AI-Based Speech Assessment of Cognitive Impairment Disorders

by

R'mani Haulcy

B.S., Yale University (2017)
S.M., Massachusetts Institute of Technology (2019)

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: R'mani Haulcy
Department of Electrical Engineering and Computer Science
May 15, 2023

Certified by: James Glass
Senior Research Scientist
Computer Science Artificial Intelligence Laboratory
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

AI-Based Speech Assessment of Cognitive Impairment Disorders

by

R'mani Haulcy

Submitted to the Department of Electrical Engineering and Computer Science
on May 15, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Previous research has shown that speech can be used to detect cognitive impairment in patients with dementia and other neurodegenerative diseases. These diseases produce cognitive deficits that lead to changes in the acoustic and linguistic content of the speech produced by the patients.

In this thesis, we analyze the speech of subjects with Alzheimer's Disease (AD), Frontotemporal Dementia (FTD), and Primary Progressive Aphasia (PPA). We show that AD subjects can be distinguished from healthy controls with 85.4% accuracy and that the Mini-Mental State Examination scores of the subjects can be predicted with a root mean squared error of 4.56, using sentence embeddings. We present the Crowdsourced Language Assessment Corpus (CLAC), a corpus that we created to provide the community with a collection of audio samples from various speakers that can be used to learn a general representation for speech from healthy subjects, as well as complement other health-related speech datasets.

We present a novel, language-agnostic approach for measuring the quality of repetition in a recording, a method that was inspired by the need to automatically quantify the impaired repetition abilities that characterize the speech of people with the logopenic variant of PPA (lvPPA). A subset of the CLAC corpus was used as healthy controls and we demonstrated the feasibility of our approach by using it to distinguish between healthy and lvPPA speakers with impaired repetition with 85.7% accuracy. Lastly, we compare standard linguistic features to more advanced sentence embeddings by using a variety of feature extraction methods to extract features from picture description and monologue data for four different FTD/PPA variants. We show that all variants can be distinguished from healthy controls with >= 90% accuracy using transformer-based sentence embeddings.

We hope that the work presented in this thesis will contribute to the goal of using artificial intelligence to improve human health, clinical trial design, and drug development.

Thesis Supervisor: James Glass
Title: Senior Research Scientist

Acknowledgments

Several people have been instrumental in helping me through the challenging process of getting a PhD from MIT.

I want to start by thanking John Belcaster, my high school teacher and mentor over the past several years. He has followed my journey from a young age and has supported me in ways that have deeply touched my heart. He has written my favorite recommendation that anyone has ever written me to this day and was one of the first to support me on the start of my academic journey post-high school. His generosity and encouragement is one of many things that have contributed to my academic success thus far.

I would like to thank Dr. Jonathan Sprinkle for writing me a glowing recommendation that helped to cement my acceptance to MIT. My first research experience was with him as a part of the 2016 CAT Vehicle Summer REU and that experience solidified my interest in research and helped propel me towards my PhD journey.

I would like to thank my first MIT advisor, Professor Berthold Horn, for his kindness and support throughout the duration of my first two years as I completed the requirements for my Master's degree. I would also like to thank him for his willingness to provide guidance as I prepared for my Research Qualifying Exam (RQE), even though he technically was no longer my advisor.

I would like to thank my current advisor, Dr. James Glass, for allowing me to join his research lab, the Spoken Language Systems (SLS) group, at the start of my third year and seeing my potential as a student. He has provided a research environment that allows me to learn from him and my peers and produce quality work that contributes to the research community. I am also thankful for the guidance and feedback that I received from my colleagues in SLS.

I would like to thank my thesis committee members, Professor Randall Davis and Professor Collin Stultz, for providing valuable feedback during the writing process. This thesis is much stronger as a result of their unique perspectives. I would also like to thank our colleagues at Takeda for their kindness, feedback, expertise, and contributions over the years. Also, thank you to Professor Adam Vogel for providing us with the Frontotemporal

Dementia (FTD) dataset needed to complete much of the work in this thesis.

I must thank the University Center for Exemplary Mentoring (UCEM) for providing me with a community and safe haven within MIT, outside of my research community. Without UCEM, I know that my PhD experience would have felt much more lonely and isolating. I don't believe that I could have completed this journey without the love and support that I received from the UCEM staff and my peers. Leslie Kolodziejski deserves to be acknowledged specifically for her kind spirit and unrelenting support. In every moment of doubt and stress throughout the past six years, Leslie has been a listening ear and has helped me to find solutions to any problem I presented her with. Thank you so much Leslie.

I would like to thank the Graduate Women of Color Support Group for giving me a space to be myself fully and unapologetically. It is one of the few places at MIT that I have felt fully supported and seen and the participants have helped to remind me of my strength and resilience in moments when I doubted my ability to cross the finish line. I would not have this accomplishment without the group and its members. I will never forget what it did for me and I am honored to be an alum.

Last but certainly not least, I would like to thank my family, who has never waivered in their faith, love, and support of me. I am who I am because of them and would not have any of my accomplishments without them.

I am grateful for all of the fellowships that I have received throughout my time at MIT, including the Alfred P. Sloan Foundation's Minority PhD program scholarship, the MIT Presidential Graduate fellowship, the Frank Quick fellowship, the Mathworks Engineering fellowship, and the Takeda fellowship. The funding has been paramount in the completion of my degree and has allowed me to complete meaningful work without the added stress of finding funding. The work in this thesis was supported by Takeda Development Center Americas, INC. (successor in interest to Millennium Pharmaceuticals, INC.).

Bibliographic Note

Portions of this thesis have appeared in peer-reviewed publications or are currently under review:

- R'mani Haulcy and James Glass. Classifying Alzheimer's disease using audio and text-based representations of speech. Frontiers in Psychology, 11:3833, 2021.
- R'mani Haulcy and James Glass. CLAC: A Speech Corpus of Healthy English Speakers. In Proc. Interspeech 2021, pages 2966-2970, 2021.
- R'mani Haulcy, Katerina Placek, Brian Tracey, Adam Vogel, and James Glass. Repetition assessment for speech and language disorders: A study of the logopenic variant of primary progressive aphasia. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6932-6936, 2022.
- R'mani Haulcy, Katerina Placek, Brian Tracey, Adam Vogel, and James Glass. Classifying Primary Progressive Aphasia and Frontotemporal Dementia from Speech and Text. Under Review For Interspeech 2023.

Contents

List of Figures

List of Tables

Glossary

Symbols | A | B | C | D | F | G | H | L | M | N | O | P | Q | R | S | T | U | V | W

Symbols

p Probability

1NN K-Nearest Neighbor Regressor/Classifier With One Neighbor

A

AD Alzheimer's Disease

ADReSS Alzheimer's Dementia Recognition through Spontaneous Speech

AMT Amazon Mechanical Turk

ASR Automatic Speech Recognition

AUC Area Under the Receiver Operating Characteristic Curve

B

B bert-base-uncased

BERT Bidirectional Encoder Representations from Transformers

bvFTD Behavioral Variant of FTD

C

CBOW Continuous Bag of Words

CHAT Codes for the Human Analysis of Transcripts

CLAC Crowdsourced Language Assessment Corpus

CLAN Computerized Language Analysis

CNN Convolutional Neural Network

COOK Cookie Theft

D

dB Decibel

dBFS dB relative to full scale

DBS diffcse-bert-base-uncased-sts

DBT diffcse-bert-base-uncased-trans

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

DiffCSE Difference-based Contrastive Learning for Sentence Embeddings

DNN Deep Neural Network

DRS diffcse-roberta-base-sts

DRT diffcse-roberta-base-trans

DT Decision Tree

F

FPR False Positive Rate

FTD Frontotemporal Dementia

G

Grad-Boost Gradient-Boosting

H

HA Harmonic Amplitude

HF Harmonic Frequency

HIT Human Intelligence Tasks

Hz Hertz

L

LDA Linear Discriminant Analysis

LIWC Linguistic Inquiry and Word Count

LOSO Leave-One-Subject-Out

LR Linear Regression

LSTM Long Short-Term Memory

lvPPA Logopenic Variant of PPA

M

MFCC Mel-frequency Cepstral Coefficient

MMSE Mini-Mental State Examination

MONL Monologue

MR Minima Range

MRCG Multi-resolution Cochleagram

MT Minima Time

MV Majority Vote

N

NLP Natural Language Processing

nvPPA Nonfluent Variant of PPA

O

OLS Ordinary Least Squares

P

PCA Principal Component Analysis

PD Parkinson's Disease

PET Positron Emission Tomography

PLM Pretrained Language Model

PPA Primary Progressive Aphasia

Q

QNLI Question-Answering Entailment

QQP Quora Question Pair

R

ReLu Rectified Linear Unit

RF Random Forest

RMSE Root Mean Squared Error

RoBERTa Robustly Optimized BERT Approach

ROC Receiver Operating Characteristic

S

SDTW Segmental Dynamic-Time-Warping

SMR Sequential Motion Rate

STS Semantic Textual Similarity

SVD Singular Value Decomposition

SVM Support Vector Machine

svPPA Semantic Variant of PPA

T

TDP-43 TAR DNA Binding Protein of 43 kDa

TE trans-encoder

TPR True Positive Rate

U

UBM Universal Background Model

V

VAD Voice Activity Detection

W

WER Word Error Rate

Chapter 1

Introduction

Previous research has shown that speech can be used to detect cognitive impairment in patients with dementia and other neurodegenerative diseases. These diseases produce cognitive deficits that lead to changes in the acoustic and linguistic content of the speech produced by the patients. Examples of these acoustic and linguistic changes include the voice tremors that are characteristic of Parkinson's disease (PD) (Perez et al., 1996), the difficulty with recalling words that Alzheimer's Disease (AD) patients experience (Martin et al., 1985), and the changes in pitch, speech timing, and struggles with agrammatical speech that frontotemporal dementia (FTD) patients experience (Poole et al., 2017). We are interested in using machine learning techniques to analyze the speech of patients with various forms of cognitive impairment to identify biomarkers that will help (1) distinguish between different types of dementia and cognitive impairment and (2) understand how the acoustic and linguistic symptoms change over time for various neurodegenerative diseases.

In this thesis, we analyze the speech of subjects with AD, FTD, and Primary Progressive Aphasia (PPA). AD is a progressive, neurodegenerative disease that affects the lives of more than 5 million Americans every year. The number of Americans living with AD is expected to be more than double that number by 2050. AD is a deadly and costly disease that has negative emotional, mental, and physical implications for those afflicted with the disease and their loved ones (Alzheimer's Association, 2019). There is currently no cure for AD (Yadav, 2019) and early detection is imperative for effective intervention to occur (De Roeck et al., 2019). Currently, AD is diagnosed using positron emission tomography (PET) imaging and cerebrospinal fluid exams to measure the concentration of amyloid plaques in the brain, a costly and invasive process (Land and Schaffer, 2020). A more cost-effective, non-invasive and easily-accessible technique is needed for detecting AD.

FTD is a neurodegenerative disease characterized by behavioral problems and aphasic issues (e.g. grammatical issues, difficulty with word meaning, reduction in the number of words produced, etc). FTD is an overarching term given to several related conditions that are the phenotypes of FTD. There is no known genetic cause of FTD but two proteins are associated with the phenotypes: TAR DNA binding protein of 43 kDa (TDP-43) and Tau (Wennberg et al., 2019). Drugs are being developed to target these proteins. While some phenotypes are associated with just one of the two proteins, it is not clear which protein is affiliated with the other phenotypes. This makes it dificult to determine which drugs should be given to which patients during clinical trials. It is known that the speech of patients with FTD varies significantly from the speech of healthy individuals (Vogel et al., 2017). There are also some distinguishing speech and language characteristics between different phenotypes. Therefore, it is possible that analyzing the speech of FTD patients can aid us in distinguishing between the phenotypes.

PPA consists of a rare set of conditions that affect speech in different ways. Some forms of PPA are also considered variants of FTD. The FTD/PPA variants that we analyze in this thesis are the logopenic variant of PPA (lvPPA), the behavioral variant of FTD (bvFTD), the semantic variant of PPA (svPPA), and the nonfluent variant of PPA (nfvPPA). Some of the characteristics for each variant can be seen in Figure 1-1 (Poole et al., 2017). The figure shows that svPPA, lvPPA, and nfvPPA subjects have a decreased speech rate. bvFTD and lvPPA subjects have altered pause timings compared to healthy controls. The speech of lvPPA subjects is also characterized by repetition impairment (Haulcy et al., 2022). svPPA subjects have word recall difficulties and often substitute pronouns for nouns (Zangrandi et al., 2021), while nfvPPA subjects have agrammatical speech and difficulties with articulation (Poole et al., 2017). In subsequent chapters, we develop methods for quantifying some of these characteristics.

In this thesis, AI-based analysis is used to improve upon traditional hand-crafted feature analysis and identify AD, FTD, and PPA-specific speech biomarkers that can be tracked

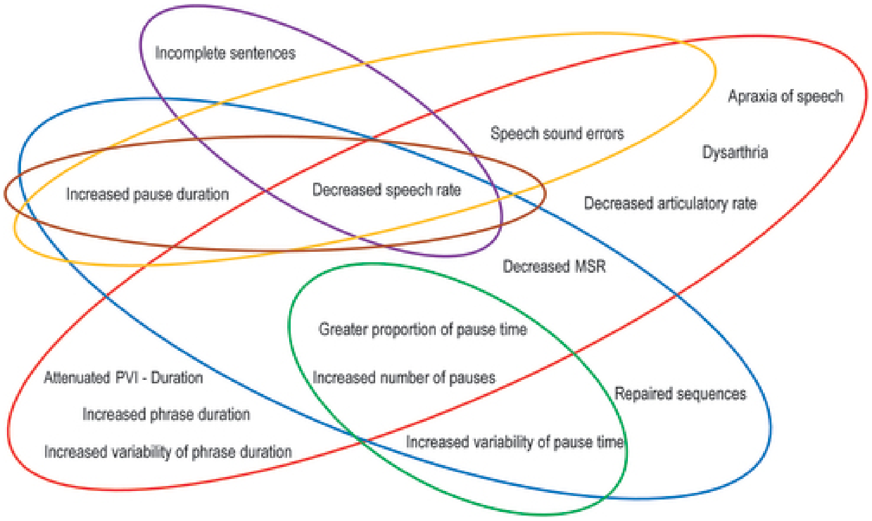

Incomplete sentences
Apraxia of speech
Speech sound errors
Dysarthria
Increased pause duration
Decreased speech rate
Decreased articulatory rate
Decreased MSR
Greater proportion of pause time
Attenuated PVI - Duration
Increased number of pauses
Repaired sequences
Increased phrase duration
Increased variability of pause time
Increased variability of phrase duration

Figure 1-1: Speech characteristics for lvPPA (blue), svPPA (purple), nfvPPA (red), and bvFTD (green) subjects. Originally published in (Poole et al., 2017).

during clinical trials. The ultimate goal is to provide biomarkers that can be used to track disease progression with greater confidence in future clinical trials.

1.1 Thesis Overview And Contributions

This thesis has the following structure and makes the following contributions:

In **Chapter 2**, we present the findings resulting from our participation in the Interspeech 2020 Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge. We extracted different types of audio and text-based features from the data before training binary classifiers and regressors. We contributed to the challenge by using features and model architectures that were not used by any of the other participants and providing insight into how those factors impact model performance.

In **Chapter 3**, we present the Crowdsourced Language Assessment Corpus (CLAC), a corpus that we created to provide the community with a collection of audio samples from various speakers that could be used to learn a general representation for speech from healthy subjects, as well as complement other health-related speech datasets (Haulcy and Glass, 2021c). CLAC consists of audio recordings and automatically-generated transcripts from 1,832 speakers located in the United States, as well as 11 other countries. Datasets with speech from FTD/PPA subjects are rare and some of the tasks completed by the subjects do not have data from healthy subjects to complement it. The CLAC corpus was used to supplement the limited data that we already had. The dataset was also publicly released so that other researchers could benefit from its creation.

In **Chapter 4**, we present the results of developing a novel, language-agnostic approach for measuring the quality of repetition in a recording (Haulcy et al., 2022), a method that was inspired by the need to automatically quantify the impaired repetition abilities that characterize the speech of people with lvPPA. A subset of the CLAC corpus was used as healthy controls and we demonstrated the feasibility of our approach by using it to distinguish between healthy and lvPPA speakers using classification. Our method is a general repetition measurement approach that can be applied to research problems in a variety of areas. It is also minimally invasive, inexpensive, and uses recordings that are quick/easy to record, making it appealing for use in clinical trials. The language-independent nature of the approach also makes it potentially useful for global clinical trials.

In **Chapter 5**, we compare standard linguistic features to more advanced sentence embeddings by using a variety of feature extraction methods to extract features from picture description and monologue data for all four FTD/PPA variants mentioned above. We present the results of training a classifier on the different feature types and comparing the performance. We perform several ablation studies to gain insight into what information the sentence embeddings may be capturing, and we show that human transcription can be replaced with automatic transcription for several variants.

In **Chapter 6**, we summarize the findings from previous chapters and present plans for future work.

We believe that the work we've done so far and continue to do is significant and beneficial to the research community, particularly the community of researchers doing work at the intersection of artificial intelligence and health. We hope that the work presented in this thesis will contribute to the goal of using artificial intelligence to improve human health, clinical trial design, and drug development.

Chapter 2

Classifying Alzheimer's Disease From Speech and Text

Alzheimer's Disease (AD) is a form of dementia that affects the memory, cognition, and motor skills of patients. Extensive research has been done to develop accessible, cost-effective, and non-invasive techniques for the automatic detection of AD. Previous research has shown that speech can be used to distinguish between healthy patients and afflicted patients. In this chapter, the ADReSS dataset, a dataset balanced by gender and age, was used to automatically classify AD from spontaneous speech. The performance of 5 classifiers, as well as a convolutional neural network and long short-term memory network, was compared when trained on audio features (i-vectors and x-vectors) and text features (word vectors, BERT embeddings, LIWC features, and CLAN features). The same audio and text features were used to train 5 regression models to predict the Mini-Mental State Examination score for each patient, a score that has a maximum value of 30. The top-performing classification models were the support vector machine and random forest classifiers trained on BERT embeddings, which both achieved an accuracy of 85.4% on the test set. The best-performing regression model was the gradient boosting regression model trained on BERT embeddings and CLAN features, which had a root mean squared error of 4.56 on the test set. The performance on both tasks illustrates the feasibility of using speech to classify AD and predict neuropsychological scores.[1]

[1] The work in this chapter was previously published in (Haulcy and Glass, 2021a).

2.1 Motivation

Previous research has shown that speech can be used to distinguish between healthy and AD patients (Pulido et al., 2020). Some researchers have focused on developing new machine learning model architectures to improve detection (Liu et al., 2020; Chen et al., 2019; Chien et al., 2019), while others have used language models (Guo et al., 2019) to classify AD. Others have focused on trying to extract acoustic and text features that capture information indicative of AD. These features include non-verbal features, such as the length of segments and the amount of silence (König et al., 2015). Other researchers have used linguistic and audio features extracted from English speech (Fraser et al., 2016; Gosztolya et al., 2019), as well as Turkish speech (Khodabakhsh et al., 2015). Prosodic features have been extracted from English speech (Nagumo et al., 2020; Ossewaarde et al., 2019; Qiao et al., 2020) and German speech (Weiner et al., 2016) to classify AD, and so have paralinguistic acoustic features (Haider et al., 2019). Other researchers have chosen to focus on the type of speech data that is used instead of the type of model or type of features and have used speech from people performing multiple tasks to improve generalizability (Balagopalan et al., 2018). This provides a brief summary of the work that has been done in the past few years. A more extensive review of the background literature can be found in the review paper of (de la Fuente Garcia et al., 2020).

Although promising research has been done, the datasets that have been used are often imbalanced and vary across studies, making it difficult to compare the effectiveness of different modalities. Two recent review papers (Voleti et al., 2019; de la Fuente Garcia et al., 2020) explain that an important future direction for the detection of cognitive impairment is providing a balanced, standardized dataset that will allow researchers to compare the effectiveness of different classification techniques and feature extraction methods. This is what the ADReSS challenge attempted to do. The ADReSS challenge provided an opportunity for different techniques to be performed on a balanced dataset that alleviated the common biases associated with other AD datasets and allowed those techniques to be directly compared.

Previous work has been done using the ADReSS dataset. Some researchers only participated in the AD classification task (Yuan et al., 2020; Edwards et al., 2020; Pompili et al., 2020), others only participated in the Mini-Mental State Examination (MMSE) prediction task (Farzana and Parde, 2020), and others participated in both tasks (Luz et al., 2020; Balagopalan et al., 2020a; Martinc and Pollak, 2020; Pappagari et al., 2020; Cummins et al., 2020; Rohanian et al., 2020; Searle et al., 2020; Sarawgi et al., 2020; Koo et al., 2020; Syed et al., 2020). The best performance on the AD classification task was achieved by (Yuan et al., 2020), who obtained an accuracy of 89.6% on the test set using linguistic features extracted from the transcripts, as well as encoded pauses. The best performance on the MMSE prediction task was achieved by (Koo et al., 2020), who obtained a root mean squared error (RMSE) of 3.747 using a combination of acoustic and textual features.

As part of our work, audio features (i-vectors and x-vectors) and text features (word vectors, BERT embeddings, LIWC features, and CLAN features) were extracted from the data and used to train several classifiers, neural networks, and regression models to detect AD and predict MMSE scores. I-vectors and x-vectors, originally intended to be used for speaker verification, have been shown to be effective for detecting AD (López et al., 2019) and other neurodegenerative diseases, such as Parkinson's Disease (Moro-Velazquez et al., 2020; Botelho et al., 2020). Word vectors have also been shown to be useful for detecting AD (Hong et al., 2019). I-vectors, x-vectors, and BERT embeddings have been used with the ADReSS dataset to classify AD (Yuan et al., 2020; Pompili et al., 2020) and predict MMSE scores (Balagopalan et al., 2020a). (Pompili et al., 2020) used the same audio features that we used and also used BERT embeddings, but they did not apply their techniques to the MMSE prediction task and their best fusion model obtained lower performance on the classification task than our best model. The difference between our work and the work of (Balagopalan et al., 2020a) and (Yuan et al., 2020) is that they finetuned a pretrained BERT model on the ADReSS data and used that model for classification and regression, whereas we used a pretrained BERT model as a feature extractor and then trained different classifiers and regressors on the extracted BERT embeddings.

CLAN features were used in the baseline paper (Luz et al., 2020) and were combined with BERT embeddings in this work to explore whether performance improved. Lastly, LIWC features have been used to distinguish between AD patients and healthy controls in the past (Shibata et al., 2016) but the dataset was very small (9 AD patients and 9 healthy controls) and, to our knowledge, literature using LIWC for Alzheimer's detection is limited. However, LIWC features have been used to analyze other aspects of mental health (Tausczik and Pennebaker, 2010) and may be useful in the field of AD. For these reasons, we wanted to further explore whether LIWC features could be useful for AD detection and MMSE prediction. Even though our results do not out-perform the best performance on the classification and MMSE prediction tasks, the approaches we employ are different than previous approaches, which provides additional insight into which techniques are best for AD classification and MMSE prediction.

2.2 Materials and Methods

2.2.1 ADReSS Dataset

The ADReSS challenge dataset consists of audio recordings, transcripts, and metadata (age, gender, and MMSE score) for non-AD and AD patients. The dataset is balanced by age, gender, and number of non-AD versus AD patients, with there being 78 patients for each class. The audio recordings are of each patient completing the cookie theft picture description task, where each participant describes what they see in the cookie theft image. This task has been used for decades to diagnose and compare AD and non-AD patients (Mueller et al., 2018b; Giles et al., 1996; Cooper, 1990; Choi, 2009; Mackenzie et al., 2007; Hernández-Domínguez et al., 2018; Mendez and Ashla-Mendez, 1991; Bschor et al., 2001), as well as patients with other forms of cognitive impairment, and was originally designed as part of an aphasia examination (Goodglass and Kaplan, 1983).

Normalized audio chunks were provided for each speaker, in which a voice activity detection (VAD) system was applied to each patient's recording to split it into several chunks. The VAD system used a log energy threshold value to detect the sections of the audio that contained speech by ignoring sounds below a certain threshold. A 65dB log energy threshold value was used, along with a maximum duration of 10 seconds per chunk. Volume normalization involves changing the overall volume of an audio file to reach a

Table 2.1: Age and gender details for **AD** patients in the **training set**, as well as the average MMSE scores, average years of education, and corresponding standard deviations (sd), for the patients in each age interval.

Age Interval	**Male**	**Female**	**MMSE (sd)**	**Educ. (sd)**
[50, 55)	1	0	30.0 (n/a)	12.0 (n/a)
[55, 60)	5	4	16.3 (4.9)	12.4 (1.7)
[60, 65)	3	6	18.3 (6.1)	12.5 (2.1)
[65, 70)	6	10	16.9 (5.8)	12.8 (2.0)
[70, 75)	6	8	15.8 (4.5)	10.4 (2.6)
[75, 80)	3	2	17.2 (5.4)	10.6 (2.7)
Full Set	24	30	17.0 (5.5)	11.9 (2.4)

certain volume level. There was some variation in the recording environment for each audio file, such as microphone placement, which lead to variation in the volume levels for different recordings. The volume of each chunk was normalized relative to its largest value to remove as much variation from the recordings as possible. Each patient had an average of 25 normalized audio chunks, with a standard deviation of 13 chunks. The CHAT coding system (MacWhinney, 2014) was used to create the transcripts.

The ADReSS dataset is a subset of the Pitt corpus (Becker et al., 1994), which is a dataset that contains 208 patients with possible and probable AD, 104 healthy patients, and 85 patients with an unknown diagnosis. The dataset consists of transcripts and recorded responses from the participants for the cookie theft picture description task, a word fluency task, and a story recall task. In order to provide additional in-domain data for training some of the feature extractors, the cookie theft data for patients not included in the ADReSS dataset was separated from the Pitt corpus and used for pretraining. Normalized audio chunks for this data were created using the steps mentioned above. The pretraining process is described in greater detail in Section 2.2.2.

The age and gender distributions, along with the average MMSE scores, average years of education, and corresponding standard deviations, for the training and test sets, can be seen in Tables 2.1, 2.2, 2.3, and 2.4. Education information was not provided with

Table 2.2: Age and gender details for **non-AD** patients in the **training set**, as well as the average MMSE scores, average years of education, and corresponding standard deviations (sd), for the patients in each age interval.

Age Interval	Male	Female	MMSE (sd)	Educ. (sd)
[50, 55)	1	0	29.0 (n/a)	12.0 (n/a)
[55, 60)	5	4	29.0 (1.3)	15.8 (2.8)
[60, 65)	3	6	29.3 (1.3)	13.1 (2.3)
[65, 70)	6	10	29.1 (0.9)	13.8 (3.1)
[70, 75)	6	8	29.1 (0.8)	14.9 (3.4)
[75, 80)	3	2	28.8 (0.4)	14.2 (3.7)
Full Set	24	30	29.1 (1.0)	14.3 (3.1)

Table 2.3: Age and gender details for **AD** patients in the **test set**, as well as the average MMSE scores, average years of education, and corresponding standard deviations (sd), for the patients in each age interval.

Age Interval	Male	Female	MMSE (sd)	Educ. (sd)
[50, 55)	1	0	23.0 (n/a)	20.0 (n/a)
[55, 60)	2	2	18.7 (1.0)	12.5 (1.0)
[60, 65)	1	3	14.7 (3.7)	13.2 (2.2)
[65, 70)	3	4	23.2 (4.0)	11.7 (1.9)
[70, 75)	3	3	17.3 (6.9)	12.8 (3.6)
[75, 80)	1	1	21.5 (6.3)	13.0 (1.4)
Full Set	11	13	19.5 (5.3)	12.8 (2.7)

the ADReSS dataset. However, the Pitt corpus did have education information and was cross-referenced with the ADReSS dataset to determine which patients overlapped and to extract each patient's education information.

A total of 108 patients (54 non-AD and 54 AD) were selected from the full dataset to create the training set, and the remaining 48 patients (24 non-AD and 24 AD) were used for the test set. For both the training and test sets, an equal number of AD and non-AD patients were included for each age group and the number of male and female AD and non-AD

Table 2.4: Age and gender details for **non-AD** patients in the **test set**, as well as the average MMSE scores, average years of education, and corresponding standard deviations (sd), for the patients in each age interval.

Age Interval	**Male**	**Female**	**MMSE (sd)**	**Educ. (sd)**
[50, 55)	1	0	28.0 (n/a)	12.0 (n/a)
[55, 60)	2	2	28.5 (1.2)	13.7 (2.1)
[60, 65)	1	3	28.7 (0.9)	12.2 (0.5)
[65, 70)	3	4	29.4 (0.7)	13.3 (1.4)
[70, 75)	3	3	28.0 (2.4)	13.2 (1.8)
[75, 80)	1	1	30.0 (0.0)	14.0 (2.8)
Full Set	11	13	28.8 (1.5)	13.2 (1.6)

patients was the same for each age group. For the training set, the average MMSE score for the AD patients was 17.0 and the average MMSE score for the non-AD patients was 29.1. The average years of education were 11.9 and 14.3 for the AD and non-AD patients, respectively. For the test set, the AD patients had an average MMSE score of 19.5 and the non-AD patients had an average MMSE score of 28.8. The average years of education were 12.8 and 13.2 for the AD and non-AD patients, respectively.

2.2.2 Feature extraction

Text features: fastText word vectors, BERT embeddings, LIWC and CLAN features

FastText is an open-source library that is used to classify text and learn text representations. A fastText model pretrained on Common Crawl and Wikipedia was used to extract word vectors (Grave et al., 2018) from the transcripts of each speaker. PyLangAcq (Lee et al., 2016), a Python library designed to handle CHAT transcripts, was used to extract the sentences from the CHAT transcript of each participant. A 100-dimensional word vector was computed for each word in each sentence, including punctuation. A dimension of 100 was chosen because this was the value recommended on the fastText website and 100 was compatible with the size of the pretrained model. The longest sentence had a total of 47 words. For this reason, every sentence was padded to a length of 47, resulting in a (47, 100) representation for each utterance.

BERT (Devlin et al., 2018) models are text classification models that have achieved state-of-the-art results on a wide variety of natural language processing tasks and they provide high-level language representations called embeddings. Embeddings are vector representations of words or phrases and are useful for representing language because the embeddings often capture information that is universal across different tasks. Keras BERT was used to load an official, pretrained BERT model and that model was used to extract embeddings of shape (x, 768) for each utterance in the transcript of each speaker, where x depends on the length of the input. After embeddings were extracted for each utterance, the largest embedding had an x value of 60. For this reason, the remaining embeddings were padded to be the same shape, resulting in a (60, 768) embedding for each utterance. For both the word vectors and the BERT embeddings, features were extracted at the utterance level, resulting in a total of 1,492 embeddings in the training set and 590 embeddings in the test set.

Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010) features were also extracted from the transcripts of each speaker. The LIWC program takes in a transcript and outputs a 93-dimensional vector consisting of word counts for different emotional and psychological categories, such as emotional tone, authenticity, and clout, to name a few. The Computerized Language Analysis (CLAN) program (MacWhinney, 2000) was also used to extract linguistic features from the transcripts of each speaker. The EVAL function was used to extract summary data, including duration, percentage of word errors, number of repetitions, etc. This extraction resulted in a 34-dimensional vector for each speaker. The CLAN features were used as linguistic features in the baseline paper (Luz et al., 2020). In this work, the CLAN features were combined with the BERT embeddings to explore whether combining the features improved performance. Both the LIWC and CLAN features were extracted at the subject-level, resulting in 108 vectors in the training set and 54 vectors in the test set.

More background information about each of the features mentioned in this section can be found in A.1 and A.2.

Audio features: i-vectors and x-vectors

VoxCeleb 1 and 2 (Nagrani et al., 2017) are datasets consisting of speech that was extracted from YouTube videos of interviews with celebrities. I-vector and x-vector systems (Snyder et al., 2017, 2018) pretrained on VoxCeleb 1 and 2 were used to extract i-vectors and x-vectors from the challenge data. The i-vector and x-vector systems were built using Kaldi (Povey et al., 2011), which is a toolkit that is used for speech recognition. The pretrained VoxCeleb models were also used to train additional extractors using the original Kaldi recipes. The original VoxCeleb models were used to initialize the i-vector and x-vector extractors and then those extractors were trained on the remaining in-domain Pitt data. I-vector and x-vector extractors were also trained on only the in-domain Pitt data to explore whether a small amount of in-domain data is better for performance than a large amount of out-of-domain data. For each type of extractor, the normalized audio chunks provided with the challenge dataset were first resampled with a sampling rate of 16kHz, a single channel, and 16 bits, to match the configuration of the VoxCeleb data. The Kaldi toolkit was then used to extract the Mel-frequency cepstral coefficients (MFCCs), compute the VAD decision, and extract the i-vectors and x-vectors. The x-vectors had a length of 512, while the i-vectors had a length of 400. There were a total of 2,834 i-vectors and 2,834 x-vectors, one i-vector and x-vector for each normalized audio chunk. More background information about i-vectors and x-vectors can be found in A.1.2.

2.2.3 Experimental approach

Classifiers

Five classifiers were trained on the text and audio features explained in Section 2.2.2: linear discriminant analysis (LDA), the decision tree (DT) classifier, the k-nearest neighbors classifier with the number of neighbors set to 1 (1NN), a support vector machine (SVM) with a linear kernel and regularization parameter set to 0.1, and a random forest (RF) classifier. The classifiers were implemented in Python using the scikit-learn library Pedregosa et al. (2011). The word vectors and BERT embeddings were averaged before being used to train the scikit-learn classifiers, resulting in utterances represented by 100-dimensional vectors and 768-dimensional vectors, respectively. When the LIWC and CLAN features were combined with the averaged BERT embeddings, the subject-level LIWC/CLAN vector was concatenated with each utterance-level BERT embedding belonging to that subject. Standard scaling is commonly applied to data before using machine learning estimators to avoid the poor performance that is sometimes seen when the features are not normally distributed (i.e. Guassian with a mean of 0 and unit variance). Because we were combining different types of features with different data distributions, standard scaling was applied to the features after the LIWC/CLAN vectors were concatenated with the BERT embeddings so that the data would be normally distributed before training and testing. More background information about each of the classifiers mentioned in this section can be found in A.3.

Regressors

Five regression models were also trained on the text and audio features explained in Section 2.2.2 for the MMSE prediction task: linear regression (LR), decision tree (DT) regressor, k-nearest neighbor regressor with the number of neighbors set to 1 (1NN), support vector machine (SVM), and a gradient-boosting regressor (grad-boost). The regression models were implemented in Python using the scikit-learn library. Just as with the classifiers, the word vectors and BERT embeddings were averaged before being used to train the scikit-learn regressors. When the LIWC and CLAN features were combined with the BERT embeddings, the subject-level LIWC/CLAN vector was concatenated with each utterance-level BERT embedding belonging to that subject, and, after the features were concatenated, standard scaling was applied. More background information about each of the regressors mentioned in this section can be found in A.3.

Dimensionality reduction

The classifiers and regressors mentioned above were trained with different dimensionality reduction techniques to see if applying dimensionality reduction improves performance. Feature sets were created with no dimensionality reduction, with LDA, and with principal component analysis (PCA), and each classifier was trained on each feature set to see what effect dimensionality reduction had on performance. The dimensionality reduction techniques were applied to all of the audio and text features. When LDA was applied, the features were reduced to 1 dimension for the classification task and 23 dimensions for the regression task. With PCA, different dimension values were selected manually. The best results and corresponding dimension values can be seen in the Results section. More background information about each of the dimensionality reduction techniques mentioned in this section can be found in A.4.

Neural networks

A bidirectional long short-term memory (LSTM) network and a convolutional neural network (CNN) were also trained on the word vectors to see if the neural networks could extract some temporal information that would lead to better performance compared to the classifiers mentioned in Section 2.2.3. The topologies of the two networks are shown in Figure 2-1. The LSTM model had 1 bidirectional LSTM layer with 8 units, a dropout rate of 0.2, and a recurrent dropout rate of 0.2. The CNN model had the following layers: 3 2D convolution layers with 32, 64, and 128 filters, respectively, rectified linear unit (ReLu) activation and a kernel size of 3, 1 2D max pooling layer with a pool size of 3, 1 dropout layer with a rate of 0.5, and 1 2D global max pooling layer. For both models, the output was passed into a dense layer with sigmoid activation. Both models were implemented in Python using Keras and were trained with an Adam optimizer. The CNN was trained with a learning rate of 0.001, and the LSTM was trained with a learning rate of 0.01.

2.3 Results

2.3.1 Classification

Cross-validation

In order to stay consistent with the baseline paper, each of the classifiers and neural networks were evaluated on the challenge training set using leave-one-subject-out (LOSO) cross-validation, where there was no speaker overlap between the training and test sets for each split. Each model was trained and tested at the utterance level, where each utterance

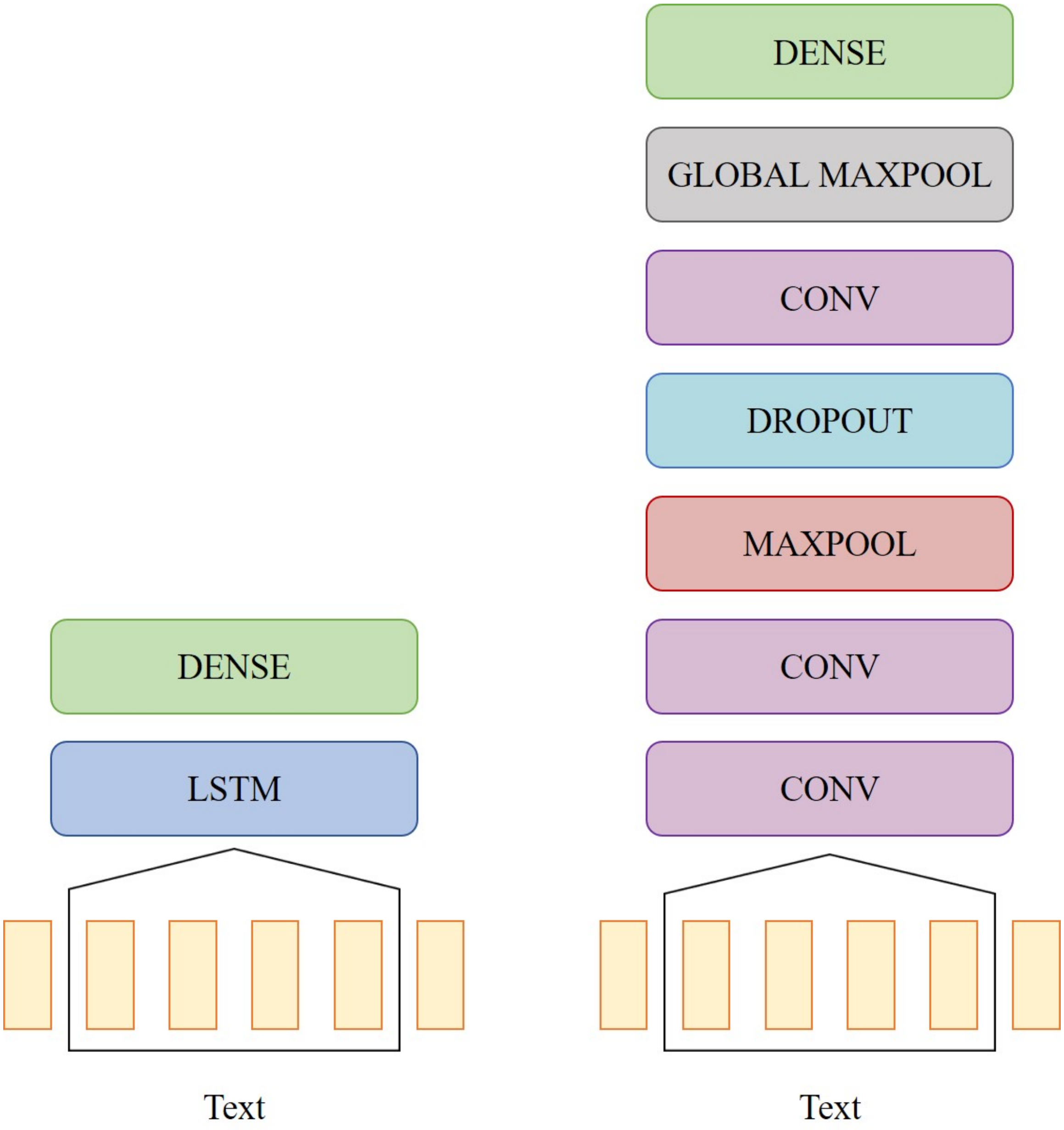


Figure 2-1: Diagrams of the network topology for the LSTM model (left) and the CNN model (right).

was classified as belonging to a patient with or without AD. Then majority vote (MV) classification was used to assign a label to each speaker based on the label that was assigned most to the speaker's utterances.

The MV classification accuracy (the number of correctly classified speakers divided by the total number of speakers), for each feature type can be seen in Table 2.5. The accuracies

Table 2.5: LOSO accuracies for each of the classifiers. The best-performing models for each feature type are red.

Features	Dim. Red. (n_comp)	LDA	DT	1NN	SVM	RF
LIWC	None	0.741	0.593	0.620	**0.833**	0.778
	LDA (1)	0.741	0.750	0.750	0.731	0.750
	PCA (20)	0.778	0.620	0.704	0.787	0.759
BERT	None	0.713	0.676	0.787	**0.796**	0.769
	LDA (1)	0.713	0.657	0.667	0.713	0.657
	PCA (2)	0.630	0.648	0.602	0.546	0.694
	PCA (20)	0.750	0.713	0.722	0.769	**0.796**
BERT + LIWC	None	0.750	0.657	0.667	**0.824**	0.806
	LDA (1)	0.750	0.731	0.731	0.741	0.731
	PCA (20)	**0.824**	0.620	0.657	**0.824**	0.796
BERT + CLAN	None	0.778	0.657	0.759	0.824	0.750
	LDA (1)	0.778	0.769	0.769	0.787	0.769
	PCA (20)	0.824	0.630	0.657	**0.898**	0.778
BERT + LIWC + CLAN	None	0.593	0.731	0.713	0.815	0.806
	LDA (1)	0.593	0.611	0.611	0.593	0.611
	PCA (20)	**0.833**	0.731	0.713	0.815	0.787
word vectors	None	0.759	0.731	0.694	0.259	0.694
	LDA (1)	0.759	0.741	0.731	0.759	0.759
	PCA (2)	0.676	0.620	0.565	0.259	0.620
	PCA (70)	**0.796**	0.648	0.759	**0.796**	0.787
i-vectors (VoxCeleb)	None	0.574	0.423	0.454	0.574	0.500
	LDA (1)	0.574	0.500	0.500	0.574	0.500
	PCA (2)	0.491	0.500	**0.602**	0.519	0.491
	PCA (10)	0.528	0.556	0.546	0.491	0.528
i-vectors (Pitt)	None	0.528	0.491	0.500	0.509	**0.593**
	LDA (1)	0.528	0.537	0.537	0.537	0.537
	PCA (2)	0.463	0.500	0.528	0.343	0.546
	PCA (20)	0.565	0.537	0.528	0.565	0.565
i-vectors (VoxCeleb + Pitt)	None	0.528	0.509	0.500	0.528	0.556
	LDA (1)	0.528	0.519	0.519	0.528	0.519
	PCA (20)	0.519	0.528	0.574	0.472	**0.620**
x-vectors (VoxCeleb)	None	0.583	0.620	0.509	0.546	0.574
	LDA (1)	0.583	0.593	0.593	0.583	0.593
	PCA (2)	0.472	0.537	0.491	0.454	0.491
	PCA (40)	**0.639**	0.583	0.528	**0.639**	0.583
x-vectors (Pitt)	None	**0.546**	**0.546**	0.472	0.528	0.481
	LDA (1)	**0.546**	0.500	0.500	0.537	0.500
	PCA (40)	0.537	0.481	0.435	0.528	0.491
x-vectors (VoxCeleb + Pitt)	None	0.639	0.602	0.519	0.620	0.509
	LDA (1)	0.639	0.509	0.509	0.630	0.509
	PCA (40)	**0.657**	0.574	0.546	0.593	0.593

are presented as decimals and are rounded to 3 decimal places to match the form of the accuracies in the baseline paper. For all of the features, the LDA classifier trained on LDA-reduced features performed the same as the LDA classifier trained on features with no dimensionality reduction. Although there was no difference in performance, results are included for completeness.

The LSTM model trained on word vectors had an average accuracy of **0.787**, while the CNN model had an average accuracy of **0.704**. The highest-performing classifier trained on text features was the SVM classifier trained on a combination of BERT embeddings and CLAN features with PCA dimensionality reduction applied, which had an average accuracy of 0.898. The highest-performing classifier trained on audio features was the LDA classifier trained on x-vectors that were extracted using a system that was pretrained on VoxCeleb and in-domain Pitt data. PCA dimensionality reduction was applied and the classifier had an average accuracy of 0.657.

The highest-performing classifiers for each feature type, except for the classifiers trained on x-vectors that were extracted from a system trained on just Pitt data, performed better than the highest-performing audio and text baseline classifiers that were evaluated using LOSO on the training set, which had an average accuracy of 0.565 and 0.768, respectively (Luz et al., 2020).

Held-out test set

The MV classification accuracies on the test set for each of the classifiers can be seen in Table 2.6. The highest-performing text classifiers were the SVM classifier with no dimensionality reduction and the RF classifier with PCA dimensionality reduction, both trained on BERT embeddings. Both classifiers had an average accuracy of 0.854. The highest-performing audio classifier was the 1NN classifier trained on i-vectors that were extracted using systems pretrained on VoxCeleb with PCA dimensionality reduction applied, which had an average accuracy of 0.563.

The highest-performing text classifiers outperformed the baseline text classifier, which was an LDA classifier trained on CLAN features with an average accuracy of 0.75. The highest-performing audio classifiers did not outperform the baseline audio classifier, which

Table 2.6: Accuracies for classifiers evaluated on the test set. The test set results for the best-performing models during cross-validation are red.

Features	**Dim. Red. (n_comp)**	**LDA**	**DT**	**1NN**	**SVM**	**RF**
LIWC	None	0.583	0.708	0.583	**0.688**	0.812
	LDA (1)	0.583	0.583	0.583	0.583	0.583
	PCA (20)	0.771	0.646	0.583	0.792	0.667
BERT	None	0.604	0.708	0.771	**0.854**	0.750
	LDA (1)	0.604	0.604	0.646	0.604	0.604
	PCA (2)	0.688	0.562	0.542	0.729	0.625
	PCA (20)	0.833	0.646	0.750	0.812	**0.854**
BERT + LIWC	None	0.583	0.667	0.688	**0.729**	0.812
	LDA (1)	0.583	0.583	0.583	0.583	0.583
	PCA (20)	**0.792**	0.708	0.771	**0.771**	0.792
BERT + CLAN	None	0.729	0.750	0.771	0.812	0.812
	LDA (1)	0.729	0.708	0.708	0.708	0.708
	PCA (20)	0.729	0.708	0.667	**0.771**	0.792
BERT + LIWC + CLAN	None	0.625	0.688	0.750	0.750	0.812
	LDA (1)	0.625	0.667	0.667	0.625	0.667
	PCA (20)	**0.812**	0.604	0.729	0.812	0.812
word vectors	None	0.813	0.688	0.667	0.500	0.833
	LDA (1)	0.813	0.750	0.771	0.813	0.750
	PCA (2)	0.729	0.542	0.500	0.500	0.667
	PCA (70)	**0.812**	0.562	0.688	**0.500**	0.771
i-vectors (VoxCeleb)	None	0.542	0.563	0.521	0.625	0.625
	LDA (1)	0.542	0.521	0.521	0.542	0.521
	PCA (2)	0.750	0.625	**0.563**	0.708	0.729
	PCA (10)	0.562	0.542	0.438	0.583	0.562
i-vectors (Pitt)	None	0.417	0.521	0.521	0.438	**0.542**
	LDA (1)	0.417	0.542	0.542	0.417	0.542
	PCA (2)	0.667	0.583	0.708	0.604	0.646
	PCA (20)	0.583	0.542	0.583	0.521	0.479
i-vectors (VoxCeleb + Pitt)	None	0.458	0.521	0.500	0.500	0.563
	LDA (1)	0.458	0.542	0.542	0.458	0.542
	PCA (20)	0.458	0.563	0.604	0.458	**0.479**
x-vectors (VoxCeleb)	None	0.604	0.500	0.500	0.563	0.521
	LDA (1)	0.604	0.604	0.604	0.604	0.604
	PCA (2)	0.625	0.563	0.563	0.625	0.542
	PCA (40)	**0.479**	0.417	0.562	**0.458**	0.479
x-vectors (Pitt)	None	**0.500**	**0.479**	0.417	0.563	0.583
	LDA (1)	**0.500**	0.542	0.542	0.500	0.542
	PCA (40)	0.521	0.563	0.521	0.458	0.542
x-vectors (VoxCeleb + Pitt)	None	0.563	0.604	0.479	0.521	0.583
	LDA (1)	0.563	0.521	0.521	0.563	0.521
	PCA (40)	**0.500**	0.458	0.646	0.479	0.563

was an LDA classifier trained on ComParE openSMILE features with an average accuracy of 0.625.

2.3.2 MMSE prediction

Cross-validation

For the MMSE prediction task, one of the speakers in the training set did not have an MMSE score and was excluded from training. Each of the regressors was evaluated on the challenge training set using LOSO cross-validation, where there was no speaker overlap between the training and test sets for each split. Each model was trained and tested at the utterance level, where an MMSE score was predicted for each utterance. Then the predicted MMSE scores of the utterances belonging to a patient were averaged to assign one MMSE score to that patient. Lastly, the RMSE between the predicted and ground truth MMSE scores was computed.

The average RMSE scores for each feature type can be seen in Table 2.7. For all of the features, the LR regressor trained on LDA-reduced features performed the same as the LR regressor trained on features with no dimensionality reduction. Although there was no difference in performance, results are included for completeness.

The best-performing regressor trained on text features was the LR regressor trained on BERT embeddings combined with LIWC and CLAN features with PCA dimensionality reduction applied, which had an RMSE score of 3.774. The best-performing regressor trained on audio features was the DT regressor trained on x-vectors that were extracted using a system pretrained on Pitt. LDA dimensionality reduction was applied and the RMSE score was 6.073.

The best-performing text regressors for every feature type, except for BERT embeddings and word vectors, performed better than the baseline text regressor that was evaluated using LOSO on the training set, which had an RMSE score of 4.38. The best-performing audio regressors for every feature type performed better than the baseline audio regressor that was evaluated using LOSO on the training set, which had an RMSE score of 7.28.

Table 2.7: LOSO RMSE scores for each of the classifiers. The results for the best-performing models for each feature type are red.

Features	**Dim. Red. (n_comp)**	**LDA**	**DT**	**1NN**	**SVM**	**GradBoost**
LIWC	None	10.067	5.766	5.626	6.083	**4.014**
	LDA (23)	8.928	8.738	5.224	6.195	7.654
	PCA (20)	4.436	5.383	5.364	6.057	4.640
BERT	None	5.111	5.984	**4.953**	6.111	5.407
	LDA (23)	5.111	6.571	5.805	6.275	6.701
	PCA (2)	6.304	5.628	5.851	6.187	6.034
BERT + LIWC	None	9.475	4.956	4.752	5.919	**4.050**
	LDA (23)	8.515	8.038	5.285	6.821	7.234
	PCA (20)	4.574	5.228	5.680	5.165	4.509
BERT + CLAN	None	4.810	6.265	4.728	6.009	4.100
	LDA (23)	4.810	5.700	4.988	6.173	5.447
	PCA (20)	3.991	5.459	4.842	5.254	**3.969**
BERT + LIWC + CLAN	None	13.877	5.533	4.420	5.846	4.190
	LDA (23)	5.243	5.398	5.482	6.477	5.031
	PCA (20)	**3.774**	5.701	5.023	4.966	4.201
word vectors	None	5.294	5.467	5.204	6.146	5.684
	LDA (23)	5.294	5.158	**4.967**	5.936	5.228
	PCA (2)	6.359	6.061	5.958	6.148	6.241
	PCA (70)	5.419	5.561	4.981	6.177	5.516
i-vectors (VoxCeleb)	None	6.323	6.477	6.612	6.444	6.461
	LDA (23)	6.323	6.366	6.384	6.279	6.443
	PCA (2)	6.576	6.431	6.361	6.290	6.421
	PCA (10)	6.412	6.507	6.524	6.265	**6.264**
i-vectors (Pitt)	None	6.545	6.850	6.239	6.281	6.513
	LDA (23)	6.545	6.524	6.307	6.244	6.499
	PCA (2)	6.624	6.606	6.484	6.323	6.598
	PCA (20)	6.523	6.575	6.577	**6.207**	6.511
i-vectors (VoxCeleb + Pitt)	None	6.298	6.363	6.545	6.243	6.445
	LDA (23)	6.298	6.399	**6.110**	6.231	6.459
	PCA (20)	6.502	6.558	6.655	6.256	6.475
x-vectors (VoxCeleb)	None	6.424	6.400	6.208	6.400	6.369
	LDA (23)	6.424	6.478	6.493	**6.162**	6.413
	PCA (2)	6.618	6.767	6.531	6.381	6.634
	PCA (40)	6.246	6.320	6.517	6.329	6.378
x-vectors (Pitt)	None	6.310	6.534	6.445	6.405	6.504
	LDA (23)	6.310	**6.073**	6.403	6.245	6.318
	PCA (40)	6.471	6.456	6.181	6.369	6.474
x-vectors (VoxCeleb + Pitt)	None	6.385	6.268	6.394	6.401	6.386
	LDA (23)	6.385	6.379	6.230	**6.170**	6.442
	PCA (40)	6.296	6.433	6.411	6.288	6.467

Held-out test set

The RMSE scores on the test set for each of the regressors can be seen in Table 2.8. The best-performing text regressor was the grad-boost regressor trained on BERT embeddings combined with CLAN features with PCA dimensionality reduction applied, which had an RMSE score of 4.560. The best-performing audio regressor was the 1NN regressor trained on i-vectors extracted using a system pretrained on VoxCeleb and Pitt with LDA dimensionality reduction applied, which had an RMSE score of 5.694.

The highest-performing text regressor outperformed the baseline text regressor, which was a DT regressor trained on CLAN features with an RMSE score of 5.20. The highest-performing audio regressor outperformed the baseline audio regressor, which was a DT regressor trained on Multi-resolution Cochleagram (MRCG) openSMILE features that had an RMSE score of 6.14.

2.3.3 Effects of education and the severity of cognitive impairment

In order to explore what effect the severity of cognitive impairment and education level had on the classification and MMSE prediction results, the best-performing text and audio models from both tasks were evaluated on smaller subsets of the test set that were split based on education level and MMSE score. An MMSE score of 20-24 corresponds to mild dementia, 13-20 corresponds to moderate dementia, and a score less than 12 is severe dementia (Alzheimer's Association, 2020). This information was used to create 4 groups of cognitive severity: healthy (MMSE score greater than 25), mild dementia (MMSE score of 20-24), moderate dementia (MMSE score of 13-19), and severe dementia (MMSE score less than or equal to 12). The ranges set by the Alzheimer's Association were slightly modified to have unique boundary values.

For education level, the majority of patients had 12 years of education (likely equivalent to completing high school). Because the test set is small, we wanted to limit our experiments to a small number of groups. For the reasons previously mentioned, one education group was for patients that had 12 years of education, another group was for patients with less than 12 years of education, and the last group included patients that had more than 12 years of

Table 2.8: RMSE scores for classifiers evaluated on the test set. The results for the best-performing models during cross-validation are red.

Features	**Dim. Red. (n_comp)**	**LDA**	**DT**	**1NN**	**SVM**	**GradBoost**
LIWC	None	36.974	7.303	6.403	6.465	**4.862**
	LDA (23)	12.286	9.657	7.388	6.313	8.365
	PCA (20)	4.422	5.967	5.990	6.431	4.383
BERT	None	5.365	5.640	**4.923**	6.169	4.883
	LDA (23)	5.365	7.515	6.017	6.253	7.373
	PCA (2)	5.661	5.858	6.287	6.067	5.691
BERT + LIWC	None	34.420	7.127	5.021	6.103	**5.037**
	LDA (23)	14.905	8.624	5.742	7.189	6.561
	PCA (20)	4.872	7.078	5.159	4.895	4.404
BERT + CLAN	None	4.991	7.218	4.515	6.097	4.901
	LDA (23)	4.991	6.523	5.600	6.422	6.660
	PCA (20)	4.764	7.577	6.413	5.218	**4.560**
BERT + LIWC + CLAN	None	15.465	6.112	4.811	6.023	4.724
	LDA (23)	8.110	6.500	5.753	6.887	6.021
	PCA (20)	**4.800**	6.196	5.532	4.794	5.087
word vectors	None	4.714	5.280	5.129	6.147	5.361
	LDA (23)	4.714	5.111	**5.344**	6.063	4.955
	PCA (2)	5.732	6.452	5.992	6.129	5.803
	PCA (70)	4.785	5.700	5.237	6.169	5.271
i-vectors (VoxCeleb)	None	6.600	6.305	6.269	6.161	6.396
	LDA (23)	6.600	7.056	6.360	6.461	6.820
	PCA (2)	6.194	6.514	6.546	5.999	6.237
	PCA (10)	6.335	6.840	6.298	6.110	**6.386**
i-vectors (Pitt)	None	6.530	6.622	6.758	6.142	6.170
	LDA (23)	6.530	6.712	6.133	5.956	6.473
	PCA (2)	6.225	6.827	6.370	6.151	6.342
	PCA (20)	6.257	6.278	6.110	**6.199**	6.252
i-vectors (VoxCeleb + Pitt)	None	6.292	6.042	7.391	6.158	6.145
	LDA (23)	6.292	6.567	**5.694**	5.905	6.407
	PCA (20)	6.316	6.439	6.607	6.168	6.431
x-vectors (VoxCeleb)	None	6.559	6.665	6.401	6.094	6.309
	LDA (23)	6.559	6.289	6.261	**6.085**	6.312
	PCA (2)	6.167	6.669	6.566	6.089	6.164
	PCA (40)	6.358	6.058	6.189	6.115	6.160
x-vectors (Pitt)	None	6.428	6.483	6.563	6.287	6.333
	LDA (23)	6.428	**6.462**	6.314	6.097	6.423
	PCA (40)	6.424	6.506	6.499	6.322	6.370
x-vectors (VoxCeleb + Pitt)	None	6.644	6.622	6.338	6.096	6.208
	LDA (23)	6.644	6.450	6.188	**6.059**	6.466
	PCA (40)	6.173	6.640	6.488	6.123	6.204

Table 2.9: Test set accuracies and RMSE scores for different levels of cognitive deficiency and education. (Feature Modality: Text)

		Classification		**MMSE Prediction**
	Group (num. patients)	**SVM**	**RF**	**GradBoost**
MMSE	Healthy (28)	0.857	0.714	3.234
	Mild Dementia (8)	0.750	0.750	3.777
	Moderate Dementia (8)	0.875	0.625	4.563
	Severe Dementia (4)	1.000	0.500	10.241
Education	<12 years (5)	0.800	0.600	7.448
	12 years (24)	0.792	0.833	4.128
	>12 years (19)	0.947	0.684	3.885

Table 2.10: Test set accuracies and RMSE scores for different levels of cognitive deficiency and education. (Feature Modality: Audio)

		Classification	**MMSE Prediction**
	Group (num. patients)	**1NN**	**1NN**
MMSE	Healthy (28)	0.500	4.679
	Mild Dementia (8)	0.625	1.801
	Moderate Dementia (8)	0.500	6.224
	Severe Dementia (4)	0.750	12.323
Education	<12 years (5)	1.000	9.329
	12 years (24)	0.458	5.080
	>12 years (19)	0.474	5.138

education.

The text and audio models were trained on the full training set and then evaluated on each MMSE and education group separately by only testing on patients in the test set that belonged to a particular group. The classification and MMSE prediction results can be seen in Tables 2.9 and 2.10. For the MMSE groups, the results showed that the best classification accuracy achieved using a text model was 1.000 and that accuracy was achieved when the SVM classifier was evaluated on patients with severe dementia. The best RMSE achieved using a text model was 3.234 and that RMSE was achieved when the GradBoost regressor was evaluated on healthy patients. For the audio models, the best classification accuracy was 0.750 and was achieved when the 1NN classifier was evaluated on patients with severe dementia. The best RMSE was 1.801 and was achieved when the 1NN was evaluated on patients with mild dementia.

For the education groups, the best classification accuracy achieved using a text model was 0.947, when the SVM classifier was evaluated on patients with more than 12 years of education. The best RMSE was 3.885 and was achieved when the GradBoost model was evaluated on patients with greater than 12 years of education. For the audio models, the best classification accuracy was 1.000 and was achieved when the 1NN was evaluated on patients with less than 12 years of education. The best RMSE was 5.080 and was achieved when the 1NN was evaluated on patients with 12 years of education.

2.4 Discussion

The held-out test set results for both tasks show that text classifiers trained on BERT embeddings and text regressors trained on BERT embeddings combined with CLAN features perform better than text classifiers/regressors trained on only CLAN features (baseline text feature set). The results also show that audio classifiers trained on x-vectors and i-vectors, extracted using systems that were pretrained on VoxCeleb and Pitt data, do not perform better than audio classifiers trained on ComParE openSMILE features (baseline audio feature set). However, audio regressors trained on x-vectors and i-vectors do perform better than audio regressors trained on MRCG openSMILE features when (1) the x-vectors are trained on only out-of-domain data or a combination of in-domain data and out-of-domain data and (2) when i-vectors are trained on a combination of in-domain and out-of-domain data.

We also note that we achieved better test set results on the classification task and equal test set results on the MMSE prediction task using a pretrained BERT model as a feature extractor as opposed to using BERT as a classifier and regressor as (Balagopalan et al., 2020a) did. We received classification test set results equal to the BERT results of (Yuan et al., 2020), who also used a BERT model as a classifier and added encoded pauses to their training regime. Our results show that BERT embeddings can be used to achieve the BERT model performance of (Yuan et al., 2020) without using the BERT model itself as a classifier and without using pause information. However, the results of (Yuan et al., 2020) suggest that we could achieve even greater performance if we include pause information in our feature set.

2.4.1 I-vector and x-vector systems

One possible explanation for the poor performance of the i-vectors and x-vectors on the classification task is the domain-mismatch between the VoxCeleb datasets and the ADReSS dataset. While the pretrained model may have learned some general representations of speech from the VoxCeleb datasets, it is possible that the type of representations that the model learned were not helpful for distinguishing between the speech of AD and non-AD patients. The VoxCeleb dataset consists of speech extracted from YouTube videos of celebrities being interviewed. While there is variety in the age, race, and accent of the speakers in the VoxCeleb dataset, which may help improve the ability of a model to distinguish between speakers that differ in these qualities, the nature of the recordings (i.e. background noise, overlapping speech, etc.) varies significantly from the recording environment of the ADReSS data. There is also less variety in the types of speakers present in the ADReSS dataset: they are all within a certain age range and do not seem to have significantly different accents. Therefore, the benefits of the VoxCeleb datasets are not likely to help with the AD classification task and the difference in recording environments likely intensifies the domain-mismatch problem, leading to lower performance. It is possible that i-vectors and x-vectors pretrained on a different dataset with less of a domain-mismatch would perform better.

The i-vectors extracted from a system that was only trained on Pitt data did not improve performance on the classification task compared to the i-vectors extracted from a system that was trained on VoxCeleb but did improve performance on the MMSE prediction task. Conversely, the x-vectors extracted from a system that was only trained on Pitt did improve performance on the classification task but did not improve performance on the MMSE prediction task. The i-vector and x-vector extractors that we pretrained on a combination of VoxCeleb and Pitt data led to an improvement in performance on the MMSE prediction task, compared to the performance for i-vectors and x-vectors extracted from a system trained on VoxCeleb. The x-vector performance also improved on the classification task. This shows that a small amount of in-domain data can improve i-vector and x-vector performance for the MMSE prediction task. When choosing between training i-vector and x-vector extractors on a large amount of out-of-domain data, a small amount of in-domain data, or a combination of both, the results suggest that it is best to train on a combination of both.

2.4.2 Pros and cons of linguistic features

The highest-performing models for both tasks were trained on linguistic features (BERT embeddings). One benefit of using linguistic features is that punctuation can be included. This allows the model to use semantic and syntactical information, such as how often speakers are asking questions ('?' present in the transcript). Also, because the BERT model was pretrained on BooksCorpus and English Wikipedia, the data that the pretrained model saw was likely much more general than the VoxCeleb data and using text data meant that the model did not face the issue of the recording-environment mismatch.

However, there are some disadvantages associated with linguistic features. As discussed in the review paper of (de la Fuente Garcia et al., 2020), transcript-free approaches to AD detection are better for generalizability and for protecting the privacy of the participants. In order to use linguistic features, the speech must be transcribed, meaning that linguistic features are worse for model generalizability and patient privacy. Using linguistic features depends on the use of automatic speech recognition (ASR) methods, which often have a low level of accuracy, or transcription methods, which can be costly and time-consuming.

Some linguistic features are also content- and language-dependent. There are linguistic features that are not content-dependent, such as word frequency measures, but it is difficult to automate the extraction of content-independent linguistic features (de la Fuente Garcia et al., 2020). For these reasons, it is important that future research explore using AD classification techniques that only require acoustic features.

2.4.3 Dimensionality Reduction

For the classification task, none of the highest-performing models had LDA dimensionality reduction applied to the feature sets before training. As previously mentioned, the features were reduced to 1 dimension when LDA was applied. The results suggest that this dimensionality reduction was too extreme for the classification task and did not allow for enough information to be retained in the feature set. Conversely, the majority of the highest-performing classifiers had PCA dimensionality reduction applied to the feature sets before training. This suggests that applying PCA dimensionality reduction to the features before training can be useful for AD classification.

For the MMSE prediction task, the features were reduced to 23 dimensions when LDA was applied. Because the dimension was larger, LDA was more useful for this task. The best-perfoming audio model had LDA dimensionality reduction applied. PCA dimensionality reduction was also applied for some of the best-performing models, including the top-performing text model. This suggests that applying LDA and PCA dimensionality reduction to the features before training can be useful for MMSE prediction.

2.4.4 Group evaluation

The evaluation results for different MMSE and education groups showed that certain MMSE groups can be classified more accurately (healthy, moderate dementia, and severe dementia) while others (mild dementia) are more difficult to classify. This seems very reasonable, as it is expected that more severe forms of dementia would be more easily distinguishable from healthy patients. Also, MMSE scores are predicted least accurately when evaluated on patients with severe dementia, regardless of the type of features used (text or audio).

The education results for the best-performing text-based model showed that patients with more than 12 years of education can be classified with high accuracy (0.947), while patients with exactly 12 years (0.792) and less than 12 years (0.800) of education are more difficult to classify and are classified with similar accuracy. The MMSE scores of patients with greater than 12 years of education were predicted with the most accuracy.

These results provide some insight into which types of features are best for different levels of dementia and education for the classification and MMSE prediction tasks. However, it is important to note that the evaluation set is small, with as little as 4 speakers in certain groups (severe dementia). Therefore, these findings may not translate well to larger datasets.

2.5 Chapter Summary

Audio and text-based representations of speech were extracted from the ADReSS dataset for the AD classification and MMSE prediction tasks. Different dimensionality reduction techniques were applied to the data before training and testing the classification and regression models to explore whether applying dimensionality reduction techniques improved performance on those tasks. LOSO cross-validation was used to evaluate each of the classifiers and regressors and the models were also evaluated on a held-out test set.

The best-performing text models outperform the baseline text models on both tasks and the best-performing audio models outperform the baseline on the MMSE prediction task. The audio results suggest that, given access to a large amount of out-of-domain data and a small amount of in-domain data, it is best to use a combination of both to train i-vector and x-vector extractors. The comparison of the dimensionality reduction techniques shows that applying PCA dimensionality reduction to the features before training a classifier can be helpful for this particular AD classification task and possibly for other similar health-related classification tasks. Also, applying LDA and PCA dimensionality reduction to the features before training a regressor can be helpful for MMSE prediction tasks. Lastly, the evaluation results on different MMSE and education groups show that patients with more severe forms of dementia (moderate and severe) and healthy patients are easier to classify than patients with mild dementia, whereas the MMSE scores of severe dementia patients are the most difficult to predict. Patients with more than 12 years of education are the easiest to classify and the MMSE scores of patients with greater than 12 years of education are the easiest to predict.

In the next chapter, we present CLAC, a dataset consisting of healthy speakers that can be used to augment datasets with speech from non-healthy subjects, like the datasets used in this chapter and subsequent chapters.

Chapter 3

The Crowdsourced Language Assessment Corpus

This chapter introduces the Crowdsourced Language Assessment Corpus (CLAC), a speech corpus consisting of audio recordings and automatically-generated transcripts for several speech and language tasks, as well as metadata for each of the speakers. CLAC was created to provide the community with a collection of audio samples from various speakers that could be used to learn a general representation for speech from healthy subjects, as well as complement other health-related speech datasets, which tend to be limited. In this chapter, we describe the data collection protocol and summarize the contents of the dataset. We also extract timing metrics from the recordings of each task to explore what those metrics look like for a large, English-speaking population. Lastly, we provide an example of how the dataset can be used by comparing the metrics to those extracted from a small sample of Frontotemporal Dementia (FTD) subjects. We hope that this dataset will help advance the state of the art in the health and speech domain.[1]

[1] The work in this chapter was previously published in (Haulcy and Glass, 2021c). The dataset is publicly available at https://groups.csail.mit.edu/sls/downloads/clac/.

3.1 Motivation

Speech has been shown to be a useful modality for diagnosing subjects with various forms of cognitive impairment, including Parkinson's disease (Moro-Velazquez et al., 2021, 2020; Botelho et al., 2020), Alzheimer's disease (Haulcy and Glass, 2021a; de la Fuente Garcia et al., 2020; López et al., 2019; Pompili et al., 2020; Balagopalan et al., 2020a), FTD (Vogel et al., 2017; Poole et al., 2017; Zimmerer et al., 2020), Huntington's disease (Grimstvedt et al., 2021; Vogel et al., 2012; Skodda et al., 2014; Hinzen et al., 2018; Chan et al., 2019), and more. For this reason, datasets consisting of speech from healthy subjects and subjects diagnosed with various neurocognitive disorders have been collected and used to distinguish healthy subjects from cognitively impaired subjects. However, these datasets tend to be limited in size and often are not publicly available (Cummins et al., 2018; Novikova and Balagopalan). More speech data is needed for subjects with cognitive impairment for researchers to be able to generalize their findings.

While speech from impaired subjects is needed, speech from healthy subjects is also necessary and can be useful for learning what the speech profile of a healthy population looks like. In a recent review paper, Voleti et al. (Voleti et al., 2019) acknowledged that the characterization of the variability of the speech in healthy populations is a critical research area for advancing the state of the art. We hope to contribute to this research area by providing a dataset of healthy speakers, primarily from the United States, that can be used to gain a more complete understanding of what variability looks like in healthy populations. In this paper, we present a speech dataset consisting of audio recordings and automatically-generated transcripts from speakers that were presumed healthy. They completed several simple language tasks that are present in other health-related speech datasets (Vogel et al., 2017; Henry and Grasso, 2018; Nassif et al., 2019), including common picture description tasks like the cookie theft task (Becker et al., 1994; Kokkinakis et al., 2018; Mueller et al., 2018a), which has been used to classify numerous cognitive disorders (Mueller et al., 2018a; Giles et al., 1996; Cooper, 1990; Choi, 2009; Mackenzie et al., 2007; Hernández-Domínguez et al., 2018; Mendez and Ashla-Mendez, 1991; Bschor et al., 2001). In addition to exploring what speech looks like in a healthy population, the dataset presented in this paper can be used to compare the speech of healthy, English-speaking populations in different countries, and/or supplement the data in other health-related datasets.

As far as we know, this is the largest collection of healthy English speakers completing language comprehension tasks and we believe that the scientific community can benefit from the public release of this dataset.

3.2 Data collection

The audio recordings in the dataset were collected through Amazon Mechanical Turk (AMT), a crowdsourcing website that allows workers to complete tasks created by businesses and researchers (Requesters) for a set cost. The tasks that the workers complete are called Human Intelligence Tasks (HITs). In order to qualify to complete the HIT we created, workers were required to be in the United States (a small subset of the workers were located in different countries) and the percentage of assignments that were submitted by the workers and approved by previous Requesters had to be 90% or higher. Each worker had a unique worker ID. The worker IDs for approved submissions were used to ensure that each worker was only allowed to complete the HIT one time.

3.2.1 Task selection

The HIT used to collect the data described in this paper consisted of several tasks. Screenshots of the HIT can be seen in Appendix B. Each worker was first asked to select their gender (“Male”, “Female”, or “Other”) and age (a number between 18 and 90, or “Over 90”). Some workers were also asked to select the number of years of education they completed, with 12 years being equivalent to completing high school. There is no education information for 250 workers because the education question was added after data collection began. Each worker was also asked to tell us whether they had a cold, allergy, or other health-related symptoms that might affect their speech the day they completed the HIT (“Yes” or “No”). After that, the workers were asked to complete several simple tasks, all of which can be seen in Table 3.1, along with the corresponding prompts and the number of audio files in the dataset for each task. These tasks were selected because they have been used to assess and diagnose subjects with impaired speech in previous research (Vogel et al., 2017) and they could be easily implemented and completed by the workers without a proctor present.

Table 3.1: The tasks workers were asked to complete, the corresponding prompts, and the number of audio files in the dataset for each task.

Task	Prompt	Audio Files
Counting From 1 To 20	Record yourself counting from 1 to 20.	1,816
Days Of The Week	Record yourself saying the days of the week, starting with Monday.	1,829
Cookie Theft	Record yourself describing everything that you see in the picture below using complete sentences.	1,832
Picnic	Record yourself describing everything that you see in the picture below using complete sentences.	808
Grandfather	Record yourself reading the following passage: “You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an old black frock coat, usually several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a bit. Twice each day he plays skillfully and with zest upon a small organ. Except in the winter when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, ‘Banana oil!’ Grandfather likes to be modern in his language.”	1,832
Rainbow	Record yourself reading the following passage: “The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon.”	1,832
Repeat 5 Times	Record yourself repeating the following words 5 times each in the same recording: artillery, catastrophe, impossibility.	250
Repeat 5 Times Artillery	Record yourself repeating the word “artillery” 5 times.	1,582
Repeat 5 Times Catastrophe	Record yourself repeating the word “catastrophe” 5 times.	1,582
Repeat 5 Times Impossibility	Record yourself repeating the word “impossibility” 5 times.	1,582
SMR	Record yourself repeating /pataka/ (pah tah kah) as fast as you can for 10 seconds.	1,832
Max Phonation	Please take a deep breath and then record yourself sustaining voicing of the vowel /a/ (ah) at a comfortable pitch and loudness level for as long as you can.	1,832

The cookie theft and picnic pictures used for the picture description tasks can be seen in Figures 3-1 and 3-2. For the “repeat 5 times” task, some workers were initially asked to record themselves saying each of the 3 words 5 times in one recording. Subsequent workers were asked to submit separate recordings for each word. As a result, there are 250 workers with one repetition recording and 1,582 workers with 3 separate repetition recordings. The picnic picture description task was also added after data collection began. As a result, 1,024 workers did not complete the picnic picture description task.

3.2.2 Validation

Transcripts were automatically generated for each of the submitted audio files using the Google Speech Recognition API (Zhang, 2017). The transcripts were then used to validate the submitted audio files by checking the number of words in the transcript and the length of the audio file. If the number of words and length of the audio file were satisfactory (different threshold values were used for different tasks), then the worker was allowed to move on to the next task. Otherwise, they were asked to complete the task again. These validation checks were added to ensure that workers did not submit incomplete assignments.

3.2.3 Summary statistics

916 speakers selected “Female” for their gender, 903 speakers selected “Male”, and 13 speakers selected “Other”. The average age of the workers was 35.7 years and the average years of education was 15.4. Histograms showing the age and education distributions can be seen in Figures 3-3 and 3-4. Workers were located in 962 unique cities, all 50 US states, and 12 unique countries. The majority of the workers (1,815) were located in the United States.

3.3 Data analysis

An audio activity detection tool called auditok (aud, 2020) was applied to each AMT recording to determine the start and end times of the speech. The tool used a log energy threshold value to detect the sections of audio that contained speech by ignoring sounds

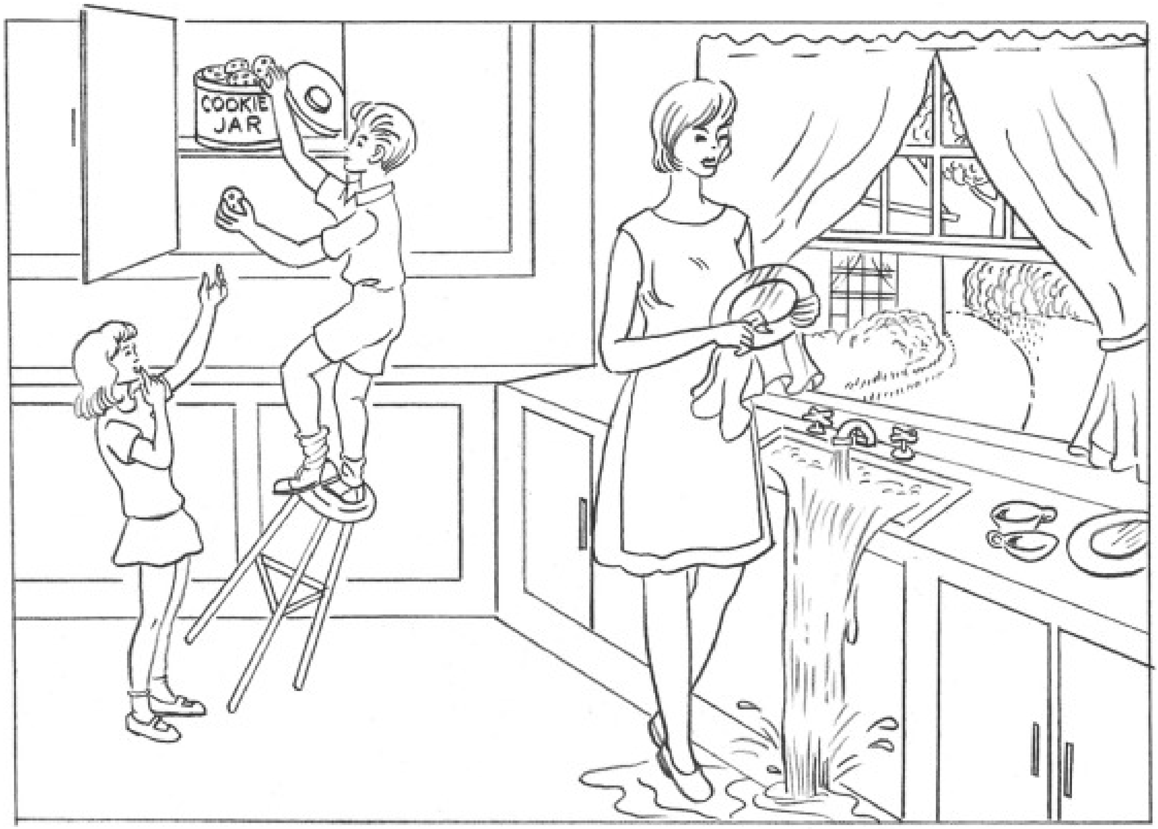


Figure 3-1: The image the workers were asked to describe for the cookie theft picture description task.

below a certain threshold. A 65dB log energy threshold value was used. The detected start and end times were used to extract several timing metrics from the recordings for each task: the total duration of each audio file (in seconds), the number of speech segments, the speech rate (speech segments per second), the number of pauses, the total duration of pauses (in seconds), and the proportion of pause time (total duration of pauses divided by the total duration of the audio file). Those metrics were used to explore what timing looks like for a general population that is presumed healthy. The average value for each of the metrics mentioned above can be seen in Table 3.2 for each of the tasks completed by the workers.

3.3.1 FTD comparison

One way that we anticipate the AMT data being used is to compare the speech of cognitively impaired individuals with that of healthy speakers. The data can also be used to explore

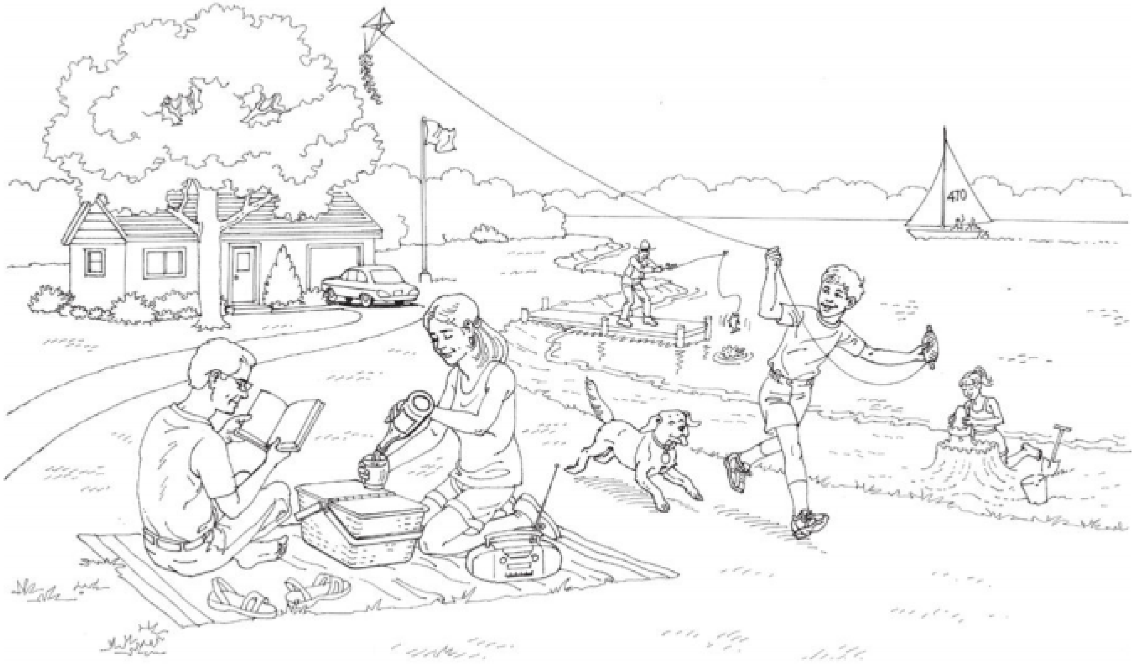


Figure 3-2: The image workers were asked to describe for the picnic picture description task.

how the speech of healthy speakers differs in different regions/countries. To illustrate this, we also extracted the timing metrics mentioned above from the cookie theft audio files of 58 healthy Australian subjects, and Australian subjects with different types of FTD. The FTD data used is a subset of a larger FTD dataset, part of which has been used in previous research to explore which speech characteristics are most salient for the detection of the behavioral variant of FTD (bvFTD) (Vogel et al., 2017). Timing metrics were extracted from 11 subjects with bvFTD, 6 subjects with the semantic variant of Primary Progressive Aphasia (svPPA), and 7 subjects with the logopenic variant of PPA (lvPPA).

The averaged metrics for each of the FTD variants and the healthy subjects can be seen in Table 3.3. The results show that each of the timing metrics are lower for the healthy Australian speakers compared to the healthy speakers in the CLAC dataset, which consists primarily of American speakers. However, due to the large difference in sample size for the two groups, we can not draw any strong conclusions from this observation. The results also show that the timing metrics are the same or higher for each of the FTD groups compared to the healthy groups. Previous research has shown that the speech of lvPPA and bvFTD

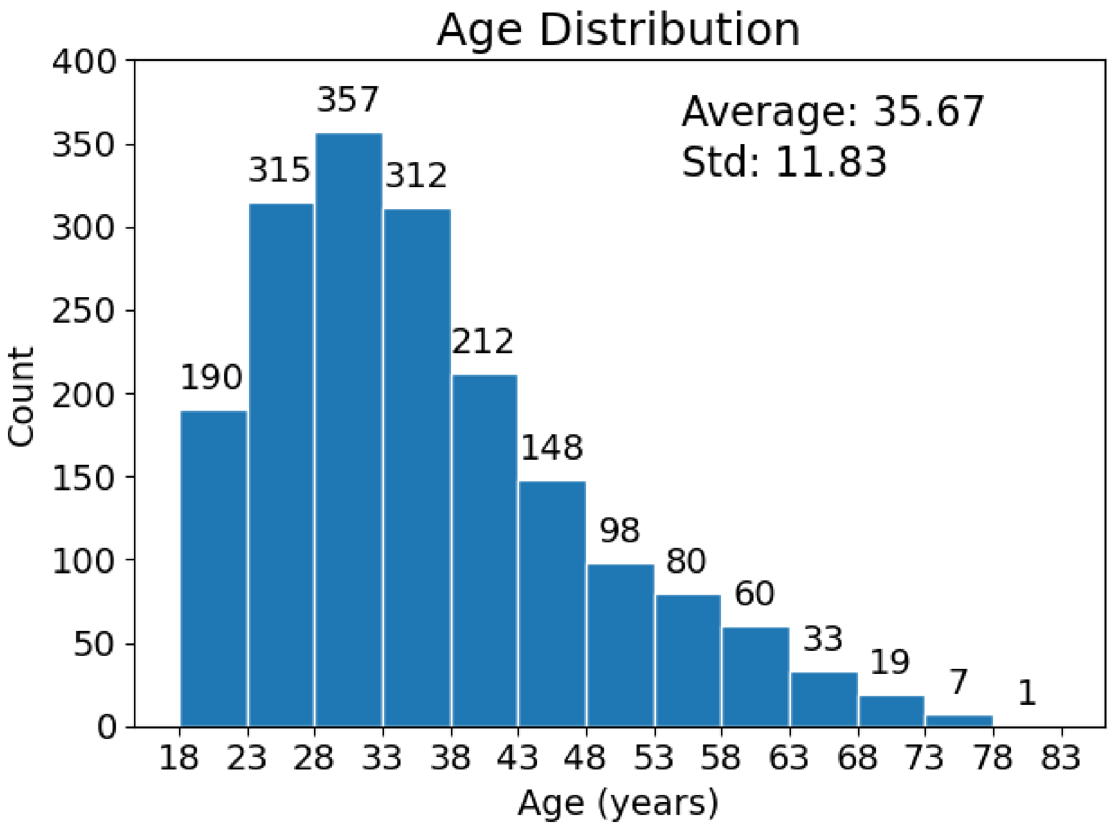


Figure 3-3: The age distribution of the data. (Minimum: 18 years, Maximum: 80 years)

subjects is characterized by a greater proportion of pause time and an increased number of pauses, the speech of lvPPA subjects is also characterized by an increase in pause duration, and the speech of svPPA subjects is characterized by a decreased speech rate (Poole et al., 2017). All of the results in Table 3.3 are consistent with the findings in previous research, except for the increase in speech rate for svPPA subjects compared to the healthy subjects. This discrepancy may be due to the limited sample size of the svPPA subjects.

The comparison of healthy speakers with FTD subjects is just one example of how CLAC can be used. Similar experiments can easily be conducted with a different kind of dataset consisting of healthy speech, impaired speech, or both.

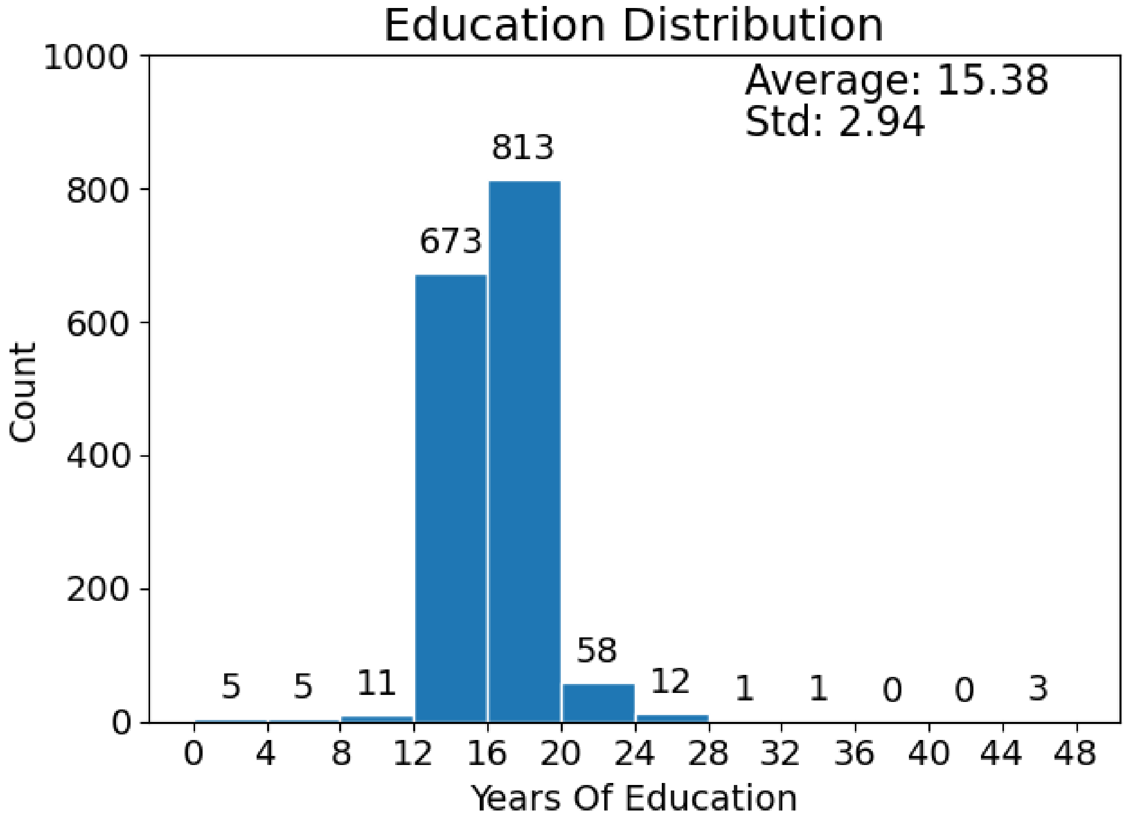


Figure 3-4: The education distribution of the data. (Minimum: 0 years, Maximum: 48 years

3.4 Limitations

While we hope that the dataset can aide researchers in understanding what the speech of the general population looks like, we acknowledge that there are some limitations associated with the dataset:

- Self-reported metadata: Each worker was allowed to report their age, gender, and years of education. The information submitted by the workers could not be verified. Therefore, some of the information may be incorrect.

- Recording environment: Since the workers were allowed to complete the tasks from wherever they were, there was a lot of variety in the type of microphones that were used and the environments that the workers were in. While the difference in recording quality may make analysis more challenging, the variety will also lead to greater generalizability.

Table 3.2: The average values for the timing metrics extracted for each task.

Task	**Speech Duration (secs)**	**Num. Speech Segments**	**Speech Rate (segments/ sec)**	**Num. Pauses**	**Pause Duration (secs)**	**Proportion Pause Duration**
Cookie Theft	30.50	13.42	0.45	12.43	6.47	0.21
Picnic	43.04	19.21	0.45	18.22	10.18	0.24
Counting 1 To 20	19.37	17.21	0.92	16.23	6.87	0.33
Days Of The Week	7.30	6.01	0.85	5.04	2.04	0.26
Grandfather	48.45	21.00	0.44	20.00	8.56	0.17
Rainbow	12.49	5.41	0.45	4.42	1.61	0.13
Repeat 5 Times	20.35	12.59	0.63	11.61	6.02	0.28
Repeat 5 Times Artillery	6.38	4.57	0.74	3.59	1.86	0.26
Repeat 5 Times Catastrophe	6.65	4.62	0.73	3.65	2.00	0.28
Repeat 5 Times Impossibility	7.45	4.59	0.64	3.61	1.81	0.22
Smr	9.94	6.07	0.64	5.10	1.19	0.12
Max Phonation	10.79	4.22	0.48	3.27	2.95	0.22

- Health assumption: We made the assumption that all workers were healthy and did not ask them about their previous medical history. For this reason, we cannot know for sure that each speaker is healthy and it is possible that some speakers in the dataset may have conditions that can impair their speech.

- Different accents: The majority of the workers were located in the United States. However, there is still a variety of different accents and dialects due to differences in the locations and backgrounds of the workers. While this makes the dataset less "clean", it can also be good for generalizability.

- Duplicate worker submissions: Each worker has a unique worker ID and that information was used to ensure that a worker with a particular worker ID was not allowed to complete our HIT more than once. However, there was no way to check whether someone had multiple AMT accounts. Therefore, we can not rule out the possibility that the same speaker completed the HIT multiple times from different AMT accounts.

Table 3.3: The average values for the timing metrics extracted for each FTD variant and healthy category on the cookie theft task.

Group	**Speech Duration (secs)**	**Num. Speech Segments**	**Speech Rate (segments/ sec)**	**Num. Pauses**	**Pause Duration (secs)**	**Proportion Pause Duration**
CLAC (n = 1,832)	30.50	13.42	0.45	12.43	6.47	0.21
Healthy (n = 58)	26.97	11.29	0.41	10.29	4.79	0.16
bvFTD (n = 11)	60.88	25.63	0.45	24.63	26.18	0.41
svPPA (n = 6)	55.82	22.56	0.60	21.56	34.06	0.51
lvPPA (n = 7)	116.35	51.38	0.433	50.38	67.04	0.58

- Transcript quality: The quality of the automatically-generated transcripts varies significantly depending on accent and recording quality. Therefore, some transcripts have high accuracy while others may have incorrect words or may be missing some words completely. Future releases of the dataset will include a corrected version of the ASR transcripts.

- Age range: The majority of the participants are not within the age range of subjects that are typically diagnosed with cognitive disorders. However, it can still be useful to have speech from younger speakers that can be used to possibly examine how speech differs between speakers of different ages in our general population.

While there are some limitations, there are also some benefits, including the fact that (1) diarization is not needed for the recordings in this dataset because only the voice of the worker is present in the recording and (2) we are not aware of any other datasets of this magnitude with speech from healthy subjects completing cognitive tasks that can complement other health-related speech data, making this dataset a significant contribution to the field.

3.5 Chapter Summary

In this chapter, we presented CLAC, a speech dataset consisting of audio recordings and automatically-generated transcripts from 1,832 speakers located in the United States, as well as 11 other countries. We demonstrated how the dataset can be used to characterize the speech of a healthy, English-speaking population and distinguish between healthy subjects and subjects with some form of cognitive impairment. We discussed the limitations of the dataset and believe that the dataset is a valuable contribution to the scientific community, despite those limitations.

In the next chapter, we use CLAC to augment the data for patients with lvPPA so that we can perform binary classification for lvPPA subjects and healthy controls.

Chapter 4

Repetition Assessment for Speech and Language Disorders: A Study of the Logopenic Variant of Primary Progressive Aphasia

Impaired repetition is a characteristic of several speech and language disorders, including certain variants of Primary Progressive Aphasia (PPA). People with the logopenic variant of PPA (lvPPA) can present with impaired repetition abilities and repetition tasks can be used to distinguish lvPPA speakers from healthy controls. In this chapter, we propose a novel technique for quantifying the quality of repetition in speech recordings and demonstrate the utility of the technique by using it to distinguish between healthy speakers and lvPPA speakers. We train several classifiers on features extracted from the repetition recordings. The best classifier distinguishes the lvPPA speakers with impaired repetition from the healthy speakers with 85.7% accuracy and classifies all healthy speakers with perfect accuracy. Although we evaluate the method on lvPPA detection, we believe that the method has potential utility for a range of tasks and speech disorders where repetition occurs.[1]

[1] The work in this chapter was previously published in (Haulcy et al., 2022).

4.1 Motivation

It is no secret that speech can be used to detect cognitive impairment in patients with dementia and other neurocognitive disorders (Moro-Velazquez et al., 2021; Botelho et al., 2020; Haulcy and Glass, 2021a; de la Fuente Garcia et al., 2020; Vogel et al., 2017; Grimstvedt et al., 2021; Vogel et al., 2012). Spoken repetition tasks have been used for decades to detect cognitive and language impairment in children and adults. Non-word repetition tasks have been used to detect language impairment in children (Dollaghan and Campbell, 1998; Estes et al., 2007; Gutiérrez-Clellen and Simon-Cereijido, 2010; Weismer et al., 2000) and word/sentence repetition tasks have been used to distinguish healthy controls from patients with various forms of PPA and Alzheimer's Disease (Leyton et al., 2014; Lukic et al., 2019; Bonner et al., 2010; Gorno-Tempini et al., 2011; Mesulam et al., 2012). In previous research, repetition scores have been manually assigned to each speaker (e.g. by computing how many syllables a speaker said correctly) by clinicians and graduate students that were trained to assess the performance of the patients on certain tasks. This process can be biased by the scorer's expectations (Clark et al., 2021) and can be tedious to complete. For this reason, an automatic way of quantifying repetition quality could be beneficial to the research and clinical communities.

Previous research has explored ways of detecting repetition in audio. Early work on unsupervised pattern discovery in speech were able to find repeating word-like sequences in speech signals without prior knowledge of the words or the language being spoken (Park and Glass, 2007; Jansen et al., 2010). These techniques involved using spectral distance matrices, segmental dynamic-time-warping (SDTW) and graph-based clustering methods to identify reoccurring sequences. We were motivated by this work to develop a repetition detection method that could be used to quantify the quality of repetition in a speech recording.

In this chapter, we describe our repetition detection method and evaluate its ability to distinguish lvPPA subjects from healthy subjects. The speech of lvPPA subjects is characterized by a decline in word retrieval, poor repetition, and phonemic paraphasias (Poole et al., 2017). Speakers with poor repetition abilities are unable to repeat other people's speech correctly, often producing phonological speech errors. For this reason, recordings of lvPPA subjects completing repetition tasks are particularly illuminating when compared to the recordings of healthy speakers completing the same tasks. Our work differs from previous work by providing an approach for computing a metric for repetition quality without the need for manual evaluation and without needing to know how many syllables/words are present beforehand. In the following sections, we explain and demonstrate the feasibility of our approach by training classification models using the extracted metrics as inputs.

4.2 Datasets

The first dataset used for analysis and classification was the Crowdsourced Language Assessment Corpus (CLAC) (Haulcy and Glass, 2021c), as described in Chapter 3. The repetition recordings for a subset of the speakers in the CLAC corpus were used as healthy controls during analysis and classification.

The second dataset consisted of speech from subjects with various forms of Frontotemporal Dementia (FTD) and PPA completing speech tasks similar to those found in the CLAC corpus, including the repetition tasks. The FTD dataset has been used in previous research to explore which speech characteristics are most salient for the detection of the behavioral variant of FTD (bvFTD) (Vogel et al., 2017). The repetition recordings associated with the lvPPA subjects were used for analysis and classification.

4.3 Signal Processing For Repetition Detection

Each repetition waveform was processed in several steps. First, silence was detected and removed from the beginning and end of each recording. This was accomplished with Pydub (Robert et al., 2018) with a minimum silence length of 250 ms, and a silence threshold set to the volume of the recording in dB relative to full scale (dBFS) minus 16.

The next step was to extract 13 Mel Frequency Cepstral Coefficients (MFCCs) from each waveform, using a window length of 25 ms and analysis rate of 10 ms. The MFCCs were used as the basis for computing the distance matrices described in the next section. More background information about MFCCs can be found in A.2.1.

4.3.1 Self-Distance Matrices

In the general SDTW framework, distance matrices correspond to frame-level distances between two speech waveforms, with each element being a distance between two individual frames. In the repetition task however, each waveform contains multiple occurrences of the same word or syllable, so a (square) self-distance matrix is an appropriate representation to capture the self-similarity between successive repetitions. For this work, we use the Euclidean distance metric to compute frame-level MFCC distances. In Figure 4-1, an example of a self-distance matrix for a recording with unimpaired repetition (top) versus the matrix for a recording with impaired repetition (bottom) can be seen. In both cases, the matrix diagonal is zero (blue), since we are comparing a recording to itself. However, in the unimpaired repetition matrix, we can also see diagonal-like stripes at regular intervals. These "stripes" correspond to low-distance alignments of successive repetitions. The first off-diagonal corresponds to matching successive repetitions in the waveform (e.g., 1v2, 2v3, 3v4, 4v5), the second off-diagonal corresponds to matching repetitions spaced two repetitions apart (1v3, 2v4, 3v5), and so on. In this way, the off-diagonal structure neatly summarizes the distances between each pair of repetitions.

In contrast to the unimpaired repetition, a poorly spoken repetition recording will not exhibit the same degree of off-diagonal structure. Aside from the main diagonal, we would expect to see more random distribution of distances, corresponding to Euclidean distances between random speech frames. In the impaired repetition example, we see some degree of off-diagonal structure, but it is clearly weaker and more sporadic than the unimpaired repetition example.

The self-distance matrix is a useful representation of the self-similarity between repeating speech patterns. It has the advantage of being agnostic to the chosen word, syllable or even language being spoken, and requires no task-specific training.

4.3.2 Normalized Diagonal Sum Profile

In order to determine the optimal alignment between successive repetitions, an algorithm such as SDTW should be used (Park and Glass, 2007). In doing so, we can establish an

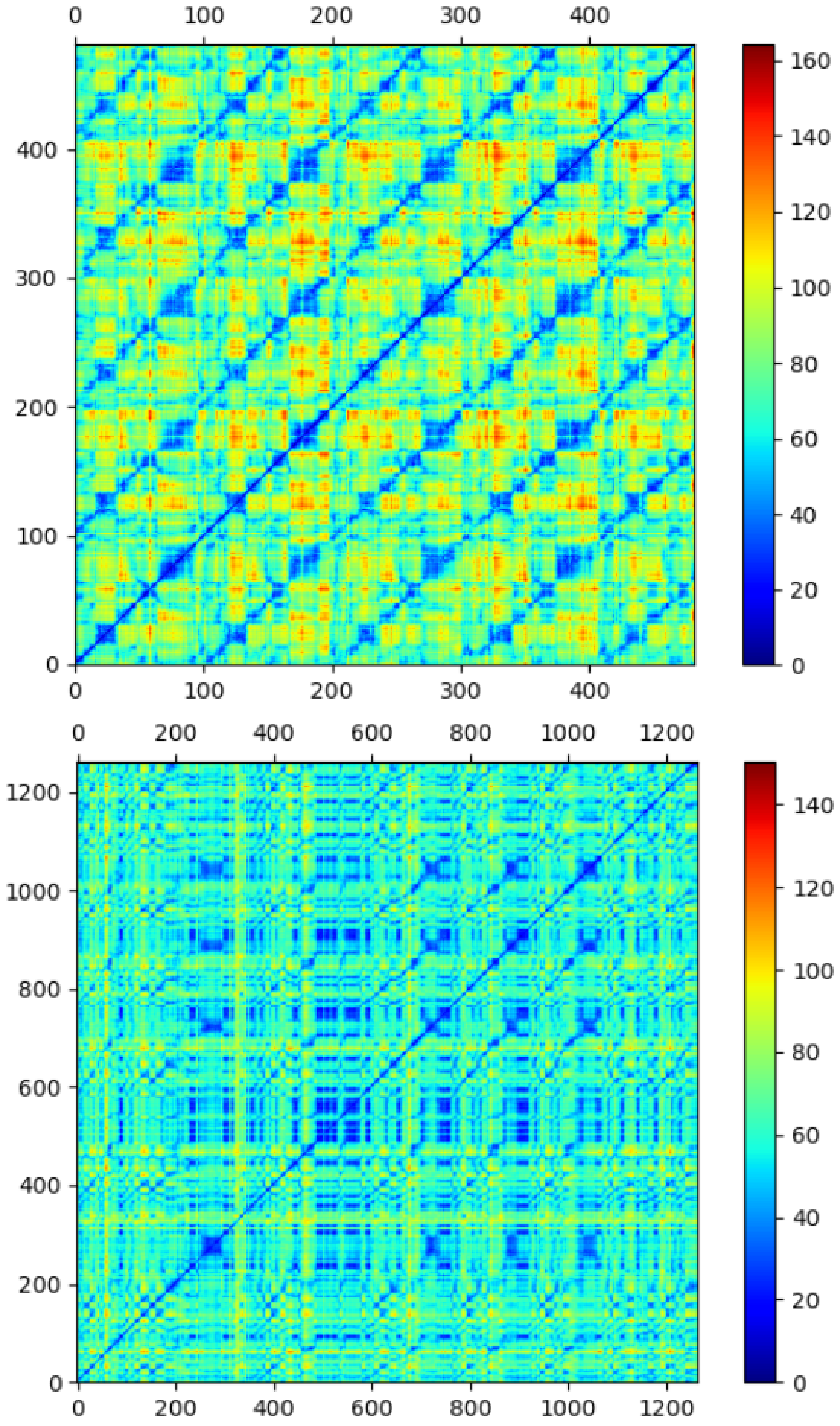


Figure 4-1: Self-distance matrices for an unimpaired (top) and impaired (bottom) repetition of five instances of the word "catastrophe".

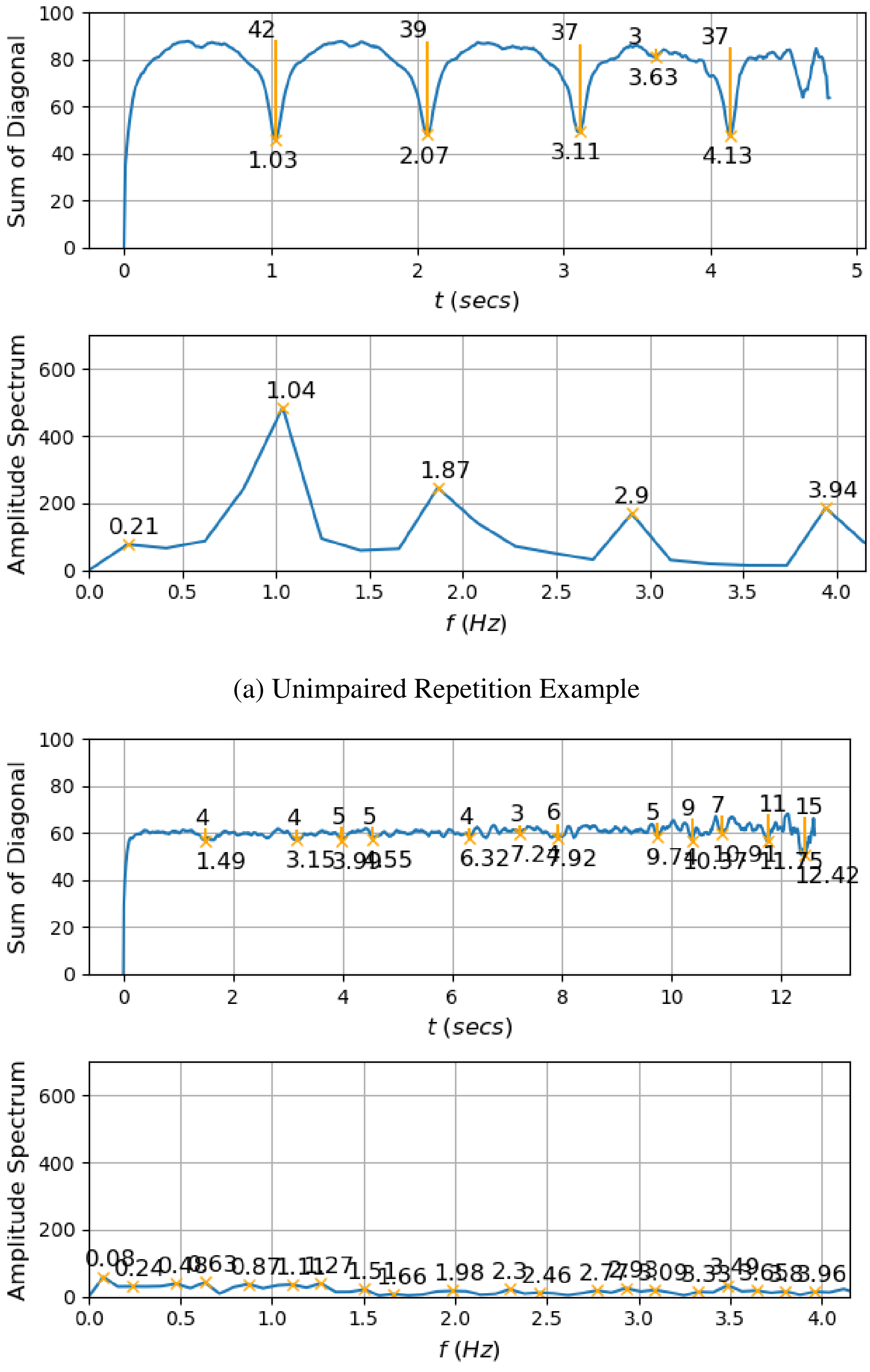
optimal warping path and an associated alignment cost. In our initial work however, we chose to approximate the warp by assuming an unimpaired repetition alignment would be nearly diagonal, and that the alignment cost could be reasonably represented by the normalized sum of distances along the diagonal.

As a way of characterizing the overall self-distance matrix, we transformed it into a one-dimensional profile, with each element corresponding to a normalized sum of distances along a particular diagonal. An example of a diagonal sum profile is shown in Figure 4-2, which shows the profile for the unimpaired repetition and impaired repetition examples from Figure 4-1. The profile is plotted as a function of time, which represents the offset in seconds between two alignments (1s corresponds to 100 frames) in the self-distance matrix. From the sum profile, we can easily identify good alignments, as they correspond to local minima. For an unimpaired repetition (Figure 4-2a), the minima occur at regular intervals and have consistent magnitude and drop from the average profile value. For an impaired repetition (Figure 4-2b), the minima are more sporadic and are rather insignificant compared to the average profile value.

Sum Profile Features

Although the self-distance matrix or the sum profile could have been used directly as a feature representation for the repetition classification model, we wanted to extract a more compact set of features due to the lack of data from lvPPA speakers. Since information about the local minima in the sum profile seemed important, we extracted information about the offset time and magnitude of the local minima. To accomplish this, the find_peaks function in Scipy's signal processing library (Virtanen et al., 2020) was used to find the minima in each sum profile. An example of the detected minima can be seen in Figure 4-2. The detected time offsets are shown beneath each minima and the associated range (the magnitude drop of the local minima) are shown above the lines representing the magnitude.

For each recording, the first three local minima were found and the minima with the largest range was used to represent the recording, along with the corresponding offset time. Each speaker had at least two repetition recordings. After each recording had an associated minima range and time, the minima with the smallest range was selected as the representation for the speaker. In other words, each speaker was represented by their worst repetition, since we expected that healthy speakers would perform well in all their



(a) Unimpaired Repetition Example

(b) Impaired Repetition Example

Figure 4-2: The diagonal sum profile and associated spectral amplitudes for the unimpaired (a) and impaired (b) repetitions of Figure 4-1.

recordings whereas struggling speakers might not.

Fourier Analysis Features

As an alternative to the direct feature extraction of local minima information on the sum profile, we also examined the use of a Fourier analysis. We hypothesized that applying a discrete Fourier Transform (DFT) to the sum profile could help us identify the regularity of the local minima structure in the sum profile.

Example DFTs can be seen in Figures 4-2a and 4-2b (bottom plots). In Figure 4-2a, there are clear DFT peaks at several frequencies. Scipy's find_peaks function was also used to find the peaks in the DFT. As can be seen in the Figure, the first harmonic occurs at 1.04 Hz, which, given the DFT resolution, approximately corresponds to the regular minima in the sum profile.

The DFT in Figure 4-2a also shows that significant harmonics have higher amplitude values (in contrast to Figure 4-2b). This suggests that the frequency and amplitude values associated with the harmonics in the DFT may provide a useful representation for analyzing repetitions. For this reason, we extracted the frequency and amplitude for the largest harmonic in the DFT and used that to represent the repetition quality in each recording. As before, we represented each speaker by the smallest amplitude from all of their recordings (representing the speaker's worst performance) and used the frequency and amplitude of that peak to represent the speaker.

More background information about DFTs can be found in A.5.

4.4 Analysis

The processing steps described in Section 4.3 were applied to the audio recordings of healthy and lvPPA speakers and used to extract four features that were described previously (sum profile local minima time and range, DFT first harmonic frequency and amplitude). Since the CLAC corpus contains a wide demographic, a subset of speakers older than 44 was selected as the healthy subset in order to match the lvPPA speaker age range. The data thus consisted of recordings from 354 healthy speakers (aged 53.3 ± 8 years; 199 female

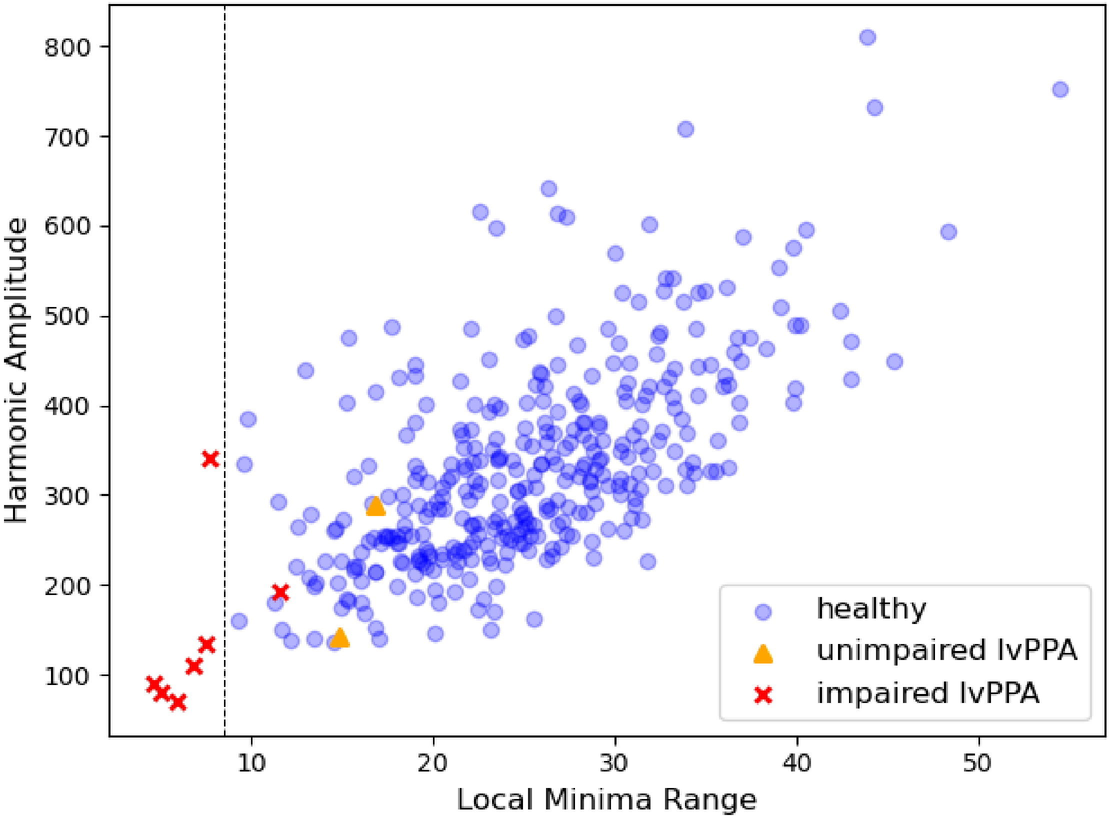


Figure 4-3: A scatter plot of the sum profile local minima range and DFT harmonic amplitude for healthy and lvPPA speakers.

speakers) and 9 lvPPA speakers (aged 64.1 ± 7 years; 5 female speakers).

We experimented with different combinations of the four features to see which were the most salient for distinguishing between speakers with unimpaired and impaired repetitions. One approach that we used to make predictions about which combination of features would be most salient was to plot one feature as a function of another for each speaker and visualize the separation. Figure 4-3 shows one such visualization that plots the local minima range as a function of the harmonic amplitude for lvPPA and healthy speakers.

While lvPPA speakers are known to have difficulty with repetition, there are some that perform the task well. To better analyze the results, a speech and language pathologist independently scored all lvPPA speakers on the multisyllabic word repetition task in terms of speed, precision/accuracy and consistency, before assigning each speaker an overall score. These scores were used to divide the lvPPA speakers into those with relatively unimpaired

Features	Acc.	Spec.	Sens.	Imp. Sens.
[MR]	0.978	1	0.111	0.143
[HA]	0.975	1	0	0
[MT, MR]	0.981	1	0.222	0.286
[MT, HF]	0.975	1	0	0
[MR, HA]	0.978	1	0.111	0.143
[HF, HA]	0.975	1	0	0
[HF, MR]	**0.983**	**1**	**0.333**	**0.429**
[MT, MR, HF, HA]	0.981	1	0.222	0.286

Table 4.1: LDA LOSO results for healthy vs. lvPPA speakers with different feature combinations used as inputs (MT: Minima Time, MR: Minima Range, HF: Harmonic Frequency, HA: Harmonic Amplitude).

repetition abilities, and those with relatively impaired repetition abilities.

In Figure 4-3, the unimpaired lvPPA speakers tend to fall among the cluster of healthy speakers while all but one impaired lvPPA speaker is separated from the healthy speakers. The one impaired lvPPA speaker that is among the healthy speakers is dysfluent throughout the recordings to a lesser degree than the other impaired lvPPA speakers. For this reason, we think it's reasonable for that speaker to be close to the border between the clusters.

4.4.1 Classification

In order to explore how useful the features we extracted were for distinguishing between unimpaired and impaired repetition recordings, we trained five classification models on several different combinations of features. The five classifiers that we trained were the Linear Discriminant Analysis (LDA) classifier, the Decision Tree (DT) classifier, the K-Nearest Neighbors (KNN) classifier, the Random Forest (RF) classifier, and the Support Vector Machine (SVM) classifier. More background information about each of the classifiers mentioned in this section can be found in A.3. Because of the small size of the data, we used Leave-One-Subject-Out (LOSO) cross-validation to train the models. We share the classification results for all five classifiers trained on all the different combinations of features (Tables 4.1, 4.2, 4.3, 4.4, and 4.5,), and we also share the best LOSO results for all five models and the combination of features that the models were trained on (Table 4.6).

Each model output a prediction of "healthy" or "lvPPA". Since there was a large disparity

Features	**Acc.**	**Spec.**	**Sens.**	**Imp. Sens.**
[MR]	0.978	0.986	0.667	0.857
[HA]	0.981	0.992	0.556	0.714
[MT, MR]	0.978	0.986	0.667	0.857
[MT, HF]	0.956	0.975	0.222	0.286
[MR, HA]	0.975	0.986	0.556	0.714
[HF, HA]	**0.986**	**0.994**	**0.667**	**0.857**
[HF, MR]	0.978	0.986	0.667	0.857
[MT, MR, HF, HA]	0.981	0.989	0.667	0.857

Table 4.2: DT LOSO results for healthy vs. lvPPA speakers with different feature combinations used as inputs (MT: Minima Time, MR: Minima Range, HF: Harmonic Frequency, HA: Harmonic Amplitude).

Features	**Acc.**	**Spec.**	**Sens.**	**Imp. Sens.**
[MR]	**0.992**	**1**	**0.667**	**0.857**
[HA]	0.989	1	0.556	0.714
[MT, MR]	**0.992**	**1**	**0.667**	**0.857**
[MT, HF]	0.975	1	0	0
[MR, HA]	0.989	1	0.556	0.714
[HF, HA]	0.989	1	0.556	0.714
[HF, MR]	0.989	1	0.556	0.714
[MT, MR, HF, HA]	0.989	1	0.556	0.714

Table 4.3: KNN LOSO results for healthy vs. lvPPA speakers with different feature combinations (MT: Minima Time, MR: Minima Range, HF: Harmonic Frequency, HA: Harmonic Amplitude).

Features	**Acc.**	**Spec.**	**Sens.**	**Imp. Sens.**
[MR]	0.978	0.986	0.667	0.857
[HA]	0.981	0.992	0.556	0.714
[MT, MR]	**0.992**	**1**	**0.667**	**0.857**
[MT, HF]	0.970	0.992	0.111	0.143
[MR, HA]	0.989	0.997	0.667	0.857
[HF, HA]	0.986	0.997	0.556	0.714
[HF, MR]	**0.992**	**1**	**0.667**	**0.857**
[MT, MR, HF, HA]	0.989	0.997	0.667	0.857

Table 4.4: RF LOSO results for healthy vs. lvPPA speakers with different feature combinations used as inputs (MT: Minima Time, MR: Minima Range, HF: Harmonic Frequency, HA: Harmonic Amplitude).

in the number of healthy speakers versus lvPPA speakers, we report the specificity (the percentage of healthy speakers that were correctly classified as healthy) and sensitivity (the

Features	Acc.	Spec.	Sens.	Imp. Sens.
[MR]	**0.992**	**1**	**0.667**	**0.857**
[HA]	0.989	1	0.556	0.714
[MT, MR]	0.989	0.997	0.667	0.857
[MT, HF]	0.975	1	0	0
[MR, HA]	0.989	1	0.556	0.714
[HF, HA]	0.989	1	0.556	0.714
[HF, MR]	**0.992**	**1**	**0.667**	**0.857**
[MT, MR, HF, HA]	0.989	1	0.556	0.714

Table 4.5: SVM LOSO results for healthy vs. lvPPA speakers with different feature combinations used as inputs (MT: Minima Time, MR: Minima Range, HF: Harmonic Frequency, HA: Harmonic Amplitude).

Classifier	Features	Acc.	Spec.	Imp. Sens.
LDA	[HF, MR]	0.983	1	0.429
DT	[HF, HA]	0.986	0.994	0.857
KNN	**[MR], [MT, MR]**	**0.992**	**1**	**0.857**
RF	**[MT, MR], [HF, MR]**	**0.992**	**1**	**0.857**
SVM	**[MR], [HF, MR]**	**0.992**	**1**	**0.857**

Table 4.6: The best LOSO results for each classifier.

percentage of lvPPA speakers that were correctly classified as lvPPA) results in addition to the average accuracy across all splits. For sensitivity, we were particularly interested in the sensitivity for the impaired lvPPA speakers, as we expected the unimpaired lvPPA speakers to be misclassified as healthy. For this reason, we report the impaired lvPPA sensitivity (the percentage of impaired lvPPA speakers that were classified as lvPPA) in both tables. Table 4.6 shows that the best-performing classifiers had an average accuracy of 0.992, a specificity of 1, and an impaired lvPPA sensitivity of 0.857.

4.5 Discussion

The classification results demonstrate the feasibility of our approach by distinguishing between speakers with unimpaired and impaired repetition (healthy and impaired lvPPA speakers). Only one impaired lvPPA speaker was misclassified as healthy. The misclassified speaker is the impaired lvPPA speaker that is among the healthy speakers in Figure 4-3, which accounts for the misclassification. The misclassified speaker was also early in their disease progression (0 years since diagnosis when the task was completed), which likely made their repetition impairment less severe compared to the other impaired lvPPA speakers who were all further along in the progression of the disease (the other impaired lvPPA speakers had an average of 3 years since diagnosis).

4.5.1 Combination Of Features

Table 4.6 shows that the MR feature is present in the feature combinations of all but one classifier when the best performance for that classifier is achieved. For two classifiers (KNN and SVM), using only the MR feature achieves the best performance (Table 4.6). We can conclude therefore that MR is an important feature for quantifying repetition in audio, perhaps the most important feature. However, only two classifiers achieved the best performance using only MR features. The other classifiers used a combination of MR and MT or HF. This suggests that having a feature that represents the time the repetition occurred, in addition to the minima range, is best for good performance across classifiers.

4.5.2 Benefits Of Our Approach

There are several benefits associated with our approach. The first is that our approach is agnostic to accent and language. Since an utterance is being compared to itself when the distance matrix is computed, the approach does not need to be augmented based on what language someone speaks or what accent they have. Our approach simply tries to capture and measure repeating sounds. This means that the word the speaker repeats does not matter.

Our approach can also potentially be used to identify repeating sounds that are not words, like repeating environmental sounds or non-word repetition tasks, like the popular pataka task that is used to study patients with Parkinson's disease (Gómez-Vilda et al., 2021). Our approach can be applied to a variety of problems in addition to the detection of language impairment.

Our method is also agnostic to the data collection method that is used. Recordings can be captured in a lab setting, where a proctor is present (e.g. FTD dataset). Recordings can also be collected by the speaker themselves without a proctor present, via phone or computer (e.g. CLAC dataset). The task is simple enough for people to complete on their own from anywhere.

Variation in the recording environment is not a problem. The concern associated with having recordings that have been collected in vastly different environments using different microphones is alleviated by the fact that self-distance matrices are collected and distances are not being computed between the recordings with vastly different recording environments. Lastly, our approach does not depend on the amount of input data we have and can be applied to a single recording. This is useful in situations where data tends to be limited (e.g. when working with patient data in the health domain).

4.5.3 Sensitivity/Specificity Analysis

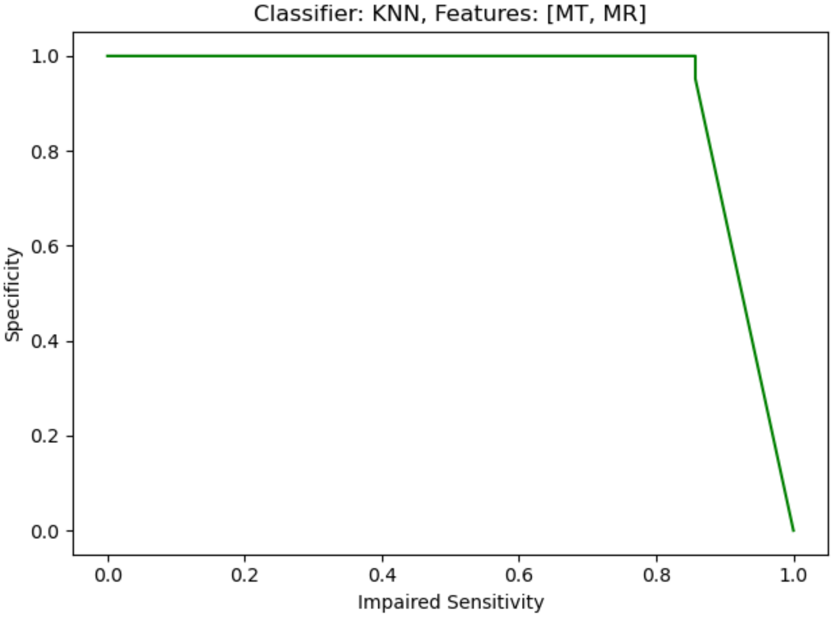


Figure 4-4: A plot of the impaired sensitivity vs. specificity values for different threshold probabilities when training the best KNN model.

For health-related classification, it is important to design systems that prioritize identifying subjects that have the condition, as we do not want anyone with the disease to be missed. Therefore, we are interested in designing systems that have high sensitivity, even at the cost

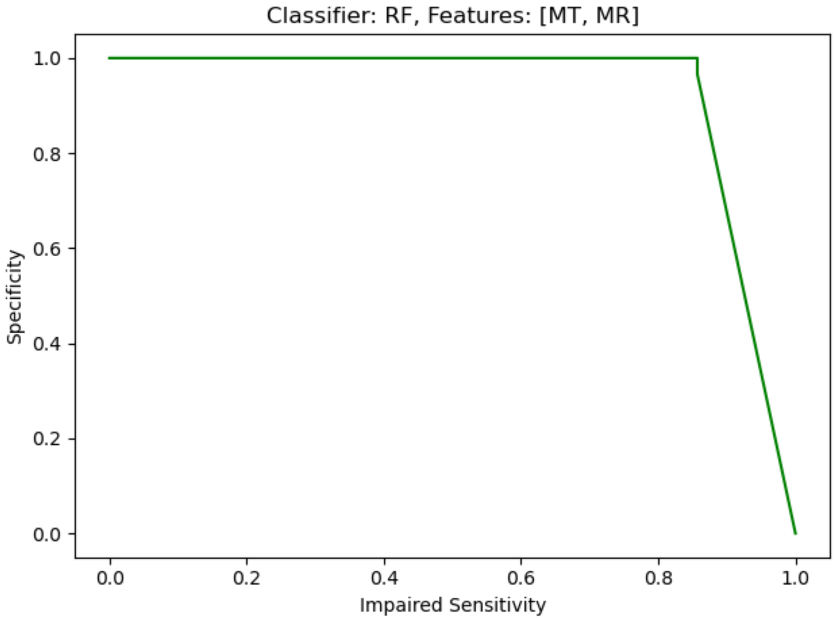


Figure 4-5: A plot of the impaired sensitivity vs. specificity values for different threshold probabilities when training the best RF model.

of having a less-than-optimal specificity, as it is preferred to misclassify healthy speakers as impaired if it means that the impaired speakers have a high likelihood of being detected.

There is currently no standard for finding the optimal sensitivity in the medical domain. While there are techniques that can be used to find an optimal cut off point when sensitivity and specificity are equally important diagnostically (e.g. Youden's index (Youden, 1950)), there is currently no approach for setting a cutoff point when sensitivity is more important, which is the case in the medical domain.

For this reason, we present impaired sensitivity and specificity values for different probability thresholds between 0 and 1 for the best models in Table 4.6, excluding the SVM model. SVM was excluded because it is documented that the prediction probabilities are calibrated using Platt scaling (Platt, 1999), which is known to have theoretical issues that result in a sample having a predicted label that is inconsistent with the assigned probability. For example, the **predict** function may label a sample as belonging to the positive class even if the output of the **predict_proba** function is less than 0.5. Because of the inconsistency, we thought it best to only include plots for the KNN and RF models. The plots can be seen in Figures 4-4 and 4-5. We hope that this approach will give readers a complete understanding of what sensitivity/specificity values can be achieved and allow them to draw their own conclusions about the optimal values and thresholds.

4.6 Chapter Summary

In this chapter, we present a novel approach for measuring the quality of repetition in a recording. We demonstrate the feasibility of this approach by using it to distinguish between healthy and lvPPA speakers using classification. We discuss the many benefits of our approach, including the fact that it is a general repetition measurement approach that can be applied to research problems in a variety of areas. We envision our approach being used to create an application that allows speakers to record themselves completing a simple repetition task (e.g. "repeat the word 'Artillery' five times") and get a repetition score in real time. Our approach is minimally invasive, inexpensive, and uses recordings that are quick/easy to record, making it appealing for use in clinical trials. The language-independent nature of the approach also makes it potentially useful for global clinical trials.

In the next chapter, we explore methods for classifying other FTD/PPA variants, in addition to lvPPA, using other tasks from the same FTD/PPA dataset.

Chapter 5

Classifying Primary Progressive Aphasia and Frontotemporal Dementia from Speech

Frontotemporal Dementia (FTD) and Primary Progressive Aphasia (PPA) are umbrella terms for a set of neurodegenerative diseases that cause problems with behavior and language. Extensive research has been conducted to study how these diseases manifest themselves in the speech of afflicted subjects. In this chapter, we explore using deep learning-based methods to distinguish between healthy subjects and subjects with some form of FTD/PPA. Specifically, we experiment with using several transformer models for feature extraction before training a Logistic Regression classifier on those embeddings for FTD/PPA detection. We compare the results of manual transcription to automatic transcription and perform ablation studies to gain insight into what may be distinguishing healthy controls from FTD/PPA speakers. We show that the Whisper speech recognizer is a viable replacement for human transcription. We also show that healthy controls can be distinguished from subjects with the logopenic variant of PPA (lvPPA), the behavioral variant of FTD (bvFTD), and the semantic variant of PPA (svPPA), with 0.91, 0.91 and 0.90 accuracy, respectively, when evaluated on a cookie theft task, and the nonfluent variant of PPA (nfvPPA) with 0.90 accuracy on a monologue task.

5.1 Motivation

Over the past several years, researchers have been applying deep learning/machine learning-based methods to several different tasks, including the detection of impaired speech. Because of the limited amount of data available for neurodegenerative diseases like FTD and PPA, some previous works have resorted to training simple machine learning classifiers on explainable features (e.g. repetition measures, word usage frequency, etc.) extracted from the data (Haulcy et al., 2022; Zimmerer et al., 2020; Kim et al., 2019; Cho et al., 2020; Vogel et al., 2017; Poole et al., 2017; Matias-Guiu et al., 2022; Neophytou et al., 2019; Fraser et al., 2013; Chlasta and Wołk, 2021; Garcia-Gutierrez et al., 2022; Javeed et al., 2023). While these methods have lead to promising results, it is possible that better performance could be achieved using more advanced deep learning architectures. In particular, the effectiveness of deep learning-based architectures have been demonstrated in several different research domains and have been used to classify several diseases, such as Alzheimer's Disease (Mahajan and Baths, 2021; Pappagari et al., 2021; Balagopalan et al., 2020b; Haulcy and Glass, 2021b; Martinc et al., 2021; Rohanian et al., 2021; Chlasta and Wołk, 2021), Parkinson's Disease (Fang et al., 2020; Chronowski et al., 2022), FTD/PPA (Rezaii et al., 2021; Themistocleous et al., 2021), and more.

The success of these deep learning-based models in previous work has lead us to explore the benefits of using more recent state-of-the-art transformers to classify FTD/PPA. While we hypothesize that the general embeddings will encode information about the speech that will allow the models to distinguish between healthy speakers and FTD/PPA speakers with high accuracy, we also acknowledge that interpretation will be a new challenge.

In this chapter, we present the classification results of using the pretrained Difference-based Contrastive Learning for Sentence Embeddings (DiffCSE) (Chuang et al., 2022) model, as well as other transformer models that are described in greater detail in Section 5.4, to extract embeddings from healthy and FTD/PPA transcripts before using those embeddings to train a classifier using leave-one-subject-out (LOSO) cross-validation. We also perform ablation studies to gain insight into the model's decision-making process before explaining future steps. To the best of our knowledge, we are the first to use DiffCSE embeddings to distinguish healthy and FTD/PPA subjects and the first to offer our interpretation of the model's predictions on this task.

5.2 Dataset

The dataset used in our experiments consists of speech recordings from Australian English-speaking subjects, some healthy and some with various forms of FTD and PPA. The subjects completed several speech tasks, such as the Cookie Theft picture description task (Mueller et al., 2018a), a word repetition task, and a monologue task, in which the subjects were prompted to talk about an event they enjoyed, a happy memory, or a topic that they liked, for approximately one minute. The dataset has been used in previous research to explore which speech characteristics are most salient for the detection of bvFTD (Vogel et al., 2017). The monologue (MONL) and cookie theft (COOK) recordings associated with the healthy, lvPPA, bvFTD, nfvPPA, and svPPA subjects were used for analysis and classification. nfvPPA results were not computed for the COOK task because the available data was severely limited (only two speakers). Demographic information about each of the groups in relation to each task (MONL, COOK) can be seen in Tables 5.1 and 5.2.

Diagnosis	**n**	**m**	**Ave. Age (+/- std)**	**Male/Female**
Healthy	55	55	63.4 (± 7.7)	26/29
lvPPA	18	19	66.3 (± 6.6)	10/8
nfvPPA	10	13	63.4 (± 6.9)	7/3
bvFTD	34	43	62.0 (± 7.4)	25/9
svPPA	13	15	64.6 (± 9.0)	7/6

Table 5.1: Demographic information summarizing the subset of speakers in the FTD/PPA dataset used for training and analysis. (n: number of subjects, m: number of recordings, task: MONL)

5.3 Data Preparation

The speech recordings of each subject completing the MONL and COOK tasks were manually transcribed at Takeda Pharmaceuticals. Two versions of the transcripts were

Diagnosis	n	m	Ave. Age (+/- std)	Male/Female
Healthy	58	58	63.2 (± 7.6)	27/31
lvPPA	6	7	65.1 (± 6.2)	3/3
bvFTD	12	17	62.9 (± 9.0)	10/2
svPPA	6	9	68.1 (± 5.1)	2/4

Table 5.2: Demographic information summarizing the subset of speakers in the FTD/PPA dataset used for training and analysis. (n: number of subjects, m: number of recordings, task: COOK)

Whisper Model	WER
medium_en_0	19.9%
medium_en_10	19.5%
large-v2_0	21%
large-v2_10	20%

Table 5.3: WER for Whisper models. Computed using the manual transcripts as the ground truth.

created by human transcribers: (1) transcripts with interviewer speech included (i.e. the speech of the person administering the test is included in the transcript) and (2) transcripts without interviewer speech (only the speech of the subject completing the task is included in the transcript). Whisper (Radford et al., 2022), a speech recognition model trained on 680,000 hours of speech for several different speech processing tasks, was also used to transcribe the MONL/COOK recordings, so that the effect of manual transcription and automatic transcription on the classification performance could be compared. Whisper was chosen because it is a recently-released model that has very good automatic speech recognition (ASR) accuracy and good robustness to unseen conditions. Several versions of the pretrained Whisper model are publicly available. Both medium and large versions of the model were used for English transcription, one of each with greedy decoding and one of each with a beam search value of 10. More background information about Whisper can be found in A.1.6.

Previous research has shown that there can be large variation in how punctuation is applied by different people (Ueffing et al., 2013). After observing some of the transcripts produced by Whisper, we noticed that punctuation/capitalization was not consistently added. For this reason, the punctuation originally added to both the manual and automatically-generated transcripts was removed and the rpunct library (http://github.com/Felflare/rpunct) was used to apply punctuation to each of the transcripts. rpunct uses a pretrained BERT (Devlin et al., 2019) model that has been finetuned on 560,000 product reviews for punctuation restoration to add punctuation to the input text. An example rpunct output can be seen below:

Before rpunct:

> looks like the mother is doing the dishes but um the sink is overflowing the kids are trying to get cookies from the cookie jar the boy has climbed on a a stool and um it's falling over and the girl is saying pass some to me please and outside there's a garden um is that enough

After rpunct:

> Looks like the mother is doing the dishes, but um, the sink is overflowing. The kids are trying to get cookies from the cookie jar, the boy has climbed on a a stool and um, it's falling over and the girl is saying, pass some to me please And outside there's a garden. Um, is that enough?

The manual transcripts were used to compute the word error rate (WER) for the Whisper transcripts. The WER was computed using the jiwer library (jiw, 2023). The library computes the number of substitutions (S), deletions (D), insertions (I), and hits (H) using the ground truth text (manual transcripts) and Whisper transcripts and the WER is computed as:

$$WER = \frac{(S+D+I)}{(H+S+D)}$$

Punctuation and capitalization were removed from all of the transcripts before the WER was computed. Table 5.3 shows the WER for the different Whisper models. The WER is similar for all the Whisper models so transcripts from all the models were used during classification and the best results were included in Section 5.5.

5.4 Feature Extraction

Several pre-trained transformer models were used to extract embeddings from the MONL/COOK data described in Section 5.2, including a BERT model (Devlin et al., 2019), four versions of the DiffCSE model (Chuang et al., 2022), and a Trans-Encoder model (Liu et al., 2022).

BERT models have been used as the basis for several more recent transformer models (e.g. DiffCSE). BERT has also been used in previous literature to classify other diseases, such as Alzheimer's Disease (Balagopalan et al., 2020b; Haulcy and Glass, 2021b), so we decided to use BERT as a transformer embedding baseline to compare to the performance achieved using the other transformer embeddings.

The trans-encoder model (full model name: trans-encoder-cross-simcse-roberta-base) is an unsupervised sentence representation model that achieved state-of-the-art results on sentence similarity tasks (Liu et al., 2022). The model consists of a bi-encoder and cross-encoder that is placed on top of a pre-trained language model and used to perform self-knowledge-distillation. Since the trans-encoder model is a recently released state-of-the-art transformer model, we decided to use the embeddings for classification as well.

The DiffCSE model is a recent transformer model that has achieved state-of-the-art results on semantic textual similarity tasks (Chuang et al., 2022). The model produces embeddings that are sensitive to the differences between an original sentence and an edited version of that sentence, therefore making the model useful for focusing on important words that change the meaning of a sentence. Four pretrained versions of the DiffCSE model are publicly available. Two of the models used the checkpoints of BERT to initialize the sentence encoder ("bert-base-uncased" in the model name), whereas the other two models used the checkpoints of RoBERTa (Liu et al., 2019) ("roberta-base" in the model name). Different models used different development sets for hyperparameter searching. Models with names ending with "sts" used the Semantic Textual Similarity (STS) Benchmark validation set (Cer et al., 2017) and models with names ending with "trans" used the SentEval (Conneau and Kiela, 2018) development set.

For a standard, non transformer-based baseline, Blabla (Shivkumar et al., 2020) linguistic features were also extracted from the MONL/COOK transcripts and used for training. The

Blabla features are a set of linguistic features consisting of metrics like the noun to verb ratio, the total number of words, the pronoun rate, and more. A total of 39 linguistic features were extracted from the full transcript and used to create a 1-dimensional input vector used to train the logistic regression model.

During feature extraction, both full transcripts (transcript-level) and transcripts split into a list of sentences using punctuation (sentence-level) were used as inputs for the transformer models. Each model output a tensor of shape [number of sentences, sequence length, hidden dimension]. Each model had a token appended to the beginning of each sentence ("[CLS]" or "<s>"). The embedding associated with that token was used to represent each sentence. The sentence vectors were then averaged to create a single vector of length 768, which was used during training.

More background information about each of the pretrained models mentioned in this section can be found in A.1.

5.5 Experiments

A logistic regression model was used for LOSO binary classification (healthy vs. each FTD/PPA variant). The model was trained with a liblinear solver, an L1 penalty, balanced class weights, and an inverse regularization strength value of 0.3. The model output a prediction of "healthy" or the diagnosis ("bvFTD", "lvPPA", "nfvPPA", or "svPPA"). Due to the imbalance between the number of healthy speakers and the other diagnoses, we report Area Under the Curve (AUC) results from the corresponding Receiver Operating Characteristic (ROC) plots, in addition to the classification accuracy.

A few subjects in the dataset completed the tasks more than once (at different times in their disease progression). For this reason, we needed a way of assigning a single label to a speaker with multiple data samples. For accuracy and AUC, we scored every recording sample independently. We also tried additional approaches for AUC, including producing a single score per subject by using the mean, max, or median probability associated with each scored recording for a given subject.

Several experiments were conducted to explore the effect of changing different aspects of the experimental design. The best independent accuracy and AUC results for each diagnostic group can be seen in the following subsections. The full tables of results can be found in Appendix C. More background information about AUC can be found in A.6.1.

5.5.1 Manual vs. Whisper Transcription

		Manual		Whisper		
Diagnosis	**Features**	**Acc.**	**AUC (Ind.)**	**Acc.**	**AUC (Ind.)**	**Model**
lvPPA	B	0.824	**0.976**	**0.838**	0.927	large-v2_10
	DBT	0.865	**0.959**	**0.892**	0.953	medium_en_0
nfvPPA	blabla	0.853	0.901	**0.882**	**0.933**	medium_en_0
bvFTD	B	**0.867**	**0.937**	0.786	0.896	medium_en_10
svPPA	B	**0.771**	**0.839**	0.743	0.766	large-v2_10

Table 5.4: LOSO results for a logistic regression model trained on blabla features and transformer embeddings extracted from sentence-level manual transcripts with interviewer speech included ("Manual"). The corresponding Whisper results are also presented. (Task: MONL, B: bert-base-uncased, DBT: diffcse-bert-base-uncased-trans)

		Manual		Whisper		
Diagnosis	**Features**	**Acc.**	**AUC (Ind.)**	**Acc.**	**AUC (Ind.)**	**Model**
lvPPA	B	**0.892**	**0.978**	0.862	0.970	large-v2_0
bvFTD	blabla	**0.853**	**0.915**	0.800	0.862	medium_en_0
	B	0.827	0.938	**0.907**	**0.973**	large-v2_0
svPPA	blabla	0.851	0.845	**0.851**	**0.862**	medium_en_0

Table 5.5: LOSO results for a logistic regression model trained on blabla features and transformer embeddings extracted from sentence-level manual transcripts with interviewer speech included ("Manual"). The corresponding Whisper results are also presented. (Task: COOK, B: bert-base-uncased)

The results associated with the highest accuracy and AUC values for each diagnosis are presented in the "Manual" column of Tables 5.4 (MONL) and 5.5 (COOK). The "Whisper" column contains the classification results for a Logistic regression model trained on the same type of features extracted from the Whisper transcripts instead of the manual transcripts, as well as the version of the Whisper model that resulted in the highest accuracy and AUC. The results are presented using this format so that the performance with manual transcription can be directly compared to the performance with Whisper transcription. The goal was to determine whether Whisper models could be used to replace human transcribers.

When comparing AUC values, Table 5.4 shows that Whisper outperforms manual transcription for nfvPPA on the MONL task, but manual transcription outperforms Whisper for lvPPA, bvFTD and svPPA, with the largest reduction in AUC being 0.073 for svPPA. For the COOK task, Table 5.5 shows that Whisper outperforms manual transcription for bvFTD and svPPA, and has comparable performance for lvPPA (a reduction in AUC of only 0.008).

For the remaining experiments in the following subsections, Blabla features were not used during training because the ablation studies were used to gain insight into what information may be being encoded in the transformer embeddings by making changes to the transcripts or input format before feature extraction to see how the results changed. For this reason, if the best performance for a particular variant was obtained using Blabla features (nfvPPA for the MONL task and bvFTD/svPPA for the COOK task), the second best results were used for comparison.

5.5.2 Transcript-Level vs. Sentence-Level

		Sentence-Level		**Transcript-Level**	
Diagnosis	**Features**	**Acc.**	**AUC (Ind.)**	**Acc.**	**AUC (Ind.)**
lvPPA	B	**0.824**	**0.976**	0.784	0.893
	DBT	0.865	0.959	**0.905**	**0.969**
nfvPPA	DBS	**0.838**	0.807	0.809	**0.820**
	DBT	**0.794**	**0.825**	0.750	0.766
bvFTD	B	**0.867**	**0.937**	0.704	0.824
svPPA	B	**0.771**	**0.839**	0.671	0.647

Table 5.6: LOSO results for logistic regression models trained on features extracted from sentence-level transcripts vs. **transcript-level** transcripts, with interviewer speech included. (Task: MONL, B: bert-base-uncased, DBT: diffcse-bert-base-uncased-trans, DBS: diffcse-bert-base-uncased-sts)

The purpose of this experiment was to determine which input format resulted in the best classification performance for each feature type. In Tables 5.6 (MONL) and 5.7 (COOK), the highest sentence-level and transcript-level accuracy and AUC values are presented for each diagnostic group. For the MONL task, the highest AUC value is always obtained when embeddings are extracted at the sentence level. For the COOK task, sentence-level embeddings result in the highest AUC for lvPPA, but transcript-level embeddings result in

		Sentence-Level		Transcript-Level	
Diagnosis	**Features**	**Acc.**	**AUC (Ind.)**	**Acc.**	**AUC (Ind.)**
lvPPA	B	**0.892**	**0.978**	0.877	0.946
bvFTD	B	0.827	0.938	**0.907**	**0.961**
svPPA	DRS	0.821	0.822	**0.896**	**0.969**
	DRT	0.776	0.835	**0.881**	**0.985**

Table 5.7: LOSO results for logistic regression models trained on features extracted from sentence-level transcripts vs. **transcript-level** transcripts, with interviewer speech included. (Task: COOK, B: bert-base-uncased, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans)

the highest AUC for bvFTD and svPPA.

5.5.3 Interviewer Speech vs. No Interviewer Speech

		Interviewer		No Interviewer	
Diagnosis	**Features**	**Acc.**	**AUC (Ind.)**	**Acc.**	**AUC (Ind.)**
lvPPA	B	**0.824**	**0.976**	0.824	0.926
	DBT	**0.865**	**0.959**	0.838	0.948
nfvPPA	DBS	0.838	0.807	**0.868**	**0.894**
	DBT	0.794	0.825	**0.897**	**0.877**
bvFTD	B	**0.867**	**0.937**	0.847	0.904
svPPA	B	0.771	0.839	**0.814**	**0.899**

Table 5.8: LOSO results for a logistic regression model trained on different features extracted from sentence-level manual transcripts with interviewer speech included vs. **without** interviewer speech included. (Task: MONL, B: bert-base-uncased, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

		Interviewer		No Interviewer	
Diagnosis	**Features**	**Acc.**	**AUC (Ind.)**	**Acc.**	**AUC (Ind.)**
lvPPA	B	**0.892**	**0.978**	0.877	0.948
bvFTD	B	**0.827**	**0.938**	0.787	0.908
svPPA	DRS	**0.821**	**0.822**	0.806	0.816
	DRT	0.776	**0.835**	**0.791**	0.732

Table 5.9: LOSO results for a logistic regression model trained on different features extracted from sentence-level manual transcripts with interviewer speech included vs. **without** interviewer speech included. (Task: COOK, B: bert-base-uncased, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans)

In this section, we present results using transcripts without interviewer speech included and compare the results to those with interviewer speech included. For the MONL task, including interviewer speech results in higher AUC values for lvPPA and bvFTD, while removing interviewer speech results in better performance for nfvPPA and svPPA. For the COOK task, including interviewer speech leads to increased AUC values for all three variants (lvPPA, bvFTD, and svPPA).

5.5.4 Removing Punctuation and Filler Words

		Punctuation		**No Punctuation**	
Diagnosis	**Features**	**Acc.**	**AUC (Ind.)**	**Acc.**	**AUC (Ind.)**
lvPPA	B	**0.824**	**0.976**	0.811	0.927
	DBT	**0.865**	**0.959**	0.824	0.955
nfvPPA	DBS	**0.838**	0.807	0.824	**0.815**
	DBT	0.794	0.825	**0.868**	**0.828**
bvFTD	B	**0.867**	**0.937**	0.776	0.823
svPPA	B	**0.771**	**0.839**	0.643	0.777

Table 5.10: LOSO results for a logistic regression model trained on features extracted from sentence-level manual transcripts with interviewer speech included, with punctuation included vs. **without punctuation** included. (Task: MONL, B: bert-base-uncased, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

		Punctuation		**No Punctuation**	
Diagnosis	**Features**	**Acc.**	**AUC (Ind.)**	**Acc.**	**AUC (Ind.)**
lvPPA	B	**0.892**	**0.978**	0.754	0.941
bvFTD	B	**0.827**	**0.938**	0.827	0.913
svPPA	DRS	**0.821**	**0.822**	0.806	0.807
	DRT	0.776	**0.835**	**0.791**	0.814

Table 5.11: LOSO results for a logistic regression model trained on features extracted from sentence-level manual transcripts with interviewer speech included, with punctuation included vs. **without punctuation** included. (Task: COOK, B: bert-base-uncased, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans)

An increased number of pauses and a greater proportion of pause time are speech characteristics of lvPPA, nfvPPA, and bvFTD subjects (Poole et al., 2017). An increased use of filler words, such as “uh” and “um”, can indicate pausing in the transcripts. svPPA is associated with a decrease in speech rate and incomplete sentences (Poole et al., 2017).

		Filler		**No Filler**	
Diagnosis	**Features**	**Acc.**	**AUC (Ind.)**	**Acc.**	**AUC (Ind.)**
lvPPA	B	**0.824**	**0.976**	0.797	0.942
	DBT	**0.865**	**0.959**	0.811	0.937
nfvPPA	DBS	**0.838**	**0.807**	0.809	0.759
	DBT	0.794	**0.825**	**0.824**	0.807
bvFTD	B	**0.867**	**0.937**	0.857	0.936
svPPA	B	0.771	0.839	**0.786**	**0.890**

Table 5.12: LOSO results for a logistic regression model trained on different features extracted from sentence-level manual transcripts with interviewer speech included, with filler words included vs. **without filler words** included. (Task: MONL, B: bert-base-uncased, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

		Filler		**No Filler**	
Diagnosis	**Features**	**Acc.**	**AUC (Ind.)**	**Acc.**	**AUC (Ind.)**
lvPPA	B	0.892	0.978	**0.908**	**0.985**
bvFTD	B	0.827	0.938	**0.840**	**0.944**
svPPA	DRS	0.821	0.822	**0.836**	**0.839**
	DRT	0.776	**0.835**	**0.791**	0.828

Table 5.13: LOSO results for a logistic regression model trained on different features extracted from sentence-level manual transcripts with interviewer speech included, with filler words included vs. **without filler words** included. (Task: COOK, B: bert-base-uncased, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans)

The placement of punctuation in the transcripts can potentially provide some insight into whether a speaker is impaired and can potentially be used to determine how complete a sentence is (e.g. an increased number of question marks or punctuation in abnormal places).

For these reasons, we wanted to conduct two ablation studies, where (1) punctuation and (2) filler words (“uh”, “um”, and “ah”) were removed from the manual transcripts with interviewer speech included, before extracting embeddings from the transformer models. We then trained the logistic regression model on the new embeddings so that we could analyze how the performance changed and gain insight into what information the transformer embeddings may be encoding.

The results in Tables 5.10 (MONL) and 5.11 (COOK) show that using punctuation results in higher AUC values for lvPPA, bvFTD and svPPA on both tasks. For nfvPPA, using transcripts with punctuation removed results in the highest AUC value on the MONL task. For the filler word experiments, Table 5.12 shows that keeping filler words in the transcript results in higher AUC values for lvPPA, nfvPPA, and bvFTD for the MONL task, compared to transcripts with filler words removed. Table 5.13 shows that removing filler words from the transcripts results in the highest AUC values from lvPPA, bvFTD, and svPPA on the COOK task.

Plots of the AUC values in Tables 5.4 - 5.13 for each FTD/PPA variant can be seen in Figures 5-1 and 5-2.

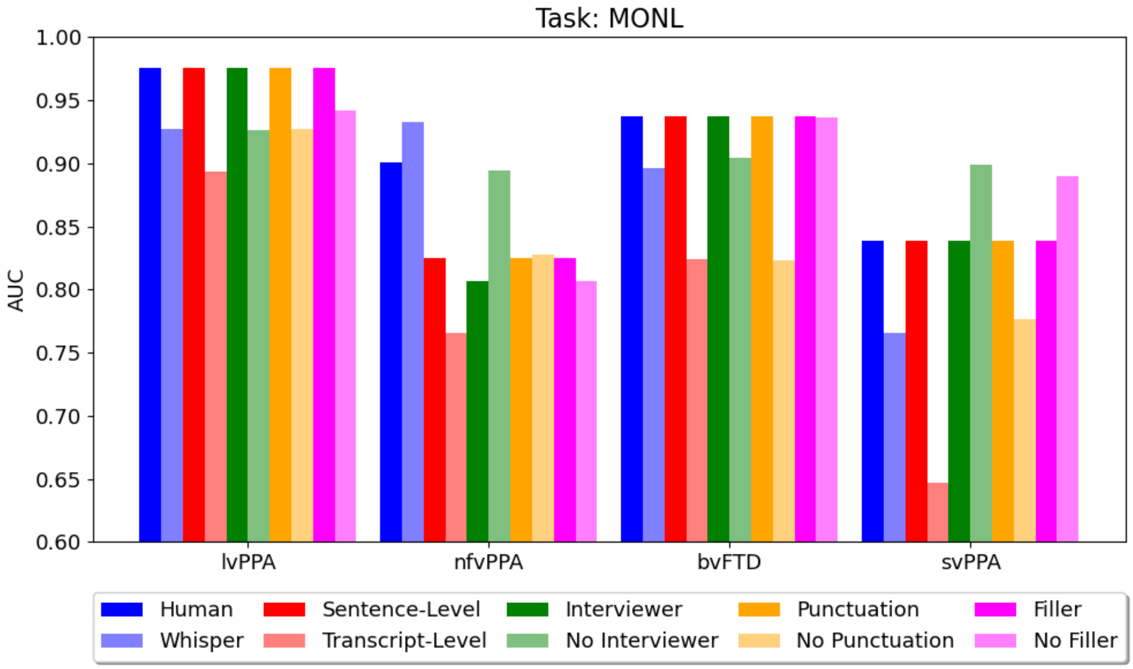


Figure 5-1: A bar plot of the AUC values for each FTD/PPA variant for each of the experiments in Sections 5.5.1 - 5.5.4, evaluated on the MONL task.

5.6 Discussion

In addition to training models that can distinguish between healthy and FTD/PPA speakers with high accuracy, it is important to determine how the models are making decisions, especially when making health-related predictions, as these predictions have significant implications and can determine what kind of treatment a subject receives. For this reason, we attempted to draw conclusions about the impact of certain choices and interpret the transformer embeddings using the experiments conducted in Section 5.5. In the following

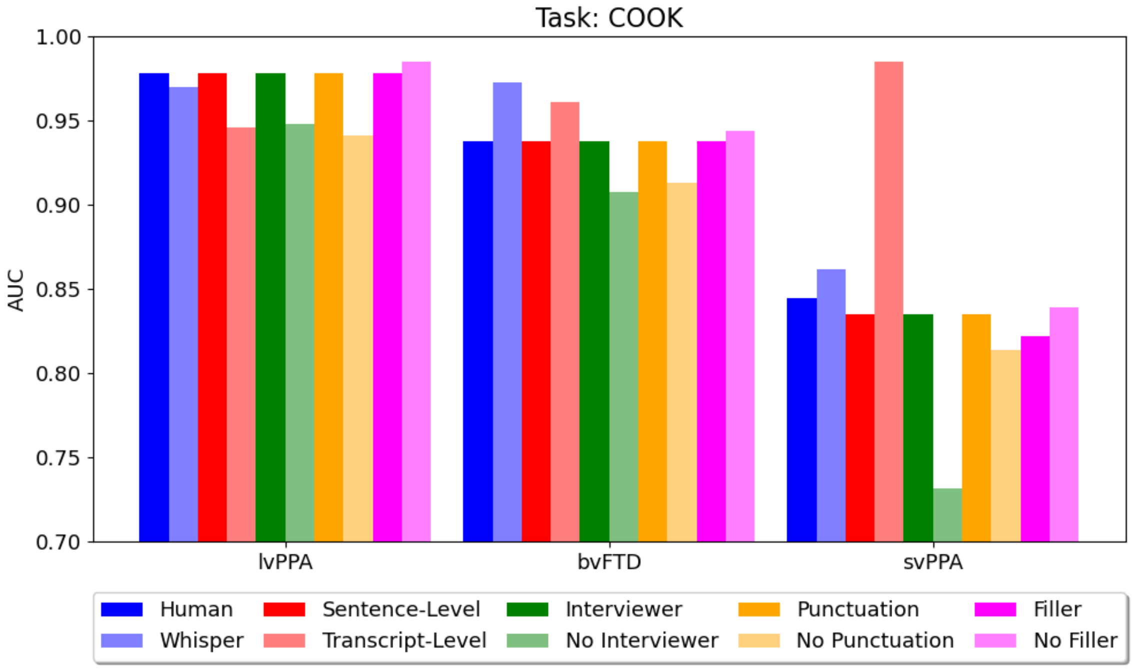


Figure 5-2: A bar plot of the AUC values for each FTD/PPA variant for each of the experiments in Sections 5.5.1 - 5.5.4, evaluated on the COOK task.

subsections, we discuss the results of those experiments.

5.6.1 Manual vs. Whisper Transcription

The results in Section 5.5.1 demonstrate the usefulness and potential of Whisper for automatic transcription. In certain cases, Whisper outperforms manual transcription and in other cases, the results are comparable and may justify using Whisper, despite the decrease in performance.

The decrease in performance for certain variants is likely due to certain errors that Whisper sometimes makes. For example, Whisper may miss words depending on how faint the voice is. For this reason, interviewer speech is not always transcribed because the voice of the subject is often louder than that of the interviewer due to proximity to the recording device. Whisper also misses filler words sometimes and does not include punctuation and capitalization consistently (e.g. some transcripts may have it while others do not). Lastly, letters are sometimes incorrectly replaced with diacritics, resulting in errors in the transcripts. However, using Whisper transcripts ultimately still results in comparable performance for most variants, if not better performance.

If there are concerns about performance decreases, Whisper could potentially be used to complement human transcription by giving an initial transcription that can then be corrected by a human transcriber. Even though the human transcriber is not fully replaced in this scenario, the time cost of transcribing the audio should be reduced. However, the results suggest that Whisper is a viable replacement for human transcription for the majority of the variants on both tasks, despite the WERs of $\sim$20% (Table 5.3).

5.6.2 Transcript-level vs. Sentence-level

The results in Section 5.5.2 suggest that sentence-level embeddings should be extracted from the MONL task, regardless of which diagnostic group is being classified, while the COOK results suggest that sentence-level embeddings should be used for lvPPA and transcript-level embeddings should be used for bvFTD and svPPA. One potential explanation for these findings is that there is more variation in the context of the MONL transcripts, since subjects are allowed to speak freely about any happy experience, compared to the COOK task, where the topic is very specific. Therefore, there is likely enough information in the content of a single sentence to provide insight into the level of a subject's impairment. Since the COOK task is more confined to a set topic (e.g. describing a specific picture), the words used by the subjects likely fall into a smaller subset of words than those of the MONL task. Therefore, less information may be present at the sentence-level for certain variants, making the transcript-level input more informative.

5.6.3 Interviewer Speech vs. No Interviewer Speech

The goal of this experiment was to see what impact the presence of interviewer speech has on the classification results. As technology progresses, more data collection protocols are being designed in a way that allows participants to submit data online, from the comfort of their own homes. For this reason, interviewer speech will likely not be present in future recordings. Therefore, it's important to know how significant the presence of interviewer speech is and whether the transformer model embeddings can encode information from the transcripts that still allows for high AUC values without the interviewer present. We also wanted to present results that align with the direction of future data collection design, so we thought it important to include results without interviewer speech present.

Before conducting the experiments, we hypothesized that removing interviewer speech may result in a decrease in performance, as the presence of interviewer speech often means that a subject is struggling to complete a task and the interviewer felt the need to provide additional guidance and instruction. In this particular dataset, the healthy subjects have very little to no interviewer speech, so it is possible that the presence of speech from a second subject in the transcript may be something that the model embeddings can encode, even though the speech from each subject is not clearly marked. The presence of certain punctuation, such as question marks, for example, may increase for subjects with more interviewer speech in the recording.

The results in Section 5.5.3 suggest that future data collection protocols without an interviewer present should not negatively affect the classification performance for nfvPPA and svPPA when subjects are completing the MONL task, as classification performance for those variants is better without the interviewer speech. While lvPPA and bvFTD have a decrease in AUC when the interviewer speech is removed, the AUC is still above 0.900, which implies that those subjects will likely still be distinguishable from healthy subjects using data collected without an interviewer present.

For the COOK task, including interviewer speech increases the AUC for all of the variants. However, similar to the MONL task, similar AUC values (0.900+) can be achieved without interviewer speech for lvPPA and bvFTD. For svPPA, the drop in AUC is more significant (-0.101). This may suggest that the COOK task is not the most salient task for distinguishing healthy subjects from svPPA subjects when interviewer speech is not present and should not be selected for data collection over other more salient tasks, like MONL. The saliency of different cognitive tasks is an area of future work that we would like to explore.

5.6.4 Removing Punctuation and Filler Words

The results in Section 5.5.4 suggest that punctuation is useful for distinguishing between healthy speakers and lvPPA, svPPA, and bvFTD speakers on both tasks. The decrease in AUC is less significant for the COOK task (<= 0.037 difference) compared to the MONL task (<= 0.113 difference). Punctuation seems to be more important for bvFTD and svPPA, as the removal of punctuation results in a larger AUC decrease compared to the other variants. For nfvPPA, the highest AUC value is obtained after removing punctuation, but the value is comparable to the AUC obtained with punctuation included (0.004 difference). We hypothesize that the amount of punctuation in the nfvPPA transcripts may be less than that of the other variants since nfvPPA subjects are characterized by disfluent speech, which may result in longer sequences of words without punctuation being inserted by rpunct because the sequences are not grammatically correct or recognizable as sentences. If punctuation is less prevalent in the nfvPPA transcripts, that may explain why removing punctuation did not signficantly change the results.

The MONL results align with the findings from previous literature that lvPPA, nfvPPA, and bvFTD subjects are associated with an increased number of pauses, including filled pauses (filler words). For the COOK task, removing filler words increases the AUC value for all three variants (lvPPA, bvFTD, svPPA), which suggests that other components of the text may be more significant for that particular task, such as which words are being used, since there is a very specific context compared to the MONL task.

These experiments provided some insight into the significance of punctuation and filler words and whether the transformer models are encoding information about those components in the embeddings. However, additional experiments are needed to try to understand exactly how punctuation and filler words are being encoded and used to make predictions, as well as what other aspects of the text are most salient.

5.6.5 Significance Testing

We used McNemar's Test (Gillick and Cox, 1989) for significance testing, as a means of quantifying the significance of the changes in the results for each experiment. Background information about McNemar's Test can be found in A.6.2. For the majority of the experiments, the results were not found to be significantly different. This is likely due to the small size of the dataset. However, there were two instances where the probability (p) of the models being the same was noticeably low: the transcript-level experiment for bvFTD ($p = 0.057$) and the no punctuation experiment for lvPPA ($p = 0.065$), both evaluated on the COOK task. These probabilities being low despite the small size of the dataset suggests that removing punctuation from lvPPA transcripts and using transcript-level inputs instead of sentence-level inputs for bvFTD is statistically significant.

5.6.6 Sensitivity/Specificity Analysis

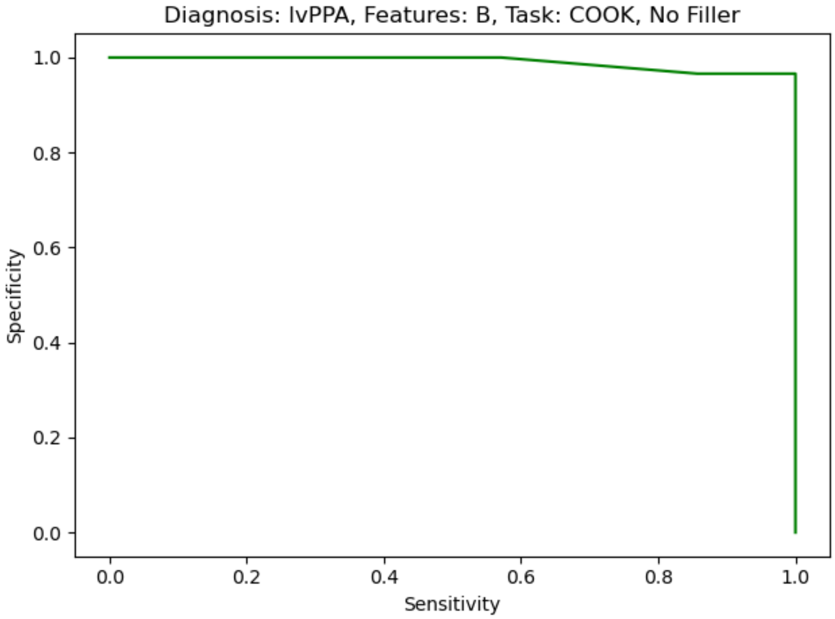


Figure 5-3: A plot of the sensitivity vs. specificity values for different threshold probabilities when training the best lvPPA vs. healthy model.

For the same reasons mentioned in Section 4.5.3, we present sensitivity and specificity values for different probability thresholds between 0 and 1 for the best models for each diagnostic group. These values can be seen in Figures 5-3, 5-4, 5-5, and 5-6. We hope that this approach will give readers a complete understanding of what sensitivity/specificity values can be achieved for each diagnostic group and allow them to draw their own conclusions about the optimal values and thresholds.

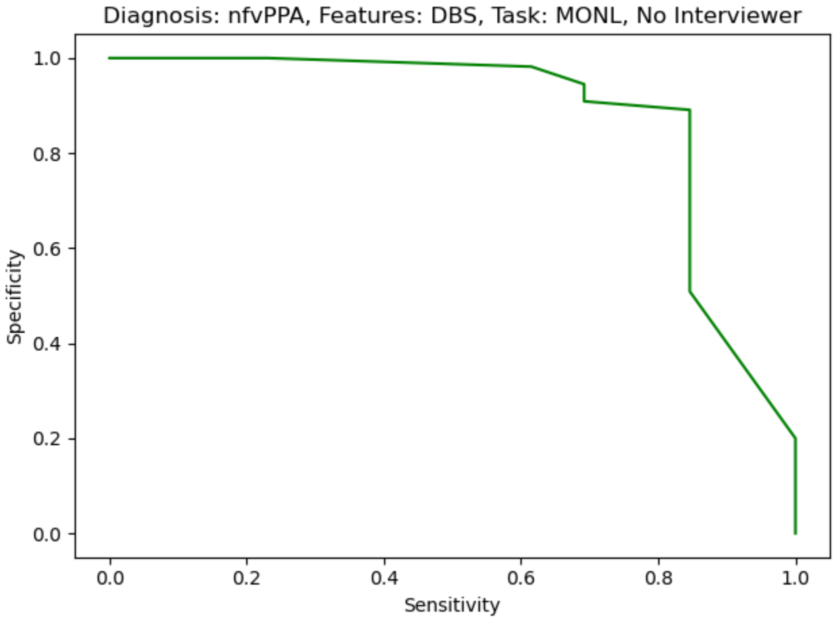


Figure 5-4: A plot of the sensitivity vs. specificity values for different threshold probabilities when training the best nfvPPA vs. healthy model.

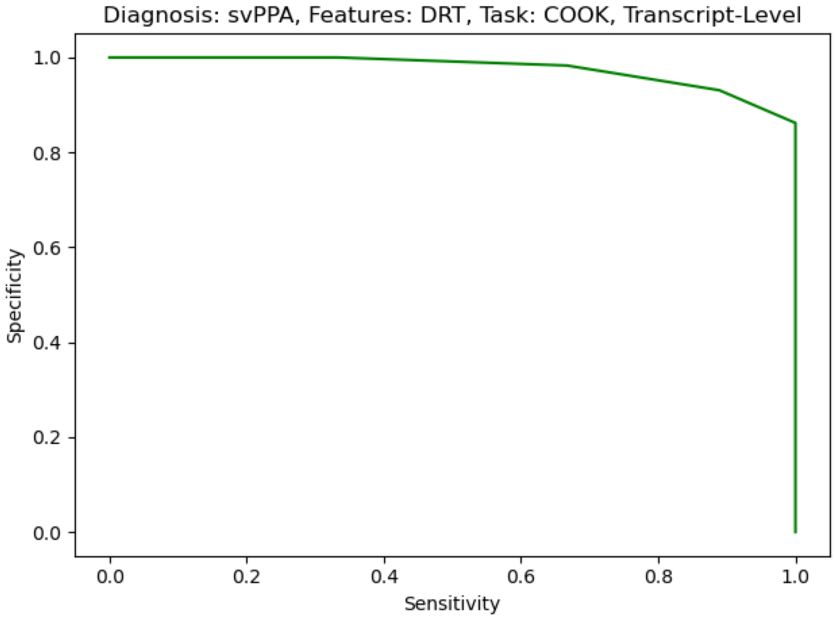


Figure 5-5: A plot of the sensitivity vs. specificity values for different threshold probabilities when training the best svPPA vs. healthy model.

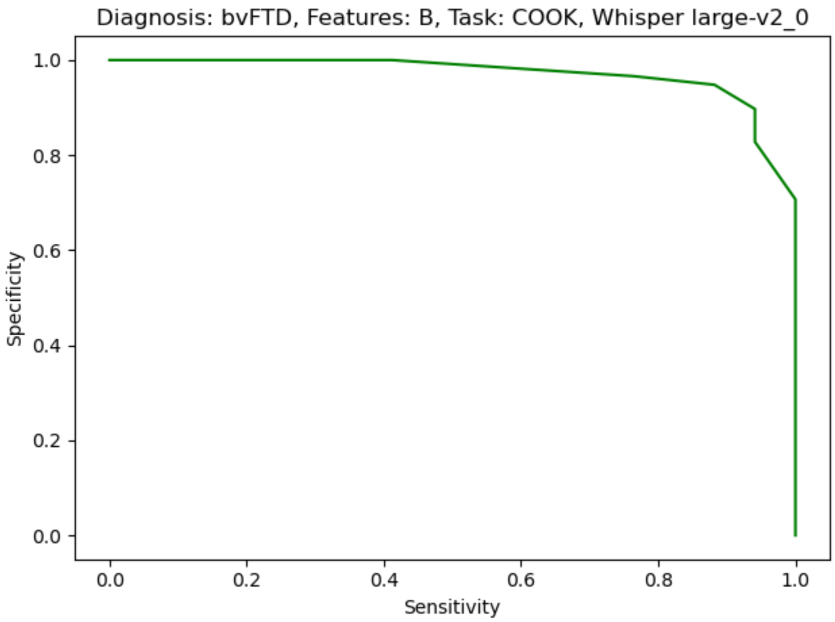


Figure 5-6: A plot of the sensitivity vs. specificity values for different threshold probabilities when training the best bvFTD vs. healthy model.

5.7 Chapter Summary

In this chapter, human-generated transcripts and automatically-generated transcripts were used during LOSO binary classification to distinguish FTD/PPA variants from healthy subjects on two tasks. We conducted several ablation studies and discussed the findings, including the fact that (1) Whisper is a viable replacement for human transcription and results in comparable if not better performance for several variants on different tasks, (2) the input format of the transcripts (transcript-level vs. sentence-level) can impact the classification performance, (3) the inclusion of interviewer speech in the transcripts has an impact but is not necessary to achieve high AUC values for most variants, meaning that future data collection protocols with only subject speech should be viable options for collecting speech in a timely, cost-effective manner, and (4) punctuation and filler words impact the classification results for some variants in statistically significant ways, suggesting that the model is paying attention to those components of the text to some degree and that information about punctuation and filler words is being encoded in the embeddings in some way.

The results show that BERT and DiffCSE embeddings outperform the standard linguistic features (Blabla) for lvPPA, nfvPPA and bvFTD on the MONL task and lvPPA on the COOK task, which justifies the use of transformer model embeddings for FTD/PPA classification. We understand the importance of being able to interpret the predictions of these models and, to do so, we must have some understanding of what information is being encoded in the embeddings. We attempted to gain some preliminary insight into what information the embeddings may encode from the text transcripts, but there is still much work to be done.

In the next chapter, we summarize all of the work described so far and present ideas for next steps.

Chapter 6

Conclusions

In this chapter, we present a summary of the findings from previous chapters and propose directions for future work, including suggesting ways in which the CLAC dataset can be expanded and improved, making suggestions for how to use the techniques we developed to measure disease progression, and proposing ideas for expanding the work to other languages and tasks.

6.1 Summary of Findings

In this section, we summarize the most significant findings from each of the previous chapters.

6.1.1 AD Classification and MMSE Prediction

In Chapter 2, text features (fastText word vectors, BERT embeddings, LIWC features, and CLAN features) and audio features (i-vectors and x-vectors) were extracted from 156 subjects, half with AD and half without. Several classification and regression models were trained for a binary classification task and an MMSE score prediction task.

We found that the SVM and RF classifiers trained on BERT embeddings resulted in the best test set accuracy of 85.4%, which outperformed the text baseline (an LDA classifier trained on CLAN features), which achieved an accuracy of 75% on the test set. For the MMSE prediction task, the best performance was achieved using a gradient boosting regressor trained on a combination of BERT and CLAN features, which achieved an RMSE of 4.560, while the baseline model (a DT regressor trained on CLAN features) obtained an RMSE of 5.20.

Different dimensionality reduction techniques were applied to the data before training and testing the classification and regression models to explore whether applying dimensionality reduction techniques improved performance on those tasks. Applying PCA dimensionality reduction to the features before training a classifier was helpful for AD classification and applying LDA and PCA dimensionality reduction to the features before training a regressor was helpful for MMSE prediction.

The best-performing text and audio models from both tasks were evaluated on smaller subsets of the test set that were split based on education level and MMSE score. This allowed us to explore what effect the severity of cognitive impairment and education level had on the classification and MMSE prediction results. We found that patients with moderate/severe forms of dementia and healthy patients are easier to classify than patients with mild dementia, whereas the MMSE scores of severe dementia patients are the most difficult to predict. Patients with more than 12 years of education are the easiest to classify and the MMSE scores of patients with greater than 12 years of education are the easiest to predict.

Overall, the work in Chapter 2 illustrated the feasibility of using speech and text to classify AD and predict neuropsychological scores.

6.1.2 CLAC Dataset

In Chapter 3, we presented CLAC, a speech dataset consisting of audio recordings and automatically-generated transcripts from 1,832 speakers. The speakers completed a total of 12 tasks, including picture description tasks and passage reading tasks, to name a few. We demonstrated how the dataset can be used to characterize the speech of a healthy, English-speaking population and distinguish between healthy subjects and subjects with some form of cognitive impairment. Specifically, we showed that there can be differences in the timing metrics extracted from the speech of different healthy populations, but the differences are more extreme when comparing the speech of healthy subjects to that of impaired subjects.

We discussed the limitations of the dataset, including the fact that the metadata is self reported and that the speakers' health status was not verified (information about medical history was not collected). However, we believe that the dataset is a valuable contribution to the scientific community, despite those limitations, and the dataset is now publicly available for anyone to use.

6.1.3 Repetition Assessment

In Chapter 4, we presented a novel approach for measuring the quality of repetition in a recording. We demonstrated the feasibility of this approach by using it to distinguish between healthy and lvPPA speakers using classification. Five classifiers were trained on features extracted from repetition recordings and we found that those features could be used to distinguish healthy subjects from lvPPA subjects with an accuracy of 0.992, a specificity of 1, and an impaired lvPPA sensitivity of 0.857. We discussed the many benefits of our approach, including the fact that it is a general repetition measurement approach that can be applied to research problems in a variety of areas, while also being language-independent, a characteristic that makes it potentially useful for global clinical trials.

6.1.4 Classifying FTD/PPA

In Chapter 5, human-generated transcripts and automatically-generated transcripts were used during LOSO binary classification to distinguish FTD/PPA variants from healthy subjects on two tasks: a monologue task and the cookie theft picture description task. We conducted several ablation studies and discussed the findings, including the fact that:

- Whisper is a viable replacement for human transcription and results in comparable if not better performance for several variants on different tasks.
- The input format of the transcripts (transcript-level vs. sentence-level) can impact the classification performance, depending on which task is being completed and which variant the subjects have been diagnosed with.

- The inclusion of interviewer speech in the transcripts has an impact but is not necessary to achieve high AUC values for most variants, meaning that future data collection protocols with only subject speech should be viable options for collecting speech in a timely, cost-effective manner.

- Punctuation and filler words impact the classification results for some variants in statistically significant ways, suggesting that the model is paying attention to those components of the text to some degree and that information about punctuation and filler words is being encoded in the embeddings in some way.

The results show that BERT and DiffCSE embeddings outperform the standard linguistic features (Blabla) for lvPPA, nfvPPA and bvFTD on the MONL task and lvPPA on the COOK task. Specifically, we show that healthy controls can be distinguished from lvPPA, bvFTD, and svPPA subjects with 0.91, 0.91 and 0.90 accuracy, respectively, when evaluated on a cookie theft task, and nfvPPA subjects with 0.90 accuracy on a monologue task. The results justify the use of transformer model embeddings for FTD/PPA classification.

6.2 Future Work

In this thesis, we presented the work that has been completed so far, but there is always more work to be done. Below, we present our ideas for next steps for the work completed in each chapter.

6.2.1 AD Classification and MMSE Prediction: Translation to Larger Datasets

As a follow-up to the work completed in Chapter 2, it would be interesting to repeat the experiments, particularly the evaluation of audio and text models on MMSE and education groups, on a larger dataset to see whether the findings translate. Another interesting future direction would be relating our findings to apathetic symptoms. Previous research has shown that patients with moderate or severe forms of AD tend to be apathetic (Lueken et al., 2007). Signs of apathy include slow speech, long pauses, and changes in facial expressions (Seidl et al., 2012). These characteristics can be measured using standardized ratings and we can explore whether our findings are consistent with the findings related to other forms of cognitive decline that affect speech. We can also explore extracting sentence embeddings from the ADReSS data using more recent pretrained state-of-the-art transformer models and compare the performance to what was obtained using BERT embeddings.

6.2.2 Expanding And Improving The CLAC Dataset

As a follow-up to the work completed in Chapter 3, we think it would be interesting to use the AMT data collection tool to expand the data collection to include other English dialects, such as British English, as well as other languages. We also plan to use this dataset to evaluate the utility of the data for augmenting experimental data for FTD subjects completing other tasks (e.g. passage reading, sustained phonation, etc.), or for subjects with conditions other than FTD. We are also interested in using more advanced speech recognizers, like Whisper (Radford et al., 2022), to generate better transcripts for the CLAC dataset.

6.2.3 Repetition Assessment: Disease Progression And Design Improvements

As a follow-up to the work completed in Chapter 4, we would like to explore how repetition score changes over time. We anticipate using this approach to measure changes in the quality of repetition over time, thus measuring disease progression. In order to complete these experiments, we would need more longitudinal data from subjects completing the repetition task. We would also like to have a baseline that directly uses speech features as input for disorder assessment while also exploring how the number of repetitions affects our method. In our work, we used a simple approximation of the optimal warping path between successive repetitions. As a next step, we could use an algorithm like SDTW to determine the optimal alignment between repetitions and compare the classification performance using our simplified approach and a more advanced approach.

6.2.4 Classifying FTD/PPA: Using Whisper to Expand To Other Languages and Tasks

As a follow-up to the work completed in Chapter 5, we are interested in using diarization techniques to remove the interviewer speech from the audio recordings so that Whisper can be used to generate transcripts without interviewer speech. We are also interested in using Whisper to transcribe other cognitive tasks or speech from subjects with other forms of cognitive impairment and using those transcripts for classification experiments. Whisper is a multilingual model and could also be used to classify impairment in different languages.

We would also like to explore using transformer embeddings for FTD/PPA vs. FTD/PPA classification, in addition to FTD/PPA vs. healthy. If provided with a larger dataset, it would be interesting to finetune the transformer models on the transcripts and explore using model explainability tools to interpret the decisions of the models directly. Lastly, we would like to conduct additional experiments to compare the effectiveness of different tasks for distinguishing between different diagnostic groups. Determining which tasks are most salient would be beneficial to clinical trial design.

Appendix A

Additional Background Information

Additional information about the technical components of the work completed in this thesis can be found in this section, including information about the pretrained models used for feature extraction, non-deep learning-based features that were extracted, classifiers that were trained, and metrics that were used to quantify classification performance.

A.1 Pretrained Models

In this section, we give a concise description of the pretrained models used for feature extraction in Chapters 2 and 5.

A.1.1 fastText Word Vectors

FastText is an open-source library that is used to classify text and learn text representations. In Chapter 2, a fastText model pretrained on Common Crawl and Wikipedia was used to extract word vectors (Grave et al., 2018) from the transcripts of each speaker. The fastText model can be trained in two ways: using Continuous Bag of Words (CBOW) or skipgram. CBOW models predict a target word based on the context of the surrounding words in a sentence or text, while skipgram models use a nearby word to predict the target word. An example of the difference between CBOW and skipgram can be seen in Figure A-1.

The fastText model that we used to extract word vectors is a CBOW model that has been

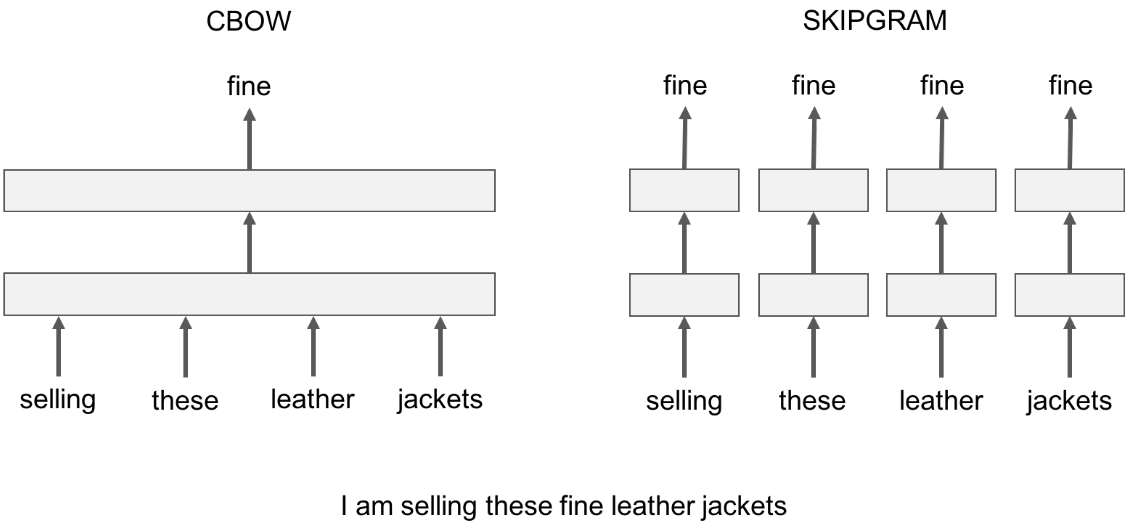


Figure A-1: An example illustrating the difference between CBOW and skipgram. This figure was used as part of a tutorial on the fastText website (fas).

modified to incorporate position weights and subword information. The model itself is a feedforward neural network that takes in a set of context words and outputs a target word. The fastText model extends the standard CBOW model by (1) representing words as bags of character n-grams and (2) multiplying each word vector by a position-dependent vector to better capture positional information for each word. The pretrained model that we used to extract English word vectors used CBOW with position weights, character n-grams of length five, and a window of size five. During pretraining, representations were learned for each of the character n-grams associated with a word and those representations were summed to compute word vectors.

The vectors learned by the model can be used for a variety of Natural Language Processing (NLP) tasks, such as sentiment analysis, language translation, and text classification. The vectors can also be used as input features for classifiers, which is the way that we used them.

A.1.2 I-vectors and X-vectors

I-vector and x-vector systems (Snyder et al., 2017, 2018) are commonly used for speaker verification, the task of verifying the identity of a speaker. Figure A-2 shows a high-level

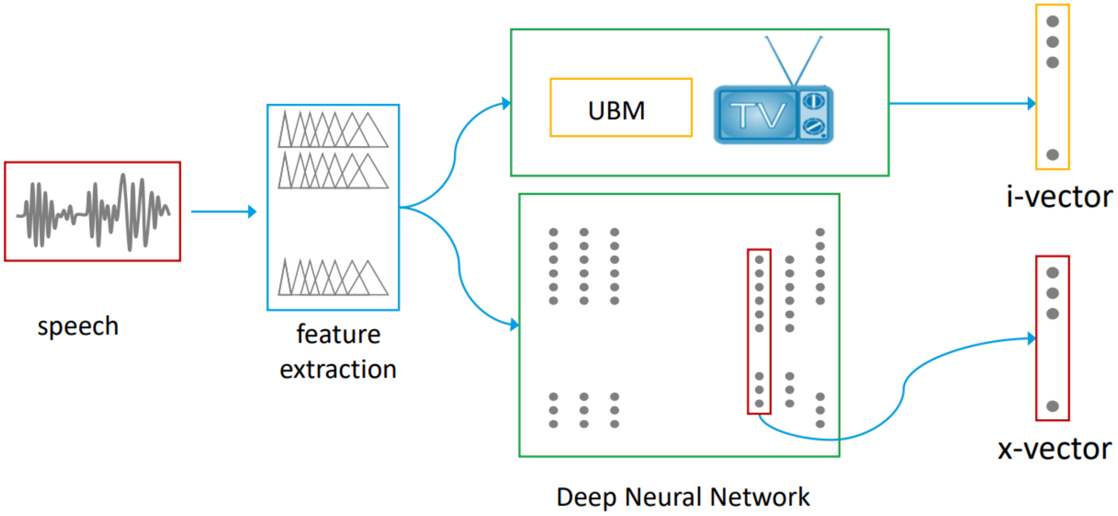


Figure A-2: I-vector and x-vector diagram illustrating the difference between the two. Originally shown in (Kelly et al., 2019).

illustration of the i-vector and x-vector systems. I-vectors are low-dimensional representations of audio signals that result from a high-dimensional representation from a Universal Background Model (UBM) being projected onto a low-dimensional space using a large projection matrix. In contrast, x-vectors are extracted from a deep neural network (DNN) that is trained to discriminate between speakers. X-vectors can be trained on larger amounts of data and have been shown to outperform i-vectors on speaker recognition tasks. Both i-vectors and x-vectors have been extracted from speech audio from impaired speakers and used to train classifiers to detect impairment. For this reason, both were extracted from the data in Chapter 2 and the impact of both on the performance of the classifiers was compared.

A.1.3 Bidirectional Encoder Representations from Transformers (BERT)

BERT (Devlin et al., 2018) is a general-purpose language representation model that is pretrained on Wikipedia and Books Corpus, a corpus consisting of passages from over ten thousand novels. BERT was designed to be trained on large amounts of data from the web and then finetuned on smaller amounts of task-specific data to address the common challenge of the lack of training data for certain tasks. BERT has shown great improvements in accuracy when compared to the performance of models that are trained on smaller

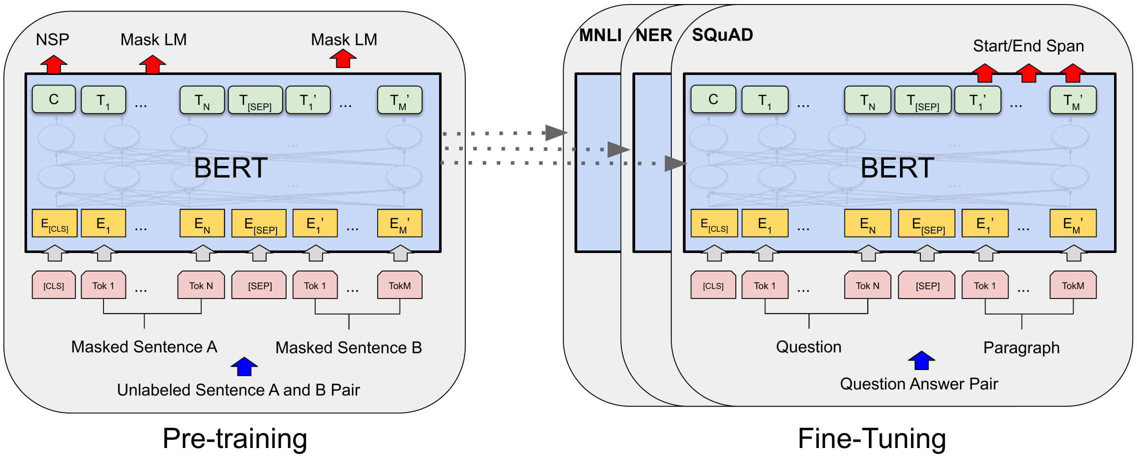


Figure A-3: The architecture diagram for the BERT model. Originally published in (Devlin et al., 2018).

task-specific datasets from scratch.

Figure A-3 illustrates the architecture for the BERT model. During pretraining, words in each input sentence are randomly masked and the model tries to predict what the original words were. For each input, the text is tokenized, a [CLS] token is added to the beginning of the first sentence, and a [SEP] token is added at the end of each sentence. A positional embedding is also added to each token to denote the position of the token in the sentence. The architecture is the same for both pretraining and fine-tuning for downstream tasks, apart from the output layers.

What makes BERT unique is that it is a language model that is bidirectionally trained, meaning that it can consider the full context of a sentence when trying to predict a masked token. At the time of publishing, BERT provided state-of-the-art results on several tasks, including natural language inference, question answering, sentiment analysis, and more.

A.1.4 Difference-based Contrastive Learning for Sentence Embeddings (DiffCSE)

The DiffCSE model is a recent transformer-based sentence embedding model that has achieved state-of-the-art results on semantic textual similarity tasks (Chuang et al., 2022). The model architecture diagram can be seen in Figure A-4. First, the tokens in an input sentence x are randomly masked, creating x'. On the left hand side, the checkpoints of a

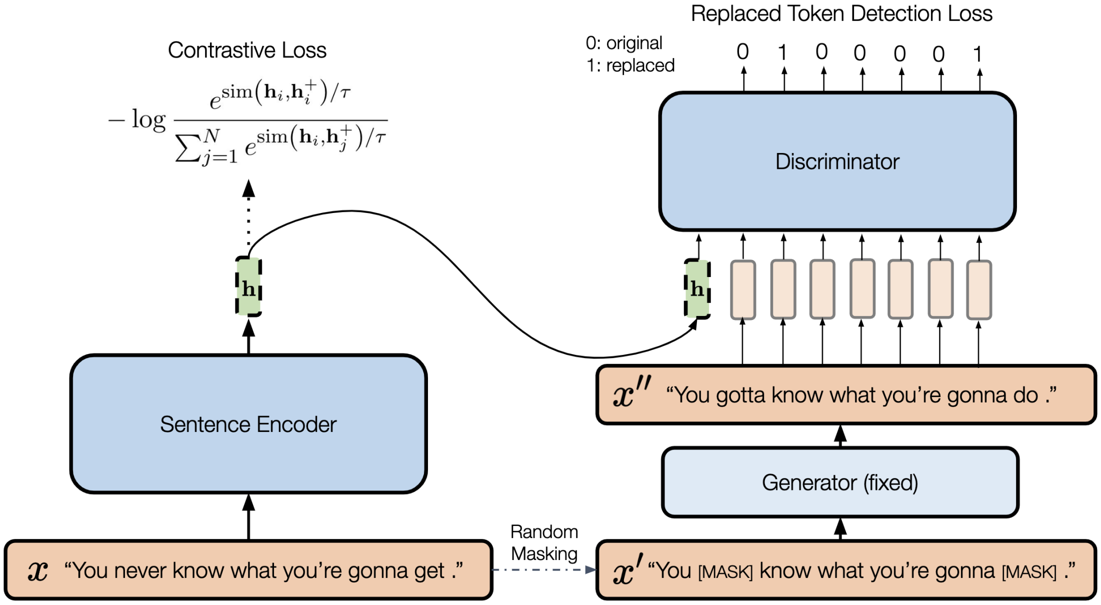

Replaced Token Detection Loss
0: original
1: replaced
0 1 0 0 0 0 1
Contrastive Loss
$-\log \frac{e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{\sum_{j=1}^N e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_j^+)/\tau}}$
Discriminator
h
h
Sentence Encoder
x'' "You gotta know what you're gonna do ."
Generator (fixed)
Random Masking
x "You never know what you're gonna get ."
x' "You [MASK] know what you're gonna [MASK] ."

Figure A-4: The architecture diagram for the DiffCSE model. Originally published in (Chuang et al., 2022).

pretrained model (BERT or RoBERTa) are used to initialize the sentence encoder, which is used to get a sentence embedding for the original sentence. On the right hand side, a pretrained masked language model (DistilBert or DistilRoBERTa (Sanh et al., 2019)) is used as the generator to get predictions for the masked tokens in x', which produces x''. The discriminator (the same pretrained masked language model that was used for the sentence encoder) is then used to determine whether the token in the original sentence x has been replaced in x'' or not.

During training, the discriminator and sentence encoder are both optimized. Due to the training design, the sentence encoder is encouraged to encode the full meaning of x into the sentence embedding so that the discriminator can recognize the differences between x and x''. After training, the discriminator is discarded and the sentence encoder is used to extract embeddings from the input data. As a result of the training design, the model produces embeddings that are sensitive to the differences between an original sentence and an edited version of that sentence, therefore making the model useful for focusing on important words that change the meaning of a sentence.

A.1.5 The Trans-Encoder Model

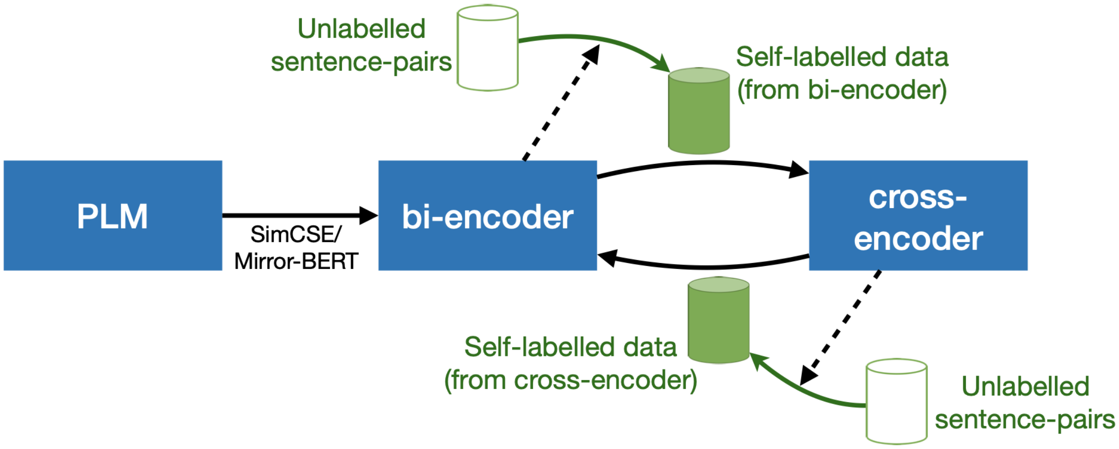


Figure A-5: The architecture diagram for the trans-encoder model. Originally published in (Liu et al., 2022).

The trans-encoder model (full model name: trans-encoder-cross-simcse-roberta-base) is an unsupervised sentence representation model that achieved state-of-the-art results on sentence similarity tasks (Liu et al., 2022). The model architecture diagram can be seen in Figure A-5. The blue boxes in the figure represent the same model architecture. The model consists of a bi-encoder and cross-encoder that is placed on top of a pretrained language model (PLM) and used to perform self-knowledge-distillation. This process involves using the weights of the PLM to initialize the bi-encoder and the cross-encoder. Self-knowledge-distillation is performed in two ways: bi- to cross-encoder and cross- to bi-encoder.

With bi- to cross-encoder self-distillation, the bi-encoder is used to get two separate embeddings for each sentence in a sentence-pair. The cosine similarity between the embeddings is computed and used as a relevance score for the sentence-pair. The two sentences and the score are used as input to the cross-encoder, which minimizes the difference between the score from the bi-encoder and the predictions from the cross-encoder. In this setup, the bi-encoder is the teacher model and the cross-encoder is the student model.

The cross-encoder that is produced as a result of bi- to cross-encoder self-distillation can be used in cross- to bi-encoder distillation, with the roles of the bi-encoder and cross-encoder reversed. The cross-encoder is used to generate the relevance score and is therefore the teacher model, while the bi-encoder is the student model and is further improved by learning from the cross-encoder.

The model is trained and evaluated on the Quora Question Pair (QQP) dataset and the question-answering entailment (QNLI) dataset. The QQP dataset consists of question pairs and the model has to determine whether the questions are duplicates. The QNLI dataset consists of question-sentence pairs and the model has to determine whether the sentence answers the question or not.

A.1.6 The Whisper Speech Recognizer

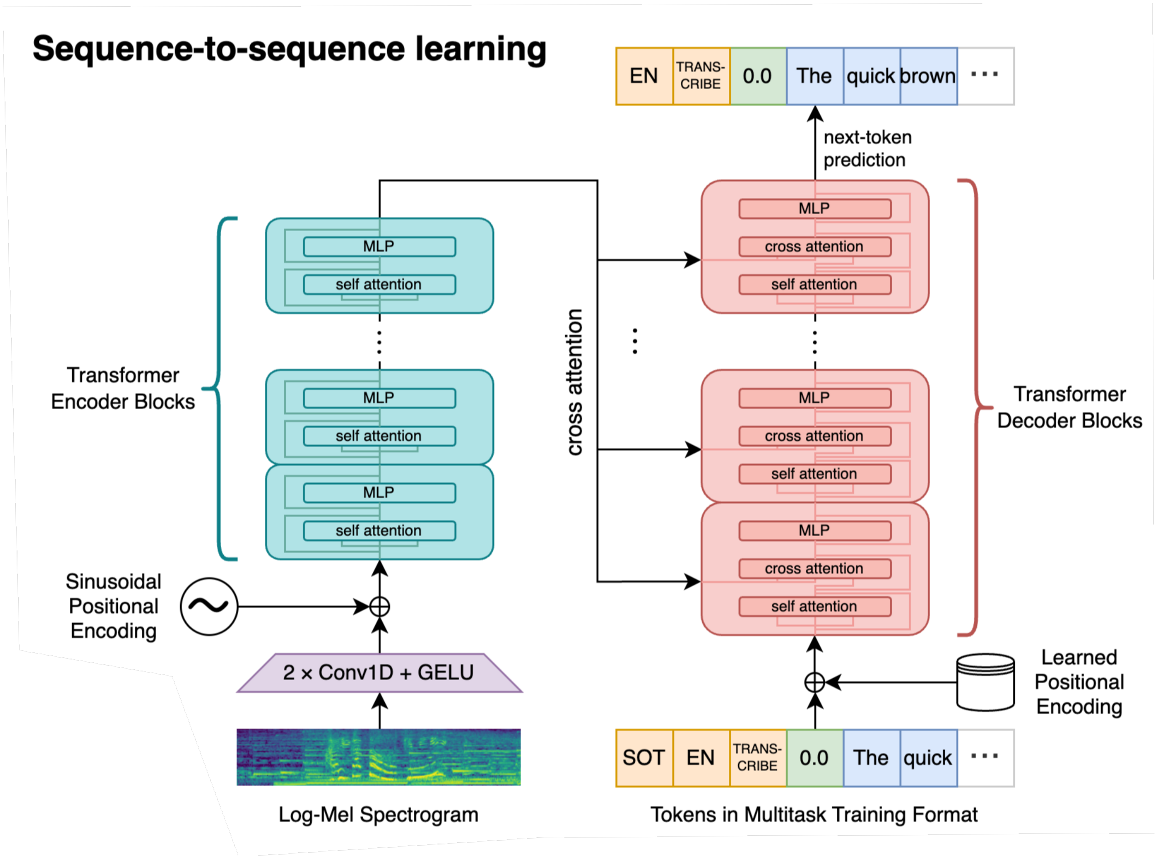

Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
···
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
Learned Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
SOT
EN
TRANS-CRIBE
0.0
The
quick
···
Tokens in Multitask Training Format

Figure A-6: The architecture diagram for the Whisper speech recognizer. Originally published in (Radford et al., 2022).

Whisper (Radford et al., 2022) is a speech recognition model trained on 680,000 hours of speech for several different speech processing tasks. Figure A-6 shows that Whisper is a transformer model with the recognizable encoder-decoder format. Thirty second audio clips are converted into spectrograms before being passed into the encoder. The decoder predicts the text associated with the input audio, as well as additional special tokens. The special tokens consist of things such as a language identification token ("EN"), a token to describe the task ("TRANSCRIBE"), the start and end times of the text tokens, and more.

Because Whisper is trained on a large, diverse dataset, it is robust to accents, background noise, sped-up speech, and more. It can be used to transcribe several languages, as well as translate from several languages to English. It can also be used for language identification and to extract phrase-level timestamps.

A.2 Feature Extraction

In this section, we give a brief description of the non-deep learning-based features that were extracted prior to training in Chapters 2, 4, and 5.

A.2.1 Mel-frequency Cepstral Coefficients (MFCCs)

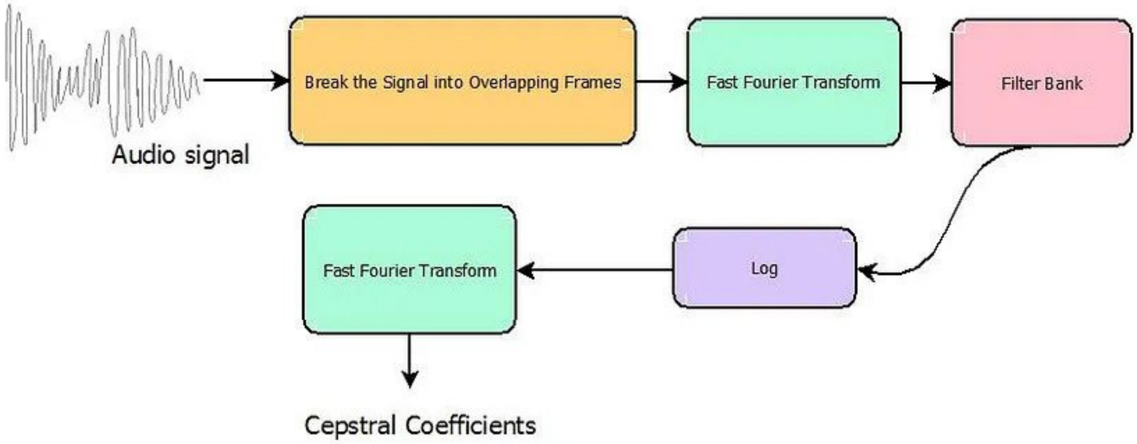


Figure A-7: A block diagram summarizing the steps for extracting MFCCs from an audio signal. Found at https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd.

MFCCs are commonly used as a feature representation of speech audio. MFCCs are often extracted before training automatic speech recognition (ASR) and speaker recognition systems. There are several steps involved in extracting the MFCCs from an audio signal (Figure A-7):

- **Chunking the audio into smaller frames**: the audio is processed at the frame level so that the audio signal does not change too much but also has enough samples to get a reliable spectral estimate

- **Calculate the power spectrum of each frame**: identifies which frequencies are present in the frame

- **Apply the mel filterbank to the power spectra and sum the energy in each filter**: determines how much energy exists in different frequency regions

- **Take the log of the filterbank energies**: compresses the features to match more closely to what humans actually hear

- **Compute the Discrete Cosine Transform (DCT) of the log filterbank energies**: decorrelates the filterbank energies

- **Keep DCT coefficients 2-13 and discard the rest**: the higher coefficients have been shown to degrade ASR performance

A full tutorial can be found at http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.

A.2.2 Linguistic Inquiry and Word Count (LIWC)

The LIWC toolbox (Tausczik and Pennebaker, 2010) was developed as a result of research that shows that people's word choice can provide insight into their emotional and psychological state, as well as their social and behavioral beliefs. LIWC consists of over 100 dictionaries that contain words associated with a particular category. These categories include positive and negative emotion words (e.g. "happy", "mad"), I-words ("I", "me", "my"), positive tone, negative tone, social words, and more.

At the time of completing the work for Chapter 2, the LIWC toolbox returned word counts for each of the categories. A more recent version of the toolbox, LIWC-22, returns the percentage of total words in the text belonging to a particular category, as well as the raw word count (the number of words in the file), the number of words per sentence, and summary measures. There are four summary measures: Analytical Thinking (based on different categories of function words), Clout (related to social status, confidence and leadership demonstrated through the text), Authenticity (tries to measure whether the speaker is honest or deceptive), and Emotional Tone (combines positive and negative tone measures). The measures are computed using algorithms that combine values associated with various LIWC variables.

A.2.3 Computerized Language Analysis (CLAN)

Feature Name	Feature Description
Duration	total time of the sample in hours:minutes:seconds
Total Utts	total number of utterances
MLU Utts	number of utterances used to compute MLU
MLU Words	MLU in words
MLU Morphemes	MLU in morphemes
FREQ types	total word types as counted by FREQ
FREQ tokens	total word tokens as counted by FREQ
FREQ TTR	type/token ratio
Words/min	words per minute (FREQ tokens/Duration converted to minutes)
Verbs/Utt	verbs per utterance
% Word Errors	percentage of words that are coded as errors
Utt Errors	number of utterances coded as errors
Density	measure of propositional idea density
% Nouns	percentage of nouns
% Plurals	percentage of plurals (we, us, our(s) they, them, their)
% Verbs	percentage of verbs
% Aux	percentage of auxiliaries
% Mod	percentage of modals

Table A.1: CLAN feature names and descriptions. (MLU: ratio of morphemes over utterances)

CLAN is a program that was designed to analyze transcripts in the Codes for the Human Analysis of Transcripts (CHAT) format (MacWhinney, 2000). CLAN can be used to extract features from CHAT transcripts using several different commands. For example, the FREQ command can be used to extract information about the frequencies of the words in one or more files, while TIMEDUR can be used to compute the time durations of the utterances

Feature Name	Feature Description
% 3S	percentage of third person singular terms
% 1S/3S	percentage of identical forms for first and third person (e.g., I was, he was)
% Past	percentage of words in the past tense
% PastP	percentage of past participles
% PresP	percentage of present participles
% prep	percentage of prepositions
% adv	percentage of adverbs
% adj	percentage of adjectives
% conj	percentage of conjunctions
% det	percentage of determiners (includes articles, demonstratives, interrogatives, numbers, and possessives)
% pro	percentage of pronouns
noun/verb ratio	total number of nouns / total number of verbs (excluding auxiliaries and modals)
open/closed ratio	total number of open class words / total number of closed class words
#open-class	total number of open class words
#closed-class	total number of closed class words
retracing	number of retracings (self-corrections or changes)
repetition	number of repetitions

Table A.2: CLAN feature names and descriptions (cont.).

and pauses in one or more files, to name a few. In Chapter 2, the CLAN EVAL program, which runs several commands, was used to extract several measures from the transcripts in the ADReSS dataset, all of which can be seen in Tables A.1 and A.2. More information about each of the features can be found in (MacWhinney, 2000), along with instructions for how to use CLAN program.

A.2.4 Blabla

Blabla (Shivkumar et al., 2020) is a Python package that is used to extract clinical linguistic features from transcripts, with support for multiple languages. The supported languages are English, Arabic, Chinese, French, German, and Spanish. The linguistic features consist of metrics like the noun to verb ratio, the total number of words, the pronoun rate, and more. A total of 39 linguistic features were extracted from the full transcript. The name of the 39 features extracted from the transcripts used in Chapter 5, along with the descriptions of the

Feature Name	Feature Description
adjective_rate	The rate of adjectives across sentences
adposition_rate	The rate of adpositions across sentences
adverb_rate	The rate of adverbs across sentences
auxiliary_rate	The rate of auxiliaries across sentences
determiner_rate	The rate of determiners across sentences
interjection_rate	The rate of interjections across sentences
noun_rate	The rate of nouns across sentences
numeral_rate	The rate of numerals across sentences
particle_rate	The rate of particles across sentences
pronoun_rate	The rate of pronouns across sentences
proper_noun_rate	The rate of proper nouns across sentences
punctuation_rate	The rate of punctuations across sentences
subordinating_conjunction_rate	The rate of subordinating conjunctions across sentences
symbol_rate	The rate of symbols across sentences
verb_rate	The rate of verbs across sentences
demonstrative_rate	The rate of demonstratives across sentences
conjunction_rate	The rate of conjunctions across sentences
possessive_rate	The rate of possessive words across sentences
noun_verb_ratio	The ratio of nouns to verbs across sentences
noun_ratio	The ratio of nouns to the sum of nouns and verbs across sentences
pronoun_noun_ratio	The ratio of pronouns to nouns across sentences
total_dependency_distance	The total distance of all dependencies across sentences
average_dependency_distance	The average distance of all dependencies across sentences
total_dependencies	The total number of unique dependencies across sentences
average_dependencies	The average number of unique dependencies across sentences
closed_class_word_rate	The proportions of determiners, pronouns, conjunctions and prepositions to all words across sentences
open_class_word_rate	The proportions of nouns, verbs, adjectives and adverbs to all words across sentences

Table A.3: Blabla feature names and descriptions.

features, can be seen in Tables A.3 and A.4. More information about how to extract Blabla features can be found at https://github.com/novoic/blabla.

Feature Name	Feature Description
content_density	The proportions of number of open class words to the number of close class words
idea_density	The proportions of verbs, adjectives, adverbs, prepositions and conjunctions to all words across sentences
honore_statistic	Calculated as R = (100*log(N))/(1-(V1)/(V)), where V is number of unique words, V1 is the number of words in the vocabulary only spoken once, and N is overall text length / number of words.
brunet_index	Calculated as N^(V^(-0.165)), where V is number of unique words and N is overall text length / number of words. Measure of lexical richness. Text-length insensitive version of TTR.
type_token_ratio	The number of word types divided by the number of word tokens
word_length	The mean length of words across the corpus
prop_inflected_verbs	The ratio of the number of inflected verbs to the number of verbs
prop_auxiliary_verbs	The ratio of the number of auxiliary verbs to the number of verbs
prop_gerund_verbs	The ratio of the number of gerund verbs to the number of verbs
prop_participles	The ratio of the number of particile verbs to the number of verbs
num_words	The total number of words
num_filler_words	The total number of filler words

Table A.4: Blabla feature names and descriptions (cont.).

A.3 Classifiers and Regressors

Several classifiers and regression models were used for classification in Chapters 2, 4, and 5. All of the models were implemented in Python using the scikit-learn library (Pedregosa et al., 2011). More information about each of the models can be found in the user guides on the scikit-learn website (https://scikit-learn.org). A brief description of each model is provided in the following subsections.

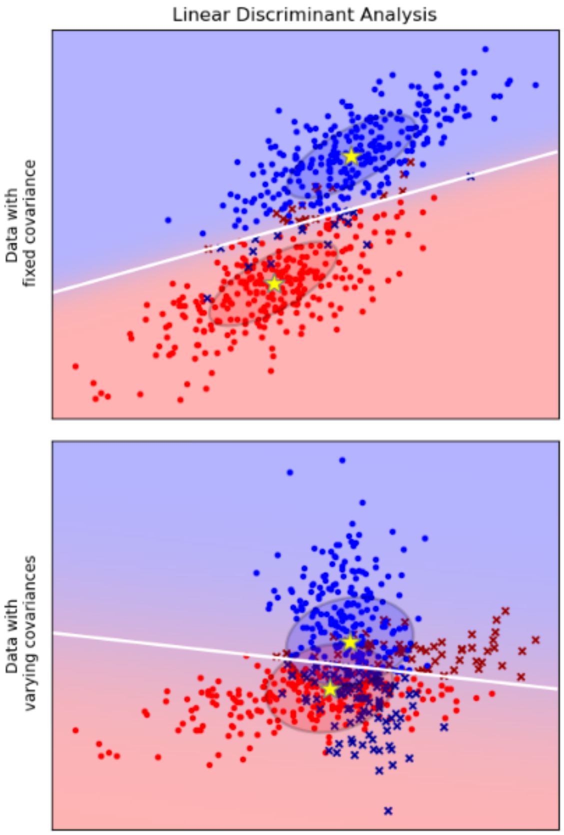


Figure A-8: Partial figure from the LDA user guide on the scikit-learn website (sci, e). The plot shows decision boundaries for the LDA classifier. The subplots show that LDA can only learn linear boundaries.

A.3.1 Linear Discriminant Analysis (LDA)

The usefulness of the LDA classifier has been demonstrated in practice for various classification problems. Some of the appealing attributes of the classifier include the ease with which solutions can be computed and the fact that hyperparameter tuning is not needed. Simple probabilistic models are used to model the distribution of the data for each class and Bayes' rule is used to get predictions for the training samples. The LDA classifier can only learn linear boundaries. Examples of linear boundaries computed using LDA can be seen in Figure A-8.

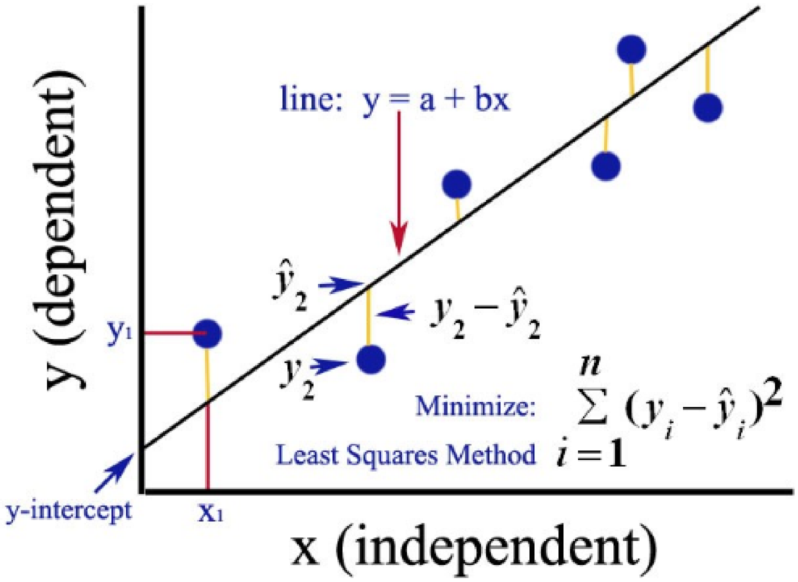


Figure A-9: Figure from (Gulve, 2020). Visually shows how linear regression fits a straight line to a set of data points.

A.3.2 Linear Regression (LR)

LR is a simple model that fits a straight line to a set of data points, as shown in Figure A-9. The solution is found by minimizing the sum of squared errors between the ground truth values and the predicted values. The LR algorithm uses the Ordinary Least Squares (OLS) method, which involves estimating the values of a set of linear coefficients that minimize the error. The OLS method involves computing the distance from each input data point to a regression line, squaring those distances to get the error values, and then computing the sum of those errors. The goal is to find the regression line that minimizes the sum of those errors.

A.3.3 Decision Tree (DT)

DT models can be used for both classification and regression. The models consist of simple decision rules (e.g. if-then-else statements) that are learned from the training data and that become more complex as the depth of the tree increases. The simplicity of the decision rules make DT models easy to interpret and the trees can be visualized. An example of a visualization of a DT model can be seen in Figure A-10.

There are some disadvantages associated with DT models, including the fact that models

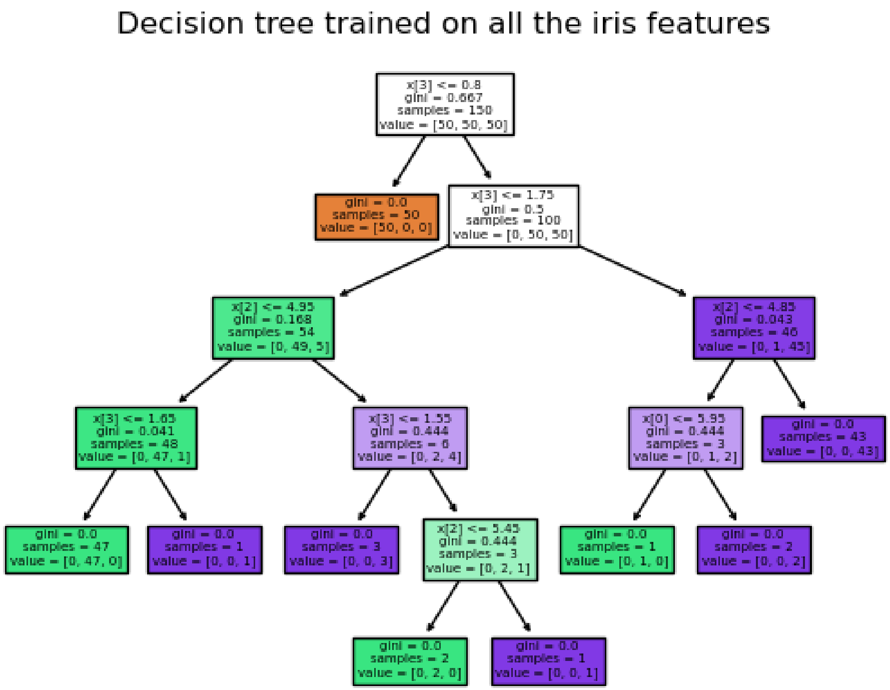


Figure A-10: Figure from the DT user guide on the scikit-learn website (sci, b). The plot shows a visualization of the tree for a model trained on features from the iris dataset, a dataset consisting of four features for different types of flowers.

with a maximum depth that is too large can overfit to the training data and not generalize well and that small changes in the data can result in completely different models (trees). Therefore, it is important to try trees with different depths and evaluate the models on unseen data.

A.3.4 K-Nearest Neighbor (KNN)

The KNN classifier and regressor both use a technique that involves using information about the closest data points (nearest neighbors) to determine the class or regression value of the target point. The distances between the current data point and the other data points are computed and the nearest neighbors are chosen based on the user-selected k value. The most common label of the k nearest neighbors is assigned to the current point for classification,

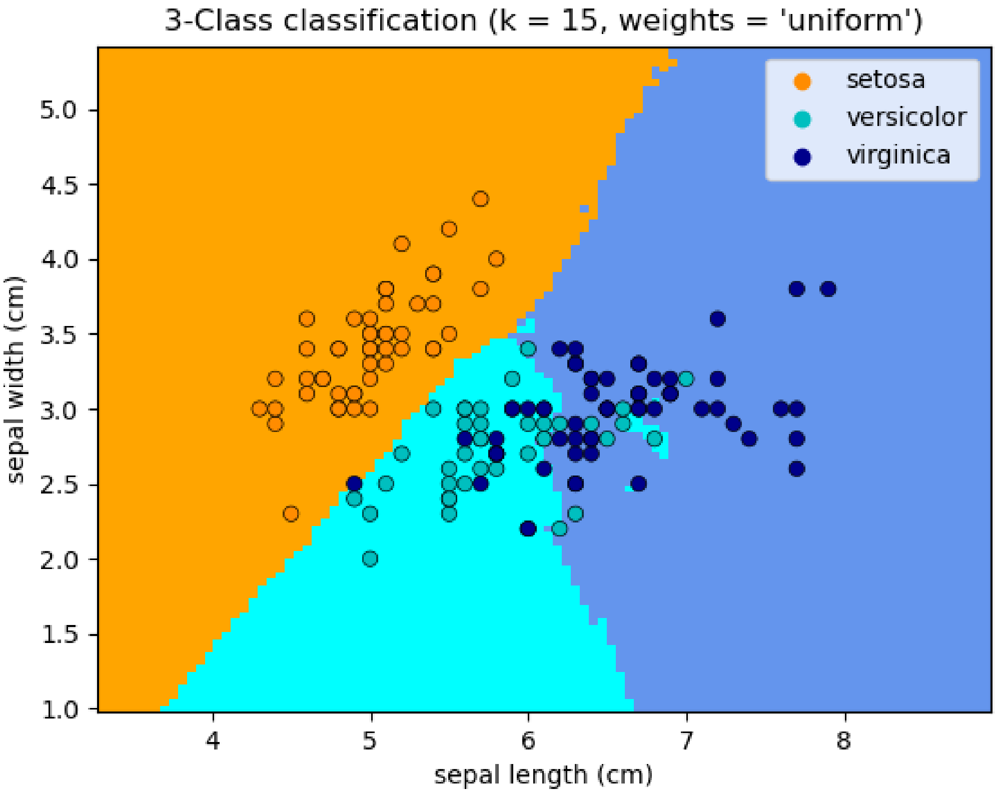


Figure A-11: Figure from the KNN user guide on the scikit-learn website (sci, c). The plot shows a visualization of the class boundaries for the iris dataset when a KNN classifier with the number of neighbors (k) set to 15 is used for classification.

and the average of the k labels is assigned to the current point for regression. The best k value is data-dependent, meaning that results using several different values should be compared to find the optimal value.

KNN is a simple algorithm that is easy to implement. The main disadvantage of the algorithm relates to its scalability: as the size of the data increases, the speed of the algorithm decreases.

A.3.5 Random Forest (RF)

The RF classifier and regressor consists of several DT models (described in section A.3.3) trained on subsets of the data. Data samples are randomly drawn from the training set (with replacement) and those samples are used to create a DT. This process is repeated for several trees and then the predictions from the individual DT models are averaged. This improves the overall accuracy of the RF model and alleviates the issue of overfitting that is sometimes encountered when using a single DT model.

A.3.6 Support Vector Machine (SVM)

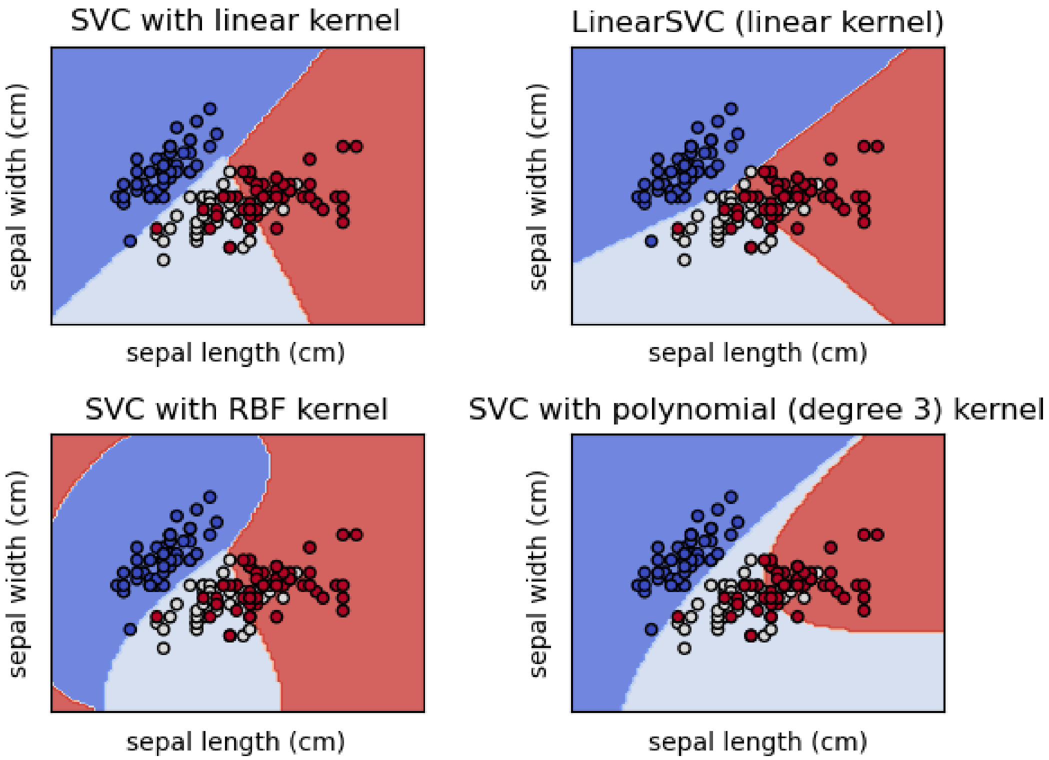


Figure A-12: Figure from the SVM user guide on the scikit-learn website (sci, d). Here, subplots with class boundaries are shown for SVM models trained and evaluated using different types of kernels.

SVMs can be used for both classification and regression. SVMs are linear models that find a line or hyperplane that separates the input data into its respective classes. For linearly separable datasets, this is achieved by first computing the support vectors, which are the points from each class that are closest to the line separating the classes. The margin, which is defined as the distance between the line and the support vectors, is maximized during optimization so that the optimal hyperplane can be found.

When a dataset is not linearly separable, the data is first projected onto a higher-dimensional space so that the data can be separated linearly. Once the optimal hyperplane is found in the higher dimensional space, the data is then mathematically transformed back to its original dimensions. Different kernels are used to find the correct mathematical transformation for a given dataset. Examples of a few kernels can be seen in Figure A-12. For the work in Chapter 2, a linear kernel was used.

A.3.7 Gradient Boosting (grad-boost)

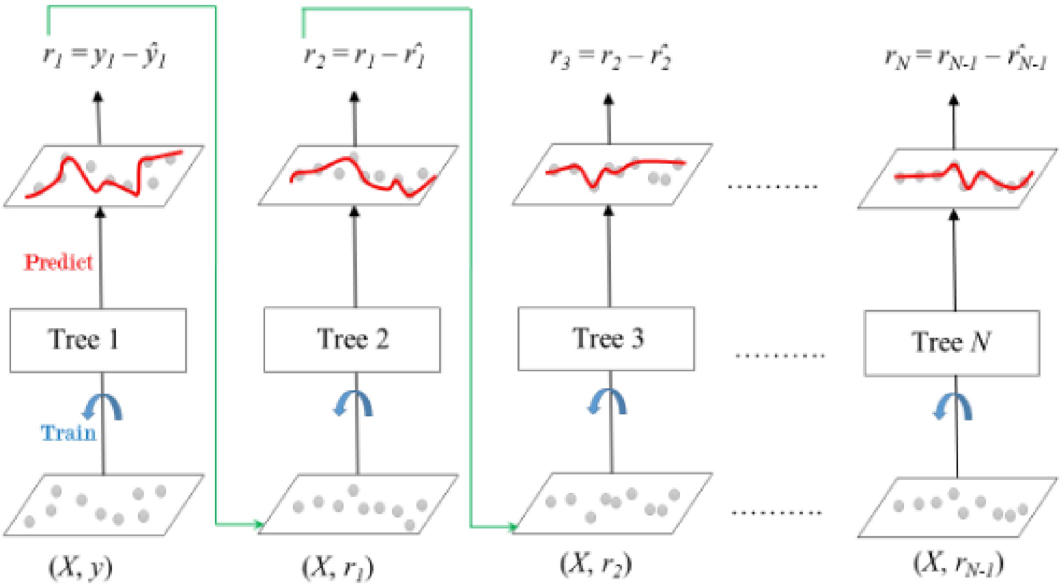


Figure A-13: Figure from (gee, 2023). The figure shows a high-level illustration of the training process for a gradient boosting model.

Gradient Boosting can be used for both classification and regression. As Figure A-13 shows, Gradient Boosting creates a model from an ensemble of weaker models. Gradient boosting models attempt to optimize a specified loss function by using decision trees as the weak models and adding one model at a time. A weak model is trained on the input data, the error/loss is computed, and gradient descent is used to add a weak model to the ensemble of trees that reduces the loss. More weight is placed on observations that are difficult to classify and the next weak model is trained on those observations to improve the overall model performance on those instances. Weak models continue to be added until a set number of trees are created or the loss reaches a certain value.

A.3.8 Logistic Regression

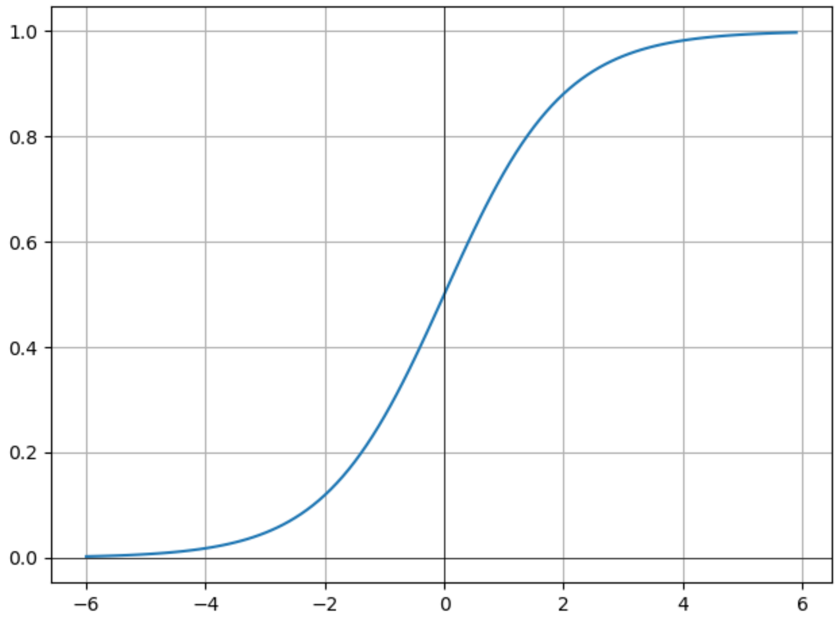


Figure A-14: The plot shows the standard logistic function, a function used to model probabilities for the logistic regression model.

The logistic regression model is a linear model that is used for classification. A logistic function (Figure A-14) is used to model the probabilities that describe possible outcomes. Similar to the LR model described in section A.3.2, the regression coefficient values are calculated from the dependent and independent variables in the dataset. However, LR models can be used to compute actual values for continuous variables, such as price or age, while the logistic regression model is used to predict a set label for classification (e.g. 0 or 1, “yes” or “no”, etc.). Regularization is commonly applied when the logistic regression model is used. Because of the computational simplicity of the model, the calculations are more transparent and interpretable compared to deep learning models.

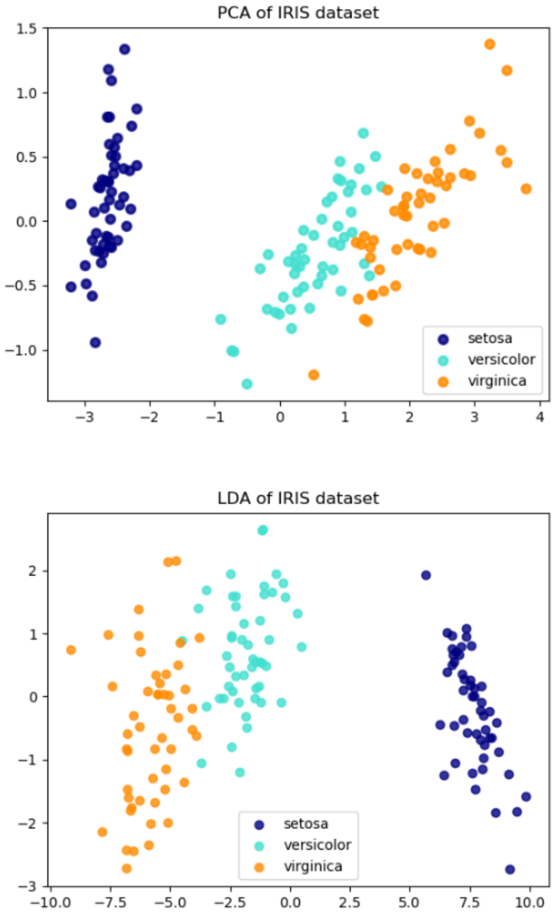


Figure A-15: Figure from the scikit-learn website (sci, a). The plot shows (1) the first two principal components after PCA is applied to the data samples, which consists of four attributes for three different kinds of flowers (top) and (2) the LDA components

A.4 Dimensionality Reduction

Both LDA and principal component analysis (PCA) were used to apply dimensionality reduction to the input features before training the classifiers in Chapter 2. The LDA classifier described in section A.3.1 can also be used for supervised dimensionality reduction. The technique involves projecting the data onto a linear subspace by identifying aspects of the data that result in the most variance between the classes, thus maximizing the separation between classes. The data has to be projected to a dimension smaller than the number of classes, making LDA a relatively strong form of dimensionality reduction.

PCA applies singular value decomposition (SVD) to the data to project it onto a subspace with a smaller dimension by finding the aspects (principal components) that result in the most variance in the data. Figure A-15 shows a comparison of the plots of the first component vs. the second, both computed using PCA (top) and LDA (bottom). In the figures, the dimensionality reduction techniques are applied to the iris dataset, which consists of four features for three different types of flowers, and the four original features are projected onto two dimensions.

The main differences between PCA and LDA, in addition to how the components are computed, are the fact that LDA is supervised and requires class labels, while PCA does not, and the fact that LDA reduces the dimensions to a value less than the number of classes, while the number of PCA components does not depend on the number of classes.

A.5 Discrete Fourier Transform (DFT)

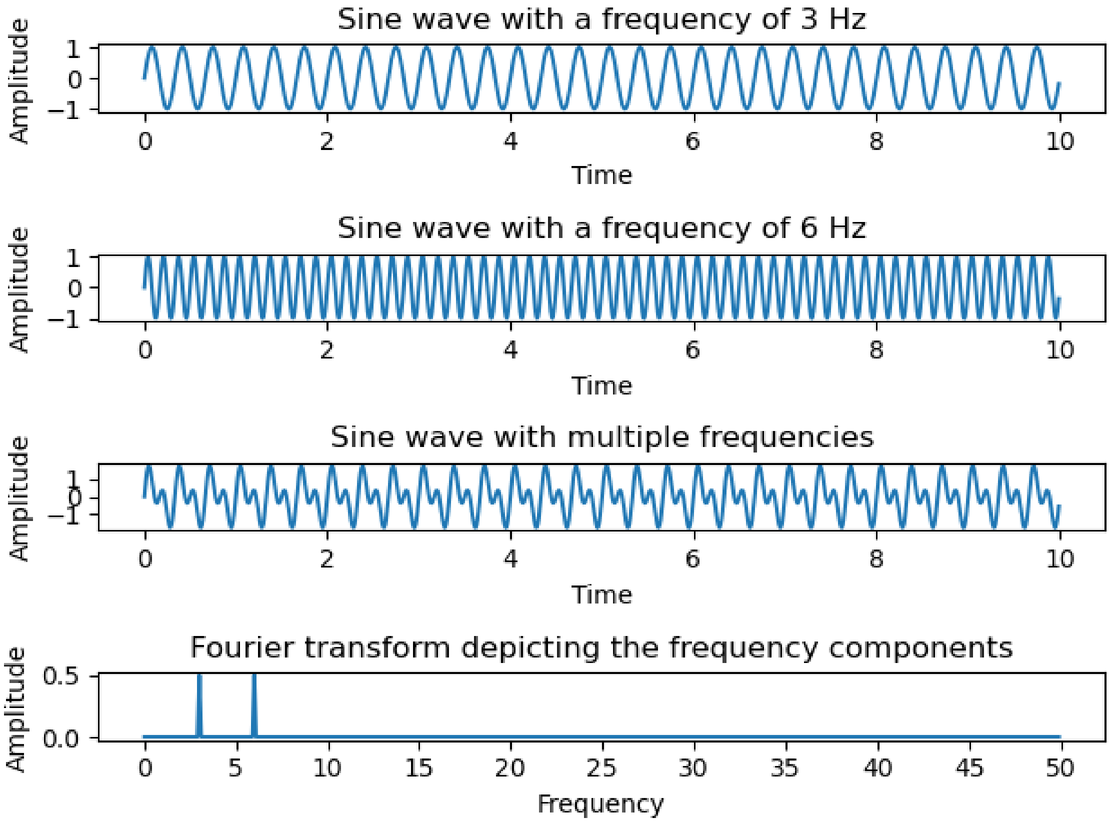


Figure A-16: The DFT of a signal that consists of the combination of two sine waves, one with a frequency of three Hertz (Hz) and one with a frequency of six Hz. The DFT (bottom subplot) shows that there are peaks at three and six Hz, as expected.

The DFT is a tool that allows you to extract information about the frequency content of a time-domain signal, meaning that the DFT of a signal is a representation of the original signal in the frequency domain. The DFT is used in many different fields for practical applications, such as digital signal processing and image processing. It can be used to perform convolutions and solve differential equations, as well as perform other operations that would be more complicated in the time domain.

The Python NumPy package (Harris et al., 2020) was used to compute the one-dimensional DFTs shown in Chapter 4. Figure A-16 shows the DFT of a signal that consists of the combination of two sine waves, one with a frequency of three Hertz (Hz) and one with a frequency of six Hz. The DFT (bottom subplot) shows that there are peaks at three and six Hz, as expected. This example illustrates the purpose of the DFT, which is to identify the frequencies present in the signal. More information about the implementation details for the NumPy DFT function can be found in the documentation on the NumPy website (num).

A.6 Metrics

In this section, a few of the metrics used in previous chapters are described in more detail.

A.6.1 AUC

The datasets for rare diseases like FTD and PPA tend to have class imbalances, with there being more healthy speakers than impaired speakers. This is reflective of the general population and can make it difficult to judge how well a model performs when using a metric like accuracy, since a model can have high accuracy simply by predicting that all speakers are healthy. For this reason, we decided to compute the AUC for each of the classifiers in Chapter 5, since AUC is recommended for imbalanced datasets and summarizes the performance of the classifier across several different probability thresholds.

In Chapter 5, the AUC was computed using the scikit-learn (Pedregosa et al., 2011) function for generating the ROC curve from ground truth labels and positive class probabilities (**roc_curve**), as well as the function for computing the AUC value using the trapezoidal rule (**auc**). The ROC function sorts the probabilities in descending order and the average

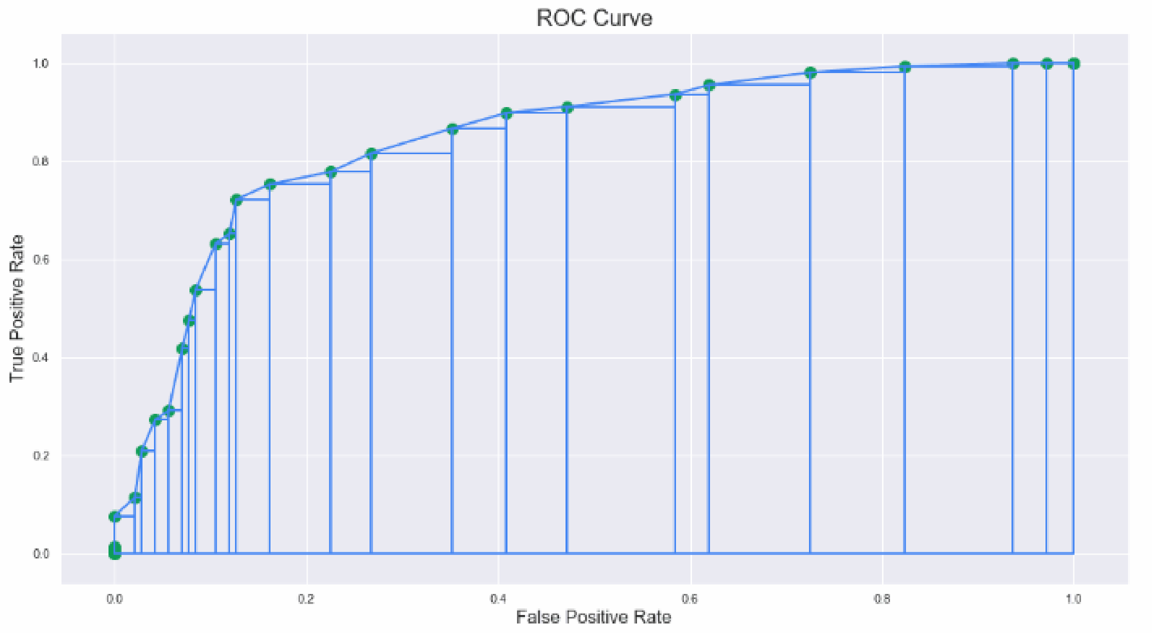


Figure A-17: Figure from (Letelier, 2021). The plot shows an ROC curve with rectangles and triangles used to approximate the area under the curve (i.e. trapezoidal method).

probabilities between consecutive points are used as the thresholds, in addition to 0 and 1. Those thresholds are used to compute the false positive rate (FPR) and true positive rate (TPR) pairs and then those pairs are returned and can be used to create the ROC curve. An example curve can be seen in Figure A-17. The AUC function uses the trapezoidal rule to compute the AUC value from the FPR/TPR pairs used to create the ROC plot. In our work, the positive class probabilities for each held-out speaker were computed during LOSO training. Those probabilities were passed into the ROC curve function, along with the ground truth labels, and used to compute the AUC results that are presented in Chapter 5.

A.6.2 McNemar's Test

C_1/C_2	Correct	Incorrect
Correct	a	b
Incorrect	c	d

Table A.5: The structure of the contingency table used for McNemar's test when comparing two classifiers, C_1 and C_2.

McNemar's Test (Gillick and Cox, 1989) is a significance test that can be used to compare the results from different classifier algorithms. The test measures the probability that two algorithms are the same by keeping track of which data samples the models disagree on (e.g. which samples one model makes correct predictions for and the other makes incorrect predictions for, and vice versa).

The first step of the test involves creating a contingency table similar to what is shown in Table A.5, where b is the number of data samples that C_1 classifies correctly that are not correctly classified by C_2, and vice versa for c. Once b and c are computed, the probability of both algorithms being the same can be computed using the following equations:

$$P = \sum_{k=0}^{l} P(k) + \sum_{k=m}^{n} P(k)$$

$$P(k) = \binom{n}{k} \left(\frac{1}{2}\right)^{n}$$

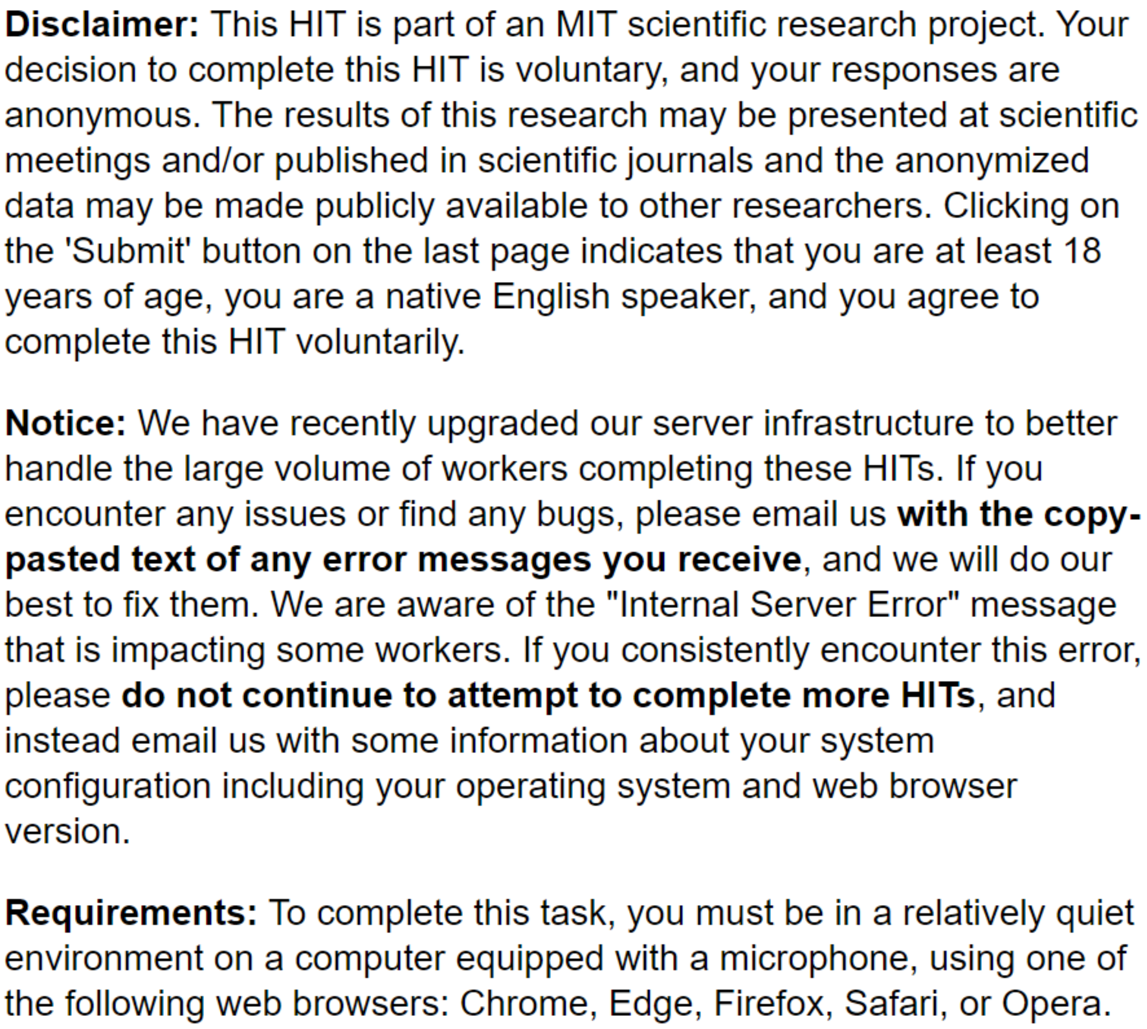
where $n = b + c$, $l = min(b,c)$, and $m = max(b,c)$.

This approach was used to compute the statistical significance of the results in Chapter 5. In our case, the classification algorithms were the same but the input data was perturbed in some way (e.g. interviewer speech was removed from the transcripts) and the predictions of the the classifier before and after the experimental change were used to compute the statistical significance of the change.

Appendix B

Supplementary Material For Chapter 3: Screenshots of the Tasks That Workers Completed For Our AMT HIT

The following figures show screenshots of the different components of the AMT HIT that workers were asked to complete. The screenshots are included here for replication purposes. The template that we used as a starting point for our AMT task has since been published as an open-source toolkit called Speak, which can be found at https://github.com/soupdtag/speak-tool (Song et al., 2022).

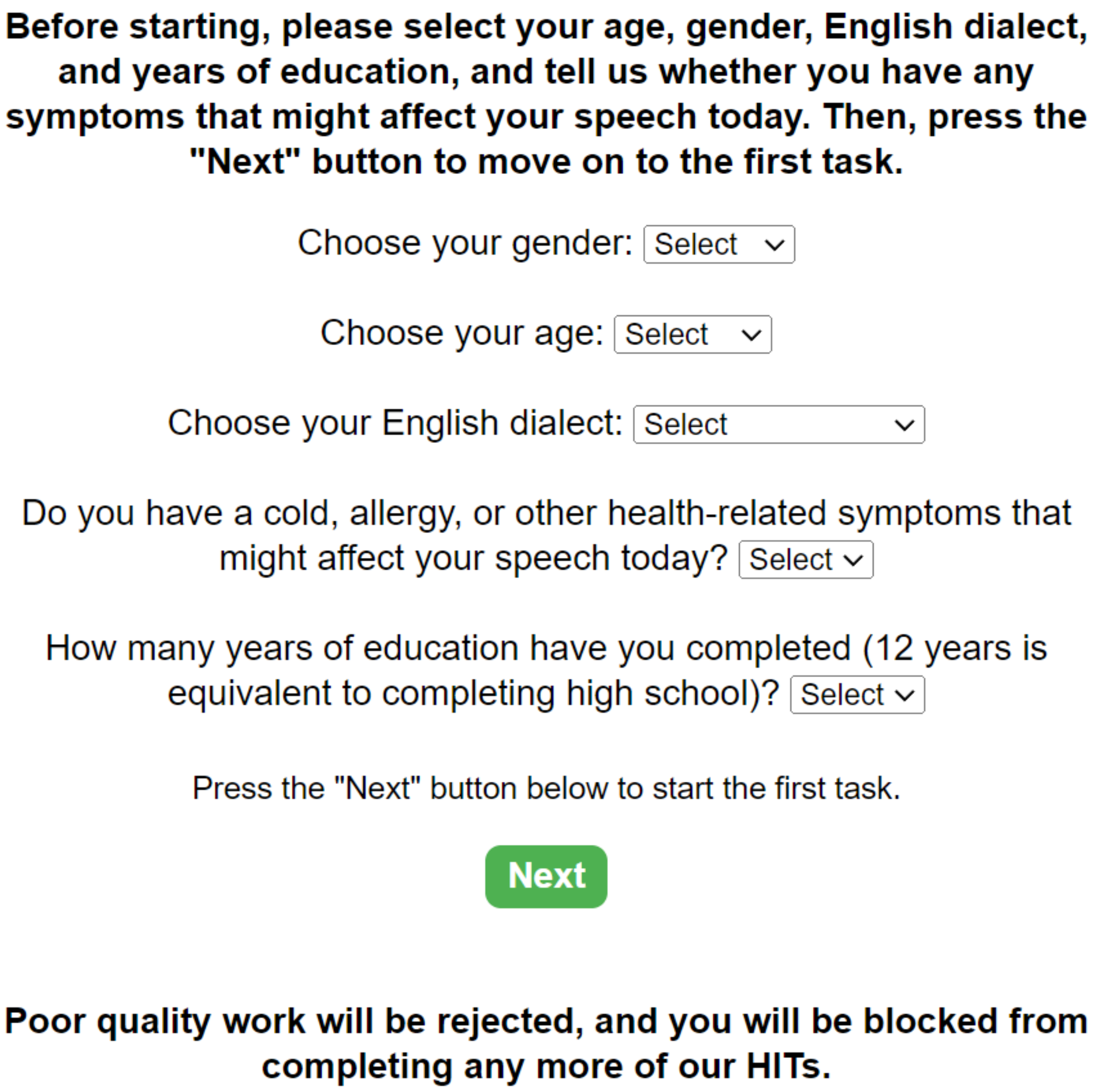


Disclaimer: This HIT is part of an MIT scientific research project. Your decision to complete this HIT is voluntary, and your responses are anonymous. The results of this research may be presented at scientific meetings and/or published in scientific journals and the anonymized data may be made publicly available to other researchers. Clicking on the 'Submit' button on the last page indicates that you are at least 18 years of age, you are a native English speaker, and you agree to complete this HIT voluntarily.

Notice: We have recently upgraded our server infrastructure to better handle the large volume of workers completing these HITs. If you encounter any issues or find any bugs, please email us **with the copy-pasted text of any error messages you receive**, and we will do our best to fix them. We are aware of the "Internal Server Error" message that is impacting some workers. If you consistently encounter this error, please **do not continue to attempt to complete more HITs**, and instead email us with some information about your system configuration including your operating system and web browser version.

Requirements: To complete this task, you must be in a relatively quiet environment on a computer equipped with a microphone, using one of the following web browsers: Chrome, Edge, Firefox, Safari, or Opera.

Figure B-1: Information that is placed at the top of every page of the HIT.



Before starting, please select your age, gender, English dialect, and years of education, and tell us whether you have any symptoms that might affect your speech today. Then, press the "Next" button to move on to the first task.

Choose your gender: Select

Choose your age: Select

Choose your English dialect: Select

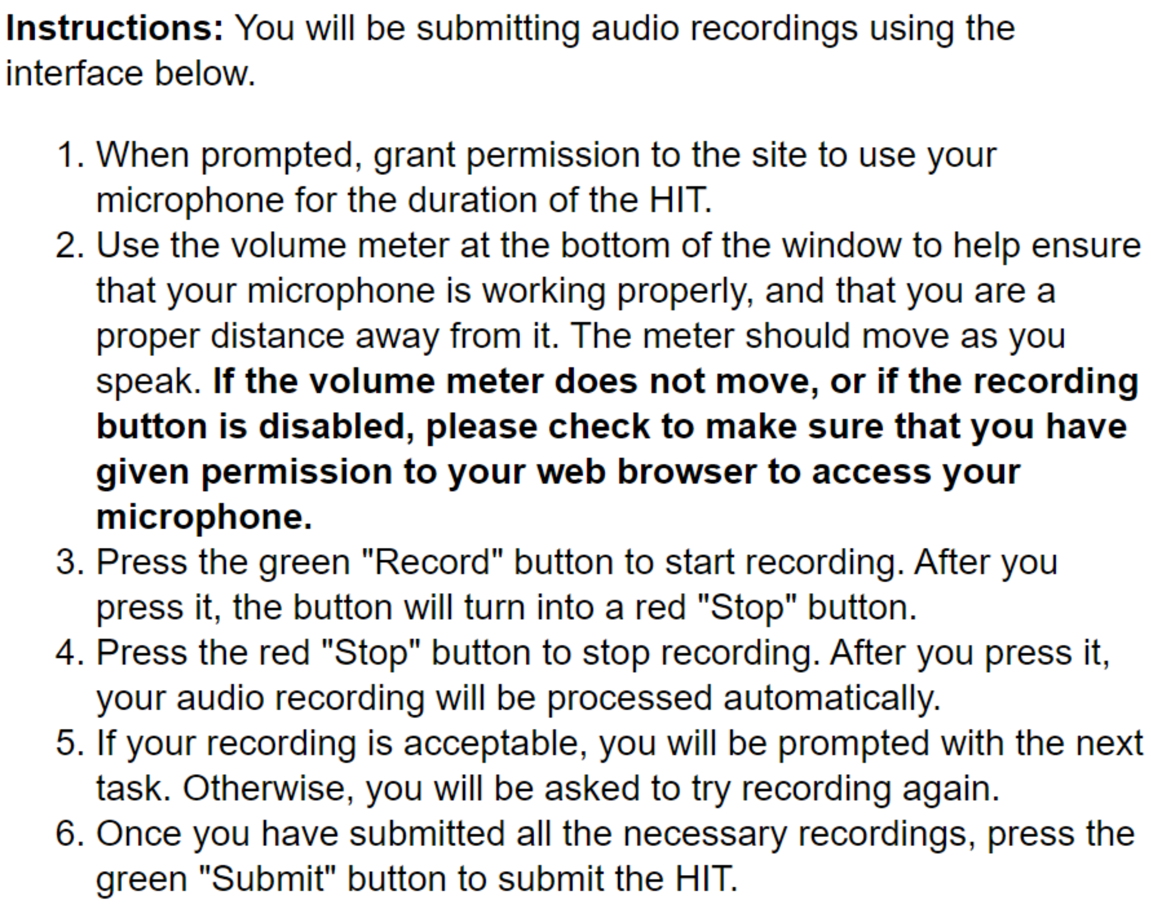
Do you have a cold, allergy, or other health-related symptoms that might affect your speech today? Select

How many years of education have you completed (12 years is equivalent to completing high school)? Select

Press the "Next" button below to start the first task.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

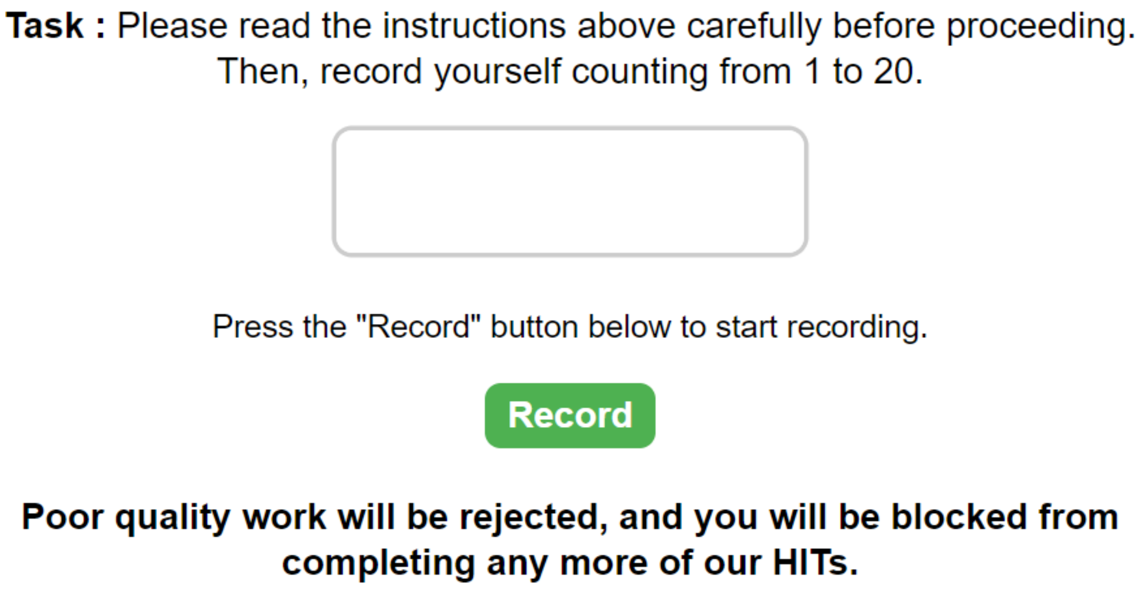
Figure B-2: The metadata task.



Instructions: You will be submitting audio recordings using the interface below.

1. When prompted, grant permission to the site to use your microphone for the duration of the HIT.
2. Use the volume meter at the bottom of the window to help ensure that your microphone is working properly, and that you are a proper distance away from it. The meter should move as you speak. **If the volume meter does not move, or if the recording button is disabled, please check to make sure that you have given permission to your web browser to access your microphone.**
3. Press the green "Record" button to start recording. After you press it, the button will turn into a red "Stop" button.
4. Press the red "Stop" button to stop recording. After you press it, your audio recording will be processed automatically.
5. If your recording is acceptable, you will be prompted with the next task. Otherwise, you will be asked to try recording again.
6. Once you have submitted all the necessary recordings, press the green "Submit" button to submit the HIT.

Figure B-3: The instructions for submitting audio recordings, which are placed on each task page.

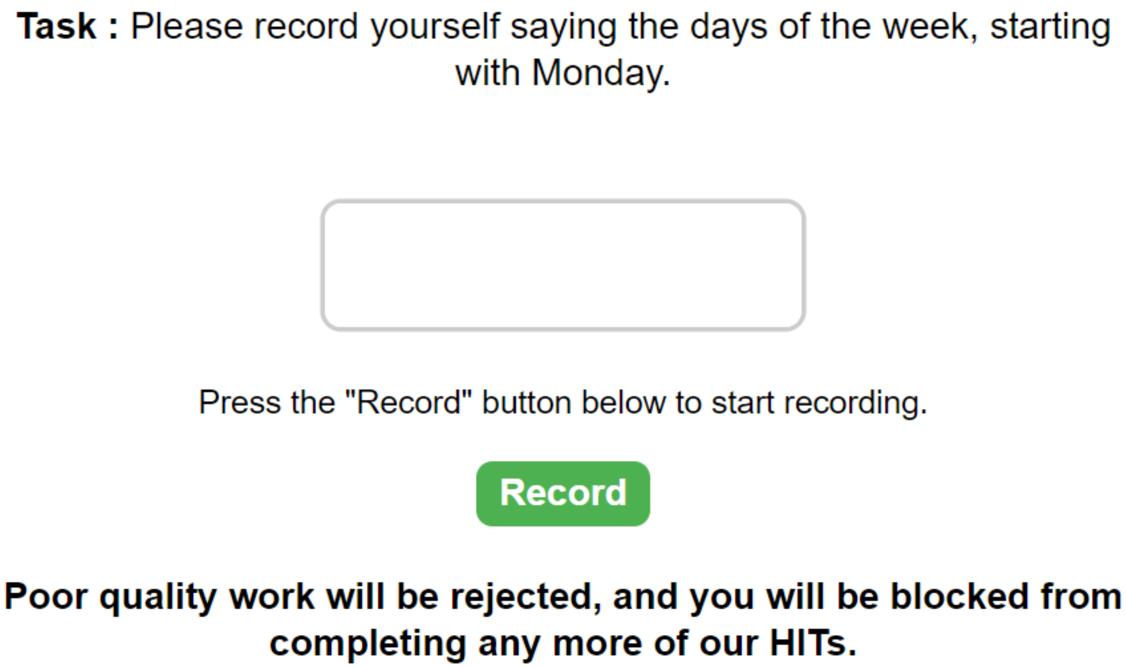


Task : Please read the instructions above carefully before proceeding. Then, record yourself counting from 1 to 20.

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-4: The counting from 1 to 20 task.

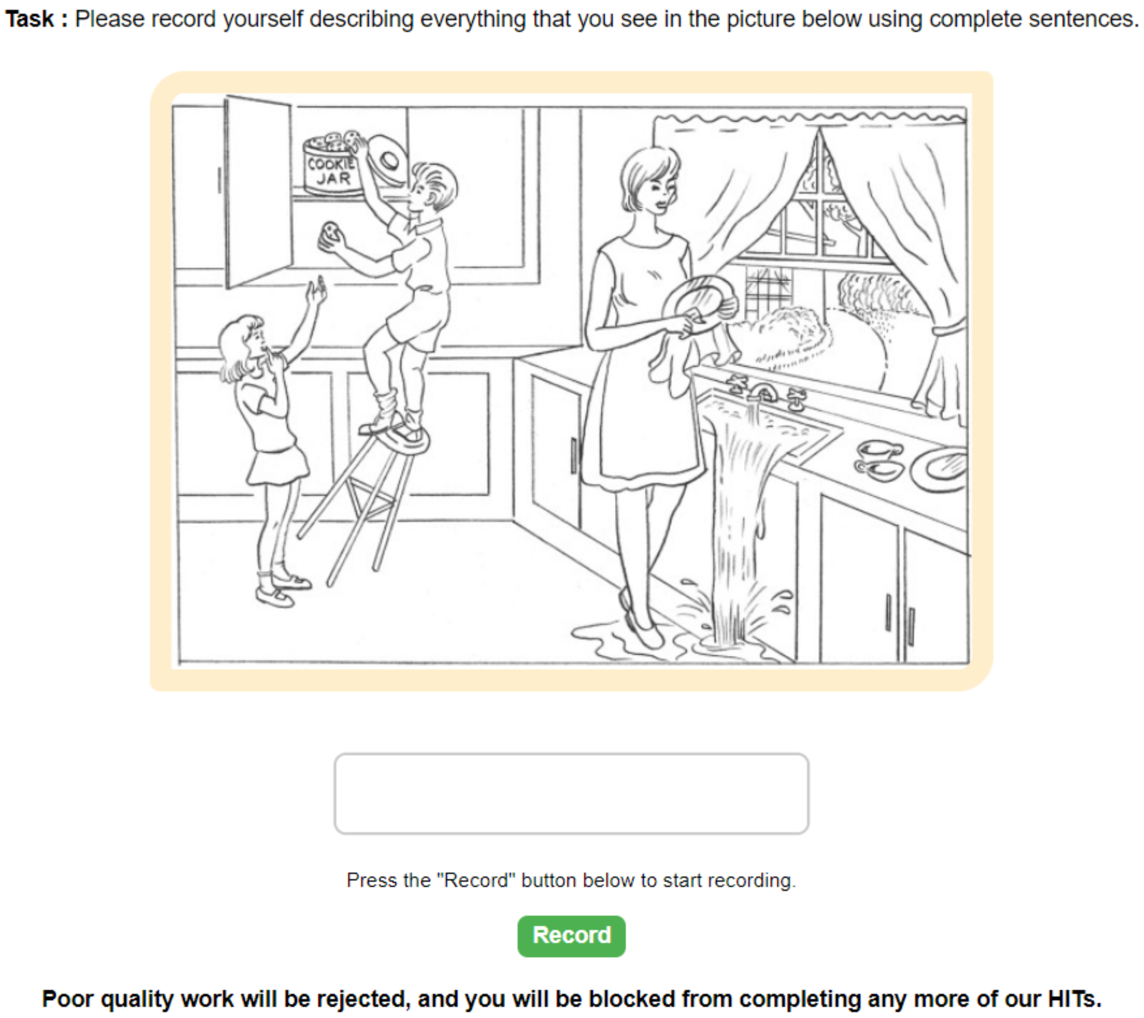


Task : Please record yourself saying the days of the week, starting with Monday.

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-5: The days of the week task.

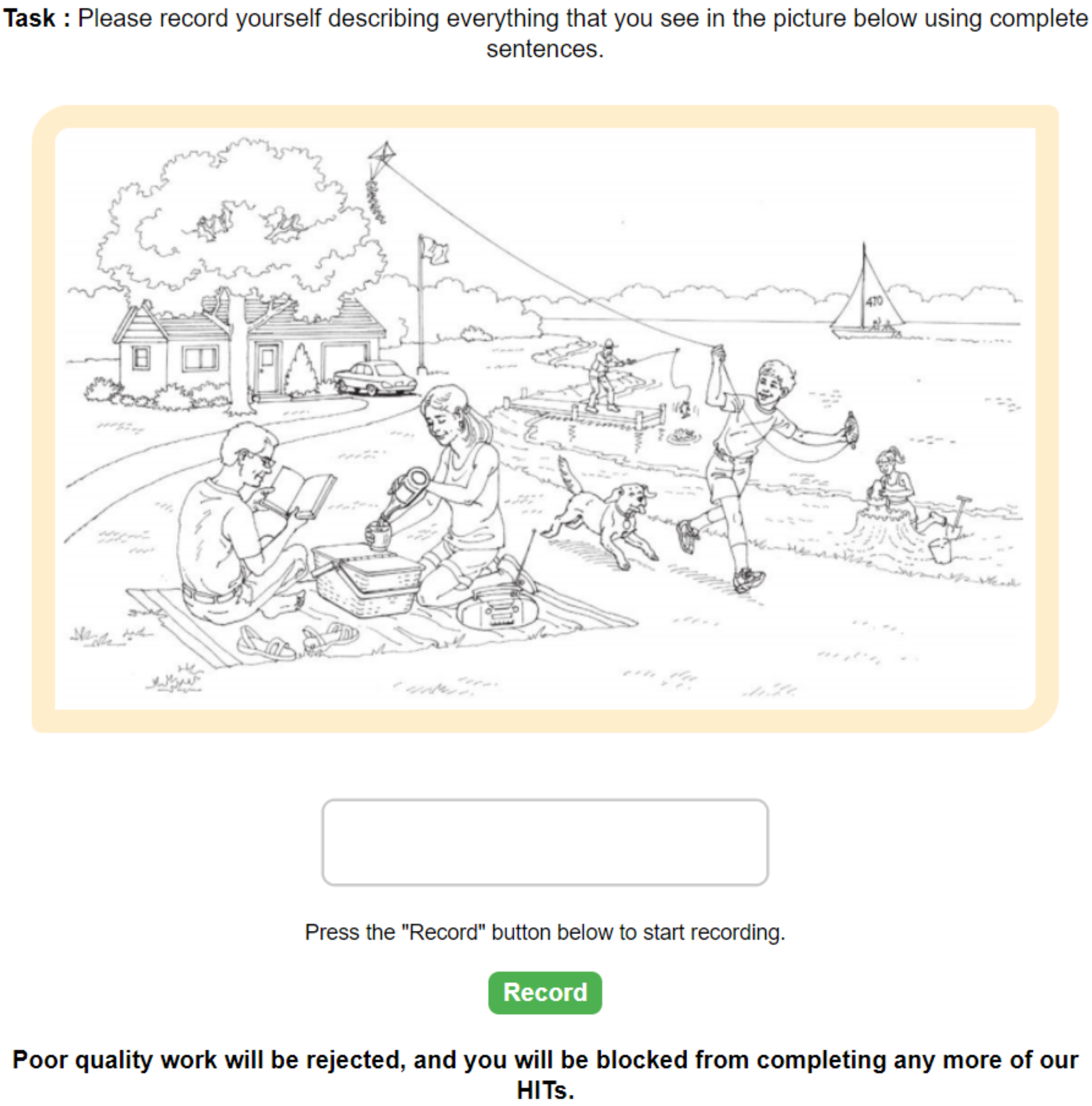


Task : Please record yourself describing everything that you see in the picture below using complete sentences.

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-6: The cookie theft picture description task.

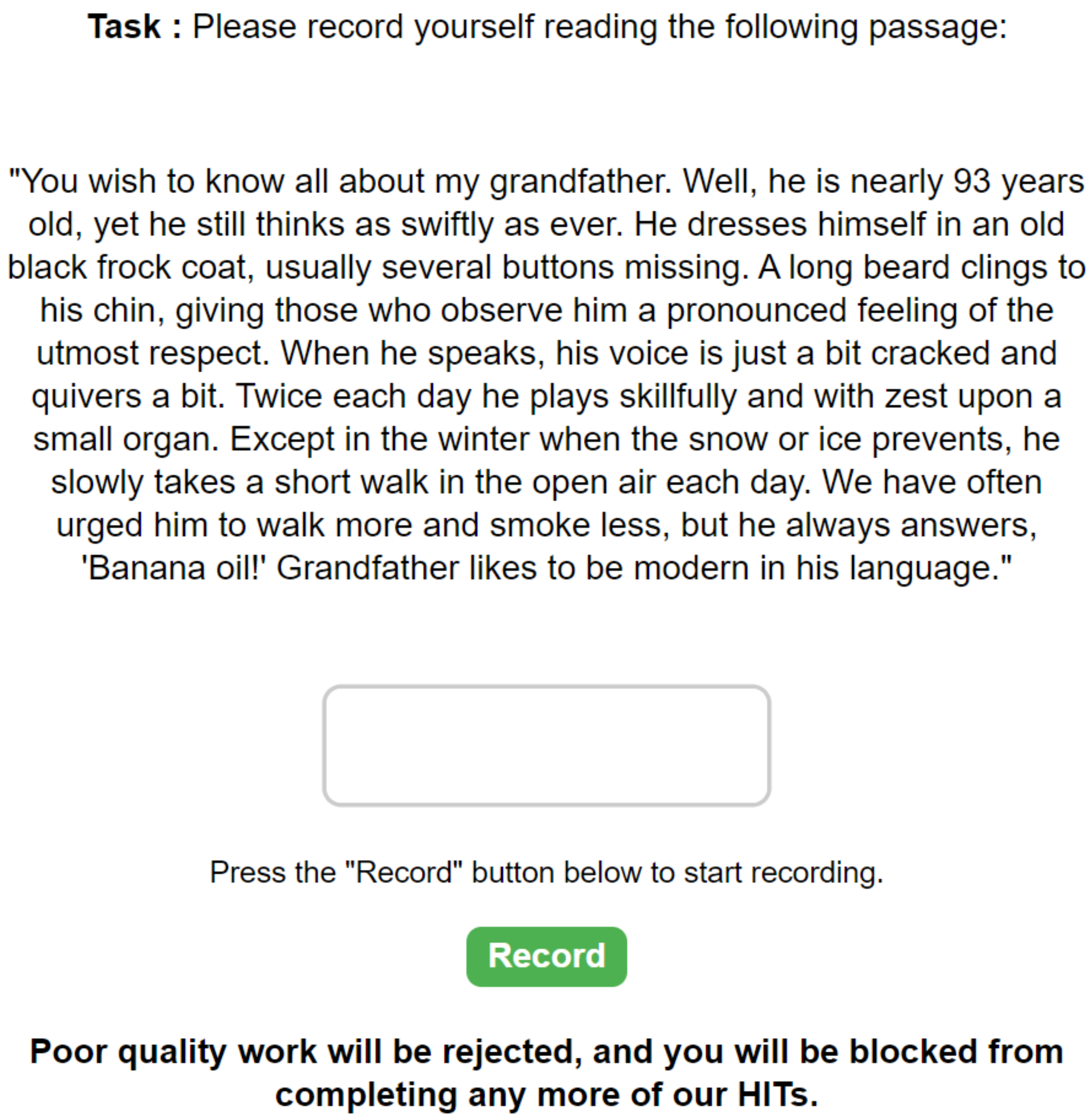


Task : Please record yourself describing everything that you see in the picture below using complete sentences.

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-7: The picnic picture description task.



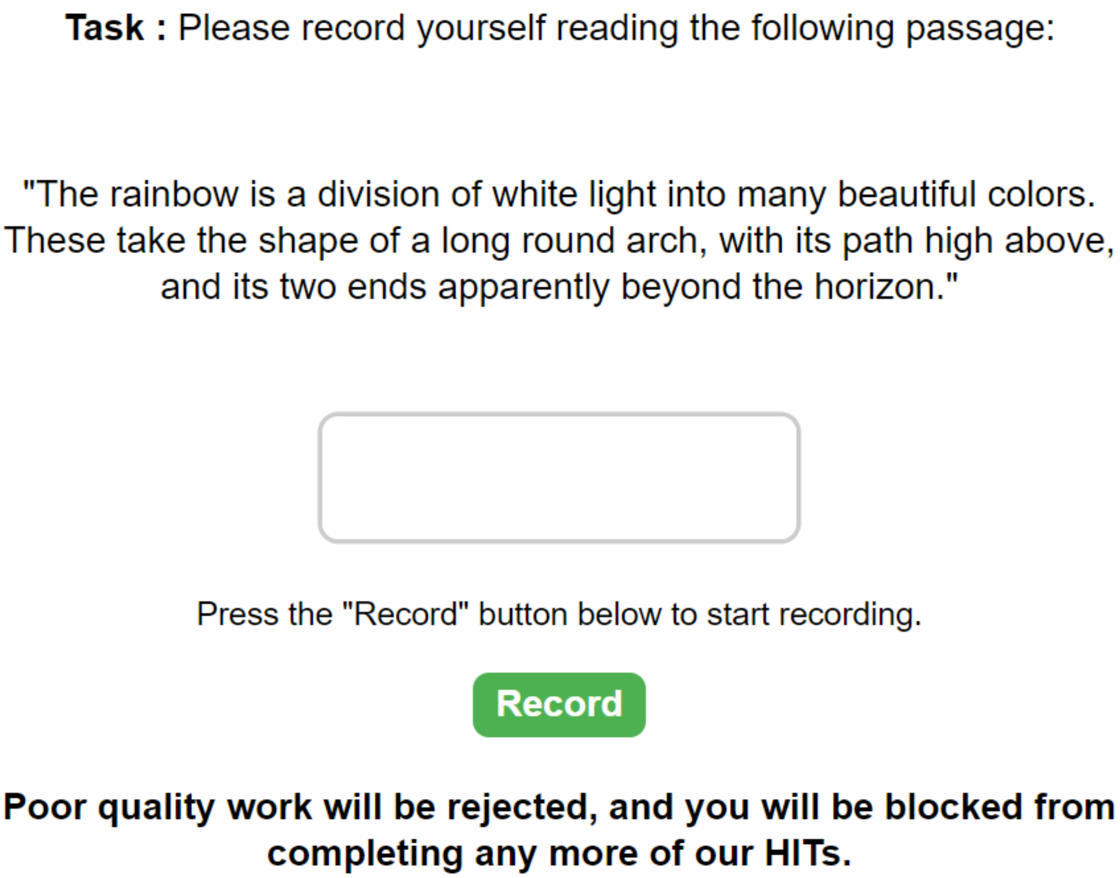
Task : Please record yourself reading the following passage:

"You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an old black frock coat, usually several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a bit. Twice each day he plays skillfully and with zest upon a small organ. Except in the winter when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, 'Banana oil!' Grandfather likes to be modern in his language."

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-8: The grandfather reading task.



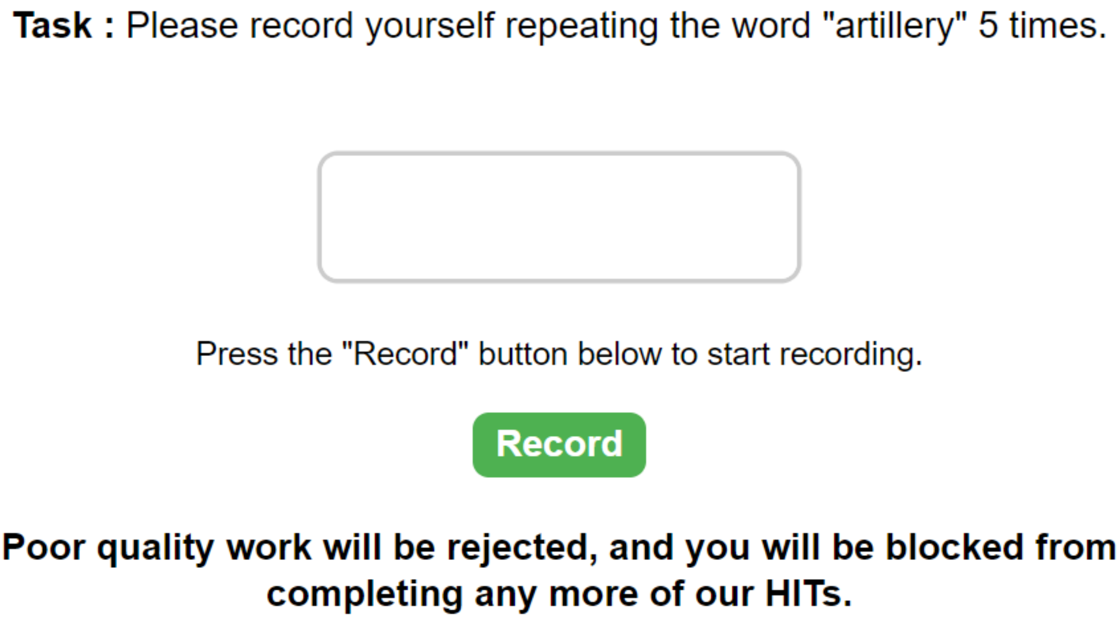
Task : Please record yourself reading the following passage:

"The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon."

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-9: The rainbow reading task.

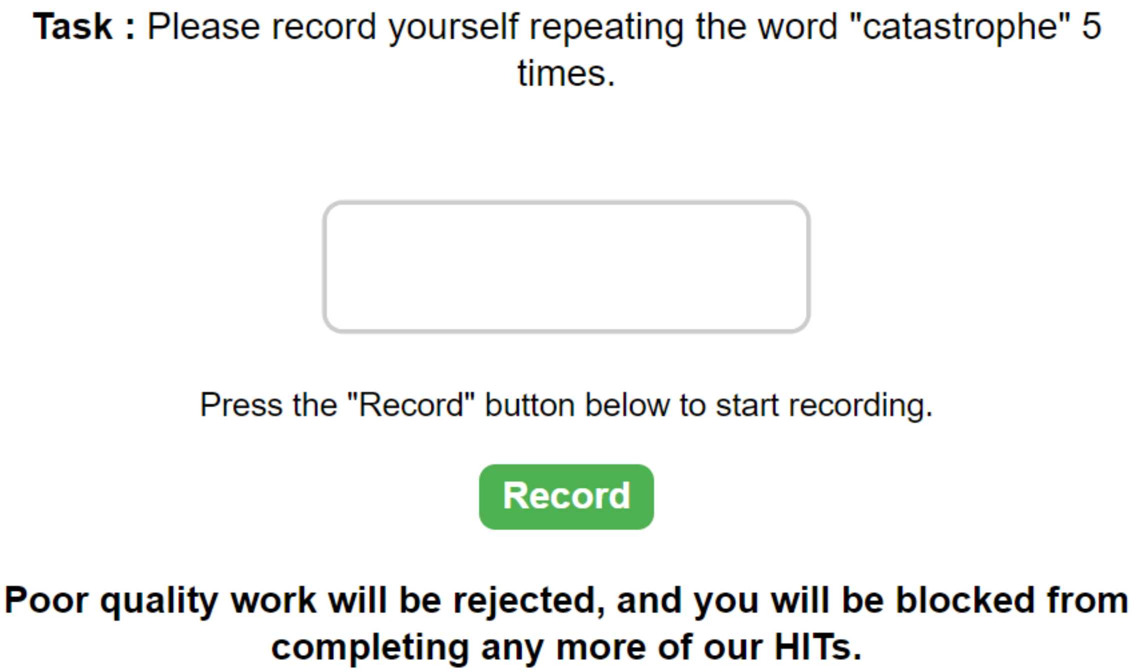


Task : Please record yourself repeating the word "artillery" 5 times.

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-10: The task for repeating the word “Artillery” five times.

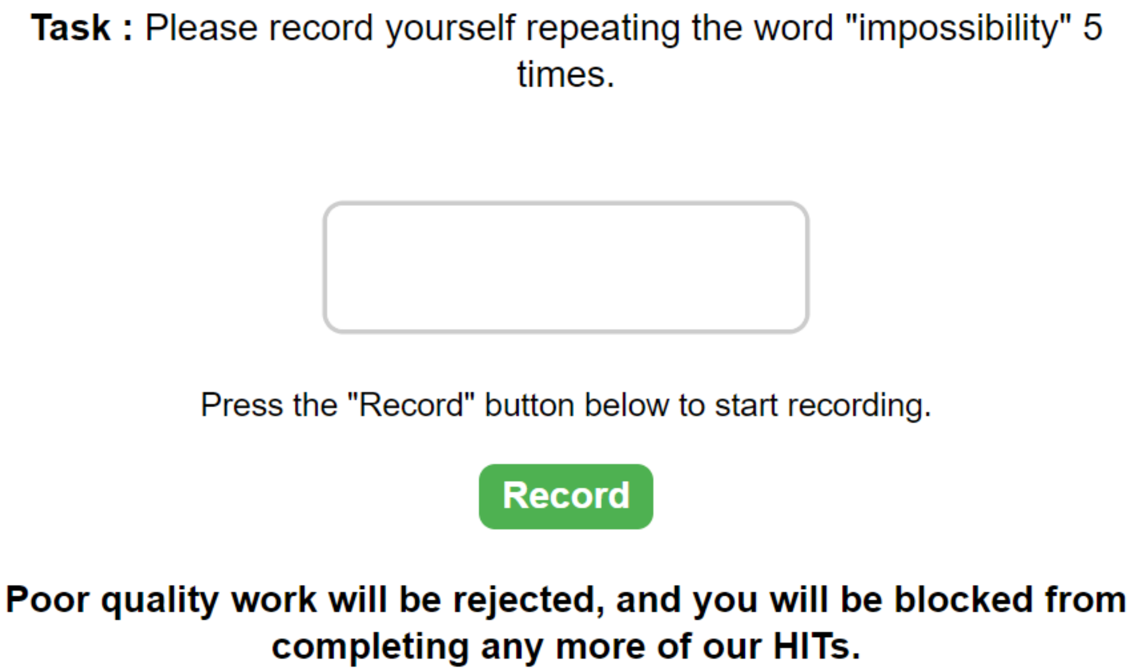


Task : Please record yourself repeating the word "catastrophe" 5 times.

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-11: The task for repeating the word “Catastrophe” five times.

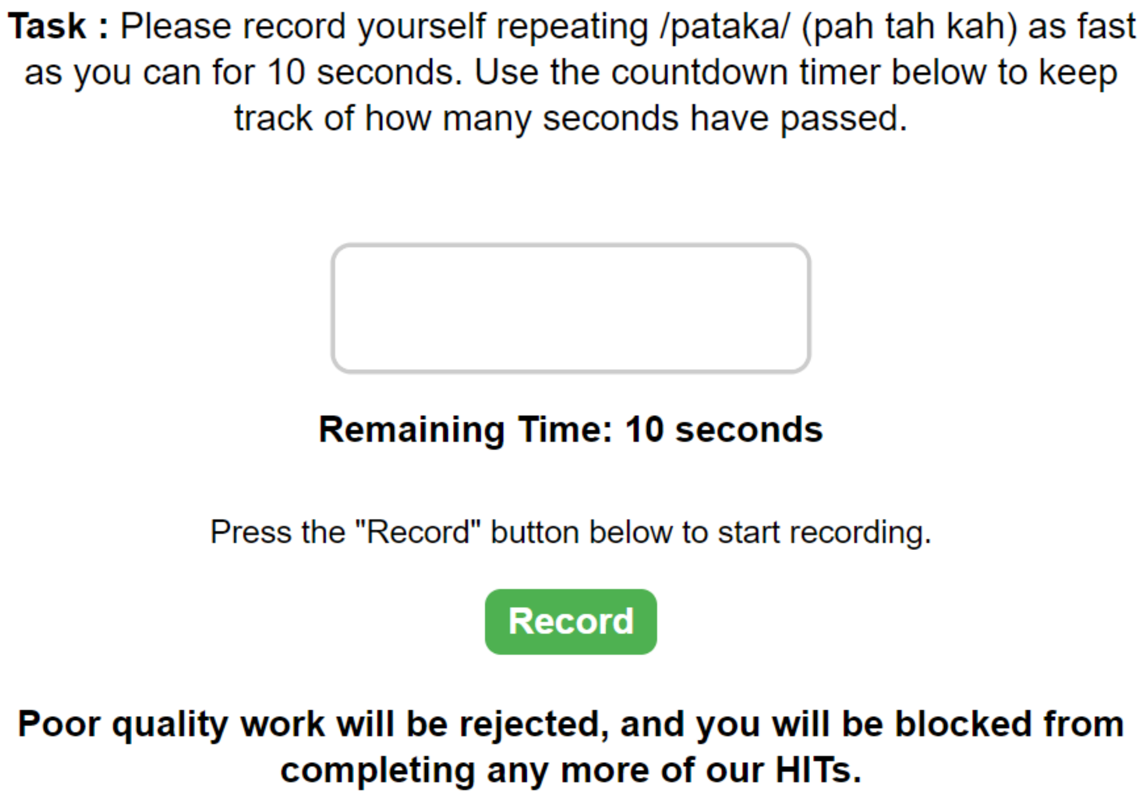


Task : Please record yourself repeating the word "impossibility" 5 times.

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-12: The task for repeating the word “Impossibility” five times.



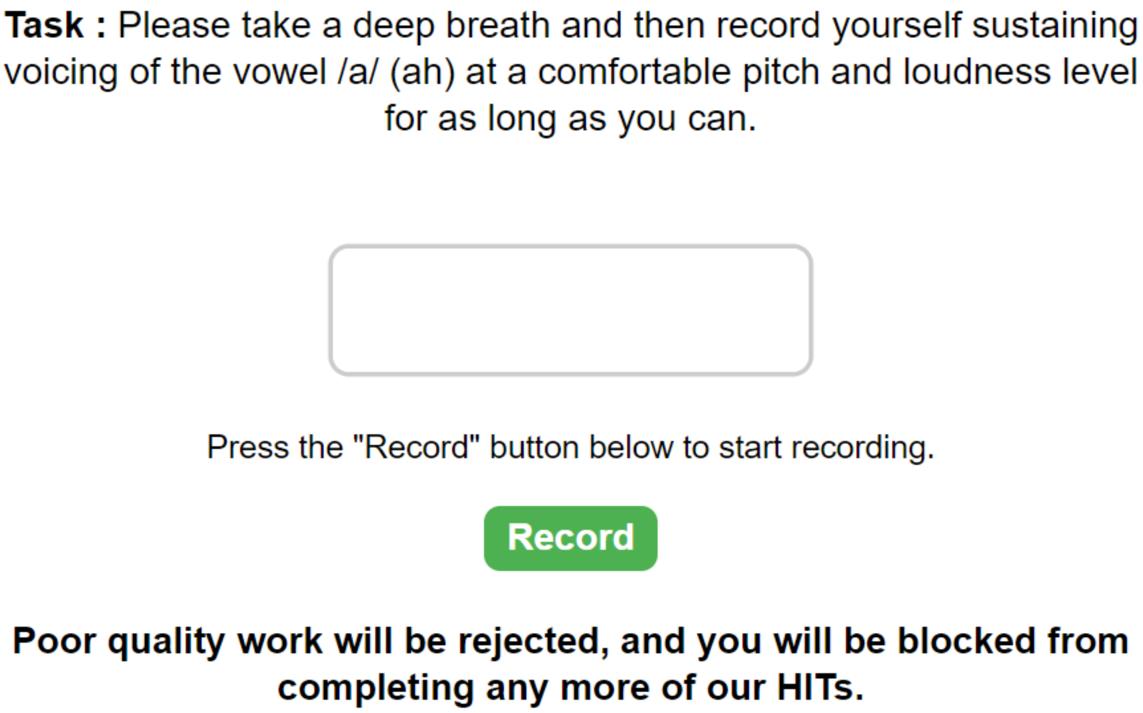
Task : Please record yourself repeating /pataka/ (pah tah kah) as fast as you can for 10 seconds. Use the countdown timer below to keep track of how many seconds have passed.

Remaining Time: 10 seconds

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-13: The task for repeating "pah tah kah" as fast as possible for 10 seconds.

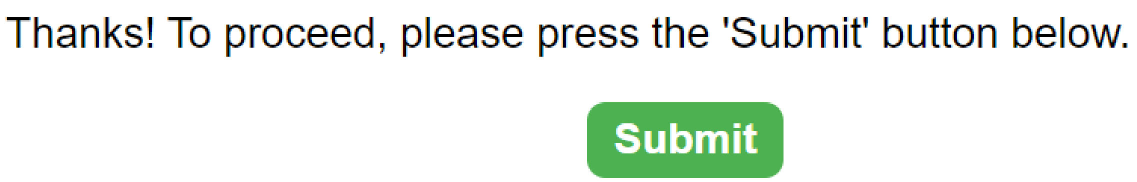


Task : Please take a deep breath and then record yourself sustaining voicing of the vowel /a/ (ah) at a comfortable pitch and loudness level for as long as you can.

Press the "Record" button below to start recording.

Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.

Figure B-14: The sustained phonation task.



Thanks! To proceed, please press the 'Submit' button below.

Figure B-15: The final page where workers submit their work.

Appendix C

Supplementary Material For Chapter 5: Full Result Tables

The full tables of results mentioned in Chapter 5. The highest values for each metric are bolded for each diagnostic group. Accuracies are computed on the independent sample level, where the predicted label for each data sample is used instead of majority voting.

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
blabla	lvPPA	0.865	0.934	0.933	0.932	0.932
	nfvPPA	**0.853**	**0.901**	**0.944**	**0.933**	**0.936**
	bvFTD	0.663	0.684	0.697	0.676	0.678
	svPPA	0.643	0.703	0.734	0.712	0.712
B	lvPPA	0.824	**0.976**	**0.976**	**0.976**	**0.976**
	nfvPPA	0.824	0.803	0.887	0.864	0.860
	bvFTD	**0.867**	**0.937**	**0.944**	**0.939**	**0.939**
	svPPA	**0.771**	**0.839**	**0.871**	**0.855**	**0.855**
TE	lvPPA	0.730	0.800	0.797	0.794	0.794
	nfvPPA	0.750	0.737	0.795	0.771	0.773
	bvFTD	0.714	0.787	0.807	0.796	0.794
	svPPA	0.700	0.804	0.828	0.803	0.803
DRS	lvPPA	0.811	0.912	0.908	0.908	0.908
	nfvPPA	0.779	0.776	0.851	0.844	0.844
	bvFTD	0.776	0.823	0.849	0.827	0.826
	svPPA	0.686	0.722	0.729	0.710	0.710
DRT	lvPPA	0.851	0.938	0.935	0.934	0.934
	nfvPPA	0.794	0.824	0.902	0.893	0.895
	bvFTD	0.663	0.749	0.773	0.747	0.748
	svPPA	0.714	0.742	0.740	0.731	0.731
DBS	lvPPA	0.838	0.958	0.957	0.957	0.957
	nfvPPA	0.838	0.807	0.898	0.867	0.862
	bvFTD	0.633	0.758	0.781	0.765	0.770
	svPPA	0.757	0.747	0.744	0.734	0.734
DBT	lvPPA	**0.865**	0.959	0.961	0.960	0.960
	nfvPPA	0.794	0.825	0.935	0.905	0.889
	bvFTD	0.776	0.842	0.863	0.849	0.847
	svPPA	0.757	0.764	0.751	0.743	0.743

Table C.1: LOSO results for a logistic regression model trained on blabla features and different transformer embeddings extracted from sentence-level manual transcripts with interviewer speech included. (Task: MONL, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
blabla	lvPPA	0.877	0.941	0.940	0.940	0.940
	bvFTD	**0.853**	0.915	0.941	0.922	0.922
	svPPA	**0.851**	**0.845**	0.802	0.796	0.796
B	lvPPA	**0.892**	**0.978**	**0.974**	**0.974**	**0.974**
	bvFTD	0.827	**0.938**	0.967	**0.948**	**0.948**
	svPPA	0.761	0.826	0.816	0.807	0.807
TE	lvPPA	0.877	0.904	0.905	0.899	0.899
	bvFTD	0.667	0.799	0.879	0.839	0.839
	svPPA	0.731	0.791	0.718	0.704	0.713
DRS	lvPPA	0.800	0.921	0.928	0.928	0.928
	bvFTD	0.787	0.854	0.935	0.902	0.902
	svPPA	0.821	0.822	0.787	0.764	0.764
DRT	lvPPA	0.877	0.938	0.937	0.937	0.937
	bvFTD	0.720	0.787	0.879	0.829	0.829
	svPPA	0.776	0.835	**0.836**	**0.819**	**0.813**
DBS	lvPPA	0.877	0.943	0.934	0.934	0.934
	bvFTD	0.787	0.889	**0.971**	0.937	0.937
	svPPA	0.761	0.755	0.761	0.718	0.718
DBT	lvPPA	0.846	0.956	0.948	0.948	0.948
	bvFTD	0.773	0.865	0.943	0.898	0.898
	svPPA	0.776	0.722	0.718	0.693	0.690

Table C.2: LOSO results for a logistic regression model trained on blabla features and different transformer embeddings extracted from sentence-level manual transcripts with interviewer speech included. (Task: COOK, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
blabla	lvPPA	0.824	0.915	0.910	0.910	0.910
	nfvPPA	**0.824**	**0.824**	**0.871**	**0.856**	**0.871**
	bvFTD	0.694	0.716	0.717	0.697	0.700
	svPPA	0.743	0.800	0.786	0.779	0.779
B	lvPPA	0.757	0.913	0.911	0.910	0.910
	nfvPPA	0.735	0.755	0.824	0.804	0.807
	bvFTD	**0.755**	**0.867**	**0.897**	**0.881**	**0.880**
	svPPA	0.657	0.699	0.697	0.681	0.681
TE	lvPPA	0.730	0.814	0.805	0.805	0.805
	nfvPPA	0.676	0.608	0.700	0.667	0.645
	bvFTD	0.694	0.746	0.769	0.755	0.754
	svPPA	**0.800**	**0.875**	**0.887**	**0.876**	**0.876**
DRS	lvPPA	0.743	0.844	0.849	0.844	0.844
	nfvPPA	0.735	0.743	0.825	0.802	0.793
	bvFTD	0.612	0.677	0.723	0.693	0.690
	svPPA	0.771	0.817	0.804	0.803	0.803
DRT	lvPPA	0.703	0.771	0.782	0.771	0.771
	nfvPPA	0.809	0.743	0.844	0.825	0.818
	bvFTD	0.643	0.693	0.730	0.704	0.706
	svPPA	0.757	0.792	0.783	0.776	0.776
DBS	lvPPA	**0.851**	**0.936**	**0.932**	**0.932**	**0.932**
	nfvPPA	0.735	0.719	0.827	0.789	0.778
	bvFTD	0.694	0.796	0.817	0.804	0.809
	svPPA	0.757	0.795	0.787	0.782	0.782
DBT	lvPPA	0.824	0.900	0.894	0.894	0.894
	nfvPPA	0.721	0.677	0.767	0.736	0.729
	bvFTD	0.704	0.787	0.798	0.785	0.787
	svPPA	0.743	0.833	0.832	0.825	0.825

Table C.3: LOSO results for a logistic regression model trained on blabla features and different transformer embeddings extracted from sentence-level Whisper transcripts. (Task: MONL, Whisper model: large-v2 with greedy decoding, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
blabla	lvPPA	0.846	0.847	0.822	0.822	0.822
	bvFTD	0.800	0.851	0.957	0.914	0.914
	svPPA	0.821	0.703	0.638	0.624	0.624
B	lvPPA	0.862	0.970	0.966	0.966	0.966
	bvFTD	**0.907**	**0.973**	**0.993**	**0.987**	**0.987**
	svPPA	0.766	0.843	0.796	0.793	0.796
TE	lvPPA	0.954	0.968	0.966	0.966	0.966
	bvFTD	0.827	0.866	0.947	0.914	0.914
	svPPA	0.851	0.774	0.693	0.684	0.684
DRS	lvPPA	**0.969**	0.975	0.971	0.971	0.971
	bvFTD	0.827	0.904	0.963	0.931	0.931
	svPPA	0.866	0.854	0.790	0.790	0.790
DRT	lvPPA	0.938	0.956	0.948	0.948	0.948
	bvFTD	0.813	0.911	0.963	0.931	0.931
	svPPA	0.836	0.822	0.773	0.761	0.770
DBS	lvPPA	0.908	**0.983**	**0.980**	**0.980**	**0.980**
	bvFTD	0.800	0.895	0.940	0.908	0.908
	svPPA	0.866	0.814	0.764	0.756	0.756
DBT	lvPPA	0.862	0.951	0.943	0.943	0.943
	bvFTD	0.813	0.932	0.961	0.945	0.945
	svPPA	**0.896**	**0.891**	**0.845**	**0.845**	**0.845**

Table C.4: LOSO results for a logistic regression model trained on blabla features and different transformer embeddings extracted from sentence-level Whisper transcripts. (Task: COOK, Whisper model: large-v2 with greedy decoding, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	Diagnosis	Acc.	AUC (Ind.)	AUC (Max)	AUC (Mean)	AUC (Med.)
blabla	lvPPA	**0.838**	**0.933**	**0.929**	**0.929**	**0.929**
	nfvPPA	**0.838**	**0.869**	**0.909**	**0.904**	**0.907**
	bvFTD	0.663	0.701	0.703	0.686	0.688
	svPPA	0.671	0.748	0.757	0.747	0.747
B	lvPPA	0.838	0.927	0.923	0.923	0.923
	nfvPPA	0.735	0.727	0.796	0.771	0.775
	bvFTD	**0.786**	**0.871**	**0.882**	**0.872**	**0.874**
	svPPA	0.743	0.766	0.766	0.758	0.758
TE	lvPPA	0.784	0.878	0.872	0.872	0.872
	nfvPPA	0.676	0.593	0.653	0.607	0.598
	bvFTD	0.745	0.792	0.796	0.788	0.790
	svPPA	0.714	0.857	0.869	0.859	0.859
DRS	lvPPA	0.716	0.819	0.816	0.813	0.813
	nfvPPA	0.750	0.715	0.778	0.753	0.753
	bvFTD	0.694	0.754	0.811	0.778	0.784
	svPPA	0.814	0.856	0.855	0.849	0.849
DRT	lvPPA	0.730	0.834	0.829	0.828	0.828
	nfvPPA	0.706	0.680	0.745	0.724	0.731
	bvFTD	0.531	0.596	0.649	0.618	0.627
	svPPA	0.829	0.851	0.845	0.842	0.842
DBS	lvPPA	0.784	0.883	0.877	0.877	0.877
	nfvPPA	0.735	0.666	0.720	0.700	0.702
	bvFTD	0.765	0.783	0.782	0.774	0.774
	svPPA	0.814	0.885	0.878	0.876	0.876
DBT	lvPPA	0.770	0.854	0.846	0.846	0.846
	nfvPPA	0.676	0.634	0.678	0.653	0.655
	bvFTD	0.786	0.856	0.863	0.851	0.852
	svPPA	**0.843**	**0.919**	**0.920**	**0.916**	**0.916**

Table C.5: LOSO results for a logistic regression model trained on blabla features and different transformer embeddings extracted from sentence-level Whisper transcripts. (Task: MONL, Whisper model: large-v2 with beam search of 10, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
blabla	lvPPA	0.754	0.709	0.716	0.710	0.710
	bvFTD	0.787	0.850	0.938	0.897	0.897
	svPPA	0.821	0.824	0.779	0.776	0.776
B	lvPPA	0.862	**0.941**	**0.931**	**0.931**	**0.931**
	bvFTD	0.800	0.858	0.918	0.891	0.891
	svPPA	0.776	0.724	0.695	0.675	0.670
TE	lvPPA	**0.892**	0.884	0.868	0.868	0.868
	bvFTD	0.800	0.882	0.915	0.882	0.882
	svPPA	0.806	**0.881**	**0.871**	**0.853**	**0.859**
DRS	lvPPA	**0.892**	0.926	0.914	0.914	0.914
	bvFTD	**0.827**	0.903	0.973	0.941	0.941
	svPPA	0.806	0.726	0.667	0.647	0.644
DRT	lvPPA	0.815	0.823	0.802	0.799	0.799
	bvFTD	0.813	**0.928**	0.970	**0.945**	**0.945**
	svPPA	0.776	0.761	0.716	0.704	0.704
DBS	lvPPA	0.846	0.872	0.859	0.856	0.856
	bvFTD	**0.827**	0.895	**0.977**	0.944	0.944
	svPPA	0.806	0.784	0.761	0.741	0.739
DBT	lvPPA	0.862	0.810	0.787	0.784	0.784
	bvFTD	**0.827**	0.907	0.967	0.937	0.937
	svPPA	**0.821**	0.826	0.807	0.793	0.796

Table C.6: LOSO results for a logistic regression model trained on blabla features and different transformer embeddings extracted from sentence-level Whisper transcripts. (Task: COOK, Whisper model: large-v2 with beam search of 10, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
blabla	lvPPA	0.784	0.918	0.915	0.915	0.915
	nfvPPA	**0.882**	**0.933**	**0.951**	**0.945**	**0.951**
	bvFTD	0.694	0.722	0.718	0.704	0.706
	svPPA	0.729	0.721	0.712	0.699	0.699
B	lvPPA	0.811	0.941	0.939	0.938	0.938
	nfvPPA	0.750	0.738	0.809	0.796	0.805
	bvFTD	0.735	0.822	**0.853**	**0.834**	**0.836**
	svPPA	0.714	0.768	0.769	0.759	0.759
TE	lvPPA	0.784	0.876	0.876	0.871	0.871
	nfvPPA	0.676	0.526	0.587	0.547	0.542
	bvFTD	0.694	0.756	0.775	0.757	0.755
	svPPA	0.757	**0.838**	**0.857**	**0.843**	**0.843**
DRS	lvPPA	0.878	**0.972**	**0.972**	**0.972**	**0.972**
	nfvPPA	0.765	0.793	0.878	0.851	0.844
	bvFTD	0.571	0.651	0.698	0.664	0.666
	svPPA	0.743	0.768	0.751	0.751	0.751
DRT	lvPPA	0.716	0.843	0.843	0.841	0.841
	nfvPPA	0.721	0.731	0.816	0.795	0.776
	bvFTD	0.592	0.631	0.672	0.643	0.647
	svPPA	0.771	0.765	0.765	0.762	0.762
DBS	lvPPA	0.865	0.927	0.926	0.924	0.924
	nfvPPA	0.750	0.730	0.758	0.749	0.756
	bvFTD	**0.755**	**0.836**	0.845	0.833	0.835
	svPPA	**0.814**	0.834	0.821	0.817	0.817
DBT	lvPPA	**0.892**	0.953	0.952	0.952	0.952
	nfvPPA	0.809	0.838	0.882	0.856	0.842
	bvFTD	0.714	0.808	0.817	0.810	0.810
	svPPA	0.771	0.762	0.779	0.771	0.771

Table C.7: LOSO results for a logistic regression model trained on different features extracted from sentence-level Whisper transcripts. (Task: MONL, Whisper model: medium with greedy decoding, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
blabla	lvPPA	0.846	0.830	0.807	0.807	0.807
	bvFTD	0.800	0.862	0.954	0.915	0.915
	svPPA	0.851	0.862	0.833	0.819	0.819
B	lvPPA	0.862	0.929	0.922	0.917	0.917
	bvFTD	0.840	**0.974**	**0.983**	**0.980**	**0.980**
	svPPA	0.806	**0.874**	**0.839**	**0.836**	**0.833**
TE	lvPPA	0.892	0.929	0.925	0.922	0.922
	bvFTD	**0.867**	0.939	0.976	0.961	0.961
	svPPA	0.821	0.774	0.690	0.684	0.684
DRS	lvPPA	0.877	0.953	**0.971**	**0.966**	**0.966**
	bvFTD	0.813	0.884	0.935	0.908	0.908
	svPPA	0.851	0.784	0.695	0.690	0.690
DRT	lvPPA	**0.923**	**0.958**	0.963	0.960	0.960
	bvFTD	0.813	0.913	0.951	0.934	0.934
	svPPA	0.821	0.791	0.707	0.704	0.704
DBS	lvPPA	0.862	0.906	0.894	0.891	0.891
	bvFTD	0.840	0.937	0.967	0.948	0.948
	svPPA	0.821	0.724	0.681	0.667	0.667
DBT	lvPPA	0.846	0.867	0.848	0.848	0.848
	bvFTD	0.853	0.931	0.976	0.957	0.957
	svPPA	**0.881**	0.852	0.805	0.790	0.793

Table C.8: LOSO results for a logistic regression model trained on different features extracted from sentence-level Whisper transcripts. (Task: COOK, Whisper model: medium with greedy decoding, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	Diagnosis	Acc.	AUC (Ind.)	AUC (Max)	AUC (Mean)	AUC (Med.)
blabla	lvPPA	0.784	0.922	0.920	0.918	0.918
	nfvPPA	**0.853**	**0.917**	**0.951**	**0.942**	**0.944**
	bvFTD	0.694	0.755	0.759	0.743	0.746
	svPPA	0.657	0.686	0.708	0.697	0.697
B	lvPPA	0.730	0.854	0.871	0.870	0.870
	nfvPPA	0.735	0.669	0.749	0.724	0.713
	bvFTD	**0.786**	**0.896**	**0.926**	**0.912**	**0.902**
	svPPA	0.700	0.756	0.787	0.764	0.764
TE	lvPPA	0.676	0.757	0.759	0.754	0.754
	nfvPPA	0.735	0.662	0.762	0.713	0.685
	bvFTD	0.602	0.685	0.699	0.686	0.683
	svPPA	0.700	0.742	0.768	0.751	0.751
DRS	lvPPA	0.811	0.940	0.942	0.941	0.941
	nfvPPA	0.750	0.715	0.798	0.784	0.776
	bvFTD	0.694	0.750	0.786	0.768	0.770
	svPPA	0.771	0.882	0.880	0.873	0.873
DRT	lvPPA	0.716	0.808	0.813	0.811	0.811
	nfvPPA	0.794	0.676	0.780	0.755	0.742
	bvFTD	0.653	0.701	0.735	0.713	0.724
	svPPA	0.714	0.812	0.814	0.804	0.804
DBS	lvPPA	**0.865**	**0.951**	**0.951**	**0.951**	**0.951**
	nfvPPA	0.779	0.769	0.831	0.824	0.829
	bvFTD	0.776	0.876	0.893	0.880	0.881
	svPPA	**0.843**	**0.936**	**0.931**	**0.931**	**0.931**
DBT	lvPPA	0.851	0.916	0.913	0.913	0.913
	nfvPPA	0.765	0.776	0.833	0.816	0.825
	bvFTD	0.776	0.871	0.889	0.879	0.878
	svPPA	0.814	0.861	0.881	0.873	0.873

Table C.9: LOSO results for a logistic regression model trained on different features extracted from sentence-level Whisper transcripts. (Task: MONL, Whisper model: medium with beam search of 10, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
blabla	lvPPA	0.831	0.788	0.756	0.753	0.753
	bvFTD	0.760	0.822	0.925	0.878	0.878
	svPPA	0.836	0.881	0.842	0.836	0.836
B	lvPPA	0.800	0.909	0.897	0.894	0.894
	bvFTD	0.813	0.841	0.876	0.845	0.845
	svPPA	0.776	0.860	0.825	0.822	0.822
TE	lvPPA	**0.908**	**0.914**	**0.902**	**0.902**	**0.902**
	bvFTD	0.840	0.898	0.927	0.904	0.904
	svPPA	0.806	0.761	0.672	0.661	0.661
DRS	lvPPA	0.908	0.911	0.897	0.897	0.897
	bvFTD	**0.853**	0.903	0.924	0.905	0.905
	svPPA	0.821	0.789	0.710	0.704	0.707
DRT	lvPPA	0.877	0.901	0.888	0.888	0.888
	bvFTD	0.840	0.888	0.908	0.885	0.885
	svPPA	0.821	0.801	0.747	0.733	0.741
DBS	lvPPA	0.862	0.884	0.865	0.865	0.865
	bvFTD	0.827	**0.944**	**0.963**	**0.947**	**0.947**
	svPPA	0.851	0.791	0.741	0.730	0.730
DBT	lvPPA	0.846	0.823	0.796	0.796	0.796
	bvFTD	0.760	0.889	0.918	0.886	0.886
	svPPA	**0.881**	**0.912**	**0.888**	**0.882**	**0.885**

Table C.10: LOSO results for a logistic regression model trained on different features extracted from sentence-level Whisper transcripts. (Task: COOK, Whisper model: medium with beam search of 10, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
B	lvPPA	0.784	0.893	0.893	0.892	0.892
	nfvPPA	0.794	0.867	**0.920**	**0.916**	**0.916**
	bvFTD	0.704	0.824	0.834	0.823	0.826
	svPPA	0.671	0.647	0.690	0.657	0.657
TE	lvPPA	0.770	0.871	0.868	0.868	0.868
	nfvPPA	**0.824**	**0.880**	0.913	0.911	0.913
	bvFTD	**0.816**	**0.890**	**0.897**	**0.889**	**0.890**
	svPPA	0.757	**0.842**	**0.848**	**0.839**	**0.839**
DRS	lvPPA	0.797	0.851	0.849	0.845	0.845
	nfvPPA	0.765	0.780	0.865	0.835	0.816
	bvFTD	0.735	0.810	0.829	0.810	0.811
	svPPA	**0.771**	0.779	0.765	0.757	0.757
DRT	lvPPA	0.811	0.842	0.841	0.838	0.838
	nfvPPA	0.824	0.822	0.913	0.880	0.860
	bvFTD	0.765	0.841	0.856	0.842	0.842
	svPPA	0.743	0.795	0.787	0.786	0.786
DBS	lvPPA	0.838	0.956	0.956	0.956	0.956
	nfvPPA	0.809	0.820	0.895	0.873	0.856
	bvFTD	0.765	0.786	0.804	0.783	0.782
	svPPA	0.686	0.713	0.703	0.695	0.695
DBT	lvPPA	**0.905**	**0.969**	**0.972**	**0.971**	**0.971**
	nfvPPA	0.750	0.766	0.855	0.820	0.813
	bvFTD	0.796	0.846	0.863	0.850	0.851
	svPPA	0.700	0.724	0.719	0.705	0.705

Table C.11: LOSO results for a logistic regression model trained on different features extracted from **transcript-level** manual transcripts with interviewer speech included. (Task: MONL, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	Diagnosis	Acc.	AUC (Ind.)	AUC (Max)	AUC (Mean)	AUC (Med.)
B	lvPPA	0.877	**0.946**	**0.937**	**0.937**	**0.937**
	bvFTD	**0.907**	**0.961**	0.970	0.960	0.960
	svPPA	0.791	0.747	0.756	0.724	0.718
TE	lvPPA	0.846	0.727	0.796	0.776	0.776
	bvFTD	0.813	0.793	0.819	0.802	0.802
	svPPA	0.896	0.952	0.960	0.951	0.951
DRS	lvPPA	0.815	0.835	0.810	0.807	0.807
	bvFTD	0.787	0.834	0.911	0.876	0.876
	svPPA	**0.896**	0.969	0.974	0.974	0.974
DRT	lvPPA	0.862	0.909	0.894	0.894	0.894
	bvFTD	0.773	0.864	0.941	0.905	0.905
	svPPA	0.881	**0.985**	**0.986**	**0.986**	**0.986**
DBS	lvPPA	0.862	0.929	0.917	0.917	0.917
	bvFTD	0.867	0.946	0.981	0.967	0.967
	svPPA	0.866	0.864	0.922	0.902	0.902
DBT	lvPPA	**0.908**	0.933	0.922	0.922	0.922
	bvFTD	0.867	0.955	**0.984**	**0.973**	**0.973**
	svPPA	0.791	0.841	0.871	0.842	0.842

Table C.12: LOSO results for a logistic regression model trained on different features extracted from **transcript-level** manual transcripts with interviewer speech included. (Task: COOK, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
B	lvPPA	0.824	0.926	0.924	0.924	0.924
	nfvPPA	0.809	0.832	0.909	0.893	0.895
	bvFTD	**0.847**	**0.904**	**0.914**	**0.904**	**0.904**
	svPPA	**0.814**	**0.899**	**0.933**	**0.919**	**0.919**
TE	lvPPA	0.757	0.836	0.834	0.833	0.833
	nfvPPA	0.765	0.792	0.813	0.811	0.807
	bvFTD	0.745	0.799	0.808	0.798	0.798
	svPPA	0.571	0.615	0.625	0.594	0.594
DRS	lvPPA	0.784	0.881	0.881	0.879	0.879
	nfvPPA	0.809	0.808	0.887	0.873	0.878
	bvFTD	0.724	0.773	0.812	0.788	0.786
	svPPA	0.757	0.785	0.797	0.785	0.785
DRT	lvPPA	0.811	0.903	0.903	0.901	0.901
	nfvPPA	0.809	0.800	0.896	0.875	0.875
	bvFTD	0.714	0.768	0.799	0.779	0.782
	svPPA	0.671	0.737	0.726	0.717	0.717
DBS	lvPPA	0.824	0.945	0.944	0.943	0.943
	nfvPPA	0.868	**0.894**	0.935	0.927	0.935
	bvFTD	0.663	0.754	0.786	0.769	0.772
	svPPA	0.771	0.832	0.828	0.821	0.821
DBT	lvPPA	**0.838**	**0.948**	**0.949**	**0.948**	**0.948**
	nfvPPA	**0.897**	0.877	**0.971**	**0.940**	**0.938**
	bvFTD	0.724	0.805	0.826	0.807	0.806
	svPPA	0.757	0.829	0.820	0.818	0.818

Table C.13: LOSO results for a logistic regression model trained on different features extracted from sentence-level manual transcripts **without** interviewer speech included. (Task: MONL, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	Diagnosis	Acc.	AUC (Ind.)	AUC (Max)	AUC (Mean)	AUC (Med.)
B	lvPPA	0.877	0.948	0.945	0.945	0.945
	bvFTD	**0.787**	**0.908**	**0.964**	**0.932**	**0.932**
	svPPA	0.687	0.726	0.770	0.736	0.739
TE	lvPPA	0.846	0.542	0.632	0.621	0.621
	bvFTD	0.693	0.850	0.927	0.907	0.907
	svPPA	0.776	0.732	0.670	0.667	0.667
DRS	lvPPA	0.862	0.884	0.894	0.885	0.885
	bvFTD	0.773	0.841	0.911	0.871	0.871
	svPPA	0.806	**0.816**	0.802	0.784	**0.793**
DRT	lvPPA	**0.908**	0.951	0.954	0.951	0.951
	bvFTD	0.707	0.756	0.899	0.826	0.826
	svPPA	0.791	0.732	0.753	0.721	0.718
DBS	lvPPA	0.908	0.943	0.948	0.948	0.948
	bvFTD	0.760	0.858	0.960	0.925	0.925
	svPPA	**0.836**	0.791	**0.833**	**0.793**	0.784
DBT	lvPPA	0.892	**0.968**	**0.963**	**0.963**	**0.963**
	bvFTD	0.707	0.807	0.915	0.853	0.853
	svPPA	0.806	0.757	0.750	0.741	0.736

Table C.14: LOSO results for a logistic regression model trained on different features extracted from sentence-level manual transcripts **without** interviewer speech included. (Task: COOK, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
B	lvPPA	0.811	0.927	0.924	0.924	0.924
	nfvPPA	0.824	**0.924**	**0.958**	**0.956**	**0.956**
	bvFTD	**0.776**	0.823	0.845	0.831	0.832
	svPPA	0.643	0.777	0.794	0.786	0.786
TE	lvPPA	0.784	0.909	0.906	0.905	0.905
	nfvPPA	0.838	0.838	0.842	0.829	0.825
	bvFTD	0.663	0.763	0.783	0.764	0.762
	svPPA	0.671	0.704	0.706	0.695	0.695
DRS	lvPPA	0.851	0.946	0.948	0.945	0.945
	nfvPPA	0.735	0.793	0.853	0.835	0.827
	bvFTD	0.704	0.822	0.834	0.818	0.813
	svPPA	0.700	**0.816**	**0.801**	**0.799**	**0.799**
DRT	lvPPA	**0.865**	0.922	0.923	0.920	0.920
	nfvPPA	0.735	0.713	0.804	0.775	0.764
	bvFTD	0.704	0.789	0.794	0.775	0.776
	svPPA	0.757	0.799	0.789	0.782	0.782
DBS	lvPPA	0.811	0.945	0.942	0.942	0.942
	nfvPPA	0.824	0.815	0.895	0.875	0.880
	bvFTD	0.694	0.756	0.786	0.771	0.770
	svPPA	**0.786**	0.779	0.773	0.766	0.766
DBT	lvPPA	0.824	**0.955**	**0.953**	**0.953**	**0.953**
	nfvPPA	**0.868**	0.828	0.909	0.889	0.885
	bvFTD	0.755	**0.840**	**0.858**	**0.844**	**0.843**
	svPPA	0.757	0.782	0.764	0.762	0.762

Table C.15: LOSO results for a logistic regression model trained on different features extracted from sentence-level manual transcripts with interviewer speech included, with **punctuation removed**. (Task: MONL, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	**Diagnosis**	**Acc.**	**AUC (Ind.)**	**AUC (Max)**	**AUC (Mean)**	**AUC (Med.)**
B	lvPPA	0.754	0.941	0.937	0.937	0.937
	bvFTD	**0.827**	**0.913**	0.955	0.938	0.938
	svPPA	0.791	**0.875**	**0.897**	**0.885**	**0.876**
TE	lvPPA	0.800	0.828	0.865	0.856	0.856
	bvFTD	0.693	0.812	0.895	0.851	0.851
	svPPA	0.701	0.724	0.681	0.667	0.661
DRS	lvPPA	**0.877**	**0.980**	**0.977**	**0.977**	**0.977**
	bvFTD	0.787	0.878	0.955	0.927	0.927
	svPPA	0.806	0.807	0.796	0.779	0.779
DRT	lvPPA	0.862	0.916	0.908	0.908	0.908
	bvFTD	0.773	0.819	0.917	0.878	0.878
	svPPA	0.791	0.814	0.828	0.802	0.802
DBS	lvPPA	0.846	0.906	0.897	0.894	0.894
	bvFTD	0.813	0.871	**0.970**	**0.940**	**0.940**
	svPPA	0.821	0.830	0.830	0.802	0.802
DBT	lvPPA	0.862	0.904	0.891	0.888	0.888
	bvFTD	0.787	0.844	0.931	0.902	0.902
	svPPA	**0.851**	0.875	0.856	0.839	0.845

Table C.16: LOSO results for a logistic regression model trained on different features extracted from sentence-level manual transcripts with interviewer speech included, with **punctuation removed**. (Task: COOK, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	Diagnosis	Acc.	AUC (Ind.)	AUC (Max)	AUC (Mean)	AUC (Med.)
B	lvPPA	0.797	0.942	0.941	0.941	0.941
	nfvPPA	0.809	**0.818**	0.915	0.895	**0.896**
	bvFTD	**0.857**	**0.936**	**0.947**	**0.940**	**0.940**
	svPPA	**0.786**	**0.890**	**0.922**	**0.909**	**0.909**
TE	lvPPA	0.743	0.784	0.783	0.781	0.781
	nfvPPA	0.750	0.769	0.831	0.816	0.820
	bvFTD	0.724	0.777	0.806	0.793	0.790
	svPPA	0.729	0.771	0.793	0.768	0.768
DRS	lvPPA	0.824	0.916	0.911	0.911	0.911
	nfvPPA	0.779	0.780	0.845	0.844	0.844
	bvFTD	0.786	0.851	0.878	0.858	0.853
	svPPA	0.671	0.731	0.726	0.710	0.710
DRT	lvPPA	**0.851**	**0.953**	**0.951**	**0.951**	**0.951**
	nfvPPA	0.779	0.801	0.884	0.867	0.869
	bvFTD	0.714	0.765	0.788	0.764	0.766
	svPPA	0.700	0.714	0.713	0.695	0.695
DBS	lvPPA	0.797	0.919	0.917	0.917	0.917
	nfvPPA	0.809	0.759	0.873	0.844	0.825
	bvFTD	0.724	0.803	0.827	0.813	0.818
	svPPA	0.729	0.727	0.726	0.716	0.716
DBT	lvPPA	0.811	0.937	0.940	0.939	0.939
	nfvPPA	**0.824**	0.807	**0.927**	**0.896**	0.878
	bvFTD	0.755	0.831	0.852	0.834	0.830
	svPPA	0.757	0.748	0.733	0.724	0.724

Table C.17: LOSO results for a logistic regression model trained on different features extracted from sentence-level manual transcripts with interviewer speech included, with **filler words removed**. (Task: MONL, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Features	Diagnosis	Acc.	AUC (Ind.)	AUC (Max)	AUC (Mean)	AUC (Med.)
B	lvPPA	**0.908**	**0.985**	**0.983**	**0.983**	**0.983**
	bvFTD	**0.840**	**0.944**	**0.970**	**0.954**	**0.954**
	svPPA	0.746	0.803	**0.796**	0.793	0.793
TE	lvPPA	0.862	0.914	0.908	0.905	0.905
	bvFTD	0.667	0.819	0.897	0.862	0.862
	svPPA	0.731	0.797	0.730	0.721	0.730
DRS	lvPPA	0.815	0.919	0.920	0.920	0.920
	bvFTD	0.773	0.863	0.937	0.908	0.908
	svPPA	**0.836**	**0.839**	0.793	0.773	0.773
DRT	lvPPA	0.877	0.948	0.945	0.945	0.945
	bvFTD	0.760	0.807	0.894	0.846	0.846
	svPPA	0.791	0.828	0.830	**0.813**	**0.810**
DBS	lvPPA	0.877	0.956	0.948	0.948	0.948
	bvFTD	0.773	0.888	0.970	0.938	0.938
	svPPA	0.731	0.784	0.796	0.759	0.759
DBT	lvPPA	0.846	0.953	0.948	0.948	0.948
	bvFTD	0.773	0.869	0.950	0.905	0.905
	svPPA	0.791	0.780	0.779	0.759	0.750

Table C.18: LOSO results for a logistic regression model trained on different features extracted from sentence-level manual transcripts with interviewer speech included, with **filler words removed**. (Task: COOK, B: bert-base-uncased, TE: trans-encoder, DRS: diffcse-roberta-base-sts, DRT: diffcse-roberta-base-trans, DBS: diffcse-bert-base-uncased-sts, DBT: diffcse-bert-base-uncased-trans)

Bibliography

Word representations. URL https://fasttext.cc/docs/en/unsupervised-tutorial.html.

Discrete fourier transform (numpy.fft). URL https://numpy.org/doc/stable/reference/routines.fft.html.

Comparison of lda and pca 2d projection of iris dataset, a. URL https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_vs_lda.html#sphx-glr-auto-examples-decomposition-plot-pca-vs-lda-py.

Decision trees, b. URL https://scikit-learn.org/stable/modules/tree.html.

Nearest neighbors, c. URL https://scikit-learn.org/stable/modules/neighbors.html#classification.

Support vector machines, d. URL https://scikit-learn.org/stable/modules/svm.html#svm-classification.

Linear and quadratic discriminant analysis, e. URL https://scikit-learn.org/stable/modules/lda_qda.html#lda-qda.

auditok (Version 0.1.8) [Software]. Available from https://pypi.org/project/auditok/, 2020.

Gradient boosting in ml, Mar 2023. URL https://www.geeksforgeeks.org/ml-gradient-boosting/#.

jiwer (Version 3.0.1) [Software]. Available from https://pypi.org/project/jiwer/, 2023.

Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimer's and Dementia*, 15(3):321–387, 2019.

Alzheimer's Association. Medical tests. *Alzheimer's Disease and Dementia.*, 2020. Last accessed December 3, 2020, from https://www.alz.org/alzheimers-dementia/diagnosis/medical_tests.

Aparna Balagopalan, Jekaterina Novikova, Frank Rudzicz, and Marzyeh Ghassemi. The effect of heterogeneous data for Alzheimer's disease detection from speech. *arXiv preprint arXiv:1811.12254*, 2018.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection. In *Proc. Interspeech 2020*, pages 2167–2171, 2020a. doi: 10.21437/Interspeech.2020-2557. URL http://dx.doi.org/10.21437/Interspeech.2020-2557.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. In *INTERSPEECH*, 2020b.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994.

Michael F Bonner, Sharon Ash, and Murray Grossman. The new classification of primary progressive aphasia into semantic, logopenic, or nonfluent/agrammatic variants. *Current neurology and neuroscience reports*, 10(6):484–490, 2010.

Catarina Botelho, Francisco Teixeira, Thomas Rolland, Alberto Abad, and Isabel Trancoso. Pathological speech detection using x-vector embeddings. *arXiv preprint arXiv:2003.00864*, 2020.

Tom Bschor, Klaus-Peter Kühl, and Friedel M Reischies. Spontaneous speech of patients with dementia of the alzheimer type and mild cognitive impairment. *International psychogeriatrics*, 13(3):289–298, 2001.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001.

Jess CS Chan, Julie C Stout, and Adam P Vogel. Speech in prodromal and symptomatic huntington's disease as a model of measuring onset and progression in dominantly inherited neurodegenerative diseases. *Neuroscience & Biobehavioral Reviews*, 107: 450–460, 2019.

Jun Chen, Ji Zhu, and Jieping Ye. An attention-based hybrid network for automatic detection of Alzheimer's disease from narrative speech. *Proc. Interspeech 2019*, pages 4085–4089, 2019.

Yi-Wei Chien, Sheng-Yi Hong, Wen-Ting Cheah, Li-Hung Yao, Yu-Ling Chang, and Li-Chen Fu. An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network. *Scientific Reports*, 9(1):1–10, 2019.

Karol Chlasta and Krzysztof Wołk. Towards computer-based automated screening of dementia through spontaneous speech. *Frontiers in Psychology*, 11:623237, 2021.

Sunghye Cho, Naomi Nevler, Sanjana Shellikeri, Sharon Ash, Mark Liberman, and Murray Grossman. Automatic classification of primary progressive aphasia patients using lexical and acoustic features. In *Proceedings of Language Resources and Evaluation Conference (LREC) 2020 Workshop on Resources and Processing Linguistic, Para-linguistic and Extra-linguistic Data from People with Various Forms of Cognitive/Psychiatric/Developmental Impairments (RaPID-3)*, pages 60–65, 2020.

Hyunjoo Choi. Performances in a picture description task in japanese patients with alzheimer's disease and with mild cognitive impairment. *Communication Sciences & Disorders*, 14(3):326–337, 2009.

Maurycy Chronowski, Maciej Klaczynski, Malgorzata Dec-Cwiek, and Karolina Porebska. Parkinson's disease diagnostics using ai and natural language knowledge transfer. *arXiv preprint arXiv:2204.12559*, 2022.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

Emma Louise Clark, Catherine Easton, and Sarah Verdon. The impact of linguistic bias upon speech-language pathologists' attitudes towards non-standard dialects of english. *Clinical Linguistics & Phonetics*, 35(6):542–559, 2021.

Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1269.

Patricia V Cooper. Discourse production and normal aging: Performance on oral picture description tasks. *Journal of Gerontology*, 45(5):P210–P214, 1990.

Nicholas Cummins, Alice Baird, and Björn W Schuller. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54, 2018.

Nicholas Cummins, Yilin Pan, Zhao Ren, Julian Fritsch, Venkata Srikanth Nallanthighal, Heidi Christensen, Daniel Blackburn, Björn W. Schuller, Mathew Magimai-Doss, Helmer Strik, and Aki Härmä. A Comparison of Acoustic and Linguistics Methodologies for Alzheimer's Dementia Recognition. In *Proc. Interspeech 2020*, pages 2182–2186, 2020. doi: 10.21437/Interspeech.2020-2635. URL http://dx.doi.org/10.21437/Interspeech.2020-2635.

Sofia de la Fuente Garcia, Craig Ritchie, and Saturnino Luz. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, (Preprint):1–27, 2020.

Ellen Elisa De Roeck, Peter Paul De Deyn, Eva Dierckx, and Sebastiaan Engelborghs. Brief cognitive screening instruments for early detection of alzheimer's disease: a systematic review. *Alzheimer's research & therapy*, 11(1):21, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Chris Dollaghan and Thomas F Campbell. Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, 41(5):1136–1146, 1998.

Erik Edwards, Charles Dognin, Bajibabu Bollepalli, and Maneesh Singh. Multiscale System for Alzheimer's Dementia Recognition Through Spontaneous Speech. In *Proc. Interspeech 2020*, pages 2197–2201, 2020. doi: 10.21437/Interspeech.2020-2781. URL http://dx.doi.org/10.21437/Interspeech.2020-2781.

Katharine Graf Estes, Julia L Evans, and Nicole M Else-Quest. Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. 2007.

Hao Fang, Chen Gong, Chen Zhang, Yanan Sui, and Luming Li. Parkinsonian chinese speech analysis towards automatic classification of parkinson's disease. In *Machine Learning for Health*, pages 114–125. PMLR, 2020.

Shahla Farzana and Natalie Parde. Exploring MMSE Score Prediction Using Verbal and Non-Verbal Cues. In *Proc. Interspeech 2020*, pages 2207–2211, 2020. doi: 10.21437/Interspeech.2020-3085. URL http://dx.doi.org/10.21437/Interspeech.2020-3085.

Kathleen C Fraser, Frank Rudzicz, and Elizabeth Rochon. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *Interspeech*, pages 2177–2181, 2013.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2016.

Fernando Garcia-Gutierrez, Alfonso Delgado-Alvarez, Cristina Delgado-Alonso, Josefa Díaz-Álvarez, Vanesa Pytel, Maria Valles-Salgado, María Jose Gil, Laura Hernández-Lorenzo, Jorge Matías-Guiu, José L Ayala, et al. Diagnosis of alzheimer's disease and behavioural variant frontotemporal dementia with machine learning-aided neuropsychological assessment using feature engineering and genetic algorithms. *International Journal of Geriatric Psychiatry*, 37(2), 2022.

Elaine Giles, Karalyn Patterson, and John R Hodges. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information. *Aphasiology*, 10(4):395–408, 1996.

Laurence Gillick and Stephen J Cox. Some statistical issues in the comparison of speech recognition algorithms. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 532–535. IEEE, 1989.

Pedro Gómez-Vilda, Andrés Gómez-Rodellar, Daniel Palacios-Alonso, and Athanasios Tsanas. Performance of monosyllabic vs multisyllabic diadochokinetic exercises in evaluating parkinson's disease hypokinetic dysarthria from fluency distributions. 2021.

Harold Goodglass and Edith Kaplan. *Boston diagnostic aphasia examination booklet*. Lea & Febiger, 1983.

Maria Luisa Gorno-Tempini, Argye E Hillis, Sandra Weintraub, Andrew Kertesz, Mario Mendez, Stefano F Cappa, Jennifer M Ogar, JD Rohrer, Steven Black, Bradley F Boeve, et al. Classification of primary progressive aphasia and its variants. *Neurology*, 76(11): 1006–1014, 2011.

Gábor Gosztolya, Veronika Vincze, László Tóth, Magdolna Pákáski, János Kálmán, and Ildikó Hoffmann. Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features. *Computer Speech & Language*, 53:181–197, 2019.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Thea Nygaard Grimstvedt, Jeanette Ullmann Miller, Marleen Regina van Walsem, and Kristin J Billaud Feragen. Speech and language difficulties in huntington's disease: A qualitative study of patients' and professional caregivers' experiences. *International Journal of Language & Communication Disorders*, 2021.

Aishwarya Gulve. Ordinary least square (ols) method for linear regression, Aug 2020. URL https://medium.com/analytics-vidhya/ordinary-least-square-ols-method-for-linear-regression-ef8ca10aadfc.

Zhiqiang Guo, Zhenhua Ling, and Yunxia Li. Detecting alzheimer's disease from continuous speech using language models. *Journal of Alzheimer's Disease*, 70(4):1163–1174, 2019.

Vera F Gutiérrez-Clellen and Gabriela Simon-Cereijido. Using nonword repetition tasks for the identification of language impairment in spanish-english-speaking children: Does the language of assessment matter? *Learning Disabilities Research & Practice*, 25(1):48–58, 2010.

Fasih Haider, Sofia De La Fuente, and Saturnino Luz. An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 2019.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

R'mani Haulcy and James Glass. Classifying Alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology*, 11:3833, 2021a. ISSN 1664-1078. doi: 10.3389/fpsyg.2020.624137. URL https://www.frontiersin.org/article/10.3389/fpsyg.2020.624137.

R'mani Haulcy and James Glass. Classifying alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology*, 11:624137, 2021b.

R'mani Haulcy and James Glass. CLAC: A Speech Corpus of Healthy English Speakers. In *Proc. Interspeech 2021*, pages 2966–2970, 2021c. doi: 10.21437/Interspeech.2021-1810.

R'mani Haulcy, Katerina Placek, Brian Tracey, Adam Vogel, and James Glass. Repetition assessment for speech and language disorders: A study of the logopenic variant of primary progressive aphasia. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6932–6936, 2022. doi: 10.1109/ICASSP43922.2022.9746627.

Maya L Henry and Stephanie M Grasso. Assessment of individuals with primary progressive aphasia. In *Seminars in speech and language*, volume 39, page 231. NIH Public Access, 2018.

Laura Hernández-Domínguez, Sylvie Ratté, Gerardo Sierra-Martínez, and Andrés Roche-Bergua. Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:260–268, 2018.

Wolfram Hinzen, Joana Rosselló, Cati Morey, Estela Camara, Clara Garcia-Gorro, Raymond Salvador, and Ruth de Diego-Balaguer. A systematic linguistic profile of spontaneous narrative speech in pre-symptomatic and early stage huntington's disease. *Cortex*, 100: 71–83, 2018.

Sheng-Yi Hong, Li-Hung Yao, Wen-Ting Cheah, Wei-Der Chang, Li-Chen Fu, and Yu-Ling Chang. A novel screening system for alzheimer's disease based on speech transcripts using neural network. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 2440–2445. IEEE, 2019.

Aren Jansen, Kenneth Church, and Hynek Hermansky. Towards spoken term discovery at scale with zero resources. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Ashir Javeed, Ana Luiza Dallora, Johan Sanmartin Berglund, Arif Ali, Liaqata Ali, and Peter Anderberg. Machine learning for dementia prediction: A systematic review and future research directions. *Journal of medical systems*, 47(1):1–25, 2023.

Finnian Kelly, Anil Alexander, Oscar Forth, and DVD Vloed. From i-vectors to x-vectorsâĂŤa generational change in speaker recognition illustrated on the nfi-frida database. In *Proc. 25th Int. Assoc. Forensic Phonetics Acoust.(IAFPA)*, pages 1–28, 2019.

Ali Khodabakhsh, Fatih Yesil, Ekrem Guner, and Cenk Demiroglu. Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):9, 2015.

Jun Pyo Kim, Jeonghun Kim, Yu Hyun Park, Seong Beom Park, Jin San Lee, Sole Yoo, Eun-Joo Kim, Hee Jin Kim, Duk L Na, Jesse A Brown, et al. Machine learning based hierarchical classification of frontotemporal dementia and alzheimer's disease. *NeuroImage: Clinical*, 23:101811, 2019.

Dimitrios Kokkinakis, Kristina Lundholm Fors, Kathleen C Fraser, and Arto Nordlund. A swedish cookie-theft corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H Robert, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124, 2015.

Junghyun Koo, Jie Hwan Lee, Jaewoo Pyo, Yujin Jo, and Kyogu Lee. Exploiting Multi-Modal Features from Pre-Trained Networks for Alzheimer's Dementia Recognition. In *Proc. Interspeech 2020*, pages 2217–2221, 2020. doi: 10.21437/Interspeech.2020-3153. URL http://dx.doi.org/10.21437/Interspeech.2020-3153.

Walker H Land and J David Schaffer. Alzheimer's disease and speech background. In *The Art and Science of Machine Intelligence*, pages 107–135. Springer, 2020.

Jackson L. Lee, Ross Burkholder, Gallagher B. Flinn, and Emily R. Coppess. Working with chat transcripts in python. Technical Report TR-2016-02, Department of Computer Science, University of Chicago, 2016.

Mauricio Letelier. Roc curve and auc from scratch in numpy (visualized!), Jan 2021. URL https://towardsdatascience.com/roc-curve-and-auc-from-scratch-in-numpy-visualized-2612bb9459ab.

Cristian E Leyton, Sharon Savage, Muireann Irish, Samantha Schubert, Olivier Piguet, Kirrie J Ballard, and John R Hodges. Verbal repetition in primary progressive aphasia and alzheimer's disease. *Journal of Alzheimer's Disease*, 41(2):575–585, 2014.

Fangyu Liu, Serhii Havrylov, Yunlong Jiao, Jordan Massiah, and Emine Yilmaz. Trans-encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations. *ArXiv*, abs/2109.13059, 2022.

Lin Liu, Shenghui Zhao, Haibao Chen, and Aiguo Wang. A new machine learning method for identifying Alzheimer's disease. *Simulation Modelling Practice and Theory*, 99: 102023, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

José Vicente Egas López, László Tóth, Ildikó Hoffmann, János Kálmán, Magdolna Pákáski, and Gábor Gosztolya. Assessing alzheimer's disease from speech using the i-vector approach. In *International Conference on Speech and Computer*, pages 289–298. Springer, 2019.

Ulrike Lueken, Ulrich Seidl, Lena Völker, Elisabeth Schweiger, Andreas Kruse, and Johannes Schröder. Development of a short version of the apathy evaluation scale specifically adapted for demented nursing home residents. *The American journal of geriatric psychiatry*, 15(5):376–385, 2007.

Sladjana Lukic, Maria Luisa Mandelli, Ariane Welch, Kesshi Jordan, Wendy Shwe, John Neuhaus, Zachary Miller, H Isabel Hubbard, Maya Henry, Bruce L Miller, et al. Neurocognitive basis of repetition deficits in primary progressive aphasia. *Brain and language*, 194:35–45, 2019.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Proc. Interspeech 2020*, pages 2172–2176, 2020. doi: 10.21437/Interspeech.2020-2571. URL http://dx.doi.org/10.21437/Interspeech.2020-2571.

Catherine Mackenzie, Marian Brady, John Norrie, and Ninik Poedjianto. Picture description in neurologically normal adults: Concepts and topic coherence. *Aphasiology*, 21(3-4): 340–354, 2007.

Brian MacWhinney. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press, 2000.

Brian MacWhinney. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press, 2014.

Pranav Mahajan and Veeky Baths. Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech. *Frontiers in Aging Neuroscience*, 13:623607, 2021.

Alex Martin, Pim Brouwers, Christiane Cox, and Paul Fedio. On the nature of the verbal memory deficit in alzheimer's disease. *Brain and Language*, 25(2):323–341, 1985.

Matej Martinc and Senja Pollak. Tackling the ADReSS Challenge: A Multimodal Approach to the Automated Recognition of Alzheimer's Dementia. In *Proc. Interspeech 2020*, pages 2157–2161, 2020. doi: 10.21437/Interspeech.2020-2202. URL http://dx.doi.org/10.21437/Interspeech.2020-2202.

Matej Martinc, Fasih Haider, Senja Pollak, and Saturnino Luz. Temporal integration of text transcripts and acoustic features for alzheimer's diagnosis based on spontaneous speech. *Frontiers in Aging Neuroscience*, 13:642647, 2021.

Jordi A Matias-Guiu, Paz Suárez-Coalla, Miguel Yus, Vanesa Pytel, Laura Hernández-Lorenzo, Cristina Delgado-Alonso, Alfonso Delgado-Álvarez, Natividad Gómez-Ruiz, Carmen Polidura, María Nieves Cabrera-Martín, et al. Identification of the main components of spontaneous speech in primary progressive aphasia and their neural underpinnings using multimodal mri and fdg-pet imaging. *Cortex*, 146:141–160, 2022.

Mario F Mendez and Mary Ashla-Mendez. Differences between multi-infarct dementia and alzheimer's disease on unstructured neuropsychological tasks. *Journal of Clinical and Experimental Neuropsychology*, 13(6):923–932, 1991.

M-Marsel Mesulam, Christina Wieneke, Cynthia Thompson, Emily Rogalski, and Sandra Weintraub. Quantitative classification of primary progressive aphasia at early and mild impairment stages. *Brain*, 135(5):1537–1553, 2012.

Laureano Moro-Velazquez, Jesus Villalba, and Najim Dehak. Using x-vectors to automatically detect parkinson's disease from speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1155–1159. IEEE, 2020.

Laureano Moro-Velazquez, Jorge A Gomez-Garcia, Julian D Arias-Londoño, Najim Dehak, and Juan I Godino-Llorente. Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects. *Biomedical Signal Processing and Control*, 66:102418, 2021.

Kimberly D Mueller, Bruce Hermann, Jonilda Mecollari, and Lyn S Turkstra. Connected speech and language in mild cognitive impairment and alzheimer's disease: A review of picture description tasks. *Journal of clinical and experimental neuropsychology*, 40(9): 917–939, 2018a.

Kimberly D. Mueller, Bruce Hermann, Jonilda Mecollari, and Lyn S. Turkstra. Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology*, 40(9): 917–939, October 2018b. ISSN 1744411X. doi: 10.1080/13803395.2018.1446513. URL https://www.tandfonline.com/action/journalInformation?journalCode=ncen20. Publisher: Routledge.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

Ryosuke Nagumo, Yaming Zhang, Yuki Ogawa, Mitsuharu Hosokawa, Kengo Abe, Takaaki Ukeda, Sadayuki Sumi, Satoshi Kurita, Sho Nakakubo, Sangyoon Lee, et al. Automatic detection of cognitive impairments through acoustic analysis of speech. *Current Alzheimer Research*, 17(1):60–68, 2020.

Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7: 19143–19165, 2019.

Kyriaki Neophytou, Robert W Wiley, Brenda Rapp, and Kyrana Tsapkini. The use of spelling for variant classification in primary progressive aphasia: Theoretical and practical implications. *Neuropsychologia*, 133:107157, 2019.

Jekaterina Novikova and Aparna Balagopalan. On speech datasets in machine learning for healthcare.

Roelant Ossewaarde, Roel Jonkers, Fedor Jalvingh, and Roelien Bastiaanse. Classification of spontaneous speech of individuals with dementia based on automatic prosody analysis using support vector machines (svm). In *The Thirty-Second International Flairs Conference*, 2019.

Raghavendra Pappagari, Jaejin Cho, Laureano Moro-Velázquez, and Najim Dehak. Using State of the Art Speaker Recognition and Natural Language Processing Technologies to Detect Alzheimer's Disease and Assess its Severity. In *Proc. Interspeech 2020*, pages 2177–2181, 2020. doi: 10.21437/Interspeech.2020-2587. URL http://dx.doi.org/10.21437/Interspeech.2020-2587.

Raghavendra Pappagari, Jaejin Cho, Sonal Joshi, Laureano Moro-Velázquez, Piotr Zelasko, Jesús Villalba, and Najim Dehak. Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios. In *Interspeech*, pages 3825–3829, 2021.

Alex S Park and James R Glass. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):186–197, 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Kathe S Perez, Lorraine Olson Ramig, Marshall E Smith, and Christopher Dromey. The parkinson larynx: tremor and videostroboscopic findings. *Journal of Voice*, 10(4):354–361, 1996.

J Platt. Probabilistic outputs for svms and comparisons to regularized likehood methods. In *Advances in Large Margin Classifiers*. MIT Press, 1999.

Anna Pompili, Thomas Rolland, and Alberto Abad. The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge. In *Proc. Interspeech 2020*, pages 2202–2206, 2020. doi: 10.21437/Interspeech.2020-2833. URL http://dx.doi.org/10.21437/Interspeech.2020-2833.

Matthew L Poole, Amy Brodtmann, David Darby, and Adam P Vogel. Motor speech phenotypes of frontotemporal dementia, primary progressive aphasia, and progressive apraxia of speech. *Journal of Speech, Language, and Hearing Research*, 60(4):897–911, 2017.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

María Luisa Barragán Pulido, Jesús Bernardino Alonso Hernández, Miguel Ángel Ferrer Ballester, Carlos Manuel Travieso González, Jiří Mekyska, and Zdeněk Smékal. Alzheimer's disease and automatic speech analysis: a review. *Expert Systems with Applications*, page 113213, 2020.

Yuan Qiao, Xin-Yi Xie, Guo-Zhen Lin, Yang Zou, Sheng-Di Chen, Ru-Jing Ren, and Gang Wang. Computer-assisted speech analysis in mild cognitive impairment and alzheimer's disease: A pilot study from shanghai, china. *Journal of Alzheimer's Disease*, (Preprint): 1–11, 2020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Neguine Rezaii, Nicole Carvalho, Michael Brickhouse, Emmaleigh Loyer, Phillip Wolff, Alexandra Touroutoglou, Bonnie Wong, Megan Quimby, and Brad C Dickerson. Neuroanatomical mapping of artificial intelligence-based classification of language in ppa. *Alzheimer's & Dementia*, 17:e055340, 2021.

James Robert, Marc Webbie, et al. Pydub, 2018. URL http://pydub.com/.

Morteza Rohanian, Julian Hough, and Matthew Purver. Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech. In *Proc. Interspeech 2020*, pages 2187–2191, 2020. doi: 10.21437/Interspeech.2020-2721. URL http://dx.doi.org/10.21437/Interspeech.2020-2721.

Morteza Rohanian, Julian Hough, and Matthew Purver. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. *arXiv preprint arXiv:2106.15684*, 2021.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Utkarsh Sarawgi, Wazeer Zulfikar, Nouran Soliman, and Pattie Maes. Multimodal Inductive Transfer Learning for Detection of Alzheimer's Dementia and its Severity. In *Proc. Interspeech 2020*, pages 2212–2216, 2020. doi: 10.21437/Interspeech.2020-3137. URL http://dx.doi.org/10.21437/Interspeech.2020-3137.

Thomas Searle, Zina Ibrahim, and Richard Dobson. Comparing Natural Language Processing Techniques for Alzheimer's Dementia Prediction in Spontaneous Speech. In *Proc. Interspeech 2020*, pages 2192–2196, 2020. doi: 10.21437/Interspeech.2020-2729. URL http://dx.doi.org/10.21437/Interspeech.2020-2729.

Ulrich Seidl, Ulrike Lueken, Philipp A Thomann, Andreas Kruse, and Johannes Schröder. Facial expression in alzheimer's disease: impact of cognitive deficits and neuropsychiatric symptoms. *American Journal of Alzheimer's Disease & Other Dementias®*, 27(2): 100–106, 2012.

Daisaku Shibata, Shoko Wakamiya, Ayae Kinoshita, and Eiji Aramaki. Detecting japanese patients with alzheimer's disease based on word category frequencies. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 78–85, 2016.

Abhishek Shivkumar, Jack Weston, Raphael Lenain, and Emil Fristed. Blabla: Linguistic feature extraction for clinical analysis in multiple languages. In *INTERSPEECH*, 2020.

Sabine Skodda, Uwe Schlegel, Rainer Hoffmann, and Carsten Saft. Impaired motor speech performance in huntington's disease. *Journal of Neural Transmission*, 121(4):399–407, 2014.

David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, pages 999–1003, 2017.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

Christopher Song, David Harwath, Tuka Alhanai, and James Glass. Speak: A toolkit using Amazon Mechanical Turk to collect and validate speech audio recordings. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7253–7258, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.787.

Muhammad Shehram Shah Syed, Zafi Sherhan Syed, Margaret Lech, and Elena Pirogova. Automated Screening for Alzheimer's Dementia Through Spontaneous Speech. In *Proc. Interspeech 2020*, pages 2222–2226, 2020. doi: 10.21437/Interspeech.2020-3158. URL http://dx.doi.org/10.21437/Interspeech.2020-3158.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54, 2010.

Charalambos Themistocleous, Bronte Ficek, Kimberly Webster, Dirk-Bart den Ouden, Argye E Hillis, and Kyrana Tsapkini. Automatic subtyping of individuals with primary progressive aphasia. *Journal of Alzheimer's Disease*, 79(3):1185–1194, 2021.

Nicola Ueffing, Maximilian Bisani, and Paul Vozila. Improved models for automatic punctuation prediction for spoken and written text. In *Interspeech*, pages 3097–3101, 2013.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Adam P Vogel, Christopher Shirbin, Andrew J Churchyard, and Julie C Stout. Speech acoustic markers of early stage and prodromal huntington's disease: a marker of disease onset? *Neuropsychologia*, 50(14):3273–3278, 2012.

Adam P Vogel, Matthew L Poole, Hugh Pemberton, Marja WJ Caverlé, Frederique MC Boonstra, Essie Low, David Darby, and Amy Brodtmann. Motor speech signature of behavioral variant frontotemporal dementia: Refining the phenotype. *Neurology*, 89(8): 837–844, 2017.

Rohit Voleti, Julie M Liss, and Visar Berisha. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):282–298, 2019.

Jochen Weiner, Christian Herff, and Tanja Schultz. Speech-based detection of Alzheimer's disease in conversational german. In *INTERSPEECH*, pages 1938–1942, 2016.

Susan Ellis Weismer, J Bruce Tomblin, Xuyang Zhang, Paula Buckwalter, Jan Gaura Chynoweth, and Maura Jones. Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research*, 43(4):865–878, 2000.

Alexandra M Wennberg, Jennifer L Whitwell, Nirubol Tosakulwong, Stephen D Weigand, Melissa E Murray, Mary M Machulda, Leonard Petrucelli, Michelle M Mielke, Clifford R Jack Jr, David S Knopman, et al. The influence of tau, amyloid, alpha-synuclein, tdp-43, and vascular pathology in clinically normal elderly individuals. *Neurobiology of aging*, 77:26–36, 2019.

Vikramaditya G Yadav. The hunt for a cure for alzheimer's disease receives a timely boost. *Science Translational Medicine*, 11(509):eaaz0311, 2019.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. doi: https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/1097-0142%281950%293%3A1%3C32%3A%3AAID-CNCR2820030106%3E3.0.CO%3B2-3.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. In *Proc. Interspeech 2020*, pages 2162–2166, 2020. doi: 10.21437/Interspeech.2020-2516. URL http://dx.doi.org/10.21437/Interspeech.2020-2516.

Andrea Zangrandi, Alessandro Mioli, Alessandro Marti, Enrico Ghidoni, and Federico Gasparini. Multimodal semantic battery to monitor progressive loss of concepts in the semantic variant of primary progressive aphasia (svppa): an innovative proposal. *Aging, Neuropsychology, and Cognition*, 28(3):438–454, 2021.

Anthony Zhang. Speech Recognition (Version 3.8) [Software]. Available from https://github.com/Uberi/speech_recognition#readme, 2017.

Vitor C Zimmerer, Chris JD Hardy, James Eastman, Sonali Dutta, Leo Varnet, Rebecca L Bond, Lucy Russell, Jonathan D Rohrer, Jason D Warren, and Rosemary A Varley. Automated profiling of spontaneous speech in primary progressive aphasia and behavioral-variant frontotemporal dementia: an approach based on usage-frequency. *Cortex*, 133: 103–119, 2020.