university of
groningen
campus fryslân

Master Thesis

The Role of Speech Elicitation Methods and Disease Factors in Dysartrhric ASR System Development

by

Spyretta Leivaditi

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Voice Technology

Supervisor: Asst. Prof. Dr. V. Verkhodanova
Second examiner: Asst. Prof. Dr. S. Nayak

University of Groningen,
Campus Fryslân,
Leeuwarden, Netherlands

July 17, 2023

CONTENTS

Abstract

Despite significant advancements in automatic speech recognition (ASR) technology, the performance of ASR systems on dysarthric speech is still inadequate for widespread use. A reason for this is the lack of sufficiently rich and diverse dysarthric speech datasets to train machine learning models that could handle all types and varieties of such speech. Motivated by the data scarcity problem, this thesis investigates whether developers of Dutch ASR systems can take advantage of particular characteristics of dysarthric speech and increase their models' performance by selecting their training data in a strategic way. More specifically, the thesis hypothesizes a) that fine-tuning an ASR model with differently elicited speech data would lead to improved performance for the respective elicitation method, and b) that fine-tuning an ASR model with speech data affected from a specific disease would enhance model's performance on speech affected by that disease. Both hypotheses are experimentally tested by fine-tuning and evaluating a state-of-the-art self-supervised dysarthric ASR system on a new Dutch dysarthric speech dataset. The results of the experiments do not provide adequate evidence that either the elicitation method or the underlying disease of the dysarthric speakers plays a significant role in the performance of a dysarthric ASR system.

1

INTRODUCTION

Dysarthria refers to a group of divergent motor speech disorders that make a speaker lose their ability to articulate words normally[1]. It's a condition that can be caused by Parkinson's disease (PD), Multiple Sclerosis (MS), and other conditions, and affects the control of speech-related organs. Speech produced by individuals with dysarthria is characterized by poor articulation, monotonous intonation, breathy voice, and other phenomena that make its automatic recognition a challenging task (Duffy, 2013).

Indeed, despite significant advancements in automatic speech recognition (ASR) technology, the performance of ASR systems on disordered and impaired speech is still not adequate for widespread use (Gupta et al., 2016, Moore et al., 2018). This happens partially because there are no sufficiently rich and diverse dysarthric speech datasets to train machine learning models that could handle all types and varieties of such speech. As speakers with dysarthria exhibit highly variable speech patterns, both within and across individuals, it is difficult to characterize these patterns and ensure that they are sufficiently and proportionally represented in ASR corpora and datasets (Rowe et al., 2022).

Addressing the dysarthric data scarcity problem is important because dysarthric ASR is a crucial component of a broader effort to develop inclusive ASR systems. In recent years, there has been a growing interest in addressing the challenges faced by individuals with speech impairments and disorders, as demonstrated by high profile projects such as Google's Euphonia[2] and the Speech Accessibility Project[3]. These projects try to provide these individuals with improved accessibility to speech recognition technologies and, thus, to equal opportunities for participation and engagement in various aspects of life.

In general, there are three main approaches that have been followed by the dysarthric ASR community to address the data scarcity problem: transfer learning, self-supervised learning, and data augmentation. Transfer Learning (Torrey and Shavlik, 2009) is a ma-

[1] https://www.msdmanuals.com/home/brain,-spinal-cord,-and-nerve-disorders/brain-dysfunction/dysarthria

[2] https://sites.research.google/euphonia/about/

[3] https://speechaccessibilityproject.beckman.illinois.edu/

chine learning method where a model that has been pre-trained in one task is reused as the starting point for the training of a model on a new task. The assumption is that the pre-trained model has already learned knowledge that is useful for the new task. The main benefit of this method is that it can be applied to tasks with little labeled data available and produce significantly better models than the ones that would have been produced from training with only the available data. In the case of dysarthric ASR models, transfer learning is applied by pre-training models with a large amount of healthy labeled speech data and then "fine-tuning" these models with a much smaller set of labeled dysarthric speech data.

Self-supervised learning (SSL) (Liu et al., 2022) works similarly to transfer learning but the pre-trained model is derived from unlabeled data instead of labeled one. As such, this approach can be applied in tasks where there is a substantial amount of unlabeled data and little to none labeled data. In the case of dysarthric speech, self-supervised learning is applied by pre-training models with large amounts of healthy speech audio and then fine-tuning these models with dysarthric labeled data. Finally, data augmentation involves creating synthetic data by modifying existing healthy speech in a way that simulates as much as possible realistic dysarthric speech. This is done, for example, by introducing temporal and speed modifications.

In all the above approaches, which I discuss in more detail in chapter 2, a question that naturally arises is the following: when developing a dysarthric ASR system should we strive to use as much dysarthric speech data as we can, no matter their characteristics and provenance, or should we be more careful and strategic in selecting what data to use? In this thesis I aim to answer this question by investigating research questions for two dysarthric speech characteristics: a) the method used for the speech's elicitation (read vs spontaneous speech), and b) the underlying diseases that cause the speakers' dysarthria.

Regarding the speech elicitation method, several studies have found that read speech (where speakers read one or more given texts) and spontaneous speech (where speakers speak in free form given some prompt) can differ significantly in their prosody (Bunton et al., 2000, Blaauw, 1994, Laan, 1997). Spontaneous speech is more sensitive to prosodic abnormalities (Bunton et al., 2000) while read speech has lower articulation rate, more F0 variation, more F0 declination, less shimmer, and less vowel reduction (Laan, 1997). These differences can contribute significantly to the perceptual differences between spontaneous and read speech and make dysarthric symptoms more intense and distinctive in spontaneous speech. As such, the performance of an ASR system trained and evaluated on read speech might not be fully representative of its performance in a more naturalistic task such as a conversation (Leuschel and Docherty, 1996). In this thesis I am looking to establish whether this is indeed the case.

Apart from the elicitation method, dysarthric speech can have different phonemic patterns that depend on the exact subtype of the speaker's dysarthria (Rowe et al., 2022) and which might impact ASR performance. Spastic dysarthria[4], for example, is characterized by muscle stiffness and rigidity, and impairs oral constriction (Platt et al., 1980).

[4] Spastic dysarthria can be caused by Amyotrophic Lateral Sclerosis, Cerebral Palsy, Multiple Sclerosis,, Multiple Systems Atrophy, and Progressive Supranuclear Palsy.

Ataxic dysarthria[5], on the other hand, is characterized by muscle weakness and incoordination, and results into voicing contrast errors (Blaney and Hewlett, 2007). And hypokinetic dysarthria[6] is characterized by reduced range and speed of movement, leading to spirantization (Canter, 1965), articulatory undershoot, and vowel centralization (Y. Kim et al., 2009).

More than different phonemic patterns, speech intelligibility can vary greatly also due to different articulatory subsystem impairments (Lee et al., 2014, Rong et al., 2015). In a 2020 study (Rowe et al., 2020) the authors identified 24 different articulatory impairment features, grouped them into five dimensions of articulatory motor control (Coordination, Consistency, Speed, Precision, and Rate) and measured their manifestation in patients suffering from various diseases. Their results indicated a considerable variety of articulatory impairments across different diseases. For example speakers with ataxia exhibited greater impairments related to the Rate dimension than speakers with PD.

The above findings indicate that the performance of an ASR system trained and evaluated on speech from speakers with a particular disease might not be fully representative of the same ASR system performance on speech affected by different diseases. Vice versa, an ASR system that is trained on speech affected by different diseases but used on speech affected by a particular disease might not be as optimal as it could be if it had been trained on speech affected by that particular disease. That's the second thing I am looking to assess in this thesis.

1.1. Research Questions and Hypotheses

In this thesis I followed a self-supervised approach for ASR model development and I investigated two research questions and hypotheses:

Research Question 1: Does fine-tuning an ASR model with differently elicited speech data (read vs spontaneous) improve the ASR performance for the respective elicitation method?

Hypothesis 1: An ASR model that is fine-tuned with differently elicited speech data (read vs spontaneous) will have better performance for the respective elicitation method.

Research Question 2: Does fine-tuning an ASR model with speech data from different diseases improve the ASR performance for the respective disease?

Hypothesis 2: Fine-tuning an ASR model with speech data from different diseases will result into better performance for the respective disease.

In order to answer these two questions I built a new dysarthric speech dataset based on a recent Dutch dysarthric speech corpus (Verkhodanova, 2021) that contains both read and spontaneous speech from speakers with various diseases. Along with this dataset,

[5] Ataxic dysarthria can be caused by Ataxia, Multiple Sclerosis (MS), and Multiple Systems Atrophy.

[6] Hypokinetic dysarthria can be cause my Multiple Systems Atrophy, Parkinson's disease and Progressive Supranuclear Palsy.

I considered a recent Dutch dysarthric SSL ASR model as a baseline (Matsushima, 2022), and I performed three experiments:

- **Experiment 1:**
  - I fine-tuned the baseline model on **read dysarthric speech data from speakers with PD**.
  - I evaluated the fine-tuned model on **read and spontaneous speech affected by different diseases**.
  - **Expectations**:
    - If the fine-tuned model has significantly better performance on read speech than on spontaneous speech then that's evidence that hypothesis 1 holds.
    - If the fine-tuned model has significantly better performance on PD speech than on speech affected by other diseases then that's evidence that hypothesis 2 holds, at least for PD.
- **Experiment 2:**
  - I fine-tuned the baseline model on **spontaneous dysarthric speech data from speakers with PD**.
  - I evaluated the fine-tuned model on **read and spontaneous speech affected by different diseases**.
  - **Expectations**:
    - If the fine-tuned model has significantly better performance on spontaneous speech than on read speech, then that's evidence that hypothesis 1 holds.
    - If the fine-tuned model has significantly better performance in PD speech than on speech affected by other diseases then that's evidence that hypothesis 2 holds, at least for PD.
- **Experiment 3:**
  - I fine-tuned the baseline model on **spontaneous dysarthric speech data from speakers with MS**. The reason for selecting MS is that the dysarthria that accompanies it has different acoustic manifestations than hypokinetic dysarthria that PD typically causes.
  - I evaluated the fine-tuned model on **speech data from speakers with MS and other diseases**.
  - **Expectations**:
    - If the fine-tuned model has significantly better performance on MS speech than on speech affected by PD and other diseases then that's further evidence that hypothesis 2 holds.

1.2. Research Contribution

The main outcomes of this thesis are the following:

- Empirical evidence as to whether and how the elicitation method of dysarthric speech data that are used for training SSL ASR models affects the effectiveness of such models. While previous research has suggested that the elicitation method does affect ASR effectiveness, this hasn't been done by actually training and comparing different ASR models.
- Empirical evidence as to whether and how the underlying disease that causes the dysarthria affects the effectiveness of SSL ASR models. Again, previous research has suggested that the underlying disease does affect ASR effectiveness, but this suggestion hasn't been verified in actual ASR models.
- A new labeled speech dataset of both read and spontaneous speech, and of several different diseases that can be used for further research.

These outcomes contribute towards a better understanding of dysarthria diversity from an ASR perspective and help ASR developers optimize their data collection strategy based on the particular characteristics of the dysarthric speech they need to process.

1.3. Thesis Outline

The thesis is structured as follows. Chapter 2 briefly surveys the most notable resources and approaches for dysarthric speech recognition that are related to this work. Chapter 3, in turn, describes in detail the methodology that I followed in order to answer the research questions and verify or reject my hypotheses. Chapter 4 presents the results of the three experiments, while chapter 5 discusses these results with respect to the research questions and hypotheses of the thesis and outlines future research directions. Finally, chapter 6 summarizes the key points and findings of the thesis.

2

Related Work

This chapter provides a survey of the most notable resources and approaches that have been developed in the past years by the scientific community in relation to dysarthric speech recognition. This survey helps the reader a) understand the availability, diversity, and characteristics of dysarthric speech data, b) gain insights into the various techniques and methodologies used by researchers in collecting and analyzing such data, and c) establish the context for the research conducted in this thesis. In particular, section 2.1 describes the methodology I followed for this survey, while section 2.2 describes the most relevant dysarthric speech datasets and corpora. Section 2.3, in turn, outlines a number of systems and models that perform dysarthric speech recognition.

2.1. Survey Methodology

For the resources part of the survey, I have considered datasets and corpora that contain dysarthric speech, in any language, and for any task (i.e., not only for ASR). For each resource I identify, where available, information about its size, the languages it covers, the characteristics of the speakers, the types and underlying causes of the dysarthria, and the tasks or applications the dataset has been designed for. For the approaches part of the survey, I have considered relatively recent models and systems (from 2016 onwards) that primarily use machine learning techniques. For each approach I identify the languages it supports, the types of dysarthria it covers, the dysarthric and non-dysarthric data it uses for training and evaluation, the machine learning techniques it implements (e.g., self-supervised learning or transfer learning), and the results it achieves.

To find these models and datasets I searched in Google Scholar for papers with keywords like "dysarthric ASR", "dysarthric speech models", "dysarthric speech recognition" and "dysarthric speech data and corpora". I also narrowed down these searches by further specifying the language of the data or the models (e.g., Dutch, English, etc), the diseases that affect the data (e.g., Parkinson’s disease, Multiple Sclerosis, etc), or the machine learning approaches and architectures of the models (e.g., Self-Supervised Learning, end-2-end, DNN-HMM, etc).

2.2. DATASETS FOR DYSARTHRIC SPEECH

2.2.1. DUTCH CORPUS OF PATHOLOGICAL AND NORMAL SPEECH (COPAS)

The COPAS corpus[1] (Martens et al., 2011) has been developed as a means for training speech language pathology students on the various perceptual features of pathological speech, as well as for developing or enhancing speech technology tools for assessing and treating pathological speech. It is available via a BSD 2-Clause License[2] from the University of Southern California and consists of pathological and non-pathological speech in the Dutch language, recorded from 319 speakers.

The speakers belong to 8 distinct pathological categories: normal (122 speakers), dysarthria (75 speakers), hearing impairment (29 speakers), laryngectomy (30 speakers), cleft (38 speakers), articulation disorders (17 speakers), voice disorders (7 speakers) and glossectomy (1 speaker). The speakers with dysarthria are not further differentiated based on the disease that causes their dysarthria. Moreover, all speakers have performed a number of different tasks, including passage reading (at a difficulty level 7 or 8 in the Aging Voice Index), picture naming for articulation assessment, sound repetition for measuring diadochokinetic rate, formant transition, and spontaneous and semi-spontaneous storytelling. All audio samples are stored in wav format with a sampling rate of 16kHz and 16 bit linear PCM encoding.

2.2.2. EASYCALL CORPUS

The EasyCall corpus[3] (Turrisi et al., 2021) is a publicly available speech corpus that has been developed as part of a voice-controlled smartphone application that improves the ability of patients with dysarthria to communicate with their family and caregivers. It contains 21,386 audio recordings in the Italian language, coming from 24 healthy speakers (10 females, 14 males, 10,077 recordings) and 31 speakers with dysarthria (11 females, 20 males, 11,309 recordings).

All speakers were adults and their dysarthria was related to Parkinson's disease, Huntingon's Disease, Amyotrophic Lateral Sclerosis, Peripheral Neuropathy, and myopathic or myasthenic lesions. Moreover, the degree of speakers' speech impairment was assessed by neurologists through the Therapy Outcome Measure (TOM)[4], a measurement scale that ranges from 1 to 5 and corresponds to mild, mild-moderate, moderate, moderate-sever, and severe dysarthria. Speakers with aphasic syndromes, dementia or intellectual disability were excluded.

The recordings contain 37 spoken commands per speaker, related to the task of making a phone call, like typing and calling phone numbers or saving new contacts. They also contain 30 non-commands per speaker, namely words that are near or inside commands, or sentences that are phonetically close to commands.

[1] https://people.ict.usc.edu/ gordon/copa.html

[2] https://opensource.org/license/bsd-2-clause/

[3] http://neurolab.unife.it/easycallcorpus/

[4] https://natspec.org.uk/therapy/tools/therapy-outcome-measures-toms/

2.2.3. UA-Speech Database

The UA-Speech database[5] (H. Kim et al., 2008) is a large corpus of dysarthric speech in American English, available for free for university and government lab researchers. Its goal is to be used for automatic speech recognition development for people with neuromotor disability, as well as for research on articulation errors in dysarthria that can benefit clinical treatments of such people.

The corpus consists of 541 read speech recordings, from 19 adult individuals (14 males, 5 females) with Cerebral Palsy and a dysarthria diagnosis (spastic and other forms) confirmed by a certified speech-language pathologist. Its overall duration is around 102 hours. The recorded speakers were asked to repeat digits, radio alphabet letters, computer commands, and common words and uncommon words from corpora. The recordings were assessed for speech intelligibility, resulting in an overall index of dysarthria severity for each speaker. The aim of this assessment was to categorize speakers in terms of intelligibility and explore any possible relation of articulation error types and ASR architectures to intelligibility levels.

2.2.4. Domotica-3

The Domotica-3 speech database[6] (Ons et al., 2014) is a publicly available collection of recordings of Flemish Dutch dysarthric speech that contain commands related to home automation. Examples include utterances like *"turn on the kitchen light"*, *"close the blind living room door"*, *"increase heating"*, etc. The recordings have been derived from 17 participants, 15 adults aged between 14 and 61 years old, and two children. These participants suffered from multiple sclerosis, spastic quadriparesis and other similar diseases.

The total number of utterances in the collection is a bit more than 3000. The dataset contains also speech intelligibility scores for all adult speakers, obtained by applying an automated procedure to the recorded speech (for more details see Middag, 2012). The scores ranged from 64.2 to 89.4, with those greater than 85 being considered normal while those equal to or lower than 70 being considered severely pathologic.

2.2.5. TORGO Database

TORGO[7] (Rudzicz et al., 2010) is an English speech database, initially developed as a resource for developing advanced ASR models for dysarhtric speech. The database's creators focused particularly on collecting detailed physiological information that can help ASR models learn hidden articulatory parameters. The database is free for academic and non-profit purposes and consists of aligned acoustics and measured 2D and 3D articulatory features, derived from 14 speakers, 7 without any disorder (control group) and 7 with different levels of dysarthria (4 male, 3 female).

The dysarthric speakers had either cerebral palsy (spastic, athetoid, or ataxic) or amyotrophic lateral sclerosis, and were between the ages of 16 and 50 years old. Moreover, they covered a wide range of intelligibility. The speakers were asked to read single words or sentences and to describe the content of some photos. A total of 5980 and 2762

[5] http://www.isle.illinois.edu/sst/data/UASpeech/
[6] https://www.esat.kuleuven.be/psi/spraak/downloads/
[7] http://www.cs.toronto.edu/ complingweb/data/TORGO/torgo.html

utterances were produced from dysarthric and non-dysarthric speakers, producing approximately three hours of speech.

2.2.6. EST Dutch Dysarthric Speech Database

The EST Dutch dysarthric speech database (Yilmaz et al., 2016) has been developed as a resource for conducting research on dysarthric speech and for building speech-to-text systems which can be incorporated in various assistive applications for neurological patients in the Netherlands. It consists of recordings of words and sentences uttered by 16 speakers with mild to moderate dysarthria due to Parkinson's disease, traumatic brain injuries and cerebrovascular accident. In particular, the speakers suffered from non-progressive, (asymmetrical) hypokinetic, ataxic, spastic and flaccid dysarthria.

The recordings were collected in both face-to-face speech therapy sessions as well through an interactive web application. The speakers were asked to read aloud written material that includes Dutch numbers, phonetically rich sentences and frequent utterances from the Dutch Polyphone database (Damhuis et al., 1994). Moreover, the recordings have a total duration of 376 minutes with individual durations varying between 2 minutes to around 60 minutes among different speakers. They have been sampled at a frequency of 16 kHz, annotated with the orthographic transcriptions, and accompanied by detailed speaker information such as age, gender, speech intelligibility level and origin of dysarthria. Unfortunately, the database is not publicly available.

2.2.7. SSNCE Database

The SSNCE Database of Tamil Dysarthric Speech[8] (Vijayalakshmi et al., 2022) was developed by the Speech Lab at the SSN College of Engineering in India and it may only be used for non-commercial projects related to linguistic education, research and technology development. It contains around 8 hours of Tamil speech data, collected from 30 speakers (20 with dysarthria and 10 without). The non-dysarthric speakers were equally male and female while the dysarthric speaker group consisted of 7 female and 13 male persons, aged between 12 and 37 years old, with cerebral palsy. In total, each speaker recorded 365 utterances consisting of single words and of sentences that included a combination of common and uncommon Tamil phrases. The audio data is stored as 16-bit 16kHz FLAC compressed linear pcm wav files. The corpus includes also time-aligned phonetic transcripts for all collected speech data, while additional documentation includes phoneme mappings and speaker metadata.

2.2.8. TYPALOC

The TYPALOC corpus (Meunier et al., 2016) has been developed with the objective to compare phonetic variation in the speech of dysarthric and healthy speakers. The recordings contain French speech collected with the help of 28 dysarthric patients and 12 healthy speakers. The patients suffered either from Parkinson's disease (8 patients, 48-81 years old), Amyotrophic Lateral Sclerosis (12 patients, 32-77 years old) or Cereberall Ataxia (8 patients, 32-77 years old). In all cases it was ensured that none of the patients had a severe case of dysarthria that might make their produced speech fully unintelligible.

[8] https://catalog.ldc.upenn.edu/LDC2021S04

Both dysarthric and healthy speakers produced two styles of speech: read and spontaneous. For the read style a children's story of 172 words was used. For the spontaneous style, dysarthric and healthy senior speakers were asked to talk about their everyday life, their personal history or their work. The healthy junior speakers, on the other hand, talked about particular events or situations in an interactive conversation with a single interviewer. Moreover, the spontaneous speech recordings were much longer for the healthy speakers than for the non-healthy ones. Unfortunately, the database is not publicly available.

2.2.9. Nemours database

The Nemours database (Menendez-Pidal et al., 1996) was designed to test the effect of different enhancing signal processing methods in the intelligibility of dysarthric speech. It is a collection of 814 short nonsense sentences, spoken by 11 male speakers with different degrees of dysarthria. Additionally, it contains two connected-speech paragraphs produced by each of the 11 speakers. The database has been labeled at both the word and phoneme levels. Word-level labels were assigned manually, however, the phoneme-level labels were assigned using a Discrete Hidden Markov Model (DHMM) labeler followed by manual inspection and correction. The recordings were digitized using a 16 kHz sampling rate at 16-bit sample resolution with appropriate low pass filtering. Unfortunately, the database is not publicly available.

2.2.10. homeService Corpus

The homeService corpus (Nicolao et al., 2016) is a database of realistic English dysarrthric speech, available from the University of Sheffield under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The corpus has 10 hours of dysarthric speech elicited from 5 speakers with severe dysarthria (3 male and 2 female). The speakers were recorded interacting with voice-controlled devices in their real home environments.

2.2.11. Cantonese Dysarthric Speech Corpus

The Cantonese Dysarthric Speech Corpus (Wong et al., 2015) has been developed as a resource for investigating articulatory and prosodic characteristics of Cantonese dysarthric speech with particular focus on speaking rate and pitch and loudness control. It contains around 10 hours of speech produced by both healthy and dysarthric speakers. 7.5 hours of dysarthric speech were produced by 6 male and 5 female speakers while 2.5 hours of non-dysarthric speech came from 3 male and 2 female speakers. All speakers were native speakers of Cantonese from Hong Kong, while the dysarthric speakers were diagnosed with cerebellar degeneration. The stimuli for the speech generation included a range of speaking styles like single word, short sentence, paragraph and conversation, as well as articulatory tasks. The resulting audio was sampled at 44.1 kHz and quantized at 16 bits. Unfortunately, the corpus is not publicly available.

2.2.12. Corpus of Children's Disordered Speech

The Corpus of Children's Disordered Speech (Saz et al., 2008) has been built to help the development of environment control systems based on oral interfaces for physically disabled children, as well as for providing computer-aided speech and language therapy. It contains Spanish speech recorded from 14 young speakers (ages from 11 to 21 years old) with various developmental impairments (including Down's syndrome) and/or neuromuscular disorders like cerebral palsy or ataxia. These speakers have uttered several sessions over a 57-word Spanish vocabulary, producing 3192 thousand isolated-word utterances, 459 short-sentence utterances and 30 long-sentence utterances. The overall duration of the produced speech was more than 3 hours. In addition, a parallel corpus of speech from 168 unimpaired young speakers has been recorded with more than 6 hours of speech with the same vocabulary. Unfortunately, the corpus is not publicly available.

2.2.13. PC-GITA

The PC-GITA database (Orozco et al., 2014) contains speech recordings of 50 people with PD and 50 healthy people that had no symptoms associated to PD or any other neurological disease. Both groups consist of 25 men and 25 women, all Colombian Spanish. The recordings were collected following a protocol that included tasks related to phonation, articulation and prosody, such as sustained phonations of vowels, diadochokinetic evaluation, reading of words, sentences, and dialogues, and production of a spontaneous monologue. Unfortunately, the database is not publicly available.

2.3. APPROACHES AND MODELS FOR DYSARTHRIC SPEECH RECOGNITION

2.3.1. TRANSFER LEARNING APPROACHES

The existing body of research on ASR for dysarthric speech includes several works that follow a transfer learning approach. A relatively recent model is described in Xiong et al., 2020. This is an acoustic model that is first trained on healthy data from the UA-Speech database and then it is fine-tuned on the dysarthric speech of the same corpus. It uses a hybrid neural network architecture, known as CNN-TDNN-F, that combines Convolutional Neural Networks (CNN) with Time-Delay Neural Networks (TDNN) and Feed-Forward Neural Networks (FNN). The CNN component extracts local features from input acoustic frames, the TDNN component models temporal dependencies, and the FNN component performs classification or regression tasks. A novelty of this model is that is applies a data selection strategy for each dysartric speaker based on the intuition that data from speakers with similar dysarthria severity could mutually benefit from each other in transfer learning. The evaluation of the model on the UA-Speech database indicated a relative recognition improvement of 11.6% in comparison to the conventional speaker-dependent training.

Another interesting transfer learning model is that of Green et al., 2021. This is an end-to-end ASR model based on the Recurrent Neural Network Transducer (RNN-T) architecture, pretrained on around 162,000 hours of typical speech (from Google's internal production dataset), and fine-tuned (overall and per speaker) with the recordings of 432 dysarthric speakers. The overall model Word Error Rate (WER) across all 432 speakers was 29.4%, while the median personalized WER was 4.6%.

A transfer learning approach is also used in Shor et al., 2019 where the authors develop two personalized models for Amyotrophic lateral sclerosis (ALS) speech, one Bidirectional RNN Transducer and one based on the Listen, Attend and Spell paradigm. The pre-training of the models is performed on the Librispeech[9] dataset, while the fine-tuning is performed with 36.7 hours of audio from 67 ALS patients. The fine-tuning, according to the experiments, manages to bring down the WER score to 10% for mild dysarthria and to 20% for more severe cases.

For the Dutch language, two interesting dysarthric ASR models are described in Yılmaz et al., 2016, and Yılmaz et al., 2017. Both models are based on a Deep Neural Network-Hidden Markov Model (DNN-HMM) architecture and use for pre-training the CGN corpus (Oostdijk, 2000), which contains representative collections of contemporary standard Dutch and Flemish. The first model uses the COPAS corpus for fine-tuning, achieving a WER reduction between 13.0% and 14.7% in sentence reading tasks and 56.1% and 61.0% in word reading tasks. The second model is fine-tuned with the EST Dutch dysarthric speech database, reducing WER bu a range of 11.0% to 13.6%.

Finally, in Wang et al., 2021, an ASR model for Dutch dysarthric speech is developed as part of a Spoken Language Understanding (SLU) system, consisting of a Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) for aligning audio features with context dependent phonemes, and a TDNN acoustic model. As with other models described in this section, the initial acoustic model is built with the CGN corpus and the fine-tuned model is built with the dysarthric COPAS data. The latter manages to increase accuracy by 5%.

2.3.2. Self-Supervised Learning Approaches

The usefulness of self-supervised speech representations for training dysarthric ASR systems is explored in several works in the literature. For example, the authors in Violeta et al., 2022 developed two dysarthric ASR models using two different self-supervised learning framewors, namely wav2vec 2.0 (Baevski et al., 2020) and WavLM (Chen et al., 2022). The wav2vec model was pretrained on 60k hours of normal speech while the WavLM model on 94k hours. Both models were fine-tuned for two types of pathological speech: dysarthric (using the UA-Speech database) and electrolaryngeal [10] (using an in-house recorded dataset of Japanese electrolaryngeal speech). Then they were compared with fully supervised models, trained on the Librispeech dataset (Panayotov et al., 2015). The comparison showed that the best supervised setup outperformed the best self-supervised setup by 13.9% character error in electrolaryngeal speech and 16.8% word error rate in dysarthric speech.

Another work that applies self-supervised learning in developing dysarthric ASR systems is that of Hernandez et al., 2022. The authors trained acoustic models for dysarthric speech by first extracting speech features with a) the base and large wav2vec 2.0 models, b) the multilingual XLSR[11] model (Conneau et al., 2021), and c) the Hubert model (Hsu et al., 2021). Then they trained a model for English speakers with cerebral palsy using the UA-Speech database, a model for Spanish speakers with Parkinson's disease using

[9] https://www.openslr.org/12

[10] Electrolaryngeal speech refers to a method of generating speech sounds using an artificial device called an electrolarynx. This technique is commonly used by individuals who have undergone laryngectomy, a surgical procedure in which the larynx (voice box) is removed, typically due to cancer or other medical conditions.

[11] XLSR stands for Cross-Language Speech Representation.

the PC-GITA corpus, and a model for Italian speakers with paralysis using the EasyCall corpus. In all three cases, improvements were achieved, with the use of XLSR features resulting in lower WER scores than wav2vec or Hubert.

Pretrained wav2vec XLSR models were also used in Krishna et al., 2021 to develop dysarthric ASR models for three Indian low-resource languages, namely Telugu, Tamil, and Gujarati. These models achieved an average relative reduction in WER of 2.88% compared to the previous state-of-the-art supervised method. The authors also analyzed the generalization capability of multilingual pre-trained models on languages already contained in their training data, as well as on languages that were not contained. In both cases, they found that fine-tuning with only 25% of the training data gives competitive WER to the state-of-the-art supervised methods.

2.3.3. Data Augmentation Approaches

As mentioned in chapter 1, a third approach for dealing with the data scarcity problem is the generation of synthetic speech data. An application of this philosophy can be seen in Geng et al., 2020a where the authors synthesize dysarthric speech by modifying the tempo and speed (i.e., the audio duration and the spectral envelope) of healthy speech, and then use this speech to re-train an DNN-HMM ASR system that was previously trained only on healthy data. Tempo-based augmentation achieved an absolute WER improvement of 4.24% , while speed-based augmentation managed to decrease WER by 2% in WER.

A more recent approach (Geng et al., 2020b) also used tempo perturbation and speed perturbation, along with vocal tract length perturbation, to augment data from the UA-Speech database. Speed perturbation gave the highest absolute improvement in WER (2.92%).

Finally, in Shahamiri, 2021, the authors address the scarcity of dysarthric data problem both in a visual and an acoustic way. More specifically, the system they have developed, called Speech Vision, extracts and uses word-level voicegrams for given speech signals and visualizes these voicegrams as RGB images that highligh the words' shapes. Then it utilizes visual-data augmentation (Perez and Wang, 2017) to create modified versions of the voicegrams by shifting their width, sheering and zooming through them. In addition to the visual augmentation, Speech Vision also uses the Deep Convolutional Text-To-Speech (DC-TTS) system (Tachibana et al., 2018) to produce synthetic dysarthric speech. The overall data augmentation improved the system's word recognition accuracy for almost all dysarthric speakers of the evaluation dataset, with the minimum improvement being 0.53% and the maximum one 6.13%.

3

Methodology

This chapter provides a detailed description of the methodology that I followed in order to test my hypotheses and answer the research questions I defined in chapter 1. In particular, section 3.1 provides a high-level overview of the experiments I conducted and section 3.2 describes the speech recognition model I used as a baseline. Section 3.3, in turn, describes the data I used, while section 3.4 provides some technical details of the way I trained the different models in my experiments.

3.1. Overview of experiments

As mentioned in chapter 1, in this thesis I performed three different experiments. In the first experiment I took an existing dysarthric speech recognition model from Matsushima, 2022 as a baseline and I fine-tuned it with read speech data coming from speakers with PD. I then evaluated both the new and the baseline model with read and spontaneous speech not only from speakers with PD but also from speakers with other diseases that cause dysarthria. I also evaluated the two models on speech data from healthy speakers. In the second experiment, I fine-tuned the same baseline model with spontaneous PD speech data instead of read speech and I evaluated it on the same data as in the first experiment. Finally, in the third experiment, I fine-tuned the baseline model on spontaneous dysarthric speech data from speakers with MS and evaluated it on the same data as in the first experiment. The fine-tuned models, the evaluation script, and instructions on how to perform the fine-tuning and the evaluation are available in a Gitlab repository[1].

3.2. Baseline Dysarthric Speech Recognition Model

The detailed description of the dysarthric speech recognition model I utilized as a baseline and fine-tuned throughout my experiments can be found in Matsushima, 2022. It is a model specifically designed to explore the efficacy of self-supervised learning in recognizing Dutch dysarthric speech, in comparison to a supervised learning method. The

[1] https://gitlab.com/spyretta.leiv/ssl_parkinson_dysarthria_asr

main reason I chose it as a baseline was that, at the moment of starting the thesis, it was the only currently available model that applied a self-supervised training strategy for Dutch dysarthric speech recognition. The foundation of the model is wav2vec 2.0 XLSR-53, a large crosslingual speech representation model that has been pre-trained with an extensive dataset of 56 thousand hours of speech across 53 languages (Conneau et al., 2021) using the wav2vec 2.0 framework for self-supervised learning of speech representations (Baevski et al., 2020). To adapt it for dysarthric Dutch speech recognition, the XLSR-53 model was fine-tuned using the COPAS dataset, which I previously described in Section 2.2.1.

3.3. TRAINING AND EVALUATION DATA

3.3.1. DATA DESCRIPTION

The primary speech data I used in all three experiments, for both training and evaluation purposes, were derived from a Dutch dysarthric speech corpus that is described in Verkhodanova, 2021 and which was constructed with the purpose of analyzing the speech characteristics of people with PD. The corpus contains recordings derived from 126 individuals, each of whom performed the following speech production tasks:

- **MMSE**: In this task participants were asked to answer questions from the Mini–Mental State Examination (MMSE)[2], a 30-point questionnaire used extensively in clinical and research settings to measure cognitive impairment.
- **Prolonged phonation**: In this task participants were asked to hold twice the sound /a/ as long as possible.
- **Prosody elicitation tasks**: These tasks targeted production of lexical stress, boundary marking, and sentence type and focus intonations.
- **Interview**: In this task participants were asked and answered interview-style questions about their first job, the place where they grew up, their hobbies and their family.
- **Video description**: In this task participants were asked to describe a short scene from Charlie Chaplin's silent film "The Idle Class" [3].
- **Picture description**: In this task participants were asked to describe two pictures. The first picture was the Cookie Theft Picture (CPT), originally used on aphasic patients but then also applied in clinical research with various disease groups. The second picture was one of the Heaton pictures (Heaton, 1972). Both pictures can be seen in appendix D.
- **Reading**: In this task participants were asked to read a Dutch translation of the Aesop's fable "The North Wind and the Sun" (see appendix E).

[2] https://www.ihacpa.gov.au/health-care/classification/subacute-and-non-acute-care/standardised-mini-mental-state-examination

[3] https://youtu.be/F5l4DGInCBE

- **DDK**: In this task, known as phonoarticulatory diadochokinesis test, participants were asked to repeat the sounds /pa/, /ta/, /ka/, /pata/, /taka/, /pataka/, two or five times, at both speaking and accelerated pace.

The tasks were performed in Dutch, with the recording sessions taking place in quiet rooms. Initially, all tasks were recorded in a single audio file per participant but then these files were split into smaller files, each containing an individual task. Also, certain speakers' data was removed from the corpus due to quality issues, including a substantial amount of the interviewer's speech.

The read speech (RS) data that I used in my experiments included the *reading* task recordings from 101 participants, while the spontaneous speech data (SpoS) included the same participants' recordings of the *video description* and *picture description* tasks. The RS recordings contained no interviewer speech, while the SpoS recordings contained, in average, between 2 and 7 words of interviewer's speech. From the 101 participants, 40 were completely healthy (HC), 43 had Parkinson's disease (PD), 4 had Multiple Sclerosis (MS), 4 had suffered a stroke, 6 had Spinocerebellar Ataxia (SCA), and 4 had some other disease that caused their dysarthria. Participants age's ranged from 31 years old to 87, while the information about the severity of their dysarthria was not available. Table 3.1 shows the number and the duration of the recordings per participant group and speech type.

	Read Speech		**Spontaneous Speech**	
Group	**No of Recordings**	**Duration (hours)**	**No of Recordings**	**Duration (hours)**
PD	43	0.67	130	2.07
MS	4	0.11	12	0.15
Stroke	4	0.08	12	0.14
SCA	6	0.14	18	0.19
Other	4	0.09	12	0.21
HC	40	0.52	120	2.48
Total	**101**	**1.61**	**304**	**5.24**

Table 3.1: Number and duration of recordings per participant group and speech type

A second dataset that I used in all three experiments, but only for evaluation purposes, is the Domotica database that I described in 2.2.4. This dataset had been also used in the evaluation of the baseline dysarthric speech recognition model.

3.3.2. Data Preparation

Read Speech

Before I could use the read speech recordings I had to pre-process them in two ways. First, I performed a normalization of their intensity, as some recordings had their intensity too low (and sounded as if the speaker was far from the microphone) while other recordings had it too high. For that purpose, I used a Praat script[4]. Second, because the text that had been read by the participants was not aligned with their speech, I performed a sentence level forced alignment between each recording and the read text using the aeneas Python library [5]. In that way, each RS recording was split into five smaller recordings, one per text sentence. In addition, the alignment library detected the exact position of each sentence in each recording, automatically removing any trailing silences in their beginning and end.

SPONTANEOUS SPEECH

Before I could use the spontaneous speech recordings I had to acquire transcriptions for them. To do that I worked as follows. First, as with read speech, I normalized the intensity of the recordings. Second, because in almost all the recordings there was a small but not negligible amount of speech uttered by the interviewer, I had to remove that speech before I started the transcription. To do that I labeled the different segments of the recordings based on who was speaking (Interviewer or Participant), and I generated a separate recording for each segment.

To perform this labeling I used the Prodigy annotation tool [6] that allowed me to load the waveform of each recording, listen to it, and annotate its different parts according to the speaker type (see figure 3.1). The annotations were stored in a jsonl file that for each recording contained its name and the exact time spans that either the participant or the interviewer spoke. Thus, when the all recordings were annotated, I generated a new recording file for each different time span using the librosa tool [7].

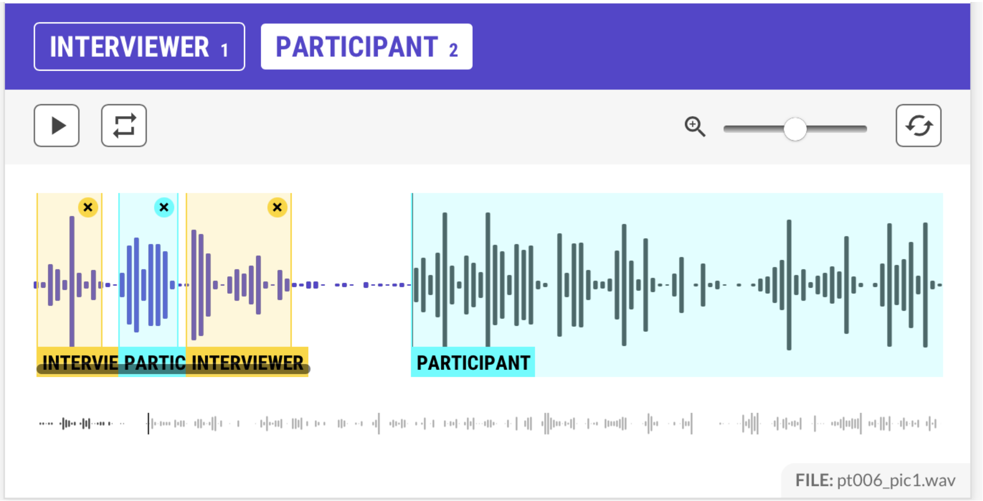


Figure 3.1: Speaker Type Annotation with Prodigy

[4] https://www.acsu.buffalo.edu/ cdicanio/scripts/Rescale_peak.praat
[5] https://www.readbeyond.it/aeneas/
[6] https://prodi.gy/
[7] https://librosa.org/doc/latest/index.html

The result of this segmentation was a new set of recordings, the number and duration of which per participant group is shown in table 3.2. In general, the duration of a single recording ranged from 30 seconds to 3 minutes.

Participant Group	**No of Recordings**	**Duration (mins)**
PD	172	111.37
MS	16	8.66
Stroke	15	8.44
SCA	20	11.04
Other	16	12.29
HC	134	145.08
Interviewer	88	3.47
Total	**461**	**173.4**

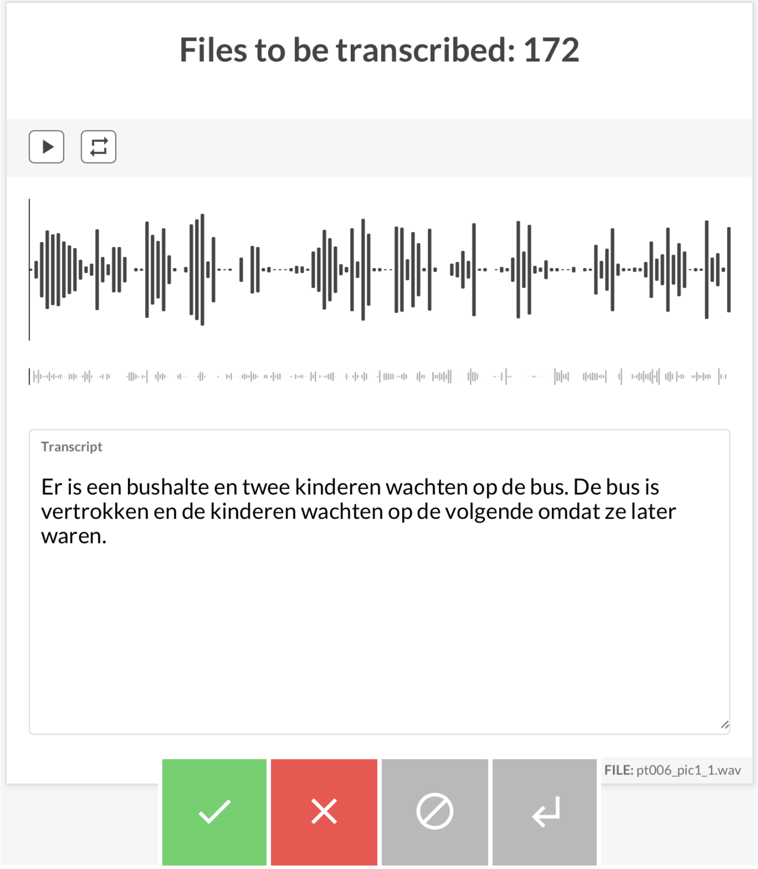
Table 3.2: Number and duration of spontaneous speech recordings per group, after removing interviewer speech.

Having these new set of recordings, I moved on to have them transcribed. For that, I hired two native Dutch transcribers, the first to transcribe the dysarthric speech recordings and the second to transcribe the healthy speech recordings. When both transcribers were done with the transcriptions, the one transcriber proofread the transcriptions of the other. To perform the transcription, the transcribers used Prodigy, listening to the waveform of each recording and writing in a textbox the text they could hear (see figure 3.2).

Beyond words, the transcribers also used certain tags to denote hesitation (tag [hes]), noise (tag [noise]), or words they could not understand (tag [unknown]). On the other hand, because of limited time, the transcribers did not mark the time span of each uttered word or sentence in the recordings; instead they just wrote the text they could hear. To fix this issue, I performed forced alignment of the final transcriptions at both word and sentence level, just as I did with read speech. Before performing the alignment, I cleaned the transcriptions by removing any trailing spaces, line changes or strange characters, as well as all tags except for [unknown]. The latter was necessary so that the trained models could deal with the inaudible parts of the recordings.

It is important to note that the two transcribers had controlled and limited access to the recordings. They could only listen to the recordings via the Prodigy interface which could access only via a password protected link to a server that ran on my local machine. This was made possible through the Ngrok[8] tunneling software that allowed me to expose my local development server to the internet in a secure way. This made sure that the recordings would not leak to third parties.

[8] https://ngrok.com/



Files to be transcribed: 172

Transcript

Er is een bushalte en twee kinderen wachten op de bus. De bus is vertrokken en de kinderen wachten op de volgende omdat ze later waren.

FILE: pt006_pic1_1.wav

Figure 3.2: An example of the transcription done with Prodigy

3.4. Training and Evaluation Set Up

For the fine-tuning process in all three experiments I utilized Fairseq [9] (Ott et al., 2019), an open-source sequence-to-sequence toolkit created by Facebook AI Research. Fairseq offers a collection of modular and adaptable components designed for training and assessing a variety of neural network models. These models can be applied to tasks such as machine translation, language modeling, text generation, and speech recognition.

For each experiment, the respective data with which the baseline model would be fine-tuned (RS-PD for experiment 1, SpoS-PD for experiment 2 and SpoS-MS for experiment 3) were split into two sets, one for the training and validation phase (80% of the data), and one for the evaluation phase (20% of the data). The other types of data that were not used for fine-tuning were used for evaluation. Also, all recordings were resampled to 16kHz in order to be compatible with the baseline model.

The technical configuration of all three experiments were the same as the one used

[9] https://github.com/facebookresearch/fairseq

in the baseline model:

- Fine-tuning with the CTC loss (Graves et al., 2006) and up to 16K updates.
- Model optimization with the Adam optimizer (Kingma and Ba, 2015) with the following tri-state learning rate schedule: 10% of updates for warming up, next 40% for annealing, and last 50% for linear decay.
- Peak learning rate set to 0.00001.
- Evaluation with the CTC beam search decoder, with beam width 50, implemented with the pyctcdecoder [10] library.

[10] https://github.com/kensho-technologies/pyctcdecode

4

Results

In this chapter I present the results of the three experiments I performed in order to answer the two research questions I defined in chapter 1. In particular, in section 4.1 I present the results related to the role of the speech elicitation method in the performance of dysarthric ASR models, while in section 4.2 I do the same for the second research question about the role of the underlying disease in dysarthric ASR performance.

4.1. The Role of the Speech Elicitation Method

The first research question of this thesis is whether fine-tuning an ASR model with differently elicited speech data (read vs spontaneous) improves the ASR performance for the respective elicitation method. To answer this question I took the dysarthric ASR model of section 3.2 and I fine-tuned it once with the PD read speech from the dataset described in section 3.3 (RS-PD) and once with the PD spontaneous speech from the same dataset (SpoS-PD). Then I evaluated the two fine-tuned models and the baseline model on read and spontaneous speech from speakers with PD, speakers with other diseases, and healthy speakers. I also evaluated the models on the Domotica dataset that consisted of read speech. The evaluation was done using Word Error Rate (WER)[1] as an ASR performance metric.

The motivation behind these first two experiments was to see a) if the RS-PD model had much better performance on read speech than on spontaneous speech and b) if the SpoS-PD model had much better performance on spontaneous speech than on read speech. If that was the case then my hypothesis that fine-tuning an ASR model with differently elicited speech data results in better performance for the respective elicitation method would be supported. Table 4.1 contains the experiment results of each model per evaluation dataset.

[1] https://huggingface.co/spaces/evaluate-metric/wer

Evaluation dataset	**Fine-tuned RS-PD**	**Fine-tuned SpoS-PD**	**Baseline**
RS-PD	**0.06**	0.31	0.62
RS-SCA	**0.05**	0.29	0.60
RS-Stroke	**0.07**	0.35	0.63
RS-MS	**0.13**	0.49	0.65
RS-Other	**0.06**	0.33	0.62
RS-HC	**0.04**	0.17	0.56
SpoS-PD	0.84	**0.39**	0.73
SpoS-SCA	0.75	**0.32**	0.69
SpoS-Stroke	0.74	**0.33**	0.70
SpoS-MS	0.83	**0.45**	0.73
SpoS-Other	0.81	**0.41**	0.71
SpoS-HC	0.69	**0.22**	0.64
Domotica - Medium	0.79	0.44	**0.36**
Domotica - Moderate	0.79	0.49	**0.42**
Domotica - Severe	0.86	0.72	**0.50**

Table 4.1: Evaluation results of the RS-PD, SpoS-PD and baseline models

4.2. THE ROLE OF THE UNDERLYING DISEASE

The second research question of this thesis is whether fine-tuning an ASR model with speech data from different diseases improves the performance for the respective disease. This question is partially answered by the first two experiments by looking if the two PD fine-tuned models had better performance on PD speech than on speech affected by other diseases. If that was the case then my hypothesis that fine-tuning an ASR model with speech data from different diseases results into better performance for the respective disease would be supported.

To find further evidence about this hypothesis I did a third experiment where I fine-tuned the baseline model on the MS read speech from the dataset described in section 3.3 (SpoS-MS). I evaluated this model on the same datasets as the other experiments, looking to see if this model had better performance in MS speech than speech affected by PD and other diseases. If that was the case then that would be further evidence that hypothesis 2 holds. The reason for selecting MS was that the dysarthria that accompanies it has different acoustic manifestations than hypokinetic dysarthria that PD typically causes. Table 4.2 contains the WER scores of the SpoS-MS model per evaluation dataset, as well as those of the SpoS-PD model for comparison purposes.

Evaluation dataset	**Fine-tuned SpoS-PD**	**Fine-tuned SpoS-MS**
RS-PD	**0.31**	0.46
RS-SCA	**0.29**	0.46
RS-Stroke	**0.35**	0.46
RS-MS	**0.49**	0.53
RS-Other	**0.33**	0.46
RS-HC	**0.17**	0.30
SpoS-PD	**0.39**	0.53
SpoS-SCA	**0.32**	0.47
SpoS-Stroke	**0.33**	0.48
SpoS-MS	**0.45**	0.49
SpoS-Other	**0.41**	0.51
SpoS-HC	**0.22**	0.37
Domotica - Medium	**0.44**	0.47
Domotica - Moderate	**0.49**	0.54
Domotica - Severe	**0.72**	0.73

Table 4.2: Evaluation results of the SpoS-MS and SpoS-PD models

5

Discussion

In this chapter I analyze and discuss the extent to which the experimental results I presented in the previous chapter provide conclusive answers to the research questions of the thesis. In particular, in section 5.1, I discuss the results with respect to the first research question regarding the role of the speech elicitation method in the performance of dysarthric ASR models. In section 5.2, I do the same for the second research question about the role of the underlying disease in dysarthric ASR performance. Section 5.3, in turn, describes additional observations and insights derived from evaluating the models on healthy speech and from comparing their average performance on all the evaluation datasets. Finally, in section 5.4 I discuss the thesis's findings with respect to the literature and chapter s 2 related work, while in section 5.5 I describe limitations of this thesis and potential directions for future research.

5.1. The Role of the Speech Elicitation Method

The first hypothesis of the thesis is that an ASR model that is fine-tuned with differently elicited speech data (read vs spontaneous) will have better performance for the respective elicitation method. As we can see in table 5.1, the RS-PD model achieves a quite low WER score in each of the RS evaluation datasets (min 0.05 in RS-SCA speech and max 0.13 in RS-MS speech) and it is better in these datasets than the SpoS-PD and baseline models. Vice versa, as table 5.2 shows, the SpoS-PD model performs better than both the RS-PD model and the baseline model when evaluated on spontaneous speech.

At first glance, these results seem to verify the first hypothesis. Nevertheless, when comparing the performance of the SpoS-PD model on read and spontaneous speech (table 5.3) we see that the average WER on read speech is 0.35 and on spontaneous speech 0.38. There is a possibility that this better performance is caused by the fact that the spontaneous evaluation datasets were bigger and more variant than the read speech data. Still, this result is against hypothesis 1 that suggests that the SpoS-PD model should have better performance on spontaneous speech.

Moreover, as table 5.4 shows, the RS-PD model performed much worse than the SpoS model and the baseline model on the Domotica dataset. As described in section 2.2.4,

Evaluation dataset	Fine-tuned RS-PD	Fine-tuned SpoS-PD	Baseline
RS-PD	**0.06**	0.31	0.62
RS-SCA	**0.05**	0.29	0.60
RS-Stroke	**0.07**	0.35	0.63
RS-MS	**0.13**	0.49	0.65
RS-Other	**0.06**	0.33	0.62

Table 5.1: Comparison of the read speech results of the RS-PD, SpoS-PD and baseline models.

Evaluation dataset	Fine-tuned RS-PD	Fine-tuned SpoS-PD	Baseline
SpoS-PD	0.84	**0.39**	0.73
SpoS-SCA	0.75	**0.32**	0.69
SpoS-Stroke	0.74	**0.33**	0.70
SpoS-MS	0.83	**0.45**	0.73
SpoS-Other	0.81	**0.41**	0.71

Table 5.2: Comparison of the spontaneous speech results of the RS-PD, SpoS-PD and baseline models.

Evaluation dataset	Fine-tuned SpoS-PD
RS-PD	0.31
RS-SCA	0.29
RS-Stroke	0.35
RS-MS	0.49
RS-Other	0.33
SpoS-PD	0.39
SpoS-SCA	0.32
SpoS-Stroke	0.33
SpoS-MS	0.45
SpoS-Other	0.41

Table 5.3: Evaluation results of the Spos-PD model on Spontaneous Speech Data

the Domotica dataset consists of read speech, so if hypothesis 1 held then the RS-PD model should have performed better. Finally, the fact that the RS dataset consists only of 6 sentences (see appendix E) increases the probability that the exceptional performance of the RS-PD model on the RS data is the result of model overfitting. All these observations indicate that the results of the first two experiments do not provide adequate evidence that the elicitation method of dysarthric speech data play a significant role in improving ASR performance for the respective method.

Evaluation dataset	**Fine-tuned RS-PD**	**Fine-tuned SpoS-PD**	**Baseline**
Domotica - Medium	0.79	0.44	**0.36**
Domotica - Moderate	0.79	0.49	**0.42**
Domotica - Severe	0.86	0.72	**0.50**

Table 5.4: Comparison of the Domotica speech results of the RS-PD, SS-PD model and baseline models

5.2. THE ROLE OF THE UNDERLYING DISEASE

The second hypothesis of the thesis is that fine-tuning an ASR model with speech data affected by different diseases will result in better performance for the respective disease. To see if this hypothesis holds, we can first observe the difference in performance between the SpoS-PD model and the Spos-MS model on PD and MS data. As table 5.5 shows, the SpoS-PD model is consistently performing better than the SpoS-MS model, even on MS data. This result may be to some extent due to the small amount of MS data the SpoS-MS model has been fine-tuned on (less than 8 min). Nevertheless, even if this is the case, it's a result that does not support the second hypothesis.

Evaluation dataset	**Fine-tuned SpoS-PD**	**Fine-tuned SpoS-MS**
RS-PD	**0.31**	0.46
RS-MS	**0.49**	0.53
SpoS-PD	**0.39**	0.53
SpoS-MS	**0.45**	0.49

Table 5.5: Comparison of PD and MS speech results of SpoS-PD and SpoS-MS models

A second observation we can make is how the performance of each individual model varies when measured against data from all the different diseases. If hypothesis 2 held, the SpoS-PD model should have had better performance on PD speech than the other diseases, and so should the SpoS-MS model on MS speech. However, as table 5.6 shows, none of the models performed best on its corresponding disease. The SpoS-PD model actually performs better on SpoS-SCA and SpoS-Stroke speech than on Spos-PD speech, while the SpoS-MS model has one of its worst WER scores on RS-MS speech (0.53). There are a couple of reasons this might have happened:

- The small size of MS data the SpoS-MS model was fine-tuned on did not help the model capture all the particular characteristics of MS speech that differentiate it from speech affected by other diseases. This, on the other hand, does not apply so much for the SpoS-PD model that has been fine-tuned with almost 2 hours of SpoS-PD speech (see table 3.2)
- Stroke typically causes unilateral upper motor neuron dysarthria which is a milder form of spastic dysarthria. The latter has several characteristics in common with

Evaluation dataset	**Fine-tuned SpoS-PD**	**Fine-tuned SpoS-MS**
RS-PD	0.31	0.46
RS-SCA	0.29	0.46
RS-Stroke	0.35	0.46
RS-MS	0.49	0.53
RS-Other	0.33	0.46
SpoS-PD	0.39	0.53
SpoS-SCA	0.32	0.47
SpoS-Stroke	0.33	0.48
SpoS-MS	0.45	0.49
SpoS-Other	0.41	0.51

Table 5.6: Comparison of evaluation results of the SpoS-PD and SpoS-MS models on read and spontaneous speech data from different diseases

hypokinetic dysarthria (that is caused by PD) such as hypernasality, reduced stress, imprecise consonants, monoloudness and monopitch Rowe et al., 2022. This can to some extent explain the fact that the SpoS-PD model performed better on stroke speech than PD speech. On the other hand, ataxic dysarthria that is typically caused by SCA is quite different than hypokinetic dysarthria, so the better performance of the SpoS-PD model on SCA speech cannot be explained in the same way.

These observations indicate that the experiments do not provide adequate evidence that fine-tuning an ASR model with speech data from different diseases will result in better performance for the respective disease.

5.3. Additional Observations

Looking again at tables 5.1, 5.2 and 5.4, we observe that all three fine-tuned models performed better than the baseline in all the evaluation datasets, except for Domotica. This is also the case for healthy speech data (table 5.7).

Evaluation dataset	**Fine-tuned RS-PD**	**Fine-tuned SpoS-PD**	**Fine-tuned SpoS-MS**	**Baseline**
RS-HC	**0.04**	0.17	0.3	0.56
SpoS-HC	0.69	**0.22**	0.37	0.64

Table 5.7: Comparison of healthy speech results of the all models

To some extent this is to be expected as the three models are further fine-tuned versions of the baseline model. The lower performance in Domotica, especially in the severe group, is most likely due to the fact that the data I used for the fine-tuning are not representative enough in terms of dysarthria severity. Unfortunately, it's hard to verify if this is the case as the severity assessments of the speech in Verkhodanova, 2021 are not comparable to those of Domotica. Another reason could be the regional provenance of the data; Domotica consists of Flemish speech, a Dutch dialect spoken in Flanders, while the dataset used for the fine-tuning of the model was acquired from speakers in the northern part of the Netherlands.

Also, if we calculate the average WER score of each model in all evaluation datasets we observe that the SpoS-PD model had the best performance (table 5.8)). Moreover, it is worth noting that the SpoS-PD data was richer in terms of quantity and variety than the RS-PD and SpoS-MS data. With all these facts combined, it is very likely that quantity and word/sentence variety of training data played a more important role in dysarthric ASR performance than the elicitation method and/or disease.

Fine-tuned RS-PD	**Fine-tuned SpoS-PD**	**Fine-tuned SpoS-MS**	**Baseline**
0.48	**0.38**	0.48	0.61

Table 5.8: Average WER scores of all models on all the evaluation datasets

5.4. Comparing the Findings with the Literature

The findings of the previous section contradict the expectations expressed in Bunton et al., 2000, Blaauw, 1994 and Laan, 1997 that spontaneous speech is more difficult to automatically recognize than read speech, and therefore that ASR systems that are to be used on spontaneous speech should better be trained on the same type of speech. They also contradict the expectations expressed in Lee et al., 2014, Rong et al., 2015, and Rowe et al., 2022 that the variety of phonemic patterns and articulatory impairments across different diseases affect the generalization and optimality of ASR systems that are trained with speech affected by only one or few different diseases.

These contradictions do not necessarily mean that dysarthria diversity does not play a role in ASR perfornance. It suggests, though, that the phonetic or other differences we observe in dysarthric speech of different type and provenance do not always affect ASR performance in the way we would expect. For this reason, we need to perform more experiments like the ones in this thesis that involve actual ASR development, and with richer datasets that contain well-documented information about their provenance and type.

To the best of my knowledge, this thesis is the first attempt to investigate the effect of the elicitation method or the underlying disease of dysarthric speech data on Dutch ASR performance. As we saw in section 2.3, related dysarthric ASR approaches have focused more on investigating the effect of different model architectures and training strategies on ASR performance, rather than the effect of targeted data selection. As such, the ASR models of this thesis are not comparable with those of other related approaches since they do not explore the same research questions.

Moreover, a key reason why in this thesis I created a new dataset were the limitations of the existing Dutch dusarthric datasets with respect to my research questions. The COPAS dataset had no explicit information about the diseases that caused its speaker's dysarthria, Domotica contained no speech with hypokinetic dysarthria, and EST was not available for public use. As such, enriching and expanding existing datasets, creating new ones, and explicitly describing their diversity characteristics should be in the agenda of dysarthric ASR research.

5.5. LIMITATIONS AND FUTURE RESEARCH

Even though the results of the three experiments did not provide adequate evidence to support the two hypotheses of the thesis, certain limitations of the dataset I used could have negatively affected the experiments' informativity.

A first limitation was the very small textual variety of the read speech data which comprised merely 6 sentences from the fable "The North Wind and The Sun". This small variety has most likely led the RS-PD model to overfit, as suggested by its very low WER scores on RS data and very high scores on SpoS data. As such its comparison with the SpoS models, which are trained in higher variety data, is not so representative. For that, as future research, it would be useful to obtain more content diverse read speech data and repeat the first experiment.

A second limitation was the relatively small representation of non-PD diseases in the data. As shown in table 3.1, the duration of the PD recordings is around 2 hours, while the duration of data from the other diseases, including MS, does not exceed 0.2 hours (12 min) per disease. As such, it is likely that the SpoS-MS model was worse than the SpoS-PD model because it was trained on less data. To check if this is actually the case, it would be useful to expand the dataset with more spontaneous speech data affected by MS and re-train and re-evaluate the SpoS-MS model. It would be also nice to train and evaluate similar models for the other diseases (SCA, Stroke, etc.). That would help gain a more accurate understanding of the role of the underlying disease in the performance of dusarthric ASR systems.

A third limitation of the dataset was the lack of dysarthria severity assessments. As we saw above, the three fine-tuned models performed worse on the Domotica dataset than the baseline model, especially in the severe subset. This indicates that the dataset probably does not contain enough severe dysartrhic speech. To verfiy if this is actually the case, one would need to conduct a proper assessment of the dysarthria severity levels in the dataset. That would make the experiments more informative and the dataset more usable for further research.

Finally, something that would also be insteresting to investigate further is how fine-tuning an ASR model with dysarthric speech affects its performance on healthy speech. The WER scores of table 5.7 show that all dysarthric models performed better in healthy speech than in dysarthric speech. We don't know, however, if these scores would have been even lower if I had not performed the dysarthric fine-tuning. And because a dysarthric ASR system could be used in a mixed speaker environment, where speakers with dysarthria interact with healthy speakers, it is important that it's performance on healthy speech is not affected negatively.

6 Conclusion

This thesis focused on investigating whether the developers of dysarthric ASR systems should prioritize the use and creation of extensive dysarthric speech data, disregarding their characteristics and provenance, or whether a more cautious and strategic approach should be taken in data selection. Specifically, the study aimed to address this question within the context of self-supervised learning and examined two significant aspects of dysarthric speech: a) the method used to elicit the speech (read vs. spontaneous speech), and b) the underlying diseases that contribute to the speakers' dysarthria. For these two aspects, I formulated two hypotheses. The first hypothesis is that fine-tuning an ASR model with differently elicited speech data would lead to improved performance for the respective elicitation method. The second hypothesis is that by fine-tuning an ASR model with speech data affected from a specific disease, it would be possible to enhance model's performance on speech affected by the corresponding disease.

To test these hypotheses, I conducted three experiments using an existing recent Dutch dysarthric SSL ASR model and a new dysarthric speech dataset that I created. The latter was based on a recent Dutch dysarthric speech corpus that comprised both read and spontaneous speech from patients with various diseases such as Parkinson's disease, Multiple Sclerosis, and others. The first two experiments involved fine-tuning the baseline model separately with read and spontaneous dysarthric speech data from patients with Parkinson's disease, and evaluating the two resulting models on read and spontaneous speech from different diseases. In the third experiment I fine-tuned the baseline model with spontaneous dysarthric speech data from patients with Multiple Sclerosis, and evaluated it on speech data from patients with Multiple Sclerosis and other diseases. The results of the three experiments did not provide adequate evidence that the elicitation method or the underlying disease of the dysarthric speakers played a significant role in the performance of a dysarthric ASR system. Instead, it was mainly the quantity of data and the content variety that seemed to affect the performance of the system.

Based on the outcome of the experiments and the limitations of the dataset I used, I identified several areas for improvement as part of future research. One such improvement is to address the limitation of the extremely limited textual variety found in the read speech data by acquiring a more diverse set of such data, with a wider range of content. By doing so, it would be possible to repeat the initial experiment with improved textual variability. A second improvement is to expand the dataset by incorporating more spontaneous speech data from individuals with non-PD diseases. This expansion would enable the extension of the third experiment and provide a more comprehensive analysis.

In overall, the thesis provided empirical evidence on how the choice of elicitation method (read vs spontaneous) for dysarthric speech data impacts the effectiveness of SSL ASR models, as well as how the underlying disease responsible for dysarthria affects the effectiveness of these models. Additionally, a valuable labeled speech dataset was created, with diverse samples of dysarthric speech associated with various diseases. These outcomes contribute to a deeper understanding of dysarthria's diversity from an ASR perspective and offer guidance to ASR developers in optimizing their data collection strategies to accommodate the specific characteristics of dysarthric speech.

Acknowledgements

This thesis is the result of a wonderful collaboration. It takes a village to raise a child, it took a team of supportive researchers to finish this work, in a challenging moment of my life.

First, I would like to express my gratitude to Asst. Prof. Dr. Vass Verkhodanova who entrusted me with her speech data corpus that was the core of this thesis. She was always there to answer my data-related questions and help me navigate the dysarthria literature landscape. Then, I would like to thank Asst. Prof. Dr. Shekhar Nayak who was also always available to answer my machine learning questions and help me take important technical and methdological decisions. They both created a friendly, pleasant, and supportive environment for me to work in, removing any obstacles that came into my way.

I also owe gratitude to my co-student and friend Tatsu Matsushima who shared his dysarthric ASR model to use as a baseline for my experiments, and warned me about various technical and methodological challenges that he had also faced in his thesis. A special thanks goes to Assoc. Prof. Dr. Matt Coler who, along with my supervisors, helped me acquire funding for the transcription of the data and resolve administrative issues regarding the hiring of transcribers. And, of course, very important was the contribution of Tessa Pino, Daisy Gallas-Gelderblom, Julia Smit, and Sterre Winter who transcribed and proofread dysarthric speech data in a highly focused and professional way.

I also appreciate the help and support of the university's High Performance Computing Cluster team who were highly responsive in identifying the root causes behind technical issues I was facing when using the cluster and in providing me with clear instructions on how to solve them. Moreover, I would like to give thanks to the my study advisors, Maaike Moltzer and Hieke Hoekstra, who offered significant emotional support throughout the duration of my studies.

Finally, I could not leave out my wonderful husband and partner Panos Alexopoulos who was my emotional and physical support when I most needed it. The only person who can decode the gibberish words when I am exhausted and who can understand me when I use four languages in one sentence. Thank you for respecting my brain thirst for Voice Tech, for following me to the other side of the Netherlands for a year, for dealing with my stress before deadlines, and for being my personal cheerleader in this journey, reminding me that as far as I am healthy I can do whatever I put my mind into.

The last year was quite challenging for me due to some health issues. I had to give priority to my health recovery and that was not an easy process. However, all the people above treated me with such kindness, joy, care, and positive energy, and I can safely say it was the greatest team I could ever ask to surround me. Studying in Campus Fryslân was a wonderful journey that I would not hesitate to repeat.

A

Fine-tuning Dataset Statistics

Dataset	Participants	Sentences	Duration (mins)
RS-PD	43	206	31.84
SpoS-PD	43	570	91.85
SpoS-MS	4	44	6.96

Table A.1: Statistics of the datasets used for fine-tuning

B

Evaluation Dataset Statistics

Dataset	Participants	Sentences	Duration (min)
RS-PD	35	52	8.26
RS-SCA	6	36	8.36
RS-Stroke	4	24	4.55
RS-MS	4	24	6.64
RS-Other	4	24	5.22
RS-HC	40	240	31.18
SpoS-PD	42	143	19.39
SpoS-SCA	6	67	11.03
SpoS-Stroke	4	56	8.44
SpoS-MS	4	11	1.68
SpoS-Other	4	62	12.32
SpoS-HC	40	1231	145.1

Table B.1: Statistics of the datasets used for evaluation

C

Fine-tuning Loss and Accuracy Movement

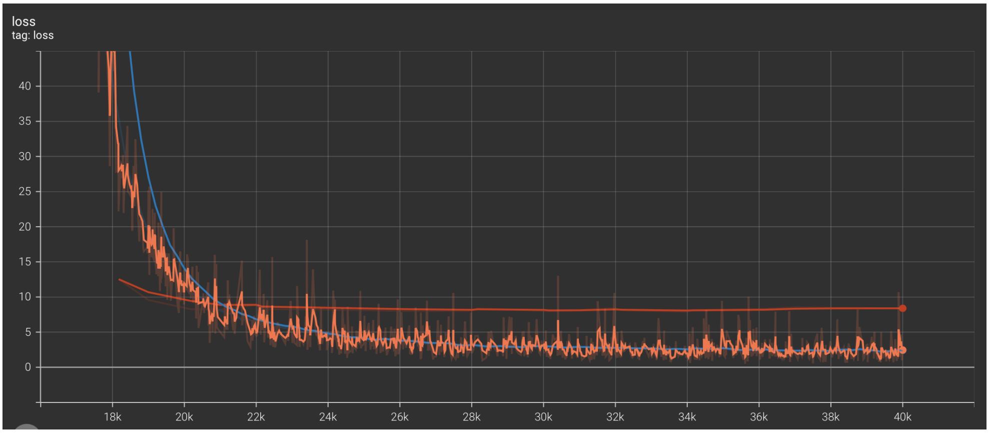


Figure C.1: RS-PD Loss

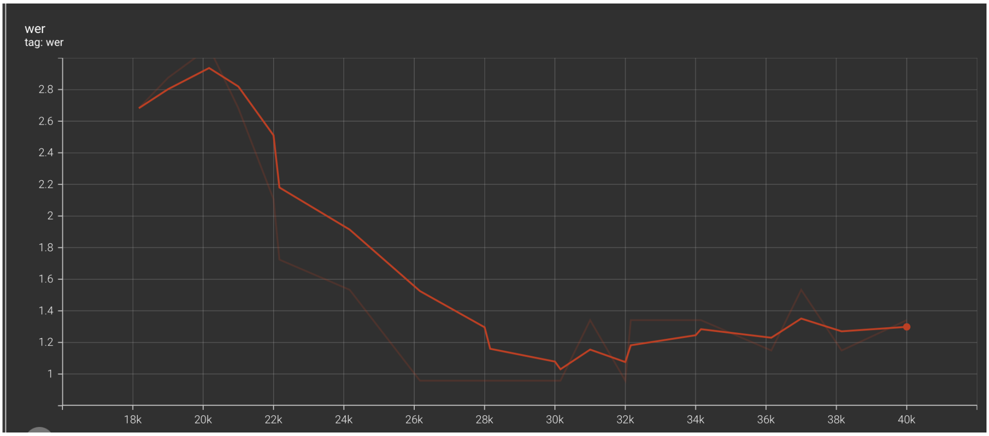


Figure C.2: RS-PD WER

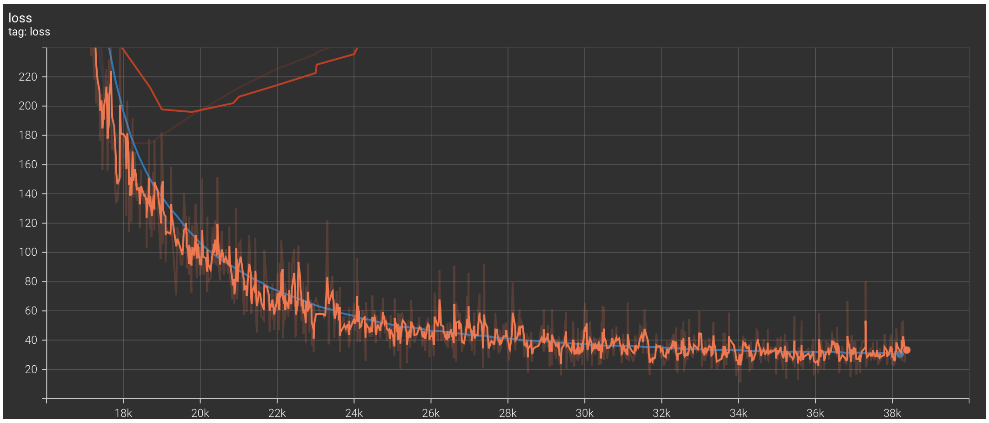


Figure C.3: SS-PD Loss

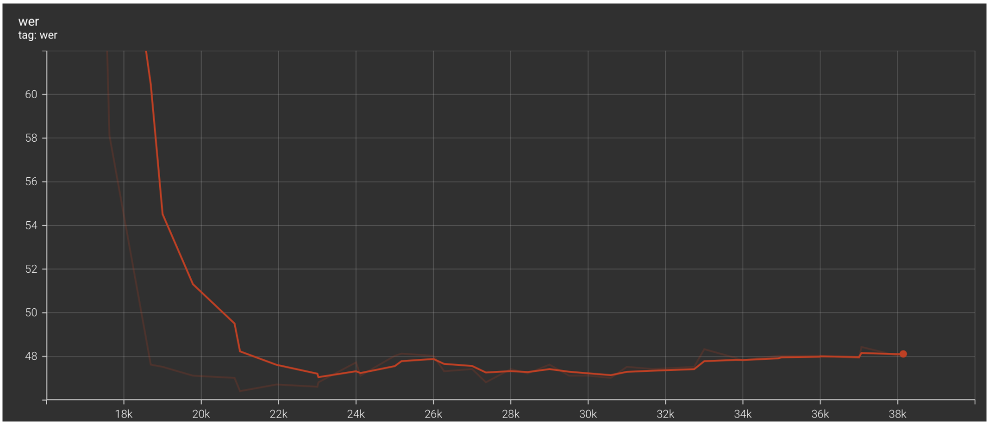


Figure C.4: SS-PD WER

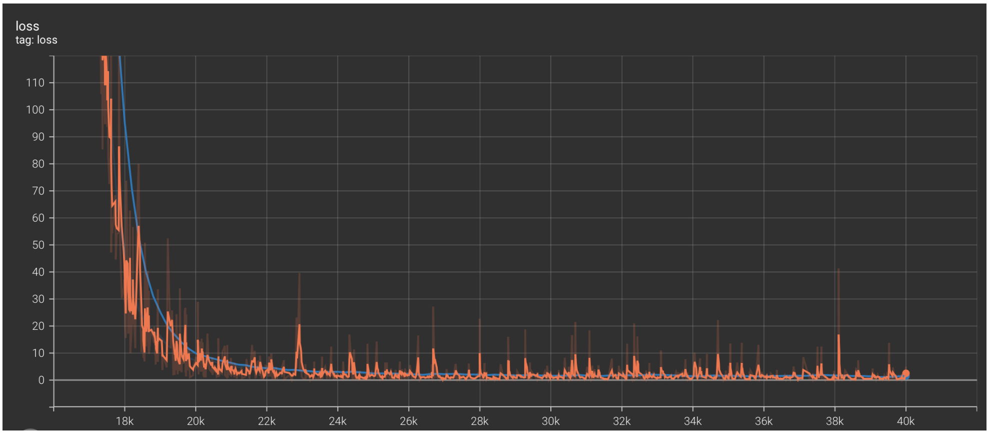


Figure C.5: SS-MS Loss

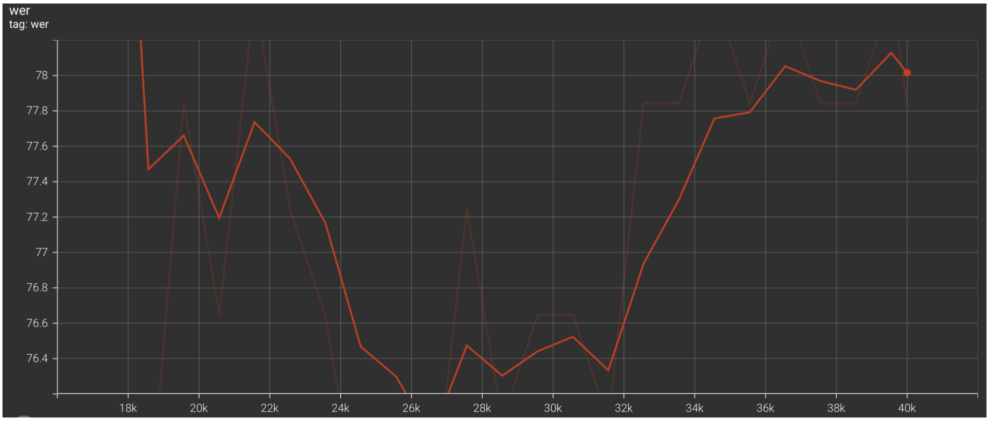


Figure C.6: SS-MS WER

D

Pictures for the Picture Description Task

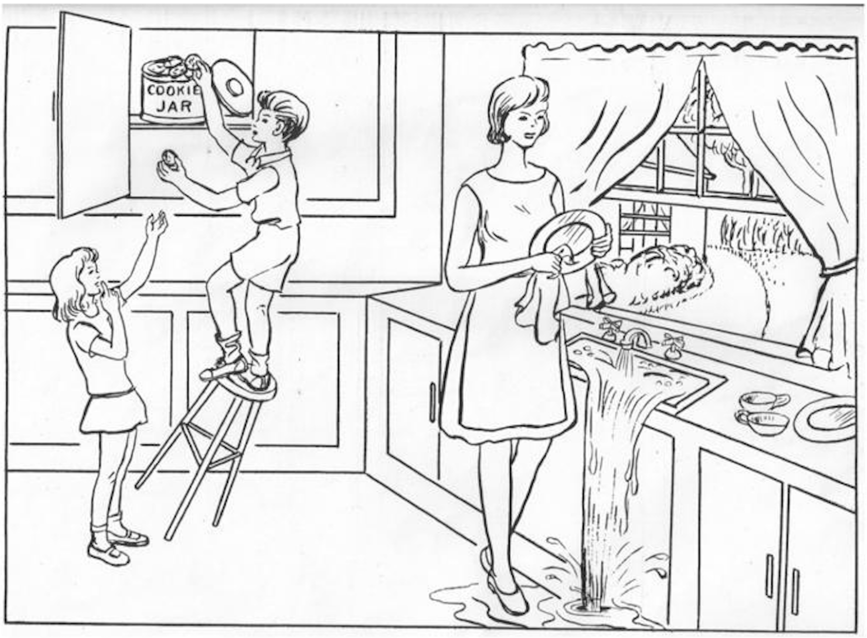


Figure D.1: The Cookie Theft Picture (CPT)

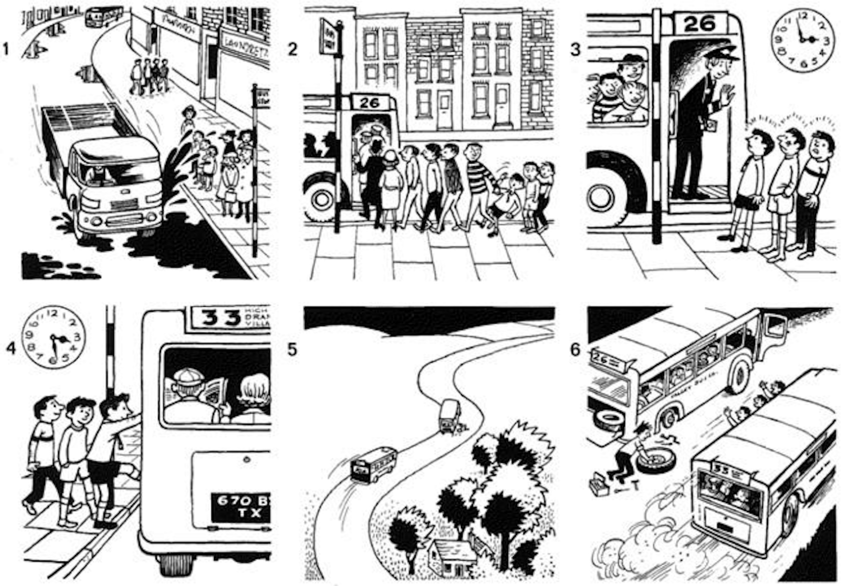


Figure D.2: Heaton Picture

E

Text for the Reading Task

De noordenwind en de zon
De noordenwind en de zon waren erover aan het redetwisten wie de sterkste was van hun beiden. Juist op dat moment kwam er een reiziger aan, die gehuld was in een warme mantel. Ze kwamen overeen dat degene die het eerst erin zou slagen de reiziger zijn mantel te doen uittrekken de sterkste zou worden geacht. De noordenwind begon toen uit alle macht te blazen, maar hoe harder hij blies, des te dichter trok de reiziger zijn mantel om zich heen, en ten lange leste gaf de noordenwind het op. Daarna begon de zon krachtig te stralen, en hierop trok de reiziger onmiddellijk zijn mantel uit. De noordenwind moest dus wel bekennen dat de zon van hun beiden de sterkste was.

Bibliography

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, neurips 2020, december 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

Blaauw, E. (1994). The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication*, *14*, 359–375. https://doi.org/10.1016/0167-6393(94)90028-0

Blaney, B., & Hewlett, N. (2007). Dysarthria and friedreich's ataxia: What can intelligibility assessment tell us? *International journal of language and communication disorders / Royal College of Speech and Language Therapists*, *42*, 19–37. https://doi.org/10.1080/13682820600690993

Bunton, K., Kent, R., Kent, J., & Rosenbek, J. (2000). Perceptuo-acoustic assessment of prosodic impairment in dysarthria. *Clinical linguistics and phonetics*, *14*, 13–24. https://doi.org/10.1080/026992000298922

Canter, G. J. (1965). Speech characteristics of patients with parkinson's disease. 3. articulation, diadochokinesis, and over-all speech adequacy. *The Journal of speech and hearing disorders*, *30*, 217–24.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Zeng, M., Yu, X., & Wei, F. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*, 1–14. https://doi.org/10.1109/JSTSP.2022.3188113

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition, 2426–2430. https://doi.org/10.21437/Interspeech.2021-329

Damhuis, M., Boogaart, T. I., in't Veld, C., Versteijlen, M., Schelvis, W., Bos, L., & Boves, L. (1994). Creation and analysis of the dutch polyphone corpus. *The 3rd International Conference on Spoken Language Processing, ICSLP 1994, Yokohama, Japan, September 18-22, 1994*. http://www.isca-speech.org/archive/icslp%5C_1994/i94%5C_1803.html

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

Geng, M., Xie, X., Liu, S., Yu, J., Hu, S., Liu, X., & Meng, H. (2020a). Investigation of data augmentation techniques for disordered speech recognition. https://doi.org/10.21437/Interspeech.2020-1161

Geng, M., Xie, X., Liu, S., Yu, J., Hu, S., Liu, X., & Meng, H. (2020b). Investigation of data augmentation techniques for disordered speech recognition. https://doi.org/10.21437/Interspeech.2020-1161

Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets. *ICML '06: Proceedings of the International Conference on Machine Learning*.

Green, J., MacDonald, R., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., Seaver, K., Ladewig, M., Tobin, J., Brenner, M., Nelson, P., & Tomanek, K. (2021). Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases, 4778–4782. https://doi.org/10.21437/Interspeech.2021-1384

Gupta, R., Chaspari, T., Kim, J., Kumar, N., Bone, D., & Narayanan, S. S. (2016). Pathological speech processing: State-of-the-art, current challenges, and future directions. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 6470–6474. https://doi.org/10.1109/ICASSP.2016.7472923

Heaton, J. B. (1972). *Composition through pictures*. Longman Group United Kingdom.

Hernandez, A., Perez-Toro, P., Nöth, E., Orozco, J. R., Maier, A., & Yang, S. (2022). *Cross-lingual self-supervised speech representations for improved dysarthric speech recognition*.

Hsu, W.-N., Tsai, Y.-H. H., Bolte, B., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: How much can a bad teacher benefit asr pre-training? *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6533–6537. https://doi.org/10.1109/ICASSP39728.2021.9414460

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Watkin, K., & Frame, S. (2008). Dysarthric speech database for universal access research. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1741–1744. https://doi.org/10.21437/Interspeech.2008-480

Kim, Y., Weismer, G., Kent, R. D., & Duffy, J. R. (2009). Statistical models of f2 slope in relation to severity of dysarthria. *Folia Phoniatrica et Logopaedica*, *61*, 329–335.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *Iclr (poster)*. http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14

Krishna, D. N., Wang, P., & Bozza, B. (2021). Using Large Self-Supervised Models for Low-Resource Speech Recognition. *Proc. Interspeech 2021*, 2436–2440. https://doi.org/10.21437/Interspeech.2021-631

Laan, G. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, *22*, 43–65. https://doi.org/10.1016/S0167-6393(97)00012-5

Lee, J., Hustad, K., & Weismer, G. (2014). Predicting speech intelligibility with a multiple speech subsystems approach in children with cerebral palsy. *Journal of speech, language, and hearing research : JSLHR*, *57*. https://doi.org/10.1044/2014_JSLHR-S-13-0292

Leuschel, A., & Docherty, G. J. (1996). Prosodic assessment of dysarthria. *D. A. Robin, K. M. Yorkston and D. R. Beukelman (Eds), Disorders of Motor Speech: Assessment, treatment, and clinical characterization (Baltimore, MD: Paul H. Brookes)*, 155–178.

Liu, S., Mallol-Ragolta, A., Parada-Cabaleiro, E., Qian, K., Jing, X., Kathan, A., Hu, B., & Schuller, B. (2022). Audio self-supervised learning: A survey. *Patterns*, *3*, 100616. https://doi.org/10.1016/j.patter.2022.100616

Martens, J.-P., Bodt, M. D., Nuffelen, G. V., & Middag, C. (2011). Corpus of pathological and normal speech (copas).

Matsushima, T. (Ed.). (2022). *Dutch dysarthric speech recognition: Applying self-supervised learning to overcome the data scarcity issue*. University of Groningen, Netherlands. https://campus-fryslan.studenttheses.ub.rug.nl/211/

Menendez-Pidal, X., Polikoff, J., Peters, S., Leonzio, J., & Bunnell, H. (1996). The nemours database of dysarthric speech. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *3*, 1962–1965 vol.3. https://doi.org/10.1109/ICSLP.1996.608020

Meunier, C., Cecile, F., Fredouille, C., Crevier-Buchman, L., Delais-Roussarie, E., Georgeton, L., Ghio, A., Laaridh, I., Legou, T., Pillot-Loiseau, C., Pouchoulin, G., & Bigi, B. (2016). *The typaloc corpus: A collection of various dysarthric speech recordings in read and spontaneous styles*.

Middag, C. (2012). *Automatic analysis of pathological speech* (Doctoral dissertation).

Moore, M., Venkateswara, H., & Panchanathan, S. (2018). Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems. *Proc. Interspeech 2018*, 466–470. https://doi.org/10.21437/Interspeech.2018-2391

Nicolao, M., Christensen, H., Cunningham, S., Green, P., & Hain, T. (2016). The homeservice corpus v. 1.0, , university of sheffield at http://mini.dcs.shef.ac.uk/resources/homeservice-corpus. https://doi.org/10.15131/shef.data.3116833

Ons, B., Gemmeke, J., & Van hamme, H. (2014). The self-taught vocal interface. *EURASIP Journal on Audio, Speech, and Music Processing, 2014*. https://doi.org/10.1186/s13636-014-0043-4

Oostdijk, N. (2000). The spoken dutch corpus: Overview and first evaluation. *Proceedings of LREC-2000, Athens, 2*.

Orozco, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J., González-Rátiva, M., & Noeth, E. (2014). New spanish speech corpus database for the analysis of people suffering from parkinsons disease. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, 342–347.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). Fairseq: A fast, extensible toolkit for sequence modeling. *Proceedings of NAACL-HLT 2019: Demonstrations*.

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

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning.

Platt, L., Andrews, G., & Howie, P. (1980). Dysarthria of adult cerebral palsy: Ii. phonemic analysis of articulation errors. *Journal of speech and hearing research*, *23*, 41–55. https://doi.org/10.1044/jshr.2301.41

Rong, P., Yunusova, Y., Wang, J., & Green, J. (2015). Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach. *Behavioural Neurology*, *2015*, 1–11. https://doi.org/10.1155/2015/183027

Rowe, H., Gutz, S., Maffei, M., & Green, J. (2020). Acoustic-based articulatory phenotypes of amyotrophic lateral sclerosis and parkinson's disease: Towards an interpretable, hypothesis-driven framework of motor control, 4816–4820. https://doi.org/10.21437/Interspeech.2020-1459

Rowe, H., Gutz, S., Maffei, M., Tomanek, K., & Green, J. (2022). Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective. *Frontiers in Computer Science*, *4*, 770210. https://doi.org/10.3389/fcomp.2022.770210

Rudzicz, F., Namasivayam, A., & Wolff, T. (2010). The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, *46*, 1–19. https://doi.org/10.1007/s10579-011-9145-0

Saz, O., Rodriguez-Dueñas, W. R., Lleida, E., & Vaquero, C. (2008). A novel corpus of children's disordered speech.

Shahamiri, S. R. (2021). Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 852–861. https://doi.org/10.1109/TNSRE.2021.3076778

Shor, J., Emanuel, D., Lang, O., Tuval, O., Brenner, M., Cattiau, J., Vieira, F., McNally, M., Charbonneau, T., Nollstadt, M., Hassidim, A., & Matias, Y. (2019). Personalizing asr for dysarthric and accented speech with limited data, 784–788. https://doi.org/10.21437/Interspeech.2019-1427

Tachibana, H., Uenoyama, K., & Aihara, S. (2018). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention, 4784–4788. https://doi.org/10.1109/ICASSP.2018.8461829

Torrey, L., & Shavlik, J. (2009). Transfer learning. *Handbook of Research on Machine Learning Applications. IGI Global.*

Turrisi, R., Braccia, A., Emanuele, M., Giulietti, S., Pugliatti, M., Sensi, M., Fadiga, L., & Badino, L. (2021). Easycall corpus: A dysarthric speech dataset, 41–45. https://doi.org/10.21437/Interspeech.2021-549

Verkhodanova, V. (2021). *More than words: Recognizing speech of people with parkinson's disease* (Doctoral dissertation). University of Groningen. University of Groningen. https://doi.org/10.33612/diss.183425053

Vijayalakshmi, P., Celin, T. A. M., & Nagarajan, T. (2022). *The SSNCE Database of Tamil Dysarthric Speech*. https://doi.org/11272.1/AB2/QXP9LM

Violeta, L. P., Huang, W.-C., & Toda, T. (2022). Investigating self-supervised pretraining frameworks for pathological speech recognition. *Interspeech*.

Wang, P., BabaAli, B., & hamme, H. V. (2021). A study into pre-training strategies for spoken language understanding on dysarthric speech. *Interspeech*.

Wong, K. H., Yeung, Y., Chan, E., Wong, P., Levow, G.-A., & Meng, H. (2015). Development of a cantonese dysarthric speech corpus, 329–333. https://doi.org/10.21437/Interspeech.2015-149

Xiong, F., Barker, J., Yue, Z., & Christensen, H. (2020). Source domain data selection for improved transfer learning targeting dysarthric speech recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7424–7428. https://doi.org/10.1109/ICASSP40776.2020.9054694

Yilmaz, E., Ganzeboom, M., Beijer, L., Cucchiarini, C., & Strik, H. (2016). A dutch dysarthric speech database for individualized speech therapy research.

Yılmaz, E., Ganzeboom, M., Cucchiarini, C., & Strik, H. (2016). Combining non-pathological data of different language varieties to improve dnn-hmm performance on pathological speech. https://doi.org/10.21437/Interspeech.2016-109

Yılmaz, E., Ganzeboom, M., Cucchiarini, C., & Strik, H. (2017). Multi-stage dnn training for automatic recognition of dysarthric speech. https://doi.org/10.21437/Interspeech.2017-303