two layers are reinitialized at the start of training, as suggested by Pasad et al. (2021).

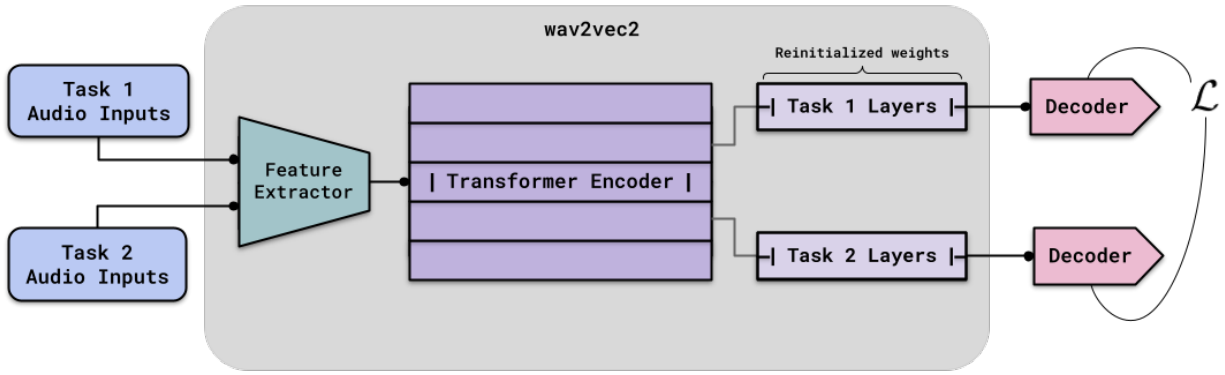



Figure 3.1: A diagram of the multitask model. The transformer encoder branches off at the last two layers before passing its outputs to separate decoders.

Tasks indices are provided as inputs to the multitask models to identify the branch to which the model needs to forward inputs. The branches then provide outputs to two separate decoders consisting of a single linear layer. The multitask models are constructed by duplicating the weights from the linear layer that obtains grapheme probabilities from Wav2vec2's encodings. A diagram of the multitask configuration can be found in Figure 3.1.

Initially, we considered passing the control data from the dysarthric speech datasets to a separate branch. Earlier experiments found that passing the control and dysarthric data to a single branch proved more effective, leading us to pass all data from dysarthric speech datasets to the same branch.

### *3.5 Chapter summary*

This chapter describes the datasets, training paradigms, and evaluation measures used in the study. The chapter describes the training protocols to finetune to models and the architecture of the multitask model. It also describes the corpora used by the study: the UA-Speech, TORGO, and L2-Arctic corpora, and the resulting balance of native English non-dysarthric, native English dysarthric, and L2 speech in the dataset. The models are trained on this data using the CTC objective, normalized by label lengths. Despite their theoretical shortcomings, models are then evaluated using word error and character error rates. Performance is compared across models using the Matched-Pair Sentence Segment Word Error statistic.

# Chapter 4

# EXPERIMENTS

## *4.1 Preprocessing*

### *4.1.1 Audio processing*

I prepare the datasets for training and evaluation tasks by performing several preprocessing and filtering procedures. The L2 Arctic dataset was resampled at 16 kHz, matching the sample rate of the other datasets. Additional preprocessing was only performed on the UA-Speech and TORGO datasets. A table summarizing the preprocessing steps can be found in Table 4.1.

In the UA-Speech dataset, there were 8 audio files for each recording, each representing a different channel of the microphone array or the head microphone. The channels were averaged so that if one sample contained a noisy segment, that noise would be reduced in the averaged recording.[1]

Table 4.1: Audio processing procedures

| Dataset | Process | Motivation |
|---|---|---|
| UA-Speech | Averaged microphone array using Librosa and Soundfile (McFee et al., 2023; Bechtold, 2023). | Averaging the channels from a microphone array should create a less noisy signal. |
| TORGO | Applied noisereduce algorithm to the data (Sainburg, 2019). | Achieve comparable sound quality to the noise-reduced UA-Speech dataset. |

Code and other materials is available on GitHub at https://github.com/hasteinmetz/transfer-learning-for-dysarthria.

[1] For speaker M16, the file `M16_B3_UW100_M8.wav` was corrupted. This channel was ignored when during the averaging procedure.

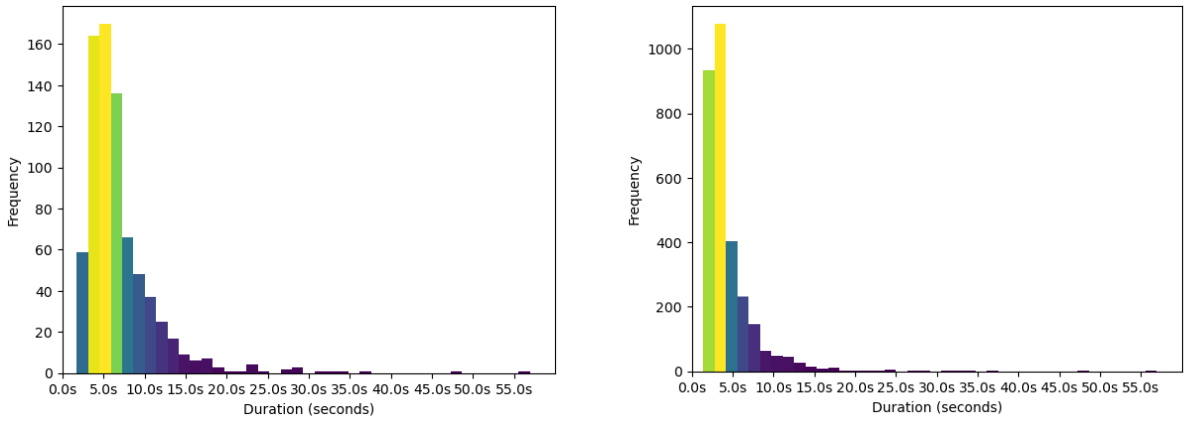
The TORGO dataset contains mono-channel audio recordings from both a head microphone and a microphone array (the audio from the microphone array is already averaged. Listening to the audio files revealed similar levels of audio quality between the head microphone and microphone array. Recordings from the array microphone were used as the audio data for the experiment since it was believed to be of higher quality than the head microphone recordings. The array microphone recordings were originally sampled at 44.1kHz. Audio data from the head microphone or array microphone was missing for some speakers. In those cases, recordings from the other microphone were used.[2] The TORGO dataset also contained speech files with repeated syllables and image-prompted speech. These recordings were removed from the dataset. The repeated syllables were removed since they were not annotated with the number of repetitions made by the speaker. The image-prompted speech was excluded since it lacked annotations. We applied the `noisereduce` algorithm (Sainburg, 2019) to the TORGO speech data to reduce the differences in sound quality between the dysarthric speech datasets since the algorithm was also applied to the UA-Speech data. A short description of the `noisereduce` algorithm can be found in Appendix A.

### *4.1.2 Filtering*

Both the TORGO and UA-Speech datasets contained several long recordings. Often these recordings contained audio of the speaker repeating the prompt several times or contained audio of a researcher guiding the speaker. Additionally, long recordings interfered with model training and evaluation because they required more memory, a problem compounded by the need to pad batches to the same input lengths. Therefore, we used recording length as a coarse heuristic for a recording's quality and removed recordings greater than 15 seconds in length. In total, 95 files were removed from the datasets, 69 from the UA-Speech dysarthric group and 26 from the TORGO dysarthric group.

---

[2] There are some audio recordings without corresponding prompts, most belonging to speaker F03. These recordings are excluded. Additionally, typos were found in one of F04's and one of MC02's prompts. These were corrected.




(a) Durations for speaker M01 from UA-Speech (b) Durations for "very low" intelligibility group

Figure 4.1: Distribution of recording durations for (a) speaker M01 and (b) speakers with "very low" intelligibility ratings within the UA-Speech dataset

The 15-second upper limit was chosen by calculating the 95th percentile of recording durations. The speaker with the highest 95th percentile was M01 of the UA-Speech dataset, at around 15 seconds. Additionally, we visually examined the distributions of recording durations for each speaker to ensure the long recordings were outliers. A histogram of recordings length for the speaker M01 from the UA-Speech dataset (shown in figure 4.1a) indicates that recordings with durations longer than 20 seconds are outliers. M01 belonged to the "very low" intelligibility group within the UA-Speech dataset. Examining the distribution of recording durations for the "very low" intelligibility group reveals that most utterances in the group were shorter than 10 seconds, supporting the decision to exclude sentences longer than 15 seconds.

## *4.2 Modified speech experiment*

We conducted two small experiments to study how modifications to acoustic features associated with dysarthric speech improve or degrade Wav2vec2 performance. One experiment simulated dysarthric speech by modifying the loudness, pitch, and speech rate of audio from

the control group of the UA-Speech dataset. The other experiment modified the speech rate of the dysarthric group from the UA-Speech dataset.

### *4.2.1 Simulating dysarthric speech*

The experiment uses the base Wav2vec2 (960h) model with a pretrained CTC modeling head. We modify native English and non-dysarthric speech from the control group of the UA-Speech dataset using Praat (Boersma and Weenink, 2023) and compare model performance to the unmodified utterances. The speech was modified along three acoustic dimensions: pitch, speech rate, and loudness to resemble several speech patterns that characterize dysarthric speech. The pitch was made more monotone, the speech rate was slowed, and loudness was manipulated to be mono-loud or linearly decrease halfway through the utterance. The implementation of these modifications is detailed below.

The pitch was monotonized by obtaining data points representing F0 measurements at different times using Praat. The data points were transformed to be closer to the mean pitch of the utterance. The distance between a point, $x_i$, and the mean pitch, $\mu$, was scaled by a factor $a$, where $a$ is the minimum of 1, and the point's Z-score is divided by a constant $c$. This scaled distance is subtracted from the point (see equation 4.2.1). The equation scales the distance term such that greater amounts are subtracted from larger distances. It also alters the pitch contour by replacing peaks with two much smaller ones on either side. The constant $c$ was used to alter the strength of the pitch monotonization. For this study, $c$ was set to 1.1, which was thought to make the speech as monotone as possible without sounding unnatural to a listener.

$$f(x_i) = x_i - a(x_i - \mu) \qquad a = \min\left(\left|\frac{x_i - \mu}{\sigma \cdot c}\right|, 1\right) \tag{4.1}$$

Speech rate was altered using Praat's `Lengthen (overlap add)` command, which lengthens an audio recording while preserving its quality and the speaker's vocal characteristics

using the TD-PSOLA algorithm (Moulines and Charpentier, 1990). Speech rates were modified by factors of 0.8, 1.2, and 1.4.

The modifications to loudness were designed to mimic different dysarthria subtypes. One modification linearly decreased loudness halfway through the utterance. This modification is intended to simulate flaccid dysarthria where utterances involve short expressions, breathiness, and mono-loudness (Darley et al., 1975). The other modification reduced variation in the loudness of speech segments. This choice was intended to simulate the mono-loudness that can occur in the speech of people with spastic and flaccid dysarthria (Darley et al., 1975). The audio was transformed by mapping intensity points louder than 65dB to the maximum audio intensity. Performing this transformation only on intensity points above 65dB should avoid applying the transformation to non-speech segments. Additionally, to avoid raising the loudness of spacebar presses (which are above 65dB in several recordings), no intensity points in the last 10% of the audio were transformed.

#### *4.2.2 Altering dysarthric speech*

It is worth examining whether we observe any immediate improvements in WER by applying speech rate transformations to dysarthric speech, given the large impact of speech rate observed on dysarthric speech recognition in Vachhani et al. (2018) and (Bhat et al., 2022). For this experiment, dysarthric speech data from the UA-Speech dataset was sped up by factors of 0.7, 0.8, and 0.9 using the TD-PSOLA algorithm (Moulines and Charpentier, 1990). The performance of Wav2vec2 on the modified data was compared to the original data, with more detailed analyses for each intelligibility group. For this experiment, the data were filtered to utterances shorter than 10 seconds to avoid long silences as a potential confound. The utterances are filtered before they are modified.

### *4.3 Finetuning Wav2vec2 experiments*

The experiments described below examine whether finetuning Wav2vec2 with L2 speech data improves the model's ability to recognize dysarthric speech. A secondary goal of these exper-

iments is determining the architecture that best finetunes Wav2vec2 to recognize dysarthric speech.

Table 4.2: Experiments

| Experiment | Baseline | Finetune | | Multitask |
|---|---|---|---|---|
| | Wav2vec2-960h | Dys | Dys + L2 | L1/L2 Branch |
| Speaker-dependent | X | X | X | X |
| Speaker-independent | X | X | X | X |
| Zero-shot | X | X | X | X |

The study involves three finetuning experiments: a speaker-dependent experiment, a speaker-independent experiment, and a zero-shot learning experiment. The speaker-dependent experiment determines whether models benefit from training on L2 speech. The speaker-independent experiment strengthens the findings of the speaker-dependent experiment by evaluating whether models trained on L2 speech also better generalize to unseen dysarthric and L2 speakers. The zero-shot experiment tests the strength of the hypothesis by determining whether a model would improve its ability to recognize dysarthric speech without any explicit training on it. In other words, it evaluates whether L2 speech alone trains a model to attend to acoustic features shared between L2 and dysarthric speech.

All experiments evaluated the performance of a baseline model, a model finetuned on just the control data, a model finetuned on both L2 speech and the control data, and a multitask model trained on both L2 speech and the control data. Comparing models finetuned on just dysarthric speech to the models trained on both dysarthric and L2 speech can identify whether including L2 speech improves dysarthric speech recognition. Comparing the multitask and finetuned models should help identify whether a multitask architecture further improves WER scores.

The experiments differ in how the datasets were split used and the hyperparameters chosen for the models (specifically, weight regularization and learning rate). They both use

the 960h Wav2vec2 model as a baseline. A summary of all finetuning experiments is found in Table 4.2.

### *4.3.1 Speaker-dependent*

The speaker-dependent experiment trained models on dysarthric speech, L2 speech, and non-dysarthric speech and evaluated their performance on a held-out dataset. The experiment is intended to determine whether L2 speech can improve dysarthric speech detection.

#### *Dataset splits*

We split the dataset into train, validation, and test sets for speaker-dependent experiments using an 80:10:10 split ratio. The datasets were stratified so speakers are equivalently distributed across the training and held-out datasets.

### *4.3.2 Speaker-independent*

The speaker-independent experiment splits the training and held-out sets by speaker instead of by utterance. In other words, the experiment trains on dysarthric speech, L2 speech, and non-dysarthric speech from one set of speakers and evaluates model performance on another set of speakers, stratified evenly across groups. The experiment should determine whether L2 speech improves model performance on unseen speakers. In other words, it helps determine whether performance improvements resulting from L2 speech are sufficiently robust to generalize to unseen speakers.

#### *Dataset splits*

The dataset was split using a 75:12.5:12.5 ratio for train, validation, and test sets. This ratio ensures that the validation and training sets contain at least one speaker from every UA-Speech intelligibility classification and L2-Arctic L1. Because TORGO ratings for intelligibility are unevenly distributed across speakers, one speaker with intelligibility ratings below

3 and one speaker with intelligibility ratings above 5 were selected for the validation and test sets for a total of 2 "simplified" intelligibility groups. The speakers were selected randomly with an arbitrarily chosen seed for the random number generator to ensure reproducibility. All dysarthric speakers in the test and validation datasets were male.

### *4.3.3 Zero-shot*

The speaker-independent experiment trained models only on non-dysarthric speech and evaluated their ability to recognize dysarthric speech. All models were trained using data from TORGO's and UA-Speech's control groups but differed in their inclusion of L2 speech in the training data. This experiment should strengthen the results obtained from the speaker-dependent experiment by showing that L2 speech facilitates models ASR models recognize dysarthric speech, even without any training on dysarthric speech.

Different hyperparameters were chosen for the zero-shot experiment. Most notably, the learning rate was halved and weight decay was set to a relatively high value of 0.005. Exploratory experiments found that higher weight decay parameters significantly improved model performance. This observation can be explained by the fact that weight regularization leads models to minimize weights representing task-specific or otherwise marginal features (Zhang and Yang, 2017).

#### *Dataset splits*

All the L2 and control data were included in the training set. The dysarthric speech samples are split evenly into validation and test sets, stratified by speaker.[3]

### *4.3.4 Addressing dataset imbalances*

The datasets are not adequately balanced across dysarthric, L2, and non-dysarthric speech. This imbalance can complicate a model's ability to learn speech representations. Part of

[3] A small sample of audio data from the control group and L2-Arctic dataset was copied from the training data to determine if the gradients were vanishing or exploding.

this imbalance is addressed by regularizing the CTC loss function (see 3.3). To address class imbalance further, we weight each task's losses according to its representation in the training set. For each task $i$, we calculate a weight $w_i = \frac{N}{n_i \cdot k}$ , where $N$ is the total number of samples in the dataset, $n_i$ is the total number of samples for task $i$ (e.g., L2 speech), and $k$ is the total number of tasks. CTC losses are multiplied by these weights before the optimizer updates the model parameters.

### *4.4 Training procedure*

All models were trained for 10 epochs with a batch size of 32 (obtained using a gradient accumulation). The learning rate was updated using the AdamW optimizer (Loshchilov and Hutter, 2017) with a warm-up ratio of 0.1 (or one epoch), where the learning rate linearly increases from 0 to the designated learning rate. Gradient norms are clipped at a value of 1.0. All models were trained on 16-bit floating-point numbers to enable larger batch sizes. PyTorch and Huggingface seeds fixed to 2022 enable replicability. Appendix B shows a table of hyperparameters chosen for each experiment.

### *4.5 Chapter summary*

This chapter details the experiments implemented in this study. It describes the study's methods to preprocess and filter audio data from the three datasets. The chapter then details two experiments that examine how modifying audio along certain acoustic dimensions alters Wav2vec2's performance. One experiment alters non-dysarthric speech to simulate dysarthric speech, while the other alters the rate. The chapter also describes three experiments that finetune Wav2vec2 with different datasets and architectures to determine whether L2 speech improves dysarthric speech detection and whether the knowledge transferred from L2 speech can generalize to unseen speakers.

# Chapter 5

# RESULTS

Results from all experiments can be found below. The tables below abbreviate "FT" for finetune and "MT" for multitask. "FT: Dys" describes the model finetuned only on the dysarthric speech datasets and "FT: Dys+L2" describes the model trained on both L2 and dysarthric speech. Datasets and speaker groups are abbreviated as well: "TOR" for TORGO, "UAS" for UA-Speech, "DYS" for dysarthria, and "CTL" for control.

## *5.1 Acoustic properties experiments*

### *5.1.1 Simulating dysarthric speech*

The experiment simulating dysarthric speech revealed a significant effect of decreased speech rate and mono-loudness on WER. It also found monotonized pitch manipulations and increased speech rate to have moderate effects on word recognition. Loudness dropoff had the smallest effect on model performance, lowering performance by only about 1%. The larger the rate increase, the worse the model performed, with a rate increase of 40% raising WER by close to 14% and CER by close to 6%. Mono-loudness also greatly affected Wav2vec2's performance, increasing WER by 13% and CER by close to 7%.

Table 5.1: Performance of wav2vec2 on modified wav files (highest values underlined)

| Original | Loudness | | Pitch | Rate | | | Original |
|---|---|---|---|---|---|---|---|
| | Monotonized | Drop-off | Monotonized | × 0.80 | × 1.20 | × 1.40 | |
| WER | 53.1 | 41.6 | 44.2 | 43.6 | 47.1 | <u>54.1</u> | 39.5 |
| CER | <u>22.3</u> | 15.6 | 17.4 | 17.8 | 18.3 | 21.0 | 14.7 |

MAPSSWE comparisons on dysarthric speech data found all modifications to be significantly different from the original audio file. Table C.1 contains pairwise comparisons of each audio modification and can be found in Appendix C.

### *5.1.2 Altering dysarthric speech*

Table 5.2: Performance of wav2vec2 on modified wav files (highest values underlined)

| Metric | Rate | | | Original |
|---|---|---|---|---|
| | × 0.70 | × 0.80 | × 0.90 | |
| WER | 115.5 | 116.6 | 117.4 | <u>119.4</u> |
| CER | 68.5 | 68.5 | 68.4 | <u>68.8</u> |

For the modified dysarthric speech data, the increased speech rate seems to improve overall WER by 4% but has a negligible effect on CER (Table 5.2). If the data is examined in greater detail, we notice that the directionality of WER changes is dependent on the speaker. Table 5.3 shows WER scores divided by speaker and intelligibility ratings. In general, increased speech rate degrades performance for speakers in higher intelligibility groups and improves performance for speakers in lower intelligibility groups. For instance, scores decrease by over 10% (or roughly 3.8% more correct transcriptions) for speaker M05 and by 9.7% for speaker F02 (or roughly 1.6% more correct transcriptions). Additionally, the magnitude of the improvements is smaller in the "very low" intelligibility group.

## ***5.2 Speaker-dependent***

The speaker-dependent experiments indicate that L2 speech improves dysarthric speech recognition compared to models trained solely on dysarthric speech, with the multitask model outperforming the finetuning paradigm. The FT: L2+Dys model improved CER by 2.8% and WER by 3.8% in the test set. Implementing the multitask architecture improved

Table 5.3: WER of modified wav files by intelligibility group.

| Intl. | Speaker | Rate: × 0.7 | Rate: × 0.8 | Rate: × 0.9 | Original |
|---|---|---|---|---|---|
| Very low | F03 | **128.6** | 133.2 | 132.1 | 137.2 |
| | M01 | **217.7** | 225.7 | 230.4 | 241.4 |
| | M04 | 131.2 | **130.8** | 131.8 | 138.0 |
| | M12 | **131.8** | 134.2 | 133.1 | 135.9 |
| Low | F02 | **159.5** | 166.8 | 167.9 | 172.1 |
| | M07 | 146.1 | **145.9** | 147.6 | 155.0 |
| | M16 | **116.1** | 119.4 | 124.4 | 125.8 |
| Mid | F04 | **101.2** | 104.0 | 104.2 | 104.4 |
| | M05 | **132.9** | 139.6 | 143.5 | 144.5 |
| | M11 | **118.9** | 123.3 | 124.4 | 124.8 |
| High | F05 | 72.0 | 66.8 | 60.6 | **54.9** |
| | M08 | 65.4 | **61.1** | 62.8 | 63.6 |
| | M09 | 82.3 | 80.9 | **80.2** | 81.5 |
| | M10 | 64.5 | 57.9 | 54.4 | **51.0** |
| | M14 | 70.4 | **66.3** | 70.3 | 67.8 |

CER by an additional 2.5% and WER by an additional 3.4% compared to the standard fine-tuning paradigm. The multitask architecture also slightly improved L2 speech recognition by a little over 1% for CER and 2% for WER. MAPSSWE comparisons on dysarthric speech data found all models to significantly differ from one another. Table C.2 contains pairwise comparisons of the models.

Table 5.4 shows each model's word and character error rates on each dataset. Table 5.5 shows the word and character error rates for the dysarthric speech and control data within each dysarthric speech dataset. The test set seems more challenging than the validation set, but the same patterns still hold across both datasets.

The models trained on L2 speech achieve lower CER and WER among all intelligibility groups. WER and CER scores for individual UA-Speech intelligibility group are found in Table 5.6. Scores for TORGO intelligibility groups are found in Table 5.7. The greatest performance improvements for models trained with L2 speech are found among speakers

Table 5.4: WER and CER by dataset for speaker-dependent models.

| | | Test | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Base | FT: Dys | FT: Dys+L2 | MT | Base | FT: Dys | FT: Dys+L2 | MT |
| WER | UA-Speech | 90.2 | 22.5 | 21.8 | **19.2** | 91.6 | 21.7 | 19.8 | **17.4** |
| | TORGO | 50.8 | 27.8 | 24.6 | **20.7** | 51.2 | 28.5 | 23.9 | **20.5** |
| | L2Arctic | 24.5 | - | 11.5 | **9.2** | 24.7 | - | 11.8 | **9.3** |
| CER | UA-Speech | 48.7 | 14.5 | 13.5 | **11.9** | 49.8 | 13.7 | 12.1 | **10.6** |
| | TORGO | 26.7 | 13.8 | 12.0 | **10.0** | 27.3 | 13.7 | 11.7 | **10.0** |
| | L2Arctic | 10.1 | - | 4.2 | **3.1** | 10.3 | - | 4.4 | **3.2** |

Table 5.5: WER for dysarthric and non-dysarthric speech for speaker-dependent models.

| | | | Test | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Base | FT: Dys | FT: Dys+L2 | MT | Base | FT: Dys | FT: Dys+L2 | MT |
| WER | All | DYS | 104.1 | 43.8 | 40.0 | **36.6** | 104.6 | 43.6 | 37.9 | **34.2** |
| | | CTL | 43.6 | 10.7 | 10.2 | **7.1** | 44.5 | 10.6 | 9.3 | **6.9** |
| | UAS | DYS | 121.3 | 37.6 | 36.4 | **32.5** | 124.5 | 37.2 | 33.5 | **30.0** |
| | | CTL | 54.4 | 5.1 | 5.0 | **3.8** | 54.1 | 4.1 | 4.3 | **3.0** |
| | TOR | DYS | 78.4 | 53.0 | 45.3 | **42.6** | 75.7 | 52.9 | 44.3 | **40.3** |
| | | CTL | 36.3 | 14.6 | 13.8 | **9.3** | 37.8 | 15.2 | 12.8 | **9.7** |
| CER | All | DYS | 61.9 | 27.0 | 24.2 | **21.7** | 62.6 | 26.3 | 22.4 | **20.1** |
| | | CTL | 19.0 | 4.2 | 3.9 | **2.7** | 20.0 | 3.9 | 3.8 | **2.6** |
| | UAS | DYS | 71.7 | 25.6 | 23.6 | **21.3** | 73.7 | 24.7 | 21.4 | **19.2** |
| | | CTL | 22.0 | 1.6 | 1.8 | **1.1** | 23.4 | 1.7 | 1.8 | **1.0** |
| | TOR | DYS | 46.1 | 29.2 | 25.1 | **22.3** | 45.6 | 28.7 | 23.8 | **21.5** |
| | | CTL | 16.9 | 6.0 | 5.4 | **3.9** | 17.4 | 5.6 | 5.2 | **3.8** |

with lower intelligibility ratings. For instance, including the FT: Dys+L2 model improved WER by 12.5% in the test set and 10.9% in the validation set for the "2.333" group in the TORGO dataset. The multitask model performs better than the finetuning models in lower,

particularly in the lowest intelligibility groups, for instance, improving WER by 7.6% in the test set and 12.7% in the validation set for the "low" intelligibility group in the UA-Speech dataset.

Table 5.6: WER for UA-Speech intelligibility scores for speaker-dependent models. The intelligibility scores are determined by grouping intelligibility ratings into quartiles. Intelligibility ratings were collected from 5 listeners with backgrounds in language disorders.

| | | **Test** | | | | **Validation** | | | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | Intl. | Base | FT: Dys | FT: Dys+L2 | MT | Base | FT: Dys | FT: Dys+L2 | MT |
| WER | High | 74.7 | 7.7 | **6.7** | 7.0 | 76.6 | 8.6 | **5.7** | **5.7** |
| | Mid | 122.1 | 30.7 | 26.8 | **22.1** | 124.1 | 25.0 | 25.9 | **21.1** |
| | Low | 146.1 | 39.2 | 40.1 | **33.6** | 160.0 | 40.9 | 36.1 | **30.4** |
| | Very low | 161.6 | 80.1 | 79.1 | **72.5** | 159.1 | 80.5 | 73.2 | **67.8** |
| CER | High | 34.6 | 3.5 | **3.0** | 3.3 | 39.6 | 3.5 | **2.4** | 2.9 |
| | Mid | 70.9 | 15.1 | 12.1 | **11.0** | 68.8 | 13.4 | 11.5 | **8.0** |
| | Low | 87.7 | 23.6 | 22.0 | **17.7** | 94.6 | 21.8 | 18.8 | **13.3** |
| | Very low | 108.5 | 64.0 | 60.6 | **55.2** | 105.4 | 63.0 | 55.8 | **53.8** |

Interestingly, the FT: Dys+L2 model has noticeably lower WER scores on the UA-Speech component test set than the validation set. Yet, it has larger improvements in WER for the TORGO component of the test set. The fluctuations indicate that the model is sensitive to the composition of the dataset. Additionally, the FT: Dys+L2 model's CER scores are lower than the FT: Dys model's CER scores across all intelligibility groups in the UA-Speech dataset, indicating that lower performance in the test set could be due to the severity of the WER metric on single-word utterances, rather than a negation of the pattern observed in the validation set.

A sensitivity analysis was performed on a different dataset split. Models trained and evaluated on this split achieved a lower performance boost from L2 speech, with lower WER of only around 3-4%. However, the models still achieve a performance boost, confirming

Table 5.7: WER for TORGO intelligibility scores for speaker-dependent models. The intelligibility scores reflect the average score of the intelligibility subsection of the Frenchay assessment (a lower score indicates lower intelligibility).

| | | **Test** | | | | **Validation** | | | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | Intl. | Base | FT: Dys | FT: Dys+L2 | MT | Base | FT: Dys | FT: Dys+L2 | MT |
| WER | 8.0 | 37.7 | 24.0 | 19.0 | **18.7** | 39.7 | 28.0 | 20.8 | **19.5** |
| | 5.333 | 114.8 | 73.9 | 67.8 | **58.3** | 105.6 | 71.3 | 59.3 | **54.6** |
| | 2.333 | 108.9 | 73.7 | 61.2 | **57.6** | 110.5 | 77.7 | 66.8 | **60.5** |
| | 1.666 | 109.8 | 84.8 | 76.1 | **75.0** | 106.2 | 75.0 | 73.4 | **62.5** |
| CER | 8.0 | 20.6 | 11.7 | 9.2 | **8.4** | 22.4 | 11.7 | 10.4 | **9.0** |
| | 5.333 | 65.7 | 41.7 | 39.2 | **33.2** | 67.1 | 45.1 | 36.5 | **35.0** |
| | 2.333 | 65.8 | 40.5 | 34.6 | **31.0** | 68.5 | 44.9 | 35.9 | **32.5** |
| | 1.666 | 74.3 | 55.8 | 46.8 | **43.0** | 70.7 | 48.3 | 42.4 | **35.9** |

the observations made above. A more detailed description of the sensitivity analysis can be found in Appendix B.

### *5.3 Speaker-independent*

Table 5.8: WER and CER by dataset for speaker-independent models.

| | | **Test** | | | | **Validation** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Base | FT: Dys | FT: Dys+L2 | MT | Base | FT: Dys | FT: Dys+L2 | MT |
| WER | UA-Speech | 88.0 | 30.2 | 29.4 | **28.2** | 89.1 | 31.0 | 29.1 | **27.6** |
| | TORGO | 47.3 | 26.9 | 26.8 | **22.4** | 53.5 | 31.0 | 30.0 | **26.5** |
| | L2Arctic | 24.9 | - | 12.4 | **11.7** | 24.8 | - | 12.5 | **11.7** |
| CER | UA-Speech | 45.3 | 19.7 | 19.8 | **18.8** | 47.4 | 20.4 | 19.8 | **19.3** |
| | TORGO | 24.3 | 14.2 | 14.3 | **12.0** | 28.2 | 17.3 | 16.5 | **14.8** |
| | L2Arctic | 10.4 | - | 4.7 | **4.2** | 10.3 | - | 4.7 | **4.2** |

The results of the speaker-independent experiment show similar patterns to the speaker-dependent experiments: the addition of L2 speech improves to improve dysarthric speech recognition. The MT model improved dysarthric speech WER scores by around 3.3% for the test set and 4.9% for the validation set. However, this experiment had some important differences compared to the speaker-dependent experiment. The most apparent difference is that the FT: Dys+L2 model did not substantially improve WER in the test split compared to the FT: Dys model. The FT: Dys+L2 model has lower CER scores than the FT: Dys model on the test set (33.4 and 33.2, respectively). The performance of each model on the dysarthric speech datasets is shown in Table 5.9.

Table 5.9: WER for dysarthric and non-dysarthric speech for speaker-independent models.

| | | | **Test** | | | | **Validation** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Base | FT: Dys | FT: Dys+L2 | MT | Base | FT: Dys | FT: Dys+L2 | MT |
| WER | All | DYS | 101.7 | 49.5 | 48.8 | **46.2** | 103.0 | 52.1 | 48.6 | **47.2** |
| | | CTL | 45.5 | 11.1 | 10.8 | **8.3** | 48.9 | 12.6 | 12.7 | **9.5** |
| | UAS | DYS | 122.7 | 53.7 | 53.3 | **51.5** | 119.9 | 55.3 | 52.1 | **50.2** |
| | | CTL | 53.4 | 6.7 | 5.6 | **4.9** | 58.3 | 6.9 | 6.2 | **5.0** |
| | TOR | DYS | 64.2 | 42.0 | 40.8 | **36.6** | 75.2 | 46.8 | 42.8 | **42.4** |
| | | CTL | 35.6 | 16.5 | 17.1 | **12.6** | 37.6 | 19.5 | 20.6 | **14.8** |
| CER | All | DYS | 58.2 | 33.2 | 33.4 | **31.4** | 58.9 | 34.5 | 32.9 | **32.3** |
| | | CTL | 19.3 | 4.2 | 4.3 | **3.1** | 22.9 | 5.8 | 6.0 | **4.6** |
| | UAS | DYS | 68.7 | 37.2 | 37.7 | **36.0** | 67.9 | 37.7 | 36.5 | **36.1** |
| | | CTL | 22.0 | 2.2 | 2.0 | **1.6** | 26.9 | 3.2 | 3.2 | **2.5** |
| | TOR | DYS | 37.3 | 25.4 | 24.9 | **22.2** | 42.8 | 28.7 | 26.4 | **25.4** |
| | | CTL | 15.7 | 6.8 | 7.4 | **5.2** | 17.8 | 9.1 | 9.5 | **7.3** |

Examining the scores by intelligibility groups, the FT: Dys+L2 and MT models did not improve WER scores or CER scores for lower intelligibility groups in the UA-Speech dataset (Table 5.10). It seems as though most of the overall improvements on the UA-Speech dataset can be attributed to improvements in the "Mid" and "High" intelligibility

Table 5.10: WER for UA-Speech intelligibility scores for speaker-independent models. The intelligibility scores are determined by grouping intelligibility ratings into quartiles. Intelligibility ratings were collected from 5 listeners with backgrounds in language disorders.

| | | **Test** | | | | **Validation** | | | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | Intl. | Base | FT: Dys | FT: Dys+L2 | MT | Base | FT: Dys | FT: Dys+L2 | MT |
| WER | High | 59.8 | 9.3 | 6.5 | **5.4** | 64.0 | 9.8 | 8.8 | **6.5** |
| | Mid | 154.3 | 58.7 | 58.7 | **52.5** | 149.9 | 62.9 | 53.5 | **50.1** |
| | Low | 136.3 | **52.8** | 54.1 | 54.1 | 125.1 | 54.5 | 51.2 | **50.4** |
| | Very low | 140.3 | **94.0** | **94.0** | **94.0** | 140.8 | 94.0 | 95.1 | **93.8** |
| CER | High | 26.0 | 3.4 | 2.5 | **1.9** | 27.3 | 3.8 | 3.0 | **2.9** |
| | Mid | 78.9 | 32.8 | 32.1 | **28.6** | 80.1 | 35.2 | **29.6** | 30.6 |
| | Low | 77.2 | 34.7 | 35.7 | **34.2** | 74.3 | 32.8 | 32.6 | **30.9** |
| | Very low | 92.1 | **78.0** | 80.6 | 79.6 | 90.6 | **79.5** | 81.2 | 80.3 |

groups. Additionally, the FT: Dys model outperformed the FT: Dys+L2 and MT models in CER scores for the "very low" intelligibility group.

The MT model improved WER and CER on the two held-out TORGO intelligibility groups for validation and test sets. The FT: Dys+L2 had a lower WER than the standard FT: Dys model on the "1.666" intelligibility rating in the test set but a higher score in the validation set. The results can be found in Table 5.11.

MAPSSWE comparisons on dysarthric speech data found the FT: Dys+L2 model to be significantly different from the FT: Dys model when evaluated on the validation set ($p < 0.001$). However, the FT: Dys+L2 model was not significantly different from the FT: Dys model when evaluated on the test set ($p = 0.373$). The comparisons found the MT model to differ significantly from the FT: Dys model when evaluated on both the validation set ($p < 0.001$) and the test set ($p < 0.001$). Table C.3 contains pairwise comparisons of the models. Interestingly, MAPSSWE comparisons between the FT: Dys+L2 model and MT model were not significant when evaluated on the dysarthric data in the validation set ($p = 0.085$) but were significant when evaluated on the dysarthric data in the test set

Table 5.11: WER for TORGO intelligibility scores for speaker-independent models. The intelligibility scores reflect the average score of the intelligibility subsection of the Frenchay assessment (a lower score indicates lower intelligibility).

| | | **Test** | | | | **Validation** | | | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | Intl. | Base | FT: Dys | FT: Dys+L2 | MT | Base | FT: Dys | FT: Dys+L2 | MT |
| WER | 8.0 | 26.1 | 12.6 | 10.0 | **8.1** | 26.3 | 10.4 | **7.4** | 8.6 |
| | 1.666 | 119.3 | 84.7 | 85.2 | **77.8** | 130.3 | 87.8 | 82.6 | **80.5** |
| CER | 8.0 | 10.8 | 4.4 | 3.6 | **3.2** | 11.6 | 4.0 | 3.0 | **2.8** |
| | 1.666 | 77.4 | 57.3 | 57.1 | **51.0** | 79.9 | 58.1 | 54.1 | **52.3** |

($p < 0.001$). The tests indicate that the FT: Dys+L2 model's improvements might be specific to a subset of the data and that the MT model is more robust to different data subsets.

### *5.4 Zero-shot*

Table 5.12: WER of dysarthric speech for zero-shot models.

| | | Test | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Base | FT: Dys | FT: Dys+L2 | MT | Base | FT: Dys | FT: Dys+L2 | MT |
| WER | All | 108.3 | 66.3 | **64.9** | 65.3 | 108.1 | 67.5 | **66.1** | 66.5 |
| | UAS | 126.4 | 67.4 | 67.1 | **66.9** | 124.0 | 68.0 | **67.4** | 67.5 |
| | TOR | 81.6 | 64.7 | **61.6** | 62.9 | 82.5 | 66.8 | **64.0** | 64.8 |
| CER | All | 63.8 | 50.1 | **47.9** | 48.6 | 64.5 | 50.4 | **48.6** | 48.8 |
| | UAS | 73.6 | 53.4 | **52.1** | **52.1** | 73.1 | 52.7 | 51.7 | **51.6** |
| | TOR | 48.0 | 44.8 | **41.0** | 43.1 | 49.6 | 46.4 | **43.2** | 44.0 |

While all models outperform the base Wav2vec2 model in the zero-shot experiments, they all achieve comparable word and character error rates. While the models trained on L2 speech

perform slightly better than those trained only on dysarthric speech, the improvements are small: about 1% - 2%. Table 5.12 shows the word and character error rates for the dysarthric speech and control data within each dysarthric speech dataset. Notably, the FT: Dys+L2 and MT models only perform marginally better than the FT: Dys model when evaluated on the UA-Speech dataset. However, on the TORGO dataset, the FT: Dys+L2 and MT models are found to have more substantial improvements over the FT: Dys model.

WER and CER scores for each UA-Speech intelligibility group are found in Table 5.6, and scores for each TORGO intelligibility group are found in Table 5.7. The zero-shot models do not differ substantially in their performance on the UA-Speech data. However, the FT: Dys+L2 and MT models consistently improve WER for speakers in the TORGO dataset with the highest intelligibility ratings ("8.0"). Improvements for speakers with other intelligibility ratings in the TORGO dataset are less predictable. Additionally, in intelligibility groups "5.333" and "2.333" in the TORGO dataset, the CER scores of the base Wav2vec2 model are quite high, suggesting that some improvements to WER might be attributable to models learning contextual character information and word spellings rather than learning new phonetic or acoustic representations.

Table 5.13: WER for UA-Speech intelligibility scores for zero-shot models. The intelligibility scores are determined by grouping intelligibility ratings into quartiles. Intelligibility ratings were collected from 5 listeners with backgrounds in language disorders.

| | | Test | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | Intl. | Base | FT: Dys | FT: Dys + L2 | MT | Base | FT: Dys | FT: Dys + L2 | MT |
| WER | High | 78.7 | 22.8 | 22.4 | **22.1** | 76.3 | 23.0 | **21.6** | 22.4 |
| | Mid | 131.5 | 71.9 | **69.8** | 71.9 | 129.5 | 75.3 | 73.6 | **73.4** |
| | Low | 155.7 | 94.1 | 94.6 | **93.2** | 152.5 | **92.7** | 94.3 | 93.7 |
| | Very low | 160.9 | 100.1 | 100.9 | **100.0** | 158.8 | 100.7 | 100.5 | **100.4** |
| CER | High | 40.0 | 10.8 | 10.2 | **10.1** | 37.6 | 10.9 | **9.8** | **9.8** |
| | Mid | 73.5 | 49.9 | **46.1** | 48.1 | 73.6 | 50.7 | **47.1** | 48.9 |
| | Low | 86.6 | 74.4 | 73.3 | **71.6** | 88.9 | 73.4 | 73.7 | **72.9** |
| | Very low | 105.6 | 92.8 | 92.8 | **92.5** | 107.2 | 92.9 | 93.1 | **91.8** |

Table 5.14: WER for TORGO intelligibility scores for zero-shot models. The intelligibility scores reflect the average score of the intelligibility subsection of the Frenchay assessment (a lower score indicates lower intelligibility).

| | | Test | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | Intl. | Base | FT: Dys | FT: Dys + L2 | MT | Base | FT: Dys | FT: Dys + L2 | MT |
| WER | 8.0 | 46.1 | 30.5 | **24.8** | 26.2 | 46.9 | 32.9 | **26.2** | 29.2 |
| | 5.333 | 108.4 | 93.5 | **90.9** | 93.7 | 103.7 | 96.0 | **93.5** | 94.6 |
| | 2.333 | 111.8 | **95.3** | 95.7 | **95.3** | 111.9 | 96.7 | 98.0 | **95.8** |
| | 1.666 | 119.2 | 92.5 | **90.4** | 93.9 | 129.3 | **92.1** | 95.9 | 94.8 |
| CER | 8.0 | 25.2 | 15.3 | **13.3** | 13.5 | 25.3 | 16.9 | **13.5** | 14.2 |
| | 5.333 | 63.8 | 71.8 | **63.3** | 70.9 | **64.2** | 72.6 | 67.4 | 70.7 |
| | 2.333 | 67.6 | 73.1 | **67.4** | 70.9 | 70.7 | 72.7 | **68.7** | 70.4 |
| | 1.666 | 76.5 | 65.6 | 64.9 | **63.4** | 80.8 | 69.6 | 71.8 | **67.0** |

The FT: L2+Dys model seems to perform better than the MT model. However, MAPSSWE comparisons on dysarthric speech data found that they do not significantly differ ($p = 0.289$ for the test set and $p = 0.401$ for the validation set), indicating that this improvement might not be significant. MAPSSWE comparisons on dysarthric speech data found both the MT and FT: Dys+L2 models to be significantly different from the FT: Dys model, with a significance level set to $\alpha = 0.05$. The pairwise comparisons can be found in Table C.4. Still, the scores across all models are quite similar.

## *5.5 Chapter summary*

Across all experiments, L2 speech seems to improve dysarthric speech recognition. The magnitude of this improvement seems to diminish in speaker-independent and zero-shot experiments. The low magnitude of the effect in zero-shot experiments means that it should be viewed cautiously and warrants further study, but it remains a promising outcome, particularly for its improvements in WER scores on speakers in the TORGO dataset. Multitask paradigms performed substantially better than the standard finetuning paradigms in speaker-

dependent and speaker-independent experiments. In zero-shot experiments, finetuned and multitask models performed equivalently. In speaker-dependent experiments, lower intelligibility groups across the TORGO and UA-Speech datasets benefit from training on L2 data. In speaker-independent, only the higher intelligibility groups in the TORGO dataset benefit from L2 data. In zero-shot experiments, no clear pattern emerged among the lower intelligibility groups in the test and validation sets.

# Chapter 6

# DISCUSSION

## *6.1 Acoustic properties experiments*

The experiment that simulated dysarthric speech found that slowed speech rate and mono-loudness greatly affect Wav2vec2's ability to recognize speech. The moderate impact of pitch shifting can perhaps be explained by the fact the altered data consists of single-word utterances. Because the Wav2vec2 model is trained to recognize words in multiple prosodic contexts, it may be robust to within-word pitch alterations. Another study would have to examine how the impact of pitch monotonization on multiword sentences before drawing definite conclusions. The effect of mono-loudness indicates that Wav2vec2 relies on changes in intensity over the course of an utterance. Additionally, as a feature that is more specific to dysarthric speech, it is worth exploring in future studies on data augmentation.

The experiment that modified dysarthric speech found that simply increasing the speech rate of some speakers with dysarthria leads to a small decrease in word error rates. A decrease in WERs was consistently observed among speakers with "mid", "low", and "very low" intelligibility ratings. The impact of increased speech rate on speakers in the "high" intelligibility group was inconsistent, increasing WERs for two of five speakers in that group. The fact that the rate multiplier results in inconsistent decreases in WER for some speakers raises interesting questions about tailoring speech rate modifications to speakers.

The two experiments underscore the impact of speech rate on ASR performance, complementing literature on data augmentation for dysarthric and L2 speech recognition, such as Fukuda et al. (2018) and Bhat et al. (2022), who both found that speech rate modifications were quite effective at improving both L2 speech recognition and dysarthric speech recognition.

## *6.2 Improved performance from L2 speech*

### *6.2.1 Speaker-dependent experiments*

The speaker-dependent experiments seem to suggest that L2 speech improves the performance of a model on dysarthric speech recognition. The effects were observed for both standard finetuning and multitask models, suggesting that the simple addition of L2 speech improves dysarthric speech recognition. It is challenging to assess whether the models are improving as a result of the speech or non-speech acoustic features in the L2-Arctic dataset. However, the fact that the models are successfully learning representations from three different domains suggests that they are at least learning to represent acoustic features common to all three domains, making it likely that some speech features are represented by the model.

The results also found that multitask learning paradigms lead to even greater improvements in WER and CER scores. The multitask architecture may enable a model to learn either feature specific to dysarthric speech (e.g., speech rate) or task-specific contextual information (e.g., the words used that appear in the UA-Speech dataset) while simultaneously learning the shared features of the L2-Arctic and dysarthric speech datasets.

### *6.2.2 Speaker-independent experiments*

The speaker-independent results show a similar pattern to the speaker-dependent experiments, but the magnitude of the effect discussed above is weakened. Most notably, the model finetuned on both dysarthric and L2 speech did not consistently outperform the model finetuned on dysarthric speech alone. When evaluated on the UA-Speech dataset, all models had roughly equivalent WER scores for speakers from the "low" and "very low" intelligibility groups. On the other hand, when evaluated on the TORGO dataset, models trained on L2 speakers tended to have lower WER scores for speakers from the two selected groups, "1.666" and "8.0". Multitask models continued to outperform other models for speakers with "High" and "Mid" intelligibility ratings from the UA-Speech datasets and for both speakers from the TORGO dataset. The results suggest that multitask learning can train models that

better generalize to unseen speakers. Overall, the experiment extends additional support to the study's hypothesis by showing that L2 speech improves dysarthric speech detection for unseen speakers.

### *6.2.3 Zero-shot experiments*

The zero-shot experiments indicate it is difficult for Wav2vec2 to learn the shared speech features between L2 and dysarthric speech through exposure to L2 speech alone. Nevertheless, MAPSSWE comparisons indicated models trained on control and L2 speech significantly differed from models trained on just the control data, showing some effect of L2 speech on model predictions. However, because the lower WER and CER of the former are minor improvements, it warrants further investigation before we can come to any conclusions.

The most noticeable and consistent improvements were found among speakers in the TORGO dataset, specifically those with intelligibility ratings of "8.0". We are unsure of why the improvements were mainly limited to this group. We speculate that L2-Arctic improves the TORGO dataset because the latter contains multiword sentences: the WER of the model trained on just dysarthric speech was lower than those trained on L2 speech (the data can be found in Section A.3). It is unclear whether this observation is a result of overfitting the model trained on dysarthric speech to single-word utterances or whether the acoustic features of L2 data improve the other models' performance. Future study should investigate both of these alternatives.

### *6.2.4 Multitask training*

The multitask architecture substantially improved model performance in the speaker-dependent and speaker-independent experiments. The improvements are most noticeable in the performance of the speaker-independent models on the test set. MAPSSWE comparisons found the multitask model to be significantly different from the other two finetuned models. In contrast, the model finetuned on L2 and dysarthric speech was not significantly different from the one trained on dysarthric speech alone.

Branching each task to separate transformer layers for the study's speech domains enabled the model to learn decision boundaries specific to the acoustic and contextual information within those domains. However, the benefit cannot only be attributed to building domain-specific decision boundaries since these models significantly outperform the models finetuned on just dysarthric speech in the speaker-dependent and speaker-independent experiments. The performance improvements are likely a result of the multitask architecture optimally learning acoustic features that are similar between L2-Arctic and dysarthric speech datasets. In other words, the model tends to learn the same set of decision boundaries for features that are represented similarly across domains and separate sets of decision boundaries for features that cannot be easily adapted across domains.

#### *6.2.5 Summary*

The speaker-dependent and speaker-independent experiments provide convincing evidence for the hypothesis that training ASR models on L2 speech improves dysarthric recognition. The effect is diminished when generalizing to unseen speakers, but it still results in performance improvements. The speaker-dependent and speaker-independent experiments are convincing evidence for L2 speech's ability to improve dysarthric speech recognition. In both these experiments, the multitask models consistently outperformed the finetuned models.

We cannot definitively conclude that the zero-shot experiment supports the study's hypothesis. The improvements from L2 speech in the zero-shot experiment are quite small, and the model trained on the dysarthria datasets might have overfitted to single-word utterances; nevertheless, the minor improvements are notable and warrant further study.

### *6.3 Data selection*

#### *6.3.1 Evaluation set*

The dataset used to evaluate the model seems to play a large role in model performance. The performance improvements that result from L2 speech are diminished in both speaker-

dependent and speaker-independent experiments. For instance, in speaker-independent experiments, the TORGO dataset portion of the test set split had substantially lower scores in all models compared to the validation set. In this case, it is unclear why performance changed so drastically: they both contain similar amounts of single-word and multi-word sentences. Similarly, model performance was lower on the test set in speaker-dependent experiments. These observations suggest that it is not sufficient to stratify datasets by speaker and that researchers should also stratify datasets by factors they hypothesize would contribute to performance, such as syllable structure, word frequency, or sentence length. It also underscores the importance of performing sensitivity analyses and documenting how datasets were split.

### *6.4 Wav2vec2 and interpretability*

Wav2vec2 is trained on standard English orthography. English orthography requires a model to learn contextual patterns specific to a language's spelling conventions, hindering our ability to examine the acoustic properties that led the model to arrive at a particular transcription. For instance, we observed a model transcribe "massage" as "massaghe".[1] We cannot know whether the model transcribed the "h" to indicate that it sounded unlike the standard pronunciation or simply because "gh" is a common enough digram in English. The issue is compounded by the fact that Wav2vec2 is thought to encode some semantic information (Pasad et al., 2021), adding an additional confound. Future research might benefit from training models on subphonemic segments, like in (Wong et al., 2015), who trained on neural networks to recognize distinctive features. Performing transfer learning studies using models trained on subphonemic units could provide new insight into a model's classification decisions that can be used to improve dysarthric speech recognition systems.

[1] This error was found in the speaker-dependent multitask model evaluated on the validation set.

### *6.5 Using UA-Speech and TORGO dataset for ASR*

There are some issues with UA-Speech and TORGO datasets. One big issue seems to be the audio quality of the datasets. In this study's experiments, Wav2vec2 has lower performances on the TORGO control data than the L2-Arctic data (Table 5.5), which we believe can be partially attributed to the quality of the recordings. Even after applying the `noisereduce` algorithm, there is noticeable noise in the background. The recordings contain non-speech sounds like keyboard taps, long silences, or researchers providing directions to participants.

Schu et al. (2023) found that models trained on Wav2vec2-extracted representations (among other representations, like mel-spectrograms) were able to correctly classify speakers as having or not having dysarthria when speech segments were removed. In other words, models can learn to correctly classify dysarthria severity simply from the recording environment. Their finding calls into question the quality of these two datasets and perhaps raises important questions about the appropriate ways to prepare and process the datasets before training a model.

### *6.6 The Curb-Cut Effect and data diversity in ASR*

By showing that L2 speech can improve dysarthric speech recognition, this study highlights the importance of training ASR models on diverse speech data. This study demonstrates that training models on data from one demographic can also improve model performance on speech for other demographics. For that reason, it is a good example of how the Curb-Cut Effect can inform the design of ASR systems. The Curb-Cut Effect describes how accessible design frequently benefits a much larger set of populations than the population that initially motivated the design. The name refers to the city of Berkeley's decision to add curbs to its sidewalks to make urban infrastructure accessible to people who use wheelchairs. The curb cuts benefitted a much larger population: parents pushing strollers, cyclists parking bikes, and movers pulling dollies (Blackwell, 2016).

This study found that, in several cases, simply including L2 data while finetuning on dysarthric speech improves dysarthric speech recognition. This finding is an example of the Curb-Cut Effect in ASR: training models on more diverse speech data—in terms of accents and intelligibility—can improve speech recognition for populations not included in the training data. It underscores that everyone benefits from efforts to develop accessible ASR technologies.

### *6.7 Chapter summary*

This chapter arrives answers this study's research questions using the results detailed in Chapter 5. It concludes that L2 speech improves Wav2vec2's ability to recognize dysarthric speech in speaker-dependent and speaker-independent settings. Although the magnitude of this improvement is diminished when generalizing to unseen speakers, the fact that the improvements apply to unseen speakers that the performance boost from training on L2 speech is robust. The results of the zero-shot experiment are inconclusive and warrant further study.

In discussing the results of these experiments, we reason that the improved performance of the multitask models is a result of the model optimally learning decision boundaries. We also discuss the limitations of the study and suggestions for future research. Finally, the chapter argues that this study is an example of the Curb-Cut Effect and underscores the importance of including diverse speech data when training ASR systems.

# Chapter 7

# CONCLUSION

This thesis investigated the use of L2 speech to improve dysarthric speech detection when finetuning large self-supervised models like Wav2vec2. We examined acoustic features that impact Wav2vec2's performance and, like previous research, found that decreased speech rate degrades performance. We also found that data modifications that increase speech rate improve Wav2vec2's ability to recognize dysarthric speech for some speakers with lower intelligibility ratings.

Our transfer learning experiments found that models trained on speakers with dysarthria, native English controls, and L2 speakers performed better than those trained on speakers with dysarthria and native English controls. The magnitude of this improved performance was the largest in speaker-dependent experiments and less pronounced in the speaker-independent experiments. The improvements in the zero-shot experiment are small but statistically significant according to the MAPSSWE test and warrant further study.

In speaker-dependent experiments, performance improvements were observed among speakers from all intelligibility groups. In speaker-independent experiments, improvements in WER were observed in all speakers except for the speakers in the UA-Speech dataset with lower intelligibility ratings. In the zero-shot learning experiments, adding L2 speech did not lead to notable improvements in model performance. Although the improvements in the zero-shot model are minor and require further investigation, they are promising.

Except for the zero-shot experiment, multitask training consistently outperforms standard finetuning. We speculate that multitask training allows the model to build the same set of decision boundaries for features similar features across the datasets while simultaneously learning a separate set of decision boundaries for each domain.

The study's experiments provide convincing evidence that transfer learning methods can use L2 speech to improve Wav2vec2's ability to recognize dysarthric speech. Future research might benefit from investigating using L2 speech in conjunction with other techniques, like domain-adversarial training or data augmentation. Additionally, given the study's comparisons between multilingual and L2 speech, future research may want to directly compare models using multilingual or L2 speech as source domains.

Finally, the study is an instance of the Curb-Cutting Effect in ASR. It provides an important example of how developing ASR technologies accessible to one population of users often benefits other populations of users. The results underscore the importance of developing ASR models trained on diverse speech data.

# BIBLIOGRAPHY

Baese-Berk, Melissa M. and Ann R. Bradlow (Feb. 2021). "Variability in Speaking Rate of Native and Nonnative Speech". In: *Second Language Speech Learning*. Cambridge University Press, pp. 312–334. DOI: 10.1017/9781108886901.013.

Baevski, Alexei et al. (2020). "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 12449–12460.

Bahdanau, Dzmitry et al. (Mar. 2016). "End-to-end attention-based large vocabulary speech recognition". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE, pp. 4945–4949. DOI: 10.1109/ICASSP.2016.7472618.

Baskar, Murali Karthick et al. (Sept. 2022). "Speaker adaptation for Wav2vec2 based dysarthric ASR". en. In: *Interspeech 2022*. ISCA, pp. 3403–3407. DOI: 10.21437/Interspeech.2022-10896.

Bates, Douglas et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Bechtold, Bastian (Feb. 2023). *Soundfile*. DOI: 10.5281/zenodo.7741801.

Bhat, Chitralekha, Ashish Panda, and Helmer Strik (Sept. 2022). "Improved ASR Performance for Dysarthric Speech Using Two-stage Data Augmentation". In: *Interspeech 2022*. ISCA. DOI: 10.21437/interspeech.2022-10335.

Blackwell, Angela Glover (2016). "The Curb-Cut Effect". en. In: *Stanford Social Innovation Review* 15. Publisher: Stanford Social Innovation Review, p. 2833. DOI: 10.48558/YVMS-CC96.

Boersma, Paul and David Weenink (May 2023). *Praat: doing phonetics by computer*.

Chen, Shijie, Yu Zhang, and Qiang Yang (2021). "Multi-task learning in natural language processing: An overview". In: *arXiv preprint arXiv:2109.09138*. DOI: https://doi.org/10.48550/arXiv.2109.09138.

Chodroff, Eleanor, Leah Bradshaw, and Vivian Livesay (2022). "Subsegmental Representation in Child Speech Production: Structured Variability of Stop Consonant Voice Onset Time in American English and Cantonese". In: *Journal of Child Language*, pp. 1–29. DOI: 10.1017/S0305000922000368.

Conneau, Alexis et al. (Dec. 2020). *Unsupervised Cross-lingual Representation Learning for Speech Recognition*. arXiv:2006.13979 [cs, eess].

Darley, Frederic L., Arnold E. Aronson, and Joe R. Brown (1975). *Motor speech disorders*. Philadelphia: Saunders.

Das, Nilaksh et al. (Aug. 2021). "Best of Both Worlds: Robust Accented Speech Recognition with Adversarial Transfer Learning". In: *Interspeech 2021*. ISCA. DOI: 10.21437/interspeech.2021-1888.

Ding, Chaoyue, Shiliang Sun, and Jing Zhao (June 2021). "Multi-Task Transformer with Input Feature Reconstruction for Dysarthric Speech Recognition". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. DOI: 10.1109/icassp39728.2021.9414614.

Enderby, Pamela M. (Pamela Mary) (1983). *Frenchay dysarthria assessment*. eng. ISBN: 0933014821 Place: San Diego, Calif Publication Title: Frenchay dysarthria assessment.

Flege, James Emil (Dec. 1981). "The Phonological Basis of Foreign Accent: A Hypothesis". In: *TESOL Quarterly* 15.4, p. 443. DOI: 10.2307/3586485.

Fukuda, Takashi et al. (Sept. 2018). "Data Augmentation Improves Recognition of Foreign Accented Speech". en. In: *Interspeech 2018*. ISCA, pp. 2409–2413. DOI: 10.21437/Interspeech.2018-1211.

Ganin, Yaroslav et al. (2016). "Domain-adversarial training of neural networks". In: *The journal of machine learning research* 17.1. Publisher: JMLR. org, pp. 2096–2030.

Gauder, Lara et al. (Aug. 2021). "Alzheimer Disease Recognition Using Speech-Based Embeddings From Pre-Trained Models". en. In: *Interspeech 2021*. ISCA, pp. 3795–3799. DOI: 10.21437/Interspeech.2021-753.

Gillick, L. and S.J. Cox (1989). "Some statistical issues in the comparison of speech recognition algorithms". In: *International Conference on Acoustics, Speech, and Signal Processing*. Glasgow, UK: IEEE, pp. 532–535. DOI: 10.1109/ICASSP.1989.266481.

Graves, Alex et al. (2006). "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". en. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. Pittsburgh, Pennsylvania: ACM Press, pp. 369–376. DOI: 10.1145/1143844.1143891.

Gutz, Sarah E. et al. (June 2022). "Validity of Off-the-Shelf Automatic Speech Recognition for Assessing Speech Intelligibility and Speech Severity in Speakers With Amyotrophic Lateral Sclerosis". In: *Journal of Speech, Language, and Hearing Research* 65.6. Publisher: American Speech Language Hearing Association, pp. 2128–2143. DOI: 10.1044/2022_jslhr-21-00589.

Hannun, Awni (Nov. 2017). "Sequence Modeling with CTC". en. In: *Distill* 2.11, e8. DOI: 10.23915/distill.00008.

Hernandez, Abner et al. (Sept. 2022). "Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition". In: *Interspeech 2022*. ISCA. DOI: 10.21437/interspeech.2022-10674.

Hertrich, Ingo, Hermann Ackermann, and Wolfram Ziegler (Mar. 2021). "Dysarthria". In: Wiley, pp. 334–367. DOI: 10.1002/9781119606987.ch16.

Jain, Abhinav, Minali Upreti, and Preethi Jyothi (Sept. 2018). "Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning". In: *Interspeech 2018*. ISCA. DOI: 10.21437/interspeech.2018-1864.

Jespersen, Otto (1904). *Lehrbuch der Phonetik*. Leipzig: Teubner.

Jurafsky, Dan and James H. Martin (2022). *Speech and Language Processing*. 3rd edition (draft). Online.

Kim, Heejin et al. (Sept. 2008). "Dysarthric speech database for universal access research". In: *Interspeech 2008*. ISCA. DOI: 10.21437/interspeech.2008-480.

Kominek, John and Black Alan W. (2004). "The CMU Arctic speech databases". In: *5th ISCA Speech Synthesis Workshop*. Pittsburgh, PA, pp. 223–224.

Korzekwa, Daniel et al. (Sept. 2019). "Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech". In: *Interspeech 2019*. ISCA. DOI: 10.21437/interspeech.2019-1206.

Lansford, Kaitlin L. and Julie M. Liss (Feb. 2014). "Vowel Acoustics in Dysarthria: Speech Disorder Diagnosis and Classification". In: *Journal of Speech, Language, and Hearing Research* 57.1. Publisher: American Speech Language Hearing Association, pp. 57–67. DOI: 10.1044/1092-4388(2013/12-0262).

Li, Jinyu (2022). "Recent Advances in End-to-End Automatic Speech Recognition". In: *APSIPA Transactions on Signal and Information Processing* 11.1. Publisher: Now Publishers. DOI: 10.1561/116.00000050.

Liss, Julie M. et al. (Oct. 2009). "Quantifying Speech Rhythm Abnormalities in the Dysarthrias". In: *Journal of Speech, Language, and Hearing Research* 52.5. Publisher: American Speech Language Hearing Association, pp. 1334–1352. DOI: 10.1044/1092-4388(2009/08-0208).

Loper, Edward and Steven Bird (2009). *NLTK: The Natural Language Toolkit*. arXiv:cs/0205028.

Loshchilov, Ilya and Frank Hutter (2017). "Decoupled Weight Decay Regularization". In: Publisher: arXiv Version Number: 3. DOI: 10.48550/ARXIV.1711.05101.

McFee, Brian et al. (Mar. 2023). *librosa*. DOI: 10.5281/zenodo.7741801.

Mnih, Andriy and Yee Whye Teh (July 2012). "A fast and simple algorithm for training neural probabilistic language models". In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. Ed. by John Langford and Joelle Pineau. ICML '12. event-place: Edinburgh, Scotland, GB. New York, NY, USA: Omnipress, pp. 1751–1758.

Morris, Andrew Cameron, Viktoria Maier, and Phil Green (2004). "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition". In: *Proc. Interspeech 2004*, pp. 2765–2768. DOI: 10.21437/Interspeech.2004-668.

Moulines, Eric and Francis Charpentier (Dec. 1990). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". en. In: *Speech Communication* 9.5-6, pp. 453–467. DOI: 10.1016/0167-6393(90)90021-Z.

Ordin, Mikhail and Leona Polyanskaya (Aug. 2015). "Acquisition of speech rhythm in a second language by learners with rhythmically different native languages". In: *The Journal of the Acoustical Society of America* 138.2. Publisher: Acoustical Society of America (ASA), pp. 533–544. DOI: 10.1121/1.4923359.

Panayotov, Vassil et al. (Apr. 2015). "Librispeech: An ASR corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

Pasad, Ankita, Ju-Chieh Chou, and Karen Livescu (Dec. 2021). "Layer-Wise Analysis of a Self-Supervised Speech Representation Model". In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Cartagena, Colombia: IEEE, pp. 914–921. DOI: 10.1109/ASRU51503.2021.9688093.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rowe, Hannah P. et al. (Apr. 2022). "Characterizing Dysarthria Diversity for Automatic Speech Recognition: A Tutorial From the Clinical Perspective". In: *Frontiers in Computer Science* 4. Publisher: Frontiers Media SA. DOI: 10.3389/fcomp.2022.770210.

Rudzicz, Frank, Aravind Kumar Namasivayam, and Talya Wolff (Mar. 2011). "The TORGO database of acoustic and articulatory speech from speakers with dysarthria". In: *Language Resources and Evaluation* 46.4. Publisher: Springer Science and Business Media LLC, pp. 523–541. DOI: 10.1007/s10579-011-9145-0.

Sainburg, Tim (June 2019). *timsainb/noisereduce: v1.0*. DOI: 10.5281/ZENODO.3243139.

Schneider, Steffen et al. (2019). "wav2vec: Unsupervised Pre-training for Speech Recognition". In: Publisher: arXiv Version Number: 4. DOI: 10.48550/ARXIV.1904.05862.

Schu, Guilherme, Parvaneh Janbakhshi, and Ina Kodrasi (June 2023). "On Using the UA-Speech and Torgo Databases to Validate Automatic Dysarthric Speech Classification Approaches". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Forthcoming. Rhodes Island, Greece: IEEE, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095981.

*SCTK* (Oct. 2021). *SCTK, the NIST Scoring Toolkit*.

Shibano, Toshiko et al. (Nov. 2021). "Speech Technology for Everyone: Automatic Speech Recognition for Non-Native English". In: *Proceedings of the Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*. Trento, Italy: Association for Computational Linguistics, pp. 11–20.

Shor, Joel et al. (Sept. 2019). "Personalizing ASR for Dysarthric and Accented Speech with Limited Data". In: *Interspeech 2019*. ISCA. DOI: 10.21437/interspeech.2019-1427.

Takashima, Yuki, Tetsuya Takiguchi, and Yasuo Ariki (May 2019). "End-to-end Dysarthric Speech Recognition Using Multiple Databases". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, pp. 6395–6399. DOI: 10.1109/ICASSP.2019.8683803.

Trouvain, Jürgen and Bernd Möbius (2014). "Sources of variation of articulation rate in native and non-native speech: comparisons of French and German". In: *Proc. 7th International Conference on Speech Prosody 2014*, pp. 275–279. DOI: 10.21437/SpeechProsody.2014-42.

Vachhani, Bhavik, Chitralekha Bhat, and Sunil Kumar Kopparapu (Sept. 2018). "Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition". In: *Interspeech 2018*. ISCA. DOI: 10.21437/interspeech.2018-1751.

Vásquez-Correa, Juan Camilo et al. (Sept. 2018). "A Multitask Learning Approach to Assess the Dysarthria Severity in Patients with Parkinson's Disease". In: *Interspeech 2018*. ISCA. DOI: 10.21437/interspeech.2018-1988.

Vásquez-Correa, Juan Camilo et al. (Oct. 2021). "Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages". In: *Pattern Recognition Letters* 150. Publisher: Elsevier BV, pp. 272–279. DOI: 10.1016/j.patrec.2021.04.011.

Vaughn, Charlotte, Melissa Baese-Berk, and Kaori Idemaru (Aug. 2019). "Re-Examining Phonetic Variability in Native and Non-Native Speech". en. In: *Phonetica* 76.5, pp. 327–358. DOI: 10.1159/000487269.

Violeta, Lester Phillip, Wen Chin Huang, and Tomoki Toda (Sept. 2022). "Investigating Self-supervised Pretraining Frameworks for Pathological Speech Recognition". In: *Interspeech 2022*. ISCA. DOI: 10.21437/interspeech.2022-10043.

Wang, Disong et al. (Aug. 2021). "Unsupervised Domain Adaptation for Dysarthric Speech Detection via Domain Adversarial Training and Mutual Information Minimization". In: *Interspeech 2021*. ISCA. DOI: 10.21437/interspeech.2021-2139.

Wolf, Thomas et al. (2020). "Transformers: State-of-the-Art Natural Language Processing". en. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

Wong, Ka-Ho et al. (2015). "Analysis of Dysarthric Speech using Distinctive Feature Recognition". In: *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*. Association for Computational Linguistics. DOI: 10.18653/v1/w15-5115.

Woszczyk, Dominika, Stavros Petridis, and David Millard (Oct. 2020). "Domain Adversarial Neural Networks for Dysarthric Speech Recognition". In: *Interspeech 2020*. ISCA. DOI: 10.21437/interspeech.2020-2845.

Xie, Xin and T. Florian Jaeger (May 2020). "Comparing non-native and native speech: Are L2 productions more variable?" In: *The Journal of the Acoustical Society of America* 147.5. Publisher: Acoustical Society of America (ASA), pp. 3322–3347. DOI: 10.1121/10.0001141.

Xu, Qiantong, Alexei Baevski, and Michael Auli (Sept. 2022). "Simple and Effective Zero-shot Cross-lingual Phoneme Recognition". In: *Interspeech 2022*. ISCA. DOI: 10.21437/interspeech.2022-60.

Yang, Xuesong et al. (Apr. 2018). "Joint Modeling of Accents and Acoustics for Multi-Accent Speech Recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. DOI: 10.1109/icassp.2018.8462557.

Zhang, Yixuan et al. (Sept. 2022). "Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems". en. In: *1st Workshop on Speech for Social Good (S4SG)*. ISCA, pp. 15–19. DOI: 10.21437/S4SG.2022-4.

Zhang, Yu and Qiang Yang (Sept. 2017). "An overview of multi-task learning". In: *National Science Review* 5.1. Publisher: Oxford University Press (OUP), pp. 30–43. DOI: 10.1093/nsr/nwx105.

Zhao, Guanlong et al. (Sept. 2018). "L2-ARCTIC: A Non-native English Speech Corpus". In: *Interspeech 2018*. ISCA. DOI: 10.21437/interspeech.2018-1110.

# Appendix A

# PARENTHETICAL DESCRIPTIONS OR EXPLANATIONS

## *A.1 Description of `noisereduce` algorithm*

The `noisereduce` algorithm applies two spectral gating algorithms in succession. The first is a stationary spectral gating algorithm that removes frequencies within a set of intervals of acceptable values (or, a mask). The mask's values are obtained by calculating the statistical properties of the spectrogram derived from the sound file. The mask is smoothed to ensure a clean sound. The second algorithm is a non-stationary spectral gating algorithm that largely follows the steps outlined in the former algorithm, however, it uses local statistics to obtain values for the spectral mask. It moves a window across the spectrogram, obtaining the statistics for a spectral mask from the parts of the spectrogram visible from the window.

## *A.2 Comparing speech data from CMU_ARCTIC and L2Arctic*

The finding in Baese-Berk and Bradlow, 2021 that L2 speech is slower and more variable than native English speakers is confirmed by comparing the duration of audio files from the L2Arctic dataset and those from the CMU_ARCTIC database (Kominek and Alan W., 2004) (see Table A.1). Within-speaker variances were also greater for L2 speech than native-English speech.

Table A.1: Average and standard deviation utterance durations by L1.

| Duration (s) | English | Arabic | Chinese | Hindi | Korean | Spanish | Vietnamese |
|---|---|---|---|---|---|---|---|
| Mean | 3.15 | 3.62 | 3.81 | 3.07 | 3.62 | 3.84 | 3.82 |
| Standard deviation | 0.88 | 1.24 | 1.35 | 1.00 | 1.20 | 1.39 | 1.34 |

It is possible to compare a coarse measurement of speech rate across the L1 and L2 speakers in the CMU_ARCTIC and L2Arctic databases since their speakers are given the same prompts. The coarse metric for speech rate was obtained by dividing the audio duration by the number of syllables in the utterance. The number of syllables in the utterance was estimated using NLTK's syllable tokenizer (Loper and Bird, 2009), which uses the Sonority Sequencing Principle (SSP) algorithm (Jespersen 1904, as cited in Loper and Bird 2009). Figure A.1 shows a boxplot of speech rates.

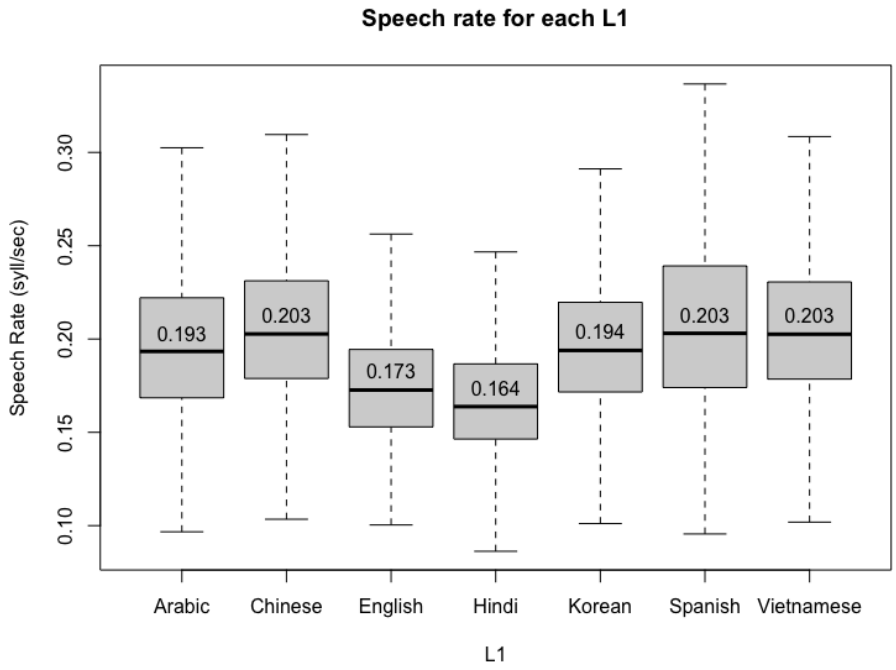



Figure A.1: Boxplot of speech rates for speakers in the CMU_ARCTIC and L2Arctic datasets

Following Baese-Berk and Bradlow (2021), I generated two mixed-effect linear models in R (R Core Team, 2023) using the `lme4` package (Bates et al., 2015). I performed a log transformation on speech rate to ensure that the residuals were evenly distributed. Both models treated the number of words in the prompt as a fixed effect and the speaker ID as

a random effect, with one model including L1/L2 speaker status as a fixed effect and comparing the models through an analysis of variance (ANOVA). While L1/L2 speaker status was only marginally more predictive of speech model accuracy ($\chi^2 = 3.428, p = 0.064$), the result could be influenced by the lower average speech rate and lower variability of Hindi speakers. Excluding Hindi speakers from the dataset, L1/L2 status significantly affects model fit ($\chi^2 = 6.829, p < 0.001$). The results bolster Baese-Berk and Bradlow (2021)'s finding that L2 speakers (largely) speak at a slower rate than native English speakers. However, the fact that Hindi speakers show different speech rate patterns raises interesting questions about whether these observations can generalize to all L1s.

### *A.3 Zero-shot improvement on multi-word utterances*

Evaluating WER on multi-word sentences in the TORGO dataset revealed that the zero-shot models trained on L2 speech had lower WER than models trained only on control speech. WER was reduced by approximately 4-5% on the test and validation splits. It is unclear whether the

Table A.2: WER on multi-word TORGO utterances for zero-shot models.

| | | Base | FT: Dys | FT: Dys+L2 | MT |
|---|---|---|---|---|---|
| WER | Validation | 73.8 | 66.4 | **61.3** | 63.0 |
| | Test | 75.0 | 64.8 | **60.0** | 61.6 |

# Appendix B

# MODEL DETAILS AND PRELIMINARY EXPERIMENTS

## *B.1 Preliminary experiments*

Preliminary experiments were run to examine the impact of different hyperparameters and model architectures on performance. This section details several of these experiments.

### *B.1.1 Separate layer for the control group*

It was unclear whether the multitask model would perform better when the control and dysarthric speech were processed by separate linear layers or the same layer, so I compared the performance of these two configurations. Early studies showed either little difference in WER for dysarthric speech–if anything, they showed better performance when the layer was shared between dysarthric and control groups (including lower WER for lower intelligibility scoring groups), perhaps because the layer acquired had more training samples of the same utterances.

## *B.2 Model hyperparameters*

Model hyperparameters used for the speaker-dependent and zero-shot experiments are shown in Table B.1.

## *B.3 Sensitivity analysis*

While determining hyperparameters, we trained models on a different train/test dataset split. Table B.2 shows the hyperparameters used to train these models. Model performance, determined by WER, is shown in Table B.3. Although the impact of L2 speech is more

Table B.1: Hyperparameters used for the speaker-dependent and zero-shot experiments.

| Hyperparameter | Speaker-dependent | Speaker-independent | Zero-shot |
|---|---|---|---|
| Training epochs | 10 | 10 | 10 |
| Optimizer | AdamW | AdamW | AdamW |
| Weight decay | 1e-6 | 1e-6 | 0.005 |
| Warmup ratio | 0.1 | 0.1 | 0.1 |
| Learning rate | 1e-4 | 1e-4 | 5e-5 |
| Loss weights | See 4.3.4 | | |

Table B.2: Hyperparameters used for sensitivity analysis

| | |
|---|---|
| Training epochs | 12 |
| Optimizer | Adagrad |
| Learning rate | 1e-4 |
| Weight decay | 0.0 |
| Warmup ratio | None |
| Weight reinitialization | None |

minor in the sensitivity analysis, it is reassuring that the pattern observed in the study holds, especially for the improvements observed with speakers with lower intelligibility ratings.

The patterns largely agree with the observation made by this study that L2 speech seems to improve the performance of wav2vec2 when it is finetuned on L2 data in addition to dysarthric speech data. However, the results are less stark than those observed in the study, with improvements in WER limited to no more than 5%. The reduced magnitude of the improvements can be attributed to two changes: 1) the improved performance of the FT: DYS model and the reduced performances of the FT: DYS+L2 and MT models. It is unclear whether the FT: DYS model's improved performance can be attributed to hyperparameter selection or the data quality within the dataset split. The model seems to have converged when examining the loss curve, making the additional epochs an unlikely

Table B.3: WER for sensitivity analysis models by dataset and group.

| | | FT: DYS | FT: DYS+L2 | MT |
|---|---|---|---|---|
| TORGO | CTL | 13.2 | 12.6 | **11.2** |
| TORGO | DYS | 47.9 | 41.2 | **40.3** |
| UA-Speech | CTL | 3.8 | 3.3 | **1.9** |
| UA-Speech | DYS | 37.8 | **34.9** | 35.4 |

explanation. The inclusion of weight reinitialization since an FT: DYS model trained without weight reinitialization on the new dataset split performed worse than one trained with weight reinitialization. The most likely explanation for the DYS+L2 model's improved performance here seems to be an advantage of the dataset split, particularly its improved performance on the TORGO dataset.

Table B.4: WER for UA-Speech intelligibility scores for sensitivity analysis models.

| | FT: DYS | FT: DYS+L2 | MT |
|---|---|---|---|
| High | 8.6 | **6.4** | 8.6 |
| Low | 42.2 | **37.8** | 39.56 |
| Mid | 26.0 | 26.0 | **24.4** |
| Very low | 80.0 | 75.6 | **74.0** |

On the other hand, we believe that the reduced magnitude of improvements is attributable to training without weight reinitialization. However, the reinitialization procedure could yield greater benefits to the DYS+L2 and MT models. Given the MT model's considerable improvements on the TORGO dataset compared to the other two models, we believe this to be the case. Moreover, we believe that the observed improvements coming from weight reinitialization are magnified in the MT model since the MT model discussed here only branched off for the final decoder and shared the last two layers of the transformer.

# Appendix C

# MAPSSWE TEST TABLES

Tables that contain pairwise comparisons of MAPSSWE tests can be found below.

Table C.1: Significance levels ($P(Z < |w|)$) of pairwise MAPSSWE comparisons of each audio modification to the UA-Speech control group data (referenced in Section 5.1). * indicates that it meets a significance level of $p < 0.01$

| | | Loudness | | Pitch | Rate | | | Original |
|---|---|---|---|---|---|---|---|---|
| | | Monotonized | Drop-off | Monotonized | × 0.80 | × 1.20 | × 1.40 | |
| Loud. | Mono. | - | <0.001* | <0.001* | <0.001* | <0.001* | 0.093 | <0.001* |
| | Drop. | - | - | <0.001* | <0.001* | <0.001* | <0.001* | <0.001* |
| Pitch | | - | - | - | 0.219 | <0.001* | <0.001* | <0.001* |
| Rate | × 0.80 | - | - | - | - | <0.001* | <0.001* | <0.001* |
| | × 1.20 | - | - | - | - | - | <0.001* | <0.001* |
| | × 1.40 | - | - | - | - | - | - | <0.001* |
| Original | | - | - | - | - | - | - | - |

Table C.2: Pairwise MAPSSWE comparisons of speaker-dependent models (referenced in Section 5.2). * indicates that it meets a significance level of $\alpha = 0.05$

| | Validation | | |
|---|---|---|---|
| | FT: Dys | FT: Dys+L2 | MT |
| FT: Dys | - | $W$=5.615, $p$<0.001* | $W$=8.606, $p$<0.001* |
| FT: Dys+L2 | - | - | $W$=3.937, $p$<0.001* |
| MT | - | - | - |

| | Test | | |
|---|---|---|---|
| | FT: Dys | FT: Dys+L2 | MT |
| FT: Dys | - | $W$=3.681, $p$<0.001* | $W$=7.235, $p$<0.001* |
| FT: Dys+L2 | - | - | $W$=3.995, $p$<0.001* |
| MT | - | - | - |

Table C.3: Pairwise MAPSSWE comparisons of speaker-independent models (referenced in Section 5.3). * indicates that it meets a significance level of $\alpha = 0.05$

| | Validation | | |
|---|---|---|---|
| | FT: Dys | FT: Dys+L2 | MT |
| FT: Dys | - | $W$=4.25, $p$<0.001* | $W$=5.894, $p$<0.001* |
| FT: Dys+L2 | - | - | $W$=1.729, $p$=0.085 |
| MT | - | - | - |

| | Test | | |
|---|---|---|---|
| | FT: Dys | FT: Dys+L2 | MT |
| FT: Dys | - | $W$=0.902, $p$=0.373 | $W$=3.993, $p$<0.001* |
| FT: Dys+L2 | - | - | $W$=3.346, $p$<0.001* |
| MT | - | - | - |

Table C.4: Pairwise MAPSSWE comparisons of zero-shot models (referenced in Section 5.4). * indicates that it meets a significance level of $\alpha = 0.05$

| | Validation | | |
|---|---|---|---|
| | FT: Dys | FT: Dys+L2 | MT |
| FT: Dys | - | $W$=2.966, $p$=0.002* | $W$=2.560, $p$=0.010* |
| FT: Dys+L2 | - | - | $W$<0.001, $p$=0.399 |
| MT | - | - | - |

| | Test | | |
|---|---|---|---|
| | FT: Dys | FT: Dys+L2 | MT |
| FT: Dys | - | $W$=3.361, $p$<0.001 | $W$=2.606, $p$=0.009* |
| FT: Dys+L2 | - | - | $W$<0.001, $p$=0.289 |
| MT | - | - | - |