Alzheimer's Dementia Detection Through Machine Learning: Analyzing Linguistic and Acoustic Features in Spontaneous Speech

by

Zehra Shah

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

Master of Science

Department of Computing Science

University of Alberta

Abstract

With the rapid aging of the world's population, the global burden of aging-related mental disorders, such as Alzheimer's dementia (AD), is also on the rise. Unfortunately global healthcare systems are vastly under-resourced, which means that many people who need mental health services are unable to receive it. There is a critical need for timely, inexpensive, objective and scalable mental health screening and monitoring methodologies to augment currently available diagnostic tools. Speech analysis has the potential to be utilized as a window into the state of the human mind, meaning it could provide support for timely, reliable, and objective screening of many psychiatric disorders including AD.

In this dissertation, we present our research on machine learned models that can diagnose AD based on linguistic and acoustic features derived from speech. We show that AD can be detected reliably from spontaneous speech samples, and that this can be done even independently of the language spoken. The first part of this thesis presents machine learned models based on linguistic and acoustic features derived from spontaneous English speech samples. We find that linguistic features alone perform well, reaching 85% balanced accuracy on a hold-out test set, and ensemble models based on linguistic and acoustic features show comparable or slightly lower accuracy. The second part presents models that use features derived from measures of speech rate, complexity and intelligibility, this time in a cross-lingual setting (training on English speech samples and testing on Greek speech samples). These learned

models, despite the significant domain shift between the training and test sets, reached a relatively high balanced accuracy of 70%, showing that AD detection from speech is possible even across two different languages. Furthermore, we provide some exploratory data analyses of the features derived in the cross-lingual experimental setting, and show that some of these features have a visibly discriminating pattern that can be successfully utilized for clustering the samples.

This work paves the way for building automated machine learned systems for detecting and monitoring AD. With further validation on larger and more diverse data sets, such systems have the potential of being deployed at scale to flag early signs of AD and monitor the progression of AD severity.

Preface

Chapter 3 "Language and acoustic models for detecting Alzheimer's dementia from English speech" reproduces a publication [56] written jointly with Jeffrey Sawalha, Mashrura Tasnim, Shi-ang Qi, Prof. Eleni Stroulia and Prof. Russell Greiner. My contributions were to develop the language-based models presented in this paper, as well as to help write the paper and also serve as project manager and coordinator. The preface and the notes and comments sections of this chapter are entirely new additions for the purpose of this thesis.

Chapter 4 "Language-agnostic representations for detecting Alzheimer's dementia from multilingual speech" reproduces a publication [55] that was written jointly with Fei Wang, Shi-ang Qi, Mahtab Farrokh, Mashrura Tasnim, Manos Plitsis, Prof. Nassos Katsamanis, Prof. Eleni Stroulia and Prof. Russell Greiner. My contributions were to serve as project manager/coordinator, liaise with challenge organizers, manage research data and modeling results, as well as develop methods based on acoustic and text embedding techniques. I also contributed towards writing the publication and developing the conference presentation. The preface and the notes and comments sections of this chapter are entirely new additions for the purpose of this thesis.

To Ali

For all your love and support always

To Mikail and Zoiya

For being my beacon of inspiration

Acknowledgements

If I have seen further it is by standing on the shoulders of giants

– Isaac Newton, 1675.

First and foremost, I would like to express my deepest gratitude to my supervisor and mentor, Professor Russell Greiner, without whose patient and unwavering support, and thoughtful guidance, my research of the last few years would not have been possible. I learned many valuable lessons from Russ, not only for research and academia but life in general, for which I am truly grateful.

I would also like to thank my research collaborators: Professor Eleni Stroulia, whose mentorship and encouragement has been extremely uplifting for me; Professors Andrew Greenshaw, Matthew Brown, Lili Mou, and Suzette Bremault-Philips, whose insightful discussions helped shape my research; Jeff Sawalha, Mashrura Tasnim, and Shi-ang Qi, with whom I had the pleasure to collaborate on several research projects related to speech analysis for psychiatry; and other collaborators including Fei Wang, Mahtab Farrokh, Nassos Katsamanis, Manos Plitsis, Steve Heisig, Raquel Norel, Guillermo Cecchi, Monica Sawchyn, and John Whitnall. I am grateful for my interactions with each of you.

I must express my gratitude to my lab partners, both in the Computational Psychiatry group as well as in the Greiner Lab. I have immensely enjoyed our weekly interactions and lab engagements. I want to especially thank Roberto Vega, Negar Hassanpour, and Jackie Harris, for their friendship over the years.

My research was funded generously by the Department of Computing Science at the University of Alberta, the Alberta Machine Intelligence Institute

(Amii), and the IBM Center for Advanced Studies. I am grateful for this support.

Lastly, I would like to thank my family for their support and encouragement. My parents and mother-in-law never really understood my passion for higher education but supported me anyway, for which I am grateful. My children, Mikail and Zoiya, have been my biggest cheerleaders and have enthusiastically listened to me talking about my research endlessly (and even chipped in with ideas once in a while!). They are my reason and my inspiration. And last but never the least, my husband Ali has always been my rock, my confidante and my greatest supporter. I am immensely grateful for his love and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Mental health is a critical issue for the global human population. In 2019, one in every eight individuals was living with some form of mental disorder [44]. Further, as a fallout of the COVID-19 pandemic, just in the year 2020 there was a significant rise (estimated at 26 to 28 percent) in the number of people living with anxiety and depression. With the rapid aging of the world's population, the global burden of aging-related mental disorders, such as dementia, is also on the rise. More than 55 million people currently live with dementia, with nearly 10 million new cases being added each year. Of these, about 60 percent live in low- and middle-income countries with limited access to mental healthcare [45]. Dementia, with the most common form known as Alzheimer's dementia (AD), is a particularly insidious mental disorder characterized by progressive degeneration of nerve cells in the brain. This degeneration results in the deterioration of cognitive function, affecting memory, thinking and the ability to perform daily activities. The societal burden of dementia is substantial, not only on patients but on caregivers and the healthcare system as well.

Unfortunately global healthcare systems are vastly under-resourced. As a result a large number of people needing mental health services are unable to receive it. There is a critical need for timely, inexpensive, objective and scalable mental health screening methodologies, with the potential to be deployed at a larger scale as well as provide continuous (or more frequent) mental health monitoring. Currently, a diagnosis of dementia, for example, is made

based on a resource-intensive combination of clinically administered cognitive assessments, brain imaging, cerebrospinal fluid and blood testing [2]. However, speech analysis has the potential to be utilized as a window into the state of the human mind, and can provide support for timely, reliable and objective screening of many psychiatric disorders including dementia [28]. The advent of smartphones and wearable technology has enabled frequent and inexpensive measurement of personal health data. These innovations, coupled with advancements in machine learning and speech signal processing, set the stage for developing automated methods based on speech for identifying and monitoring a range of psychiatric disorders.

In this dissertation, we present and discuss our research on machine learning models based on features derived from speech for the detection of Alzheimer's dementia (AD). We show that AD can be detected reliably from spontaneous speech samples, even independent of the language spoken. This work paves the way for building automated machine learned models for detecting and monitoring AD. With further validation on larger and more diverse data sets, such systems have the potential of being deployed at scale to flag early signs of AD and monitor the progression of AD severity.

Our main contributions are:

(i) Demonstrating that speech features, both language and acoustic, can effectively determine whether a subject has dementia, and also its severity;

(ii) Showing that features based on language semantics and lexical complexity are particularly important, and that these features need not be highly complex to provide significant predictive performance on the given task;

(iii) Deriving novel features that capture aspects of speech production relevant to identifying dementia *irrespective* of the language being spoken, and thus providing the potential to develop models for low-resource languages as well;

(iv) Demonstrating that machine learned models based on these derived speech features can be used to detect Alzheimer's dementia with sig-

nificant accuracy, even across *different* spoken languages; and

(v) Exploring the structure of the derived feature space in the multilingual Alzheimer's dementia setting, via data visualization and unsupervised clustering.

The rest of this thesis is organized as follows: Chapter 2 introduces the necessary background for understanding the utility of speech analysis in computational psychiatry and provides a literature review describing the use of speech technologies for the assessment of psychiatric disorders; Chapter 3 presents our work on machine learning models using language and acoustic features for detecting AD from English speech samples; Chapter 4 describes our research on a similar but harder problem, developing machine learning models using language-agnostic speech features to enable the detection of AD from both English and non-English speech samples; Chapter 5 provides further exploratory analysis of the language-agnostic speech representations from Chapter 4; and Chapter 6 concludes this thesis with a discussion and comparative analysis of the results, as well as potential future applications and research directions. Additionally, Appendix A describes a pilot study we conducted using speech analysis for detecting post-traumatic stress disorder (PTSD) in military veterans.

Chapter 2

Background and literature review

Speech has long been known to be important in the diagnosis of psychiatric disorders (e.g. depression, schizophrenia, bipolar disorder, psychosis, autism, post-traumatic stress disorder, etc.). Speech provides a rich and varied view of the state of human cognition. It enables the analysis of several different behavioral and biological signals, including language, emotion, acoustic and non-verbal paralinguistic cues. The assessment of speech is therefore a crucial component of diagnostic and prognostic assessment by clinicians [57].

The intimate linkage between cognitive processes and speech production makes speech analysis a promising non-invasive biomarker for identifying some forms of cognitive deterioration and thus supporting the diagnosis and monitoring of several types of mental illness. In recent years, there has been significant interest in developing automatic speech analysis methods for the assessment of various types of mental disorders. The rapid progress of speech and machine learning technologies has enabled the development of automated methods of speech analysis to screen for psychiatric disorders including depression, anxiety, post-traumatic stress disorder, psychosis, schizophrenia, dementia, and mild cognitive impairment (MCI) [12], [22], [28], [36], [41], [60].

Human speech and language production is a complex process involving a variety of cognitive abilities, including working and short-term memory, planning and executive function, and knowledge of lexical, semantic and syntactic concepts [27]. The cognitive decline typically associated with Alzheimer's de-

mentia can manifest in the deterioration of spoken language, which in turn affects an individual's ability to perform daily activities and interact with their environment [59]. In fact, there is evidence to suggest that speech and language impairment may occur even in the pre-clinical stage of Alzheimer's disease, known as the Mild Cognitive Impairment (MCI) stage [11]. Hence automated speech-based machine learning models to detect and monitor dementia status are very promising.

With this potential in mind, there have recently been a myriad of studies exploring the development of speech-based machine learning methods for automatically detecting Alzheimer's dementia. In order to summarize the relevant literature, we provide here a brief review of a recent survey paper [18], which we found most relevant to the work described in this thesis, and additionally point the reader to two other similar survey papers for further research [38], [47]. Also, a few more task-specific research papers have been discussed in the related work sections of the publications reproduced in Chapters 3 and 4.

The review compiled by Fuente Garcia *et al.* [18] covered 51 peer-reviewed publications on speech-based machine learning methods for Alzheimer's dementia spanning the years from 2000 to 2019. Most of the reviewed papers (41 out of 51) attempt the binary classification task of predicting the presence or absence of cognitive impairment. Of the remaining papers, seven papers attempt a multi-class classification of three or four disease stages, two explore longitudinal cognitive changes, and one attempts to discover relevant patterns via cluster analysis. Most studies reported scores on the Mini Mental State Exam (MMSE) cognitive assessment to quantify dementia severity. This is despite criticism that MMSE is biased due to ceiling and/or floor effects, reducing its sensitivity and/or specificity for detection of pre-clinical AD [17], [23]. Ceiling (and floor) effects occur when personal characteristics of the tested individuals (such as race, gender or education levels) affect their performance on a cognitive assessment tool, independently from their cognitive functioning which is actually being assessed. This can create bias in the assessment results.

In terms of speech tasks performed by study participants [18], there is sig-

nificant heterogeneity, with tasks including verbal fluency, story recall, picture description, reading passages, and conversational interviews. The authors note that, although constrained lab-based tasks have their merit for standardization and control, it is also advantageous to study spontaneous speech elicited from study participants for characterizing dementia, as this type of data can be captured in a natural setting repeatedly over time. They also point out that a picture description task can elicit relatively spontaneous speech, since participants may describe the picture in any way they want, although the content is somewhat constrained. 70% of the reviewed papers used a picture description task to elicit spontaneous speech, whereas dialogue data is used less frequently and more heterogeneously (structured, semi-structured, and conversational interviews). Additionally, more than half the studies rely on manually transcribed data, thus limiting their practical applicability. In terms of data set sizes, only 27% of papers use 100 or more study participants. Furthermore, there are only 6 studies working with data sets having 100 or more participants per experimental group (with either two, three or four groups for classification), all of which used the DementiaBank Pitt corpus [8]. Regarding data balance, only 20 out of the 51 papers studied report a class-balanced data set, and out of these only one study balanced age, gender and education within and between classes as well. This is important to consider because findings based on imbalanced data have a higher risk of bias, especially when using smaller data sets. Regarding feature engineering, most studies rely on automatic speech recognition (ASR) and voice activity detection (VAD) to preprocess the samples, then compute text-based and acoustic features such as type-token ratio, idea density, syntactic complexity, spectral features, speech rates and pause features. Most studies use conventional machine learning methods (e.g. decision trees, support vector machines), with only a few attempting deep learning based approaches, likely due to the relatively small data sets sizes. For evaluation, cross-validation (CV) is the preferred method for most studies, although many studies neglect to mention whether a nested CV is implemented for systematic hyperparameter tuning. Performance results are mostly reported in terms of classification accuracy, and range from

less than 50% to more than 90%. However, the comparability of these performance measures is relatively low, due to potential biases in terms of data set size, data imbalance and non-standardized feature generation. The authors note that even though there is a relatively large number of relevant studies spanning the last twenty years, the heterogeneity of data and methodology means that comparability of results remains low, hindering the translation of these technologies into clinical practice.

We also provide here a brief overview of the editorial [32] describing the research papers published under the Frontiers Research Topic "Alzheimer's Dementia Recognition through Spontaneous Speech". The two main challenges here were AD classification and MMSE prediction, from spontaneous speech elicited from the Boston Cookie Theft picture description task (Figure 2.1). Some studies published under this special topic adopted a different experimental design, so these are ignored in this overview for better comparability. Several of the relevant papers found that linguistic features derived from transcripts were in general more discriminative compared to acoustic features [4], [20], [21], [39], [62]. Pause and disfluency features are also shown to be important in some studies [42], [61]. Another study [37] attempts to build a multi-modal early fusion model, by fusing information from the language and acoustic features at the sentence and word level using forced alignment and a clustering-based method termed Active Data Representation (ADR); note this achieved state-of-the-art accuracy (94%). Furthermore, most of the studies found no significant improvement in results by using deep learning based methods compared to traditional machine learning methods.

Despite progress in the research on detecting Alzheimer's dementia from speech, there are certain shortcomings that can be considered as opportunities for further research. As noted in the survey papers discussed above, the field lacks standardization of data and methodology, making reproducibility a limiting factor, and is plagued by lack of data, raising concerns of generalizability. Furthermore, the multilingual and cross-lingual applicability of the proposed methods is severely limited, which makes it difficult to bridge the gap for low-resource languages. The ADReSS challenges [29]–[31] have attempted to fill

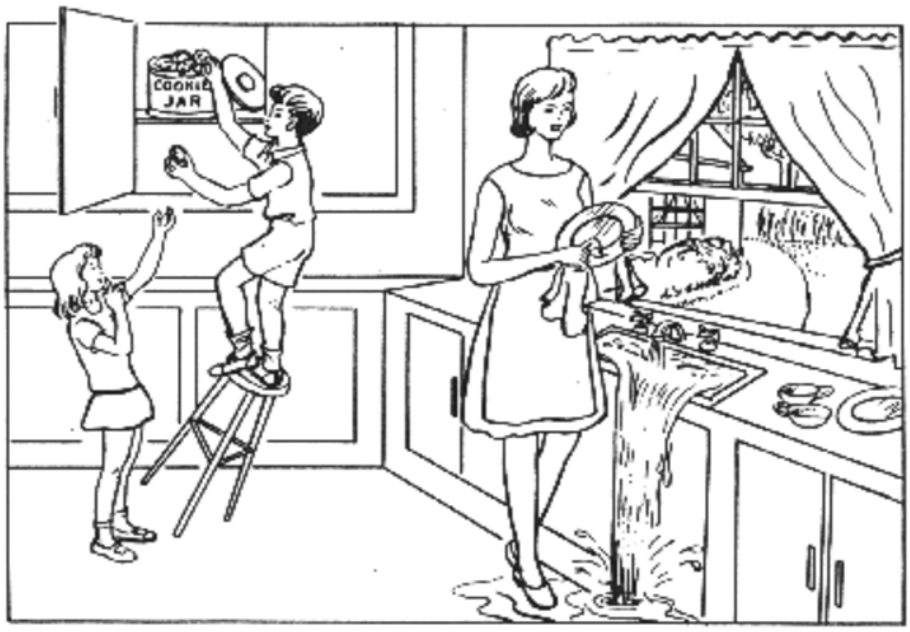

Figure 2.1: 'Cookie theft' picture used to elicit spontaneous speech response in the DementiaBank Pitt data set [8]

some of these gaps in order to move the research towards clinical adoption. Our work, presented and discussed in the next chapters of this dissertation, showcases our contributions to the field, both in terms of methods using text-based features derived from high-quality manual transcripts, as well as methods using more generalizable cross-lingual features based on multilingual ASR and pause distributions. Our hope is that the community pushes to create and share large, standardized, open, and multilingual speech data sets, so that the methods presented here can be comprehensively validated in order to enhance their scope and applicability.

Chapter 3

Language and acoustic models for detecting Alzheimer's dementia from English speech

3.1 Preface

In the next sections of this chapter, we reproduce our publication titled "Learning Language and Acoustic Models for Identifying Alzheimer's Dementia From Speech" [56] – with a few minor modifications. This preface is intended to provide additional background information and explain concepts in further detail that could only be addressed briefly in the original paper. Here we explain the Mini Mental State Examination (MMSE), an assessment tool used to measure dementia severity, and provide some more conceptual background regarding the language, fluency, and n-gram features, as well as the acoustic and prosodic speech features, used in this study. We also provide a brief description of the data set used in this study.

3.1.1 Mini Mental State Examination (MMSE)

The Mini Mental State Exam (MMSE) is a brief clinician-administered questionnaire designed to measure the severity of cognitive impairment in elderly individuals [10], [16]. The score on the MMSE can range from 0 to 30. A score of 23 or lower is considered to indicate cognitive impairment, with lower values associated with more severe cognitive decline. The items on the MMSE cover five areas of cognitive function: orientation, registration, attention and

calculation, recall and language, which are designed to test short-term memory, verbal memory and naming, visuospatial skills, and basic problem-solving skills. The MMSE is one of the most commonly used assessment tools for assessing severity of dementia, although it is not used as a standalone tool to obtain a conclusive dementia diagnosis. Further testing, including brain imaging, comprehensive neurological exam, and possibly genetic testing, would be required to establish dementia status.

3.1.2 Language features

Here we describe some of the concepts related to language modeling that we used in this work: bag-of-words and bigram models, some preprocessing steps like lemmatization and stopword removal, and the TF-IDF transform.

Language models

Language models are used to assign probabilities to sequences of words. Given a sequence of words w_i of length m, a language model assigns a probability $P(w_1, \ldots, w_m)$ to the sequence.

The **bag-of-words model** is the simplest possible language model. This model uses a drastic simplifying assumption, that each word in the sequence is completely independent of any of its preceding words. It is derived by first determining the vocabulary V of a corpus, which is just the set of unique words seen in the corpus. Then a bag-of-words vector is computed for each document in the corpus, with an element in the vector storing a frequency count of the number of occurrences of the corresponding word in that document. Taken together, these bag-of-words vectors form a fixed-dimension matrix of word counts, with each row corresponding to a single document and each column associated with a single word in the corpus vocabulary. A bag-of-words model completely disregards word context and ordering, and only depends on the *frequency* of occurrence of individual words.

The **n-gram model** is a relatively more complex language model, where an n-gram is a sequence of n words [24]. Compared to a bag-of-words model that does not consider word context, an n-gram model takes into account a context

window (or a history) of size $n-1$. So a 2-gram (or *bigram*) is a sequence of two consecutive words, like ("please", "go"), ("go", "outside"), and ("outside", "and"), and a 3-gram (or *trigram*) is a sequence of three consecutive words, like ("please", "go", "outside") and ("go", "outside", "and").

The n-gram model implicitly uses the Markov assumption, i.e. it assumes that the probability of the next word in a sequence depends only on a fixed size window of its preceding words. So a bigram model uses the single preceding word to determine the probability of the next word in the sequence, whereas a trigram model uses the last two words. In general, an n-gram model uses a context window of $n-1$ preceding words. The probability of a sequence of words $w_1, \ldots, w_m$ under the bigram model is then given by:

$$\begin{aligned} P(w_1, \ldots, w_m) &= P(w_1)P(w_2|w_1) \quad P(w_3|w_{1:2})P(w_4|w_{1:3}) \quad \ldots P(w_m|w_{1:m-1}) \\ &= P(w_1)P(w_2|w_1) \quad P(w_3|w_2)P(w_4|w_3) \quad \ldots P(w_m|w_{m-1}) \end{aligned}$$

The above equation uses the assumption that $P(w_k|w_{1:k-1}) \approx P(w_k|w_{k-1})$ for $k \in 1, \ldots, n$. The conditional probabilities can be approximated using maximum likelihood estimation (MLE). For the n-gram model, the MLE is obtained using normalized frequency counts in a given text corpus.

Preprocessing

To prepare the transcripts for computing language representations, we also performed a couple of preprocessing steps[1], namely lemmatization and stop-word removal. **Lemmatization** is the process of replacing each word in the document (or transcript in our case) with its root word. So the words "standing", "stands", and "stood" are all replaced by the single root word "stand". This step normalizes the input text, ensuring that the subsequent language model focuses on the actual semantic language content rather than the grammatical forms and parts of speech. Note that in our experiments, we did in fact build models using features based on part-of-speech tags as well, but did not find them useful for this particular task.

[1]We used the Python NLTK toolkit [7] for preprocessing

Another preprocessing step we used was **stopword removal**. The assumption here is that some generic words that occur with high frequency across almost all texts, like conjunctions and prepositions, have low information content from a predictive standpoint and therefore can be removed to normalize the input text. Some examples of such stopwords are: 'the', 'to', 'and', 'a', 'in', 'it', 'is', 'I', 'that', 'had', 'on', 'for'.

Term Frequency Inverse Document Frequency (TF-IDF)

The language models we described above represent a document corpus with a matrix of word frequency counts. However, raw frequency counts tend to overemphasize the importance of ubiquitous words such as "boy", "they" and "good". These words, that occur in a high frequency within *most* of the documents in the given corpus, are not particularly discriminative between individual documents. Hence, we would like to assign a lower weight to these high-frequency words, but at the same time preserve the weight of the more rare (and potentially more informative) words.

To do this, we use the TF-IDF normalization. It is defined as the product of the term, or word, frequency (i.e. word counts with respect to a single document) and the inverse of the document frequency (i.e. number of times the word occurs across all available documents). It ensures that a frequently occurring word in a document is given a higher weight, but *only* if that word also occurs relatively infrequently in all other documents in the corpus. In contrast, words such as "boy" and "good," that appear in essentially every document in the corpus, are seen as uninformative and therefore assigned a lower weight.

3.1.3 Acoustic and prosodic features

Besides the actual words being spoken (i.e. language content), there is also a lot of information encoded in the way in which the words are spoken. This opens the possibility of extracting further informative features from the acoustic speech signal. Speech signal processing views speech in the time domain as a time-varying waveform, and in the frequency domain as a combination

of sinusoidal components of different frequencies (using Fourier analysis [14]). These two different views of the speech signal allow us to derive many speech features that demonstrate varying levels of information content and relevance to the performance task. The **fundamental frequency** $F0$, and the **formants** ($F1$, $F2$, and so on) are speech features computed from the frequency spectrum, and are related to the natural pitch characteristics of speech. **Jitter** and **shimmer** are features related to variation in the period length and amplitude of $F0$, and are perceived as roughness, breathiness, or hoarseness in a speaker's voice [3]. Note also that delta and delta-delta features refer to the first- and second-order frame-to-frame difference between whichever speech features are being considered.

Speech analysis algorithms assume that the speech signal is stationary. This is seldom true over any meaningful time window of the speech signal. Therefore we segment the signal into overlapping windows (typically 20-30 milliseconds long), and make the assumption of stationarity within each window. Then the **spectrogram** is constructed by applying the Fourier transform separately to each window, and stacking each of the individual windows' spectra as columns of the spectrogram matrix. When these individual spectra are resampled to the Mel-frequency scale, the resulting matrix is known as the **Mel-frequency spectrogram** [14]. The Mel-frequency scale is a non-linear mapping of the audible frequency range, with the effect of expanding the detail in the lower frequencies and compressing it in higher frequencies (following typical human auditory perception, since human hearing is less sensitive at higher frequencies).

The Mel-frequency **cepstrum** is another representation computed by applying a Discrete Cosine Transform (DCT) on the log magnitude of the Mel-frequency spectrum. The word "cepstrum" is a word-play on 'spectrum', meant to indicate that it is a spectrum of a spectrum. The **Mel-frequency cepstral coefficients (MFCCs)** are a perceptually informed dimension reduction technique for speech signal processing, derived using the Mel-frequency cepstrum (see Figure 3.2 below), and as such they have proven to be important features for speech analysis tasks.

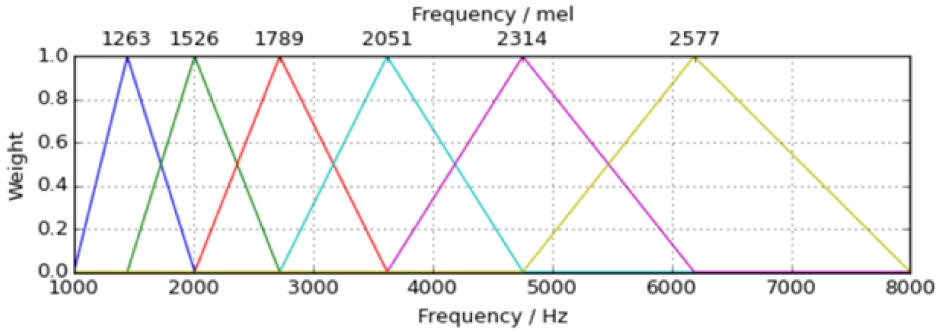

Figure 3.1: Example Mel filter bank, composed of a fixed set of logarithmically spaced triangular filters

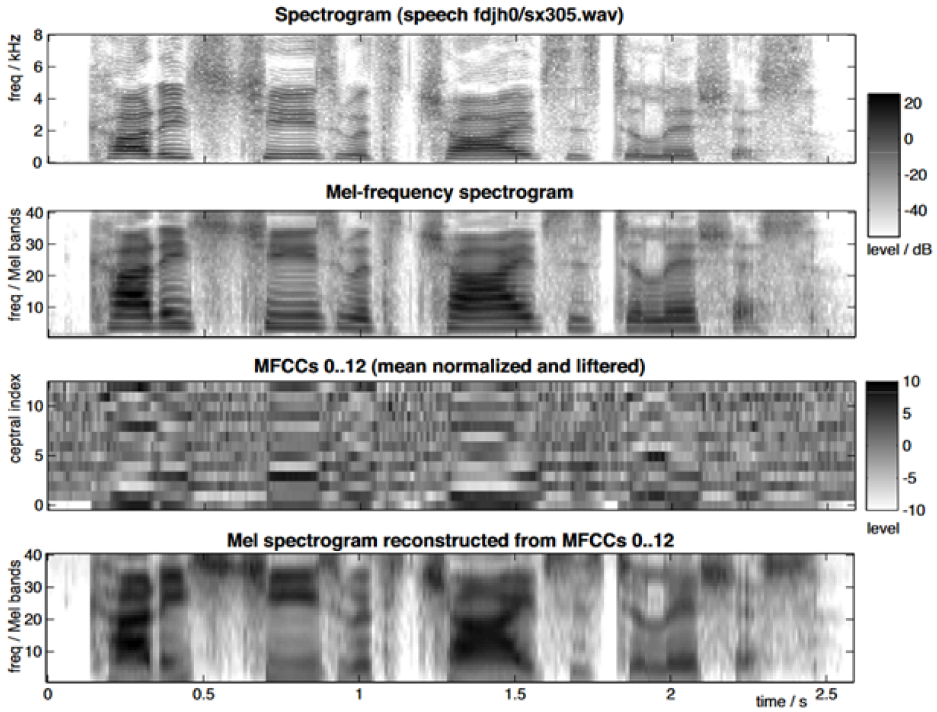

Figure 3.2: Mel-frequency spectrogram and MFCCs (this figure is Figure 12 in [14]). The top pane is a standard linear-frequency spectrogram. The second pane shows the same spectrogram, but resampled on a 40-bin Mel frequency scale (Mel spectrogram). The first 13 MFCCs, derived from the Mel spectrogram, are shown in the third pane. The fourth pane shows the Mel spectrogram reconstructed from the MFCCs. Notice that discarding higher order MFCCs effectively removes the high-frequency details of the Mel spectrogram in the second pane, while preserving the more salient frequency characteristics.

3.1.4 ADReSS Challenge data set

Table 3.1: ADReSS training set characteristics

			AD			non-AD
Age	**M**	**F**	**MMSE (sd)**	**M**	**F**	**MMSE (sd)**
[50, 55)	1	0	30.0 (n.a.)	1	0	29.0 (n.a.)
[55, 60)	5	4	16.3 (4.9)	5	4	29.0 (1.3)
[60, 65)	3	6	18.3 (6.1)	3	6	29.3 (1.3)
[65, 70)	6	10	16.9 (5.8)	6	10	29.1 (0.9)
[70, 75)	6	8	15.8 (4.5)	6	8	29.1 (0.8)
[75, 80)	3	2	17.2 (5.4)	3	2	28.8 (0.4)
Total	24	30	17.0 (5.5)	24	30	29.1 (1.0)

The data set provided for this challenge[2] is a subset of the DementiaBank Pitt corpus [6]. It was balanced in terms of age, gender and binary class labels. For details about the training and test data set characteristics, see Tables 3.1 and 3.2. The data included speech recordings of participants responding to the Cookie Theft picture description task (Figure 2.1) from the Boston Diagnostic Aphasia Exam [19], as well as high-quality transcripts annotated using the CHAT coding system [33].

Table 3.2: ADReSS test set characteristics

			AD			non-AD
Age	**M**	**F**	**MMSE (sd)**	**M**	**F**	**MMSE (sd)**
[50, 55)	1	0	23.0 (n.a.)	1	0	28.0 (n.a.)
[55, 60)	2	2	18.7 (1.0)	2	2	28.5 (1.2)
[60, 65)	1	3	14.7 (3.7)	1	3	28.7 (0.9)
[65, 70)	3	4	23.2 (4.0)	3	4	29.4 (0.7)
[70, 75)	3	3	17.3 (6.9)	3	3	28.0 (2.4)
[75, 80)	1	1	21.5 (6.3)	1	1	30.0 (0.0)
Total	11	13	19.5 (5.3)	11	13	28.8 (1.5)

Training labels were provided for both the classification task (binary AD vs non-AD) and the regression task (MMSE scores). Note that the binary classification labels are *not* obtained via a simple thresholding of the MMSE scores. The diagnosis of AD (or 'Probable AD' to be exact) was obtained by

[2]https://luzs.gitlab.io/adress/

the clinicians involved in the original study [6], through a combination of various assessment methods including extensive neuropsychological examination, laboratory tests, and (for some participants) post-mortem autopsies.

Each speech sample was segmented using voice activity detection, with a maximum duration of 10 seconds per segment, and the segmented data set was also made available (after some acoustic enhancements including noise removal and volume normalization). The average number of speech segments per participant was 24.86 (with a standard deviation of 12.84).

The next sections of this chapter reproduce the paper [56], with the final section providing some additional comments.

3.2 Introduction

This paper is motivated by the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge, hosted by the INTERSPEECH 2020 conference [30]. The data set provided in this challenge is a carefully curated subset of the larger DementiaBank corpus [6]. Among the various challenge submissions, the top-performing models analyzed both linguistic and acoustic features, and many of these top submissions used deep learning methods (including some pre-trained models) to generate their results. For example, Koo *et al.* [26] used an ensemble approach with bi-modal convolutional recurrent neural networks (cRNN), applied to a variety of feature sets from pre-trained acoustic and linguistic algorithms in addition to some hand-crafted features. They achieved an accuracy of 81.25% on their classifier evaluation and an RMSE score of 3.75. Another study by Balagopalan *et al.* [5] achieved an accuracy of 83.33% and an RMSE of 4.56 by appending a binary classification layer to a pre-trained language algorithm developed by Google: Bidirectional Encoder Representations from Transformers (BERT). The Sarawgi *et al.* [51] submission applied RNNs and multi-layered perceptrons (MLP) to various types of acoustic and linguistic features in an ensemble manner. They also used transfer learning from the classification models to the MMSE scores by modifying the last layer structure, achieving an RMSE of 4.6 and an accuracy

of 83.33%. Lastly, Searle *et al.* [54] used linguistic features only, with pre-trained Transformer-based models, and achieved their best performance using features computed from the full transcripts (including both participant and interviewer speech): a classification accuracy of 81% and an RMSE of 4.58. The commonality among these top submissions was the use of deep-learning methods along with pre-trained acoustic and/or language models.

Our study hopes to improve further by applying simple, computationally inexpensive ML techniques to natural language and acoustic information. In particular, we train models that use both acoustic and language features to distinguish AD from healthy age-matched elders and predict their MMSE scores. Our system feeds the acoustic features into one pipeline, and the linguistic ones in another. Each pipeline preprocesses the features, then uses internal cross-validation to tune the hyperparameters and select the relevant subset of features. We use ensemble methods to combine the various learned models, to produce models that can (1) label a speech sample as either AD or control, and (2) predict the associated MMSE score of that instance.

3.3 Methods

For this study, we were given a training set of 54 AD patients and an age- and gender-matched set of 54 healthy controls (this is a subset of the larger DementiaBank data set; see [6]). This subset of DementiaBank contained spontaneous speech samples of participants asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Exam ([19]). For each participant, we obtained

(1) the original recorded speech sample,

(2) the normalized speech segments extracted from the full audio sample after voice activity detection, audio normalization and noise removal,

(3) the speech transcript files annotated using CHAT (Codes for Human Analysis of Transcripts) transcription format [34], and

(4) some descriptive features about these individuals, including age, gender, binary class label (AD/control; the target for the classification task), and their

MMSE score (which we try to predict in the regression task).

The challenge organizers withheld a test set containing data from 24 AD and 24 healthy participants for final evaluation, the labels for which were subsequently provided as well. For further details of this data set, we refer the reader to [30].

We considered a set of possible base learners, applying each to some specific subset of the features – the (1), (2) and (3) mentioned above. We used internal 5-fold cross validation to identify which of these base learners was best. Due to the size of our data, we chose to use a 5-fold CV procedure. 10-fold CV or Leave-one-out CV procedure would result in small partitions, leading to possible overfitting (lower bias, higher variance). To ensure consistent and reliable comparison between our models, we defined and used a common set of folds that were balanced in terms of class labels (or MMSE scores) as well as gender. For each model, we evaluated performance metrics (average accuracy for classification, and average RMSE for regression) based on these test folds, as well as on the final hold-out test set.

3.3.1 Language and fluency features

The organizers provided transcripts that were annotated using the CHAT coding system [34]. First we extracted only the participant's speech from these transcripts, removing the interviewer's content. Then, using the CLAN (Computerized Language ANalysis) program for processing transcripts in the CHAT format, we computed the following set of global syntactic and semantic features for each transcript: type-token ratio (TTR) – the number of unique words divided by total number of words; mean length of utterance (MLU), where an utterance is a speech fragment beginning and ending with a clear pause; number of verbs per utterance; percentage of occurrence of various parts of speech (nouns, verbs, conjunctions, etc.); number of retracings (self-corrections or changes); and number of repetitions. We also computed a number of fluency features, including percent of broken words, part-word and whole-word repetitions, sound prolongations, abandoned word choices, word and phrase

repetitions, filled pauses, and non-filled pauses [35]. In total, we computed 62 such informative summary features for each transcript.

3.3.2 N-gram features

We processed the raw (unannotated) transcripts to compute bag-of-words and bigram features. First, we standardized the transcripts by converting them into a list of word tokens. Next, we used the WordNet lemmatizer [40] to find and replace each word with the corresponding lemma; for example, words like "stands", "standing" and "stood" were all replaced by the common root word "stand". Finally, we removed stopwords from each transcript, where stopwords are highly common (and presumably uninformative) words that may add noise to the data (such as "I", "am", "was", etc.), using a predefined stopwords list from the Python natural language toolkit (NLTK) package.

Next, we used the standardized transcripts to compute bag-of-words vectors (using words seen in the training set only) – that is, a vector of 514 integers for each transcript, where the k^{th} value is the number of times the k^{th} word occurred – and normalized these vectors with the Term Frequency-Inverse Document Frequency (TF-IDF) function, which is a normalization procedure that reflects how important a word is to a document in a corpus, effectively penalizing words that occur frequently in most of the documents in the corpus. For example, in our case the word "boy" might occur frequently in all transcripts (as one of the main subjects in the Boston Cookie Theft picture is a boy, see Figure 2.1), so it may not be very informative. Finally, we also computed bigram vectors in a manner similar to bag-of-words – where each bigram is a pair of words that appear adjacent to one another. Note that these bigrams are computed *after* preprocessing the transcripts, which included stopword removal. We found a set of 2,810 bigrams.

3.3.3 Acoustic features

Using the speaker timing information provided in the transcripts, we extracted the participants' utterances (removing the interviewer's voice) from the audio recordings, for a total of 1501 participant utterances from the training set, and

592 from the test set. We then normalized the audio volume across all speech segments. We computed four different sets of features from each audio segment using OpenSMILE v2.1 [15]. Note that our overall learner will consider various base-learners, each running on one of these feature sets.

(Feature Set #1) The **AVEC 2013** [58] feature set includes 2268 acoustic features including 76 low level descriptor (LLD) features and their statistical, regression and local minima/maxima related functionals. The LLD features include energy, spectral and voicing related features; delta coefficients of the energy/spectral features, delta coefficients of the voicing related LLDs and voiced/unvoiced duration based features.

(Feature Set #2) The **ComParE 2013** [53] feature set includes energy, spectral, MFCC, and voicing related features, logarithmic harmonic-to-noise ratio (HNR), voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. Statistical functionals are also computed, leading to a total of 6,373 features.

(Feature Set #3) Our third feature set consists of the following three feature sets. The **emo_large** [15] feature set consists of cepstral, spectral, energy and voicing related features, their first and second order delta coefficients as LLDs; and their 39 statistical functionals. The functionals are computed over 20 ms frames in spoken utterances. This produced 6552 acoustic features across the utterances. The **Jitter-shimmer** feature set is a subset of INTERSPEECH 2010 Paralinguistic Challenge [52] feature set, consisting of 3 pitch related LLDs and their delta coefficients. We also computed 19 statistical functionals of the LLDs on the voiced sections of the utterances, resulting in 114 features. Finally, we extracted 7 speech and articulation rate features by automatically detecting syllable nuclei [13], and used a script from the software program Praat to detect peaks in intensities (dB) followed by sharp dips. We also calculated other features, such as words per minute, number of syllables, phonation time, articulation rate, speech duration and number of pauses for each speech sample [9].

(Feature Set #4) We computed the **MFCC 1-16** features and their delta coefficients from 26 Mel-bands, which uses the fast Fourier transform

(FFT) power spectrum. The frequency range of the Mel-spectrum is set from 0 to 8 kHz. Inclusion of statistical functionals resulted in 592 features. This feature set is a subset of AVEC 2013 feature set [58].

We also added age and gender of the participants to each set of features.

3.3.4 Language based models

Given our two sets of linguistic features above (Sections 3.3.1 and 3.3.2), we explored various dimension reduction techniques and base learning algorithms to find the best performing pipeline. The dimension reduction techniques include Principal Component Analysis (PCA), Latent Semantic Analysis (LSA), and univariate feature selection using ANOVA F-values. The base learning algorithms explored for the classification task are logistic regression (LR), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB). For the regression task, the regression versions of the same algorithms are trained (except logistic regression is replaced by linear regression). Internal 5-fold cross-validation was used to tune the hyperparameters for each model based on accuracy.

The hyperparameters explored were:

Dimension reduction:

1. For classification models, dimension reduction with PCA using {10, 20, 30, 50} components, and LSA using {100, 200, 500} components;

2. For regression models, dimension reduction with PCA using {20, 30, 50} components, and LSA using {200, 500, 800} components.

Models:

1. SVM (regularization parameter C: {0.1, 1, 10, 100, 1000}, kernel: {linear, RBF, polynomial});

2. LR (regularization parameter C: 20 values spaced evenly on a log scale in the range $[10^{-4}, 10^{4}]$, loss function: {L1, L2});

3. RF (number of trees: {100, 300, 500, 700}, maximum features at each split: {5, 15, 25, 35, 45, 55}, minimum samples at leaf node: {1, 2, 3, 4}); and

4. XGB (maximum depth: {5, 6, 7, 8}, learning rate: {0.02, 0.05, 0.07, 0.1}, number of trees: {50, 100, 200, 500, 1000}).

The same hyperparameters were explored for the regression models as well (with the exception of replacing LR with linear regression).

Our internal cross-validation found the best-performing language-based classification model, which consisted of the following steps:

Step1: 5-component PCA transformation of the dense language and fluency features described in Section 3.3.1 (after standardizing using z-scores);

Step2: 50-component LSA transformation of the sparse unigram and bigram features described in Section 3.3.2 (after standardizing using TF-IDF transform); and

Step3: L1-regularized logistic regression

The best language-based regression model involved the following:

Step1: 30-component PCA transformation of the dense language and fluency features described in Section 3.3.1 (after standardizing using z-scores);

Step2: 100-component LSA transformation of the sparse unigram and bigram features described in Section 3.3.2 (after standardizing using TF-IDF transform); and

Step3: Random Forest Regressor, using 100 trees, minimum of 4 instances at each leaf node, and 25 features considered for each split.

3.3.5 Acoustic models

All acoustic features were real values and were therefore standardized using z-scores. We used PCA to reduce the dimensionality of the feature sets. For Feature Set #1 and Feature Set #2, we used PCA, and kept the minimum number of features capable of retaining 95% of the variance. In the case of Feature Set #3 and Feature Set #4, the number of principals were determined through internal 5-fold cross-validation. Therefore, the dimension of Feature

Set #1 is reduced from 2,268 to 700, Feature Set #2 from 6,373 to 1100, Feature Set #3 from 6,552 to 1000 and Feature Set #4 from 592 to 50. Next, we selected the best 50 principal components from Feature Set #1, and the best 70 from Feature Set #3, applying a univariate feature selection method based on ANOVA F-value between the label and each feature. For Feature Set #2, we calculated feature importance weights using a decision-tree regression model, and selected only the features with importance weight higher than the mean.

After this pre-processing stage, our system fed these audio features to various machine-learning algorithms, that each identify patterns of features that can distinguish dementia patients from healthy controls (the classification task), and can compute a subject's MMSE score (the regression task).

We explored several learning algorithms, including Adaboost, XGB, RF, gradient boosting (GBT), decision tree (DT), hidden Markov model (HMM) and neural network (NN). Our superlearner used internal 5-fold cross-validation to tune the hyperparameters of the classifiers and regressors. The predictions were made in two steps. In the first step, the classifiers (and resp., regressors) were trained and tested with acoustic features, age and gender to predict whether the speech segment was uttered by a healthy control or an AD patient (and resp., to predict that subject's MMSE score). Next, weighted majority vote classification was performed to assign each subject a label of health control or AD, based on the majority labels of the segment level classification. The predicted MMSE scores on all the segments of one subject were averaged to calculate the final MMSE score of that subject.

The classifiers of acoustic data, that performed best (in cross-validation on the training set) are the following (in order):

1. Neural network with 1 hidden layer, trained on Feature Set #1

2. AdaBoost Classifier with 50 estimators and logistic regression as the base estimator, trained on Feature Set #4

3. Adaboost with 100 estimators and DT as the base estimator trained on Feature Set #3.

The three regressors with the lowest root mean square error (RMSE) were

1. Gradient boosting regressor, trained on Feature Set #4
2. Decision tree with number of leaves 20, trained on Feature Set #2
3. Adaboost regressor trained on Feature Set #3 with 100 estimators.

3.3.6 Ensemble models

After obtaining our best-performing acoustic and language-based models, we computed a weighted majority-vote ensemble meta-algorithm for classification. We chose the three best-performing acoustic models along with the best-performing language model, and computed a final prediction by using weights learned on the individual model predictions. The weights assigned to each model were proportional to that model's mean cross-validation accuracy, such that the best performing model is given the highest weight in the final prediction. For regression (to predict the MMSE scores), we computed an unweighted averaging of our best language and acoustic models.

3.4 Results

3.4.1 Classification

Table 3.3 presents the results for the classification task. The model that obtained the highest average cross-validation accuracy (81% $\pm$ 1.17%) is a weighted-majority-vote ensemble of the best language-based model and three of the best acoustic-based models. The second highest accuracy (80% $\pm$ 0.00%) was obtained by the language-based logistic regression. However, a McNemar test reveals that these two models do not exhibit a statistically significant difference in performance (McNemar test statistic $= 4.0, p > 0.05$). This is also evident by the performance of these two models on the final held-out set, where the language-based logistic regression gives the highest accuracy (85%) and the weighted-majority-vote ensemble gives a slightly lower accuracy (83%). Using McNemar's test to compare these two models on the held-out test set, we

Table 3.3: Results of our best performing classification models distinguishing AD from non-AD subjects. The 'Baseline (Acoustic)' model is described in [30]. The right-most column shows accuracy on the held-out test set of 48 subjects (24 AD and 24 non-AD). The rest of the table lists model performance using 5-fold cross-validation on the training set of 108 subjects (54 AD and 54 non-AD).

Classifiers	Class	Precision	Recall	F1 Score	Accuracy	Accuracy (Hold-out set)
Logistic Regression (NLP)	AD	0.71	0.60	0.75	**80% ± 0.00%**	**85%**
	HC	1.00	1.00	0.83		
	OVR	0.80	0.80	0.79		
SVM (NLP)	AD	0.68	0.84	0.75	72% ± 1.85%	73%
	HC	0.79	0.60	0.68		
	OVR	0.73	0.72	0.72		
Majority vote (NLP + Acoustic)	AD	0.74	0.96	0.83	**81% ± 1.17%**	**83%**
	HC	0.94	0.66	0.78		
	OVR	0.84	0.81	0.81		
Majority vote (Acoustic)	AD	0.71	0.78	0.74	73% ± 1.36%	65%
	HC	0.76	0.68	0.72		
	OVR	0.73	0.73	0.73		
Baseline (Acoustic)	AD	0.57	0.52	0.54	57%	63%
	HC	0.56	0.61	0.58		
	OVR	0.57	0.57	0.56		

AD Alzheimer's dementia HC Healthy control OVR Overall rating

obtain a test statistic of 3.0, with $p > 0.05$, indicating that the performance difference between these models is not statistically significant.

Note that our ensemble model, which uses only acoustic features, performs significantly better than the "baseline model" (provided by the organizers), which also uses acoustic features only.

3.4.2 Regression

Table 3.4 shows the root mean square error (RMSE) of various regression models; columns 2 and 3 show the average RMSE and R2 scores over the 5 cross-validation folds, and columns 4 and 5, on the hold-out test set (provided by the organizers of the challenge). These results show that the language-based model obtains the best RMSE of 6.43 on the cross-validation set and 5.62 on the hold-out set. The combined language-acoustic model did not perform as well as the standalone language-based model, with an average RMSE of 6.83 on the cross-validation set and 6.12 on the hold-out set. Further, the Wilcoxon test between the RMSEs of the two best models (best acoustic + best

Table 3.4: Results of our best performing regression models predicting a subject's MMSE score (ranging from 0 to 30, with lower values indicating more severe dementia). The 'Baseline (Acoustic)' model is described in [30]. As in Table 3.3, the columns on the right show RMSE and R^2 on the held-out test set of 48 subjects (24 AD and 24 non-AD). The middle columns list RMSE and R^2 using 5-fold cross-validation on the training set of 108 subjects (54 AD and 54 non-AD).

Regressors	**RMSE**	**RMSE (Hold-out Set)**
Random Forest (NLP)	6.43 ± 0.18	**5.62**
Gradient Boosting (Acoustic)	6.89 ± 0.17	6.67
Random Forest (NLP) + Gradient Boosting (Acoustic)	6.66 ± 0.18	**6.01**
Majority vote (All models)	6.85 ± 0.16	6.12
Baseline (Acoustic)	7.30	6.14

language-based combination versus best stand-alone language-based), returns a test statistic of 66.0 with $p < 0.05$ on the hold-out set, and a test statistic of 1375.0 with $p < 0.05$ on the cross-validation set. This suggests that these two models are significantly different in performance (i.e., we cannot reject the claim that they are significantly different in performance).

We also report the coefficient of determination (R^2) for all our models: the best R^2 was 0.17 on the validation folds and 0.14 on the held-out test set. These low numbers are expected, given the relatively small size of this challenge data set and the complexity of the condition. Interpreting this statistic in an absolute sense is problematic, especially as we did not find any other study using the same data set that reported this metric. We note that models based on language features achieved the best R^2 values, which further supports our claim that language features are very important for this task.

3.4.3 Discussion

We investigated a variety of ML models, using language and/or acoustic features, to identify models that performed well at using speech information to distinguish AD from healthy subjects, and to estimate the severity of AD.

Our results, of over 85% accuracy for classification and approximately 5.6 RMSE for regression, demonstrate the promise of using ML for detecting cognitive decline from speech. In our investigation, we explored multiple different combinations of feature sets and ML algorithms; in the future, it would be interesting to delve deeper into the behavior of our best models, to determine the contribution of individual (or groups of) features to the model's ability to distinguish AD patients from healthy controls. Further, although we have currently used the full set of standard stopwords for removing noise in our language models, it may be worthwhile to see whether using a reduced set of stopwords (for example, not removing pronouns) might be more advantageous.

Our current best-performing models outperform recent results reported in the literature and provide evidence that, for discriminating between subjects with AD versus healthy controls, features based on language (semantics, fluency and n-grams) are very useful. Compared to other top ranked results, our methods do not involve complex, computationally expensive algorithms. Instead, we used an ensemble approach with simple models to produce competitive results.

Furthermore, a weighted majority vote of acoustic and language based models demonstrates competitive performance, implying that a combination of acoustic and language features also holds potential. Finally, comparing only acoustic models, we find that cross-validation performance improves significantly compared to the baseline model [30] for both the classification and regression tasks, although test set performance of the acoustic-only model on the regression task does not show improvement. Hence, given the relatively small data set size and the potential for overfitting, we cannot say conclusively that the acoustic feature sets computed here are effective by themselves for predicting MMSE scores quantifying AD severity.

Our competitive performance, obtained using simple feature engineering along with classical machine learning algorithms, indicates that putting together an efficient machine learning pipeline from basic building blocks can achieve nearly state-of-the-art results for the learning tasks explored in this study. This result suggests that, for detecting AD from speech, it may be

useful to explore traditional feature engineering and machine learning tools, especially in a limited data setting, as this will additionally provide for better interpretability and reproducibility compared to more complex deep learning based methods.

3.5 Notes and comments

In this section (which was not part of the original publication), I elaborate upon additional comments we received from journal reviewers and provide further rationale for our design choices. I also discuss some limitations of this work, and avenues for future research.

It was no surprise, given the focus on deep learning approaches in modern machine learning, that some reviewers wondered why we did not choose deep learning based feature extraction strategies, but instead used the hand-crafted features that we employed in this work. There are three main reasons for this. Firstly, we felt that the relatively small size of the available training data would likely entail a high risk of overfitting, which we wanted to avoid. Secondly, it is more difficult to analyze, validate and interpret black-box deep-learning-based features compared to features obtained via traditional feature engineering guided by subject matter expertise. Thirdly, since we were able to achieve high accuracy using traditional machine learning and feature engineering, without using highly parameterized deep learning models, we felt there was no need to add more complexity to our approach. With our systematic data-driven approach, using nested cross-validation to explore a variety of model architectures and pipelines, we developed models that demonstrated a high classification accuracy (85%) that ranked 3rd out of 34 teams globally.

There are, however, some important limitations of this work that warrant further discussion. Firstly, our highest performing models were trained using language-based features derived from the high quality transcriptions that were provided to us by the challenge organizers. Hence these models depend heavily on the quality of transcriptions available, and therefore cannot be considered fully automatic. Since good quality transcriptions are expensive to

obtain, even in the English language setting, the bottleneck this creates has a direct impact on the feasibility of deploying our models in a broader clinical context. These methods need to be rigorously validated using transcripts generated from automatic speech recognition (ASR) systems, which are generally much noisier and more error-prone, to quantify their robustness to this more challenging setting.

A second limitation of the models described in this paper is that the data set on which these models are trained is quite limited in size and scope. Even though it is balanced for age and gender, we are not provided any additional details about the income and education levels, demographics nor ethnicity of the study participants. We can make an educated guess that most participants are ethnically Caucasian, which in itself severely limits the generalizability of methods developed using this data. Furthermore, we are also not provided with information on confounding psychological disorders, such as depression status, which may be highly relevant to understand the results and their generalizability. Additionally, since the data set is a purely English language one, methods developed using it might not transfer to any other languages, especially lower-resource ones (although the study reproduced in the next chapter tries to somewhat mitigate this last limitation).

Nevertheless, this work was important because we were able to compare our methods against other competitors in a controlled environment. The data set was carefully curated so that some confounding variables such as age and gender could be ignored. This allowed us to concentrate on the speech analysis and machine learning aspects of the work, which could be used as a foundation for building automated speech-based machine learning frameworks for dementia detection.

Chapter 4

Language-agnostic representations for detecting Alzheimer's dementia from multilingual speech

4.1 Preface

In the next sections of this chapter, we reproduce our publication titled "Exploring Language-Agnostic Speech Representations Using Domain Knowledge for Detecting Alzheimer's Dementia" [55] (with small modifications for readability), which briefly described the models we developed for the ICASSP-hosted challenge "ADReSS-M: Multilingual Alzheimer's Dementia Recognition using Speech"[1]. This preface is intended to provide additional background information and explain concepts in further detail that could only be addressed briefly in the original paper. We use this preface to discuss the data set that we used, as well as the features that we found useful.

4.1.1 Challenge motivation and data set

While there has been considerable interest recently in developing automated methods for dementia detection using speech [18], a large proportion of the literature has focused on developing models for the English language only. These methods, including our work for the ADReSS challenge reproduced in

[1] https://luzs.gitlab.io/madress-2023/

Table 4.1: ADReSS-M training set characteristics (237 English speech samples)

	AD			non-AD		
Age	**M**	**F**	**MMSE**	**M**	**F**	**MMSE**
[50, 55)	1	1	29.0	1	0	23.0
[55, 60)	6	16	29.2	7	6	17.3
[60, 65)	7	13	29.0	7	8	19.0
[65, 70)	11	23	28.9	7	22	19.2
[70, 75)	10	18	28.7	9	21	17.3
[75, 80)	5	4	29.3	12	22	17.2
Total	40	75	28.9	43	79	17.9

Table 4.2: ADReSS-M test set characteristics (46 Greek speech samples)

	AD			non-AD		
Age	**M**	**F**	**MMSE**	**M**	**F**	**MMSE**
[50, 55)	2	0	28.0	0	0	-
[55, 60)	0	0	-	0	0	-
[60, 65)	0	10	29.2	1	2	20.3
[65, 70)	2	4	29.2	1	2	18.7
[70, 75)	2	1	29.0	0	7	20.7
[75, 80)	0	1	28.0	2	3	18.6
[80, 85)	0	2	28.0	1	2	24.0
[85, 90)	0	0	-	0	1	25.0
Total	6	18	28.9	5	17	20.5

the previous chapter, have found that in general, the semantic content of spontaneous speech is an important determinant of dementia status. However, such language-focused models unfortunately do not have the capacity to be directly utilized across different languages and, in that sense, are heavily language-specific. The ADReSS-M challenge took a different, potentially more globally impactful, approach to the problem of dementia detection from speech. The motivation of this challenge was to spur the development of speech-based machine learning models capable of detecting early signs of dementia *independent* of the spoken language. Hence the methods were expected to be transferable across different languages with respect to their ability to detect dementia from speech samples.

The training data set consisted of a balanced set of 237 spontaneous English speech samples from the DementiaBank corpus. It is an age-, gender-, and

label-balanced data set (see Table 4.1) of audio recordings of participants describing the Boston Cookie Theft picture [19]. No transcripts or segmented audio clips were provided. The test data set is a (similarly) balanced set of 46 audio recordings (see Table 4.2) of spontaneous speech in the Greek language, with participants describing a different picture (i.e., not the Boston Cookie Theft picture, but instead a picture of a "lion lying with a cub in the desert while eating" [29]). This test set was retained by the challenge organizers as a held-out set to compare the final performance of the submissions. Challenge participants were provided a development data set consisting of 8 Greek audio samples as well.

4.1.2 Pause rate features

Prior studies on AD [11], [59] suggest that memory deficiency and cognitive decline result in deterioration of speech fluency, due in part to diminished word-finding capacity. Therefore, in this work we hypothesized that disfluencies in speech, represented by pauses and hesitation, would be important to distinguish between dementia patients and healthy controls. Hence we derived features to represent the distribution of *pauses* in the speech samples. To achieve this, we first used the voice activity detection procedure from the OpenSMILE speech processing toolkit [15] to divide each audio sample into voiced and unvoiced segments. From the timestamps of the unvoiced segments, we determined the individual lengths of all pauses. Then we derived the following pause-related features:

(i) Sum of durations of all unvoiced segments per sample (AD vs control)
This set of features was meant to capture the total silence duration per speech sample.

(ii) Durations of individual voiced segments (AD vs control)
This set of features captures the distribution of the lengths of voiced segments (i.e. the length of speech without a pause).

For each type of feature, we computed basic statistics (like the mean sum of unvoiced segment lengths, and the mean length of voiced segments). We

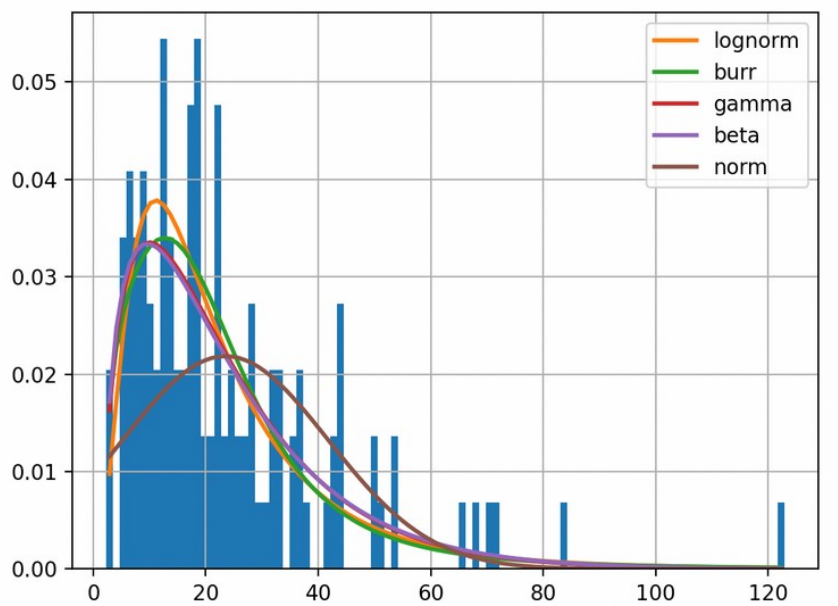

Figure 4.1: Histogram of pause distribution for AD patients, and best-fit PDFs, for several parametric models

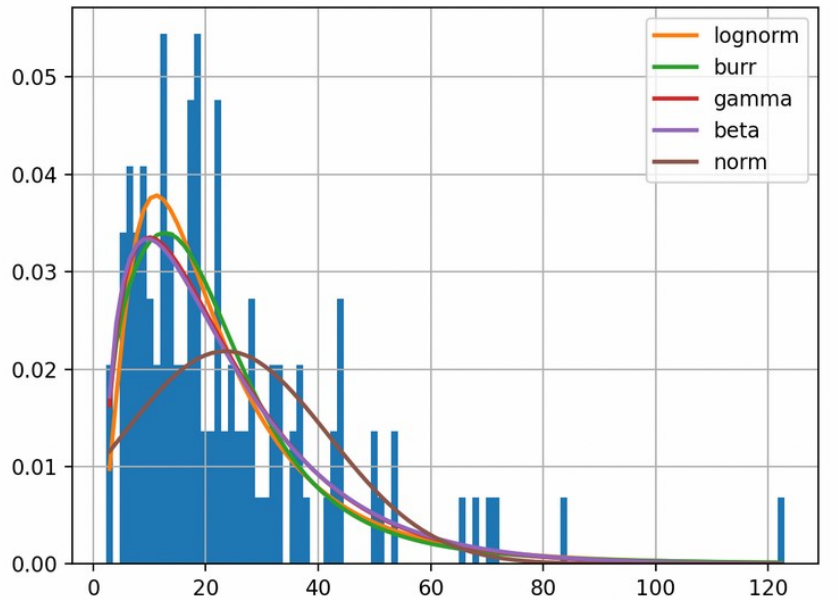

Figure 4.2: Histogram of pause distribution for healthy controls, and best-fit PDFs, for several parametric models

also computed the histograms of both types of features separately for dementia patients (Figure 4.1) and healthy controls (Figure 4.2), and found the parametric pdf curve of best fit (from the log-normal, burr, gamma, beta, and normal distributions). We then used the best fit parameters of the fitted pdf as additional features for our model.

4.1.3 Whisper model and derived features

We used the Whisper model [48] to derive language-based features that would be independent of the actual language spoken (and thus usable across *different* languages, even low-resource ones). Whisper is a large-scale model recently developed at OpenAI, capable of both speech recognition (i.e. transcribing an audio into text in the language that is spoken) as well as speech translation into the English language. It is a transformer-based weakly-supervised multitask model trained on 680,000 hours of audio and corresponding transcripts collected from the internet (out of which 117,000 hours of audio comes from 96 other non-English languages).

The Whisper model works on the premise of zero-shot transfer learning (without the need for extensive fine-tuning). It benefits from the large scale of the training data set, and builds robustness to the noisy weak supervision it is trained on. However, the authors note a clear drop in performance (in terms of Word Error Rate - WER) in proportion to the size of the data set for low-resource languages. They point out that, if good quality supervised speech data is available for a certain language, it would likely benefit further from fine-tuning [48].

Nevertheless, for the purpose of the current analysis, we are less concerned with Whisper's WER as we derive coarse word-level features specifically useful for Alzheimer's dementia classification. We use Whisper's capabilities of returning timestamps demarcating word boundaries within the generated transcript, as well as the word-level confidence scores assigned to the predicted word by the model. The word boundaries are determined using attention weights from the cross-attention mechanism and the dynamic time warping (DTW) algorithm for alignment. The word timestamps are used to compute

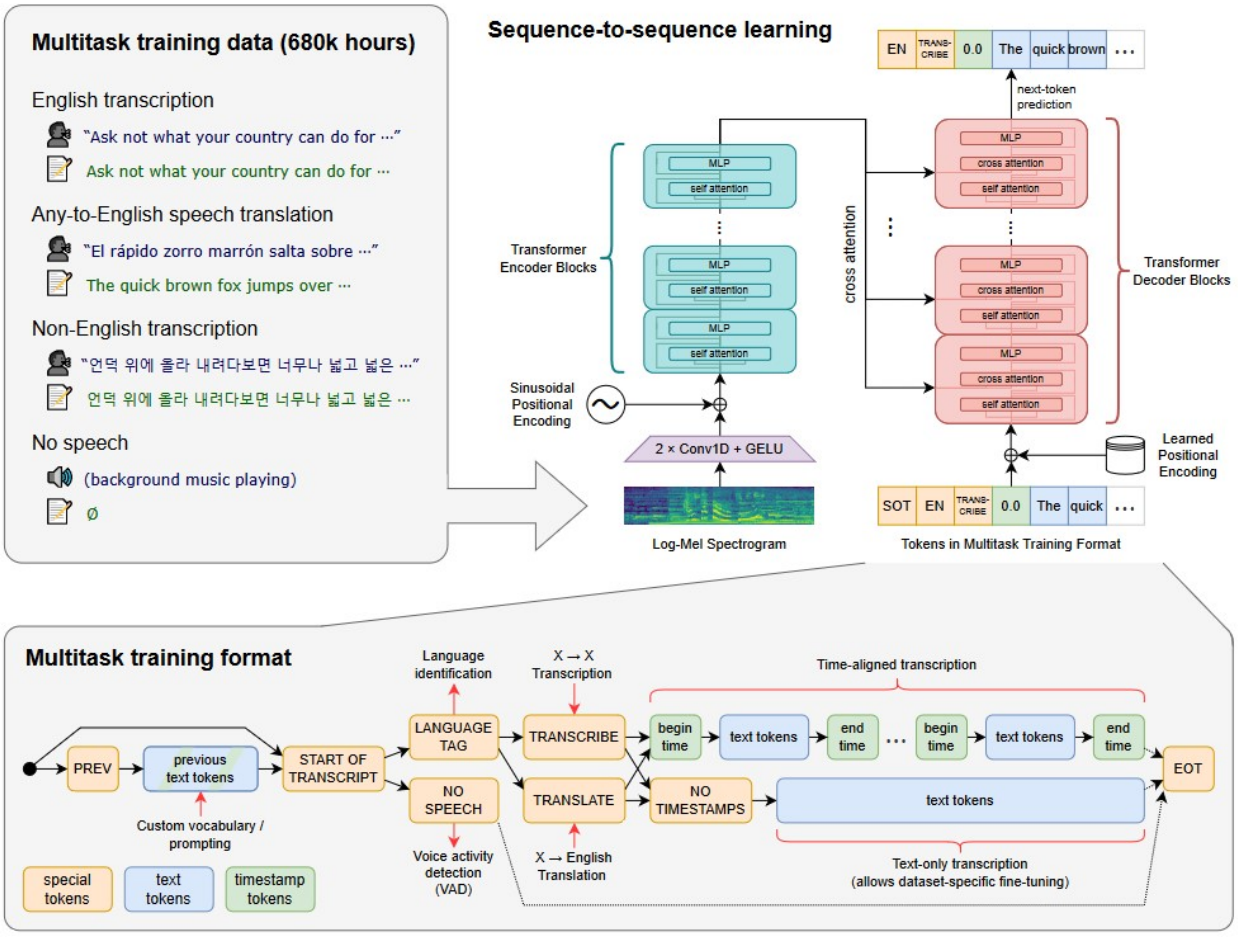

Figure 4.3: Whisper model training pipeline overview (source: Figure 1 in [48]). It uses a Transformer-based multilingual multitask training framework, for tasks including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. This enables a single model to perform the entire speech processing pipeline.

the relative *length* of each spoken word, which is used here as a proxy for *lexical complexity*. We make a simplifying assumption: typically words that are more complex are longer in spoken duration than simpler words, across different languages. For example, the phrases "I had a good time" and "I had a delightful time" are identical except for the substitution of "delightful" in place of "good". Since "delightful" is a longer word (in terms of syllables), it will generally take longer to speak than the word "good". This word-duration-based lexical complexity measure is a coarse way of quantifying (without relying on the actual transcription) the number of longer and supposedly more complex words spoken.

The other set of features derived from the Whisper model's output is intended to be a proxy for *speech intelligibility*. In general, dementia patients tend to have a greater incidence of both filled and unfilled pauses, hesitations, repetitions, broken words and other such disfluencies. The pause distribution features used in this study only account for unfilled (or silent) pauses, since they are based on audio segmentation using voice activity detection. They do not capture the other types of disfluencies mentioned here, which may also be important for discriminating dementia patients from controls. Hence we use word-level confidence scores for the predicted transcripts from the Whisper model to indirectly quantify the relative intelligibility of the spoken words. Here we make the assumption that predicted words having low confidence scores assigned by the Whisper model are less intelligible due only to the disfluencies mentioned above. We expect to see a higher incidence of such disfluencies in dementia patients compared to healthy controls.

As discussed in further detail in Section 4.6, the assumptions we have made here are quite simplistic in nature, and consequently need further validation studies on additional multilingual dementia data sets to validate them. Nevertheless, they provide a solid starting point for multilingual dementia analysis and enabled us to achieve remarkable performance on the given data. The following sections reproduce the published paper.

4.2 Introduction

With global population aging at an accelerated pace, worldwide prevalence of Alzheimer's dementia (AD) is also on the rise. This motivates today's significant interest in developing automatic, inexpensive, and scalable methods to support early diagnosis of AD. Since AD is characterized by a progressive decline in cognitive abilities, potentially leading to speech and language impairment, the analysis of speech signals for AD detection holds great potential.

In this paper, we present our methods for tackling the ICASSP 2023 Signal Processing Grand Challenge: "ADReSS-M: Multilingual Alzheimer's Dementia Recognition through Spontaneous Speech" [29], which sought methods for detecting dementia from speech signals whose predictive performance is preserved across two different languages. Given a set of English speech samples of subjects describing a specific picture, challenge participants developed models that were then tested on Greek speech samples of different subjects, describing a different picture.

Our submission explored various acoustic and language feature representations, based on pre-trained speech and language embeddings as well as conventional acoustic and linguistic feature extraction methods. Our best-performing classification model used a feature set based on language features derived from word-level attention maps, a representation of the distribution of pauses, and participant's meta-features (age, gender, and education), with PCA for dimension reduction followed by a logistic regression model, to obtain a test set accuracy of 69.57%. Our best regression model applied support vector regression to a representation of the distribution of pauses (silent segments) in the audio files and meta-features as the input features, to obtain a test set root mean squared error (RMSE) of 4.77 – leading to an overall fourth best score (out of 24 participants).

4.3 Data set and evaluation

The training data set contained 237 age- and gender-balanced audio files of English speech samples as well as associated meta-data (age, gender, education) and labels (Control vs AD[2] for classification, and MMSE scores for regression). We were also given a small set of 8 Greek audio samples, with metadata and labels. To evaluate our models and to set the hyperparameters, we used stratified 5-fold cross-validation (5SCV) on a fixed set of age-, gender-, and label-balanced folds over the English data. We used the set of 8 Greek samples as a separate hold-out test set.

4.4 Methodology

4.4.1 Feature extraction

Automatic Speech Recognition

Whisper [48] is a state-of-art Transformer-based multilingual speech-recognition model that is capable of both transcribing given audio and also translating audio context to another language – eg, Greek audio to English text. Here, we used the Whisper-Large multilingual model to transcribe all audios and created translations for each Greek speech audio using the same model. In our work, we did not directly utilize transcripts or translations. However, we found that the word-level features extracted from the model were highly beneficial.

Fluency and intelligibility

Many studies suggest that fluency and intelligibility may be important indicators of cognitive decline and AD – eg, AD patients speak at a slower rate than healthy older adults [50]; often exhibit increased levels of disfluency (pauses, hesitations, or disruptions) in their speech, increasing as the disease progresses [50]; and often exhibit reduced speech intelligibility, which can make it more difficult for them to communicate effectively with others [25]. Therefore, we compute the following three sets of features:

[2]The data set originally used the label "ProbableAD". However, for simplification we use "AD" here instead.

(1) *Word-level durations feature set* describes whether the speakers are mostly using short or long words, and how quickly they are uttering them. We obtained the duration of each word in an audio recording using a modified version of the Whisper model [48], which uses attention weights from cross-attention mechanisms and dynamic time-warping methods to determine start and end timestamps of each word. Note this method does not account for gaps between words. For each audio sample, we compute the total number of words in the speech, as well as the mean, maximum, minimum, and standard deviation of the word durations.

(2) *Pause rate feature set* describes the distributions of detected pause segments in spontaneous speech. We identified the voiced and unvoiced segments of the audio using the `openSMILE` toolkit [15], which also provided the timestamps of the onset of voiced segments, along with their duration. We used this to compute 11 features related to silence and length of audio segments. Besides some basic features like pause length means and variances, we also derived features by fitting different probability density functions (PDFs) to the histograms of total silence durations for cases versus controls over the complete audio sample (see Figures 4.1 and 4.2), and the histograms of lengths of unvoiced audio segments for cases versus controls. We then used the means and variances of these fitted PDFs as input features for our models.

(3) *Speech intelligibility feature set* describes the ease and accuracy with which a listener can comprehend the speech, which here is represented by the word-level confidence score assigned to each recognized word by a speech-recognition model. The confidence score for each word is expressed as the predicted probability of each word from the Whisper model. As with the word-level duration feature set, we compute the mean, maximum, minimum, and standard deviation of the confidence scores for every word in the speech. Additionally, we include a log-sum score of all confidence scores to represent the model's confidence in the entire transcript.

4.4.2 Modeling

To identify the optimal pipeline, we considered several dimension reduction techniques – including Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA) – and many base learning algorithms – Logistic regression, random forest, support vector machine (SVM), extreme gradient boosting, and neural networks – using different combinations of the extracted features mentioned above. We employed internal 5-fold stratified cross validation (5SCV) to fine-tune the hyperparameters of each model based on accuracy and RMSE.

4.5 Results

Our internal cross-validation identified the classification model with the best performance: Apply logistic regression with L2-regularizer, to the features corresponding to the top 10 PCA components of the union of meta-features, word-level duration, pause rate, and speech intelligibility; this yielded mean accuracies of 74.70 $\pm$ 4.90% (resp., 75%) on the 5SCV English dataset (resp., 8 Greek samples). On the competition's Greek test set, the accuracy dropped to 69.57%. For the regression task, we trained an SVM (with a radial basis function kernel and regularization parameter set to 1) on a combination of the meta-features with the pause features of the English dataset. This produced a 5SCV RMSE of 6.487 $\pm$ 0.696 (resp., 3.13) on the 5SCV English dataset (resp., 8 Greek samples). On the test set, the RMSE was 4.7693.

Note that many previous models used thousands of text embedding features and acoustic features, which makes them difficult to explain and, by implication, to trust. In contrast, both of our models focus on features thought to be relevant to diagnosing AD – here, utilizing only 24 features: 3 meta features, 5 for speech rate, 11 for pause rate, and 5 for speech intelligibility. Collectively, our outcomes show that machine-learned models can detect cognitive decline from speech, even when trained on different languages (learned from English, then used in Greek), and slightly different tasks (different pictures).

4.6 Notes and comments

There are a few important limitations of the multilingual dementia detection study presented in this chapter, that we aim to outline here. Firstly, since the Greek-based test data set is relatively small, and since the distribution shift between the English training data and Greek test data is significant, there is a chance that the results obtained on the Greek test data may not be broadly generalizable. There needs to be further validation studies to ensure that the results can be generalized to a larger Greek test set, and should additionally be validated against other non-English (and non-Greek) languages as well.

Further, the speech intelligibility features derived in this work need additional validation. Since these features are based on word confidence scores, and the authors of the Whisper model concede that this model generally performs worse on low-resource languages, there is a chance that low-resource languages might have lower word confidence scores being returned by the model in general. This would mean that speech samples from two healthy controls (one English and the other Greek), with no apparent issues with respect to intelligibility, might still have very different word confidence scores on average, with the Greek sample having lower average word confidence scores. Although this issue was not made immediately apparent in our models, we must still validate it further to ensure that this was not overlooked.

Moreover, in this study we use the assumption that longer words are more sophisticated from the viewpoint of lexical complexity. This intuition is partially supported by linguistic studies as well [43]. However, much of the available literature studying lexical complexity is focused on the English language. The languages used in this study (English and Greek) both belong to the Indo-European family of languages. In these languages, more complex words generally tend to have a greater number of syllables compared to simpler words. The correlation between word length and lexical complexity that we exploited in our research, may not hold as readily in other languages such as Chinese and Japanese. Therefore, a promising future research direction could be to investigate the predictive performance of our models on a different family of

languages such as those of the Far East.

Additionally, with respect to the multilingual dementia models presented in this chapter, we have as yet only utilized coarse features related to language use. Another future research avenue could be to explore the semantic and syntactic complexity of a given speech sample, while remaining independent of the specific language being spoken. One way of achieving this is by building semantic speech graphs and using multilingual word embeddings. Then a measure of semantic complexity and coherence could potentially be derived using these graphs, regardless of the language spoken.

Chapter 5

Data visualization and clustering of language-agnostic speech representations

5.1 Introduction and data visualization

After submitting the paper "Language-agnostic representations for detecting Alzheimer's dementia from multilingual speech" (see Chapter 4), we revisited the derived feature sets that we found useful in this publication. This time we wanted to explore these speech representations from the perspective of unsupervised clustering, with a view towards discovering any cluster structure that may help characterize the underlying data better and support further improvements in classification accuracy.

For the analysis presented in this section, we used the data set described in Chapter 4. Since the data set consisted of only 291 speech samples, with a total of 24 derived features per sample, some form of dimensionality reduction was expected to be important. We explored the use of several different dimensionality reduction methods, including Principal Component Analysis (PCA), T-distributed Stochastic Neighbor Embedding (t-SNE), univariate feature selection, and recursive feature elimination. Then, based on visual inspection of reduced feature sets, we determined the most appropriate dimensionality reduction technique that may help cluster the AD participants separately from the controls. We got the best results using PCA (with projection onto the first and second principal components). We reduced the dimensionality of

each feature set separately, to help visualize these feature sets and determine an appropriate clustering methodology.

For the actual clustering step, we based our selection of clustering algorithm on a visual inspection of potential cluster shapes, distributions and separability, as well as cluster density and the distribution of outliers. Empirically we found that k-means and k-medoids gave more meaningful results. For the evaluation, we used classification accuracy, precision, recall, and F-scores. Additionally we computed the Adjusted Rand Index (ARI) as well [49]. Here the usual challenge of cluster validation was somewhat mitigated by the fact that classification labels were available to us, and we were looking for only those clusterings that would be meaningful from the perspective of AD classification.

To help visualize the different feature sets, we used PCA and projected each feature set onto the first two principal components, obtaining a 2-D representation of the data. In Figures 5.1, 5.2, and 5.3 below, the circles represent English speech samples and the triangles represent Greek speech samples. Also, the positive label of 'ProbableAD' is shown in blue and the negative label of 'Control' is shown in green.

Figure 5.1 shows a visualization of the pause distributions feature set. Here we can see that this data follows a V-shaped pattern, with greater density near the vertex. Additionally, most of the blue 'ProbableAD' samples are on the left arm of the V shape, while most of the green 'Control' samples are on the right. This seems to be the case for both the English and the Greek samples. Although there is a significant amount of overlap between the two desired classes of 'ProbableAD' and 'Control', still this plot seems to exhibit some structure in the data that may be useful for clustering.

Figure 5.2 shows a visualization of the word feature distributions (this is a combined view of the word durations and speech intelligibility feature sets). Any cluster structure here is difficult to discern. There are significant differences in data density, with an area near the origin showing high density, and the scatter towards the right becoming sparser. Unfortunately there is little separability between the 'ProbableAD' and the 'Control' samples using this feature set.

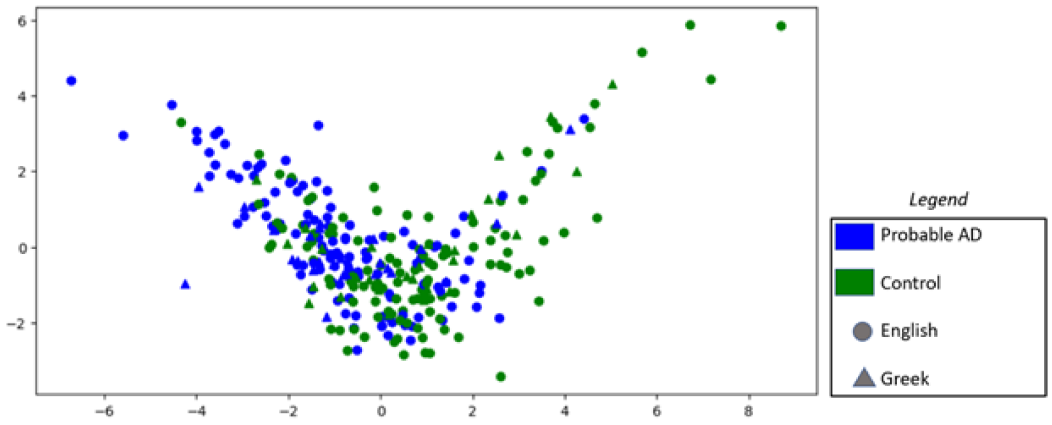

Figure 5.1: Visualizing pause distribution

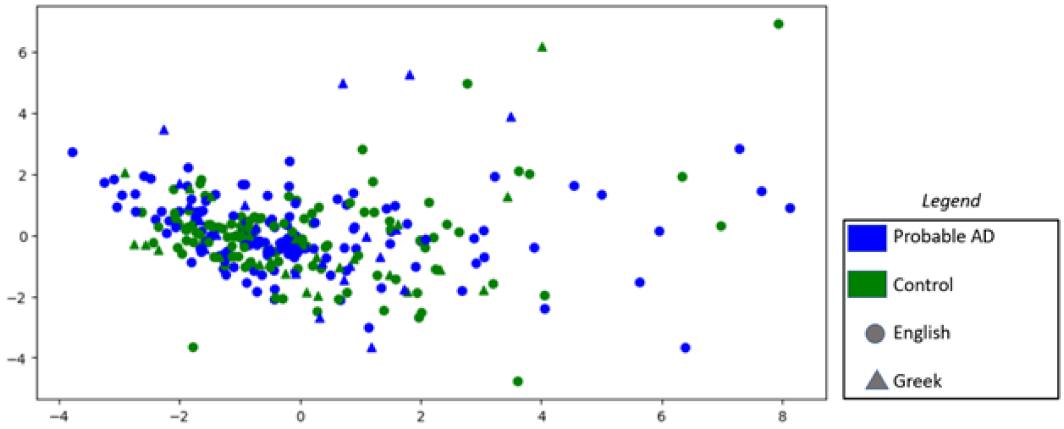

Figure 5.2: Visualizing word feature distribution

Figure 5.3 shows the metadata features (age, gender, level of education) projected into 2D using PCA. Here we can visually discern some cluster structure, with two distinct elliptical clusters easily seen. Unfortunately, upon closer inspection we find that the cluster structure here cannot be used to distinguish between the positive and negative class labels (notice that the blue and green markers are similarly distributed within the two elliptical clusters).

Summarizing some of the insights obtained from the visualization exercise described above, we note that the pause features and metadata features both exhibit good cluster structure that could potentially be utilized for our purpose. However, the metadata clusters do not help separate the positive and negative samples effectively. The pause features could, however, be used to find clusters supporting the classification task. The word features and acoustic

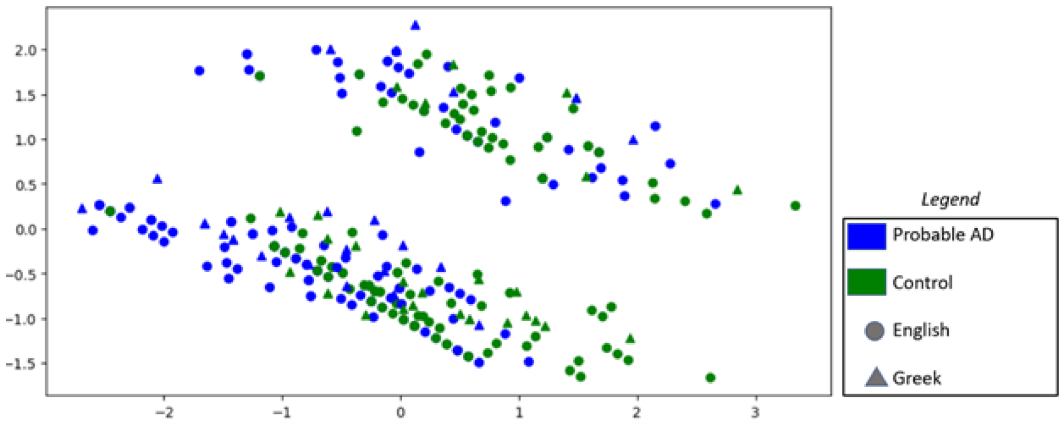

Figure 5.3: Visualizing meta feature distributions (age, gender, education level)

features have similar distributions, and there is no clear cluster structure seen in either of these distributions that could be useful based on these features alone. Therefore, our main clustering efforts focus on the pause rate feature set, and we present some relevant results below.

5.2 Unsupervised clustering

We performed k-means clustering on the 2D projection of the pause features data set (Figure 5.1) using the k-means implementation available in the Python scikit-learn cluster library [46]. We chose k-means clustering because the shape of the desired clusters (i.e. the two arms of the V shape) looks somewhat elliptical, and since we already know the number of clusters we are looking for, determining the number of clusters k is trivial in our case (i.e. $k = 2$). The scikit-learn implementation of k-means does multiple random initializations, with the best result being returned as the final clustering. This feature helps mitigate the problem of getting stuck in local optima situations, which frequently plagues the k-means algorithm. We quantified the validity of the resultant clusters by computing their classification accuracy with respect to the labels available over the entire data set. The k-means clustering achieved a classification accuracy of 68%, with precision of 70% and recall of 64% for the 'ProbableAD' class. Also, the ARI was 0.127 (see Table 5.1).

Table 5.1: Evaluation of k-means clustering of pauses feature set

	Precision	Recall	F1-score	Support
Control	0.66	0.72	0.69	143
ProbableAD	0.70	0.64	0.67	148

Table 5.2: Evaluation of k-medoids clustering of pauses feature set

	Precision	Recall	F1-score	Support
Control	0.66	0.75	0.70	143
ProbableAD	0.72	0.63	0.67	148

We also performed k-medoids clustering on the pauses feature set (using an implementation available in the `sklearn_extra` library [1]. Since the data distribution shows some sparse data points near the edges, and k-means clustering is highly prone to skewness as a result of outliers, we expected that k-medoids clustering will mitigate this bias and thus improve clustering performance. We found that k-medoids clustering did in fact slightly outperform k-means clustering in terms of classification accuracy, although the effect size was small (classification accuracy was 69%, and ARI was 0.137). Table 5.2, Figure 5.4 and Figure 5.5 show the results obtained by using k-medoids clustering.

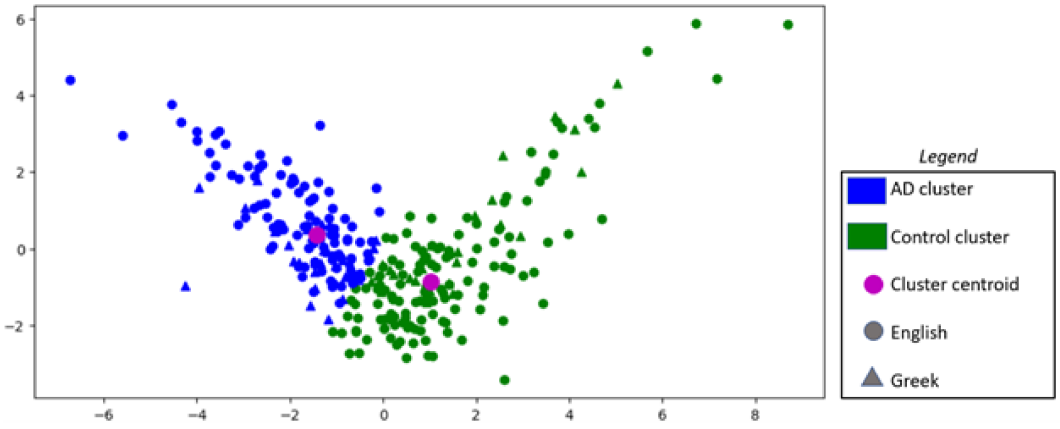

Figure 5.4: K-medoids clustering of pauses feature set

Finally, we also tried combining the different feature sets and performing cluster analysis based on these larger feature vectors (after dimension reduction using PCA). However, none of the combined feature sets gave a better result than using the pauses feature set alone.

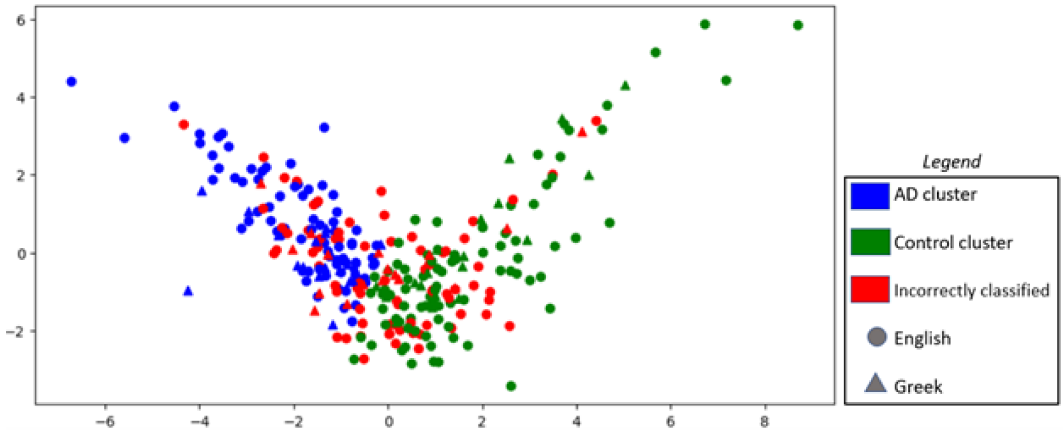

Figure 5.5: K-medoids clustering of pauses feature set, with incorrectly classified samples shown in red

5.3 Discussion

The above analysis showcased some exploratory work on speech feature representations for detecting Alzheimer's dementia from spontaneous speech, using methods that generalize across different languages. Despite the substantial domain shift between the two languages and experimental conditions being studied, we found that some of the features extracted from the speech samples were in fact useful in separating the AD samples from the control samples.

The problem being tackled here is a difficult one. Not all feature sets that were extracted showed a clear cluster structure that could be exploited. Out of the four feature sets we extracted, the one based on the distribution of pauses within the speech samples was the most informative and demonstrated the most discriminative cluster structure. Using k-means and k-medoids clustering, we were able to obtain an almost 70% classification accuracy using the pauses feature set alone. Given the complexity of this problem and the performance of other solutions available in the literature, this result is highly competitive.

Having said that, the results obtained here using clustering techniques are comparable to those obtained using classification algorithms in our earlier work [55]. We can therefore conclude that clustering itself does not significantly improve the classification performance, implying that the real gains

would likely come from more nuanced feature engineering.

Chapter 6

Conclusion

Speech is a promising biomarker to enable the early detection of a variety of psychiatric disorders. In recent years the automated analysis of speech, supported by machine learning technologies, has made remarkable advances. This rapid progress has opened up the potential for developing speech-based machine learning models for the automated detection, screening and monitoring of psychiatric disorders such as depression, dementia and post-traumatic stress disorder. Given the rising global demand for mental healthcare and the limited resources available, the search for inexpensive, non-invasive, reliable, and scalable mental health screening has become critically important. This thesis has presented our research (spanning the last three years, and resulting in two high-ranked challenge submissions and two peer-reviewed publications), on developing speech-based machine learning models for the early detection of Alzheimer's dementia (AD).

Here we summarize a few key contributions made in this thesis. Firstly, we have systematically explored both language-based and acoustic-based features from the English-only speech data set. We found that, even though both types of features demonstrated significant predictive performance, features based on language semantics and lexical complexity are particularly important for detecting early signs of dementia from English speech. Moreover, we were able to demonstrate that traditional, interpretable machine learning algorithms, coupled with appropriate feature engineering, could achieve a significantly high accuracy for discriminating AD patients from controls.

Additionally, we derived novel speech features that proved relevant to identifying dementia irrespective of the language being spoken. This enabled us to build robust machine learning models based on these derived speech features to detect Alzheimer's dementia with significant accuracy, even across *different* spoken languages. Even when the languages were different between the training and testing data sets, and additionally the picture description tasks for eliciting spontaneous speech were also different (i.e. different pictures were used for these tasks), our models were able to detect Alzheimer's dementia with high accuracy. These results pave the way for further research on speech-based models for Alzheimer's dementia detection, even in challenging multilingual low-resource settings.

Furthermore, we also provided an exploratory analysis to visualize the structure of the derived feature space in the multilingual Alzheimer's dementia setting, via data visualization and unsupervised clustering. This analysis showed that, in particular, features based on the distribution of pauses in speech demonstrated a visible structure that was useful to discriminate between AD patients and healthy controls (using simple unsupervised clustering techniques such as k-means).

Table 6.1: Summary and comparison of the two papers presented in this thesis

	ADReSS paper (Chapter 3)	**ADReSS-M paper (Chapter 4)**
Raw data types	1. Audio files 2. Good quality annotated transcripts 3. Segmented audio clips 4. Meta features	1. Audio files 2. Meta features
Data set (training)	108 English speech samples (Cookie Theft pic. descr. task - Fig. 2.1).	237 English speech samples (Cookie Theft pic. descr. task - Fig. 2.1).
Data set (test)	48 English speech samples (Cookie Theft pic. descr. task - Fig. 2.1).	46 Greek speech samples (pic. descr. task, but not Cookie Theft)
Features used	- **Language:** 62 semantic, syntactic and fluency features (Type-Token Ratio, Mean Length of Utterance, parts of speech,. . .) + Bag-of-words features + Bigram features - **Acoustic:** Four feature sets with different acoustic features related to energy, pitch, voicing, spectral, MFCCs, speech rate, etc.	- **Pause rate features:** 11 features based on distribution of silent pauses in the speech - **Word duration features:** 5 features based on word lengths derived from Whisper ASR timestamps - **Speech intelligibility features:** 5 features based on word prediction confidence scores from Whisper ASR
Modeling	Dimensionality reduction: PCA, LSA Base learners considered: Random forest, SVM, logistic regression, XGBoost, Neural nets Ensemble models considered: Majority vote (classification), averaging (Regression)	Dimensionality reduction: PCA, LSA Models: Random forest, SVM, logistic regression, XGBoost, Neural nets
Evaluation	Nested stratified 5-fold cross validation	Nested stratified 5-fold cross validation
Performance (5-CV)	Classification accuracy: 80% Regression RMSE: 6.43	Classification accuracy: 75% Regression RMSE: 6.49
Performance (Test set)	Classification accuracy: 85% Regression RMSE: 5.62	Classification accuracy: 70% Regression RMSE: 4.77
Competition ranking	Ranked 3rd out of 34 teams (for classification accuracy)	Ranked 4th out of 24 teams (over both tasks combined)

6.1 Limitations and future directions

Both of the works presented in this thesis have focused on discriminating Alzheimer's dementia from healthy controls. Neither of them have attempted the more difficult problem, of reliably discriminating the more advanced stages of dementia from Mild Cognitive Impairment (MCI). One could argue that the regression tasks in both studies are meant to develop models to predict dementia severity, and low severity likely implies an MCI diagnosis. However, such an analysis has not been made explicit, and calibration of the regression models has not been studied here.

Furthermore, the different types of dementia besides Alzheimer's dementia (including vascular dementia, fronto-temporal dementia, and Lewy body dementia) have not been considered in the works presented in this thesis. This is important because of the possibility that differences in dementia type might explain some subtle differences in speech and language processing. If appropriately harnessed, this information could potentially make our speech-based dementia detection systems more robust.

Nevertheless, the main bottleneck that limits the development of clinical-scale dementia screening systems is data. In order to create robust multilingual dementia detection systems that are capable of being deployed to a clinical setting, we must have large, diverse, multilingual, multi-site speech data sets. Ideally, these data sets should have information on relevant comorbidities as well. Additionally, longitudinal data sets would be invaluable for building models capable of identifying the progression of the cognitive decline that is associated with dementia. Furthermore, the quality of the data sets and their broad representative capacities would be crucial to ensuring that there are no bias and fairness issues when it comes to actual clinical deployment.

In conclusion, machine learning models based on speech analysis are extremely attractive for the early detection and monitoring of Alzheimer's dementia. We have demonstrated that significantly accurate models can be built using relatively simple speech features, if done in a thoughtful and systematic manner using relevant domain expertise. We have also shown that such models

can be built in the multilingual low-resource setting as well. Further studies should be conducted in order to comprehensively validate these findings on larger and more diverse datasets, with the aim of enabling large-scale clinical deployment of these multilingual speech-based machine learning models for the early detection and monitoring of Alzheimer's dementia.

References

[1] C. Aridas, J. Joswig, T. Mathieu, and R. Yurchak, *Scikit-learn-extra*. [Online]. Available: https://github.com/scikit-learn-contrib/scikit-learn-extra.

[2] A. Association. "Medical tests for diagnosing alzheimer's." (2023), [Online]. Available: https://www.alz.org/alzheimers-dementia/diagnosis/medical_tests?#mental (visited on 06/13/2023).

[3] T. Bäckström, O. Räsänen, A. Zewoudie, *et al.*, *Introduction to Speech Processing*, 2nd ed. 2022. DOI: 10.5281/zenodo.6821775. [Online]. Available: https://speechprocessingbook.aalto.fi.

[4] A. Balagopalan, B. Eyre, J. Robin, F. Rudzicz, and J. Novikova, "Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech," *Frontiers in Aging Neuroscience*, vol. 13, 2021, ISSN: 1663-4365. DOI: 10.3389/fnagi.2021.635945. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2021.635945.

[5] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[6] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[7] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[8] F. Boller and J. Becker, "Dementiabank database guide," *University of Pittsburgh*, 2005.

[9] R. Chakraborty, M. Pandharipande, C. Bhat, and S. K. Kopparapu, "Identification of dementia using audio biomarkers," *arXiv preprint arXiv:2002.12788*, 2020.

[10] J. R. Cockrell and M. F. Folstein, "Mini-mental state examination," *Principles and practice of geriatric psychiatry*, pp. 140–141, 2002.

[11] F. Cuetos, J. C. Arango-Lasprilla, C. Uribe, C. Valencia, and F. Lopera, "Linguistic changes in verbal expression: A preclinical marker of alzheimer's disease," *Journal of the International Neuropsychological Society*, vol. 13, no. 3, pp. 433–439, 2007.

[12] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[13] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[14] D. P. Ellis, "An introduction to signal processing for speech," *The Handbook of Phonetic Sciences*, pp. 755–780, 2010.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[16] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": A practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[17] F. Franco-Marina, J. J. García-González, F. Wagner-Echeagaray, *et al.*, "The mini-mental state examination revisited: Ceiling and floor effects after score adjustment for educational level in an aging mexican population," *International Psychogeriatrics*, vol. 22, no. 1, pp. 72–81, 2010. DOI: 10.1017/S1041610209990822.

[18] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, pp. 1547–1574, 2020.

[19] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[20] Y. Guo, C. Li, C. Roan, S. Pakhomov, and T. Cohen, "Crossing the "cookie theft" corpus chasm: Applying what bert learns from outside data to the adress challenge dementia detection task," *Frontiers in Computer Science*, vol. 3, 2021, ISSN: 2624-9898. DOI: 10.3389/fcomp.2021.642517. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.642517.

[21] R. Haulcy and J. Glass, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, 2021, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.624137. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.624137.

[22] D. Iter, J. Yoon, and D. Jurafsky, "Automatic detection of incoherent speech for diagnosing schizophrenia," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, New Orleans, LA: Association for Computational Linguistics, Jun. 2018, pp. 136–146. DOI: 10.18653/v1/W18-0615. [Online]. Available: https://aclanthology.org/W18-0615.

[23] X. Jia, Z. Wang, F. Huang, *et al.*, "A comparison of the mini-mental state examination (mmse) with the montreal cognitive assessment (moca) for mild cognitive impairment screening in chinese middle-aged and older population: A cross-sectional study," *BMC psychiatry*, vol. 21, no. 1, pp. 1–13, 2021.

[24] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2Nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009, ISBN: 0131873210.

[25] D. Kempler and M. Goral, "Language and dementia: Neuropsychological aspects," *Annual review of applied linguistics*, vol. 28, pp. 73–90, 2008.

[26] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, "Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition," *arXiv preprint arXiv:2009.04070*, 2020.

[27] C. Laske, H. R. Sohrabi, S. M. Frost, *et al.*, "Innovative diagnostic tools for early detection of alzheimer's disease," *Alzheimer's & Dementia*, vol. 11, no. 5, pp. 561–578, 2015.

[28] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," en, *Laryngoscope Investig. Otolaryngol.*, vol. 5, no. 1, pp. 96–116, Feb. 2020.

[29] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "Multilingual alzheimer's dementia recognition through spontaneous speech: A signal processing grand challenge," *arXiv:2301.05562*, 2023.

[30] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[31] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The adresso challenge," *medRxiv*, 2021. DOI: 10.1101/2021.03.24.21254263. eprint: https://www.medrxiv.org/content/early/2021/03/26/2021.03.24.21254263.full.pdf.

[32] S. Luz, F. Haider, S. de la Fuente Garcia, D. Fromm, and B. MacWhinney, "Editorial: Alzheimer's dementia recognition through spontaneous speech," en, *Front Comput Sci*, vol. 3, Oct. 2021.

[33] B. MacWhinney, *The CHILDES Project: Tools for analyzing talk. transcription format and programs*. Psychology Press, 2000, vol. 1.

[34] B. MacWhinney, *Tools for analyzing talk part 1: The chat transcription format*, Carnegie, 2017.

[35] B. MacWhinney, "Tools for analyzing talk part 2: The clan program," *Talkbank. Org*, no. 2000, 2017.

[36] C. R. Marmar, A. D. Brown, M. Qian, *et al.*, "Speech-based markers for posttraumatic stress disorder in US veterans," en, *Depress. Anxiety*, vol. 36, no. 7, pp. 607–616, Jul. 2019.

[37] M. Martinc, F. Haider, S. Pollak, and S. Luz, "Temporal integration of text transcripts and acoustic features for alzheimer's diagnosis based on spontaneous speech," *Frontiers in Aging Neuroscience*, vol. 13, 2021, ISSN: 1663-4365. DOI: 10.3389/fnagi.2021.642647. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2021.642647.

[38] I. Martínez-Nicolás, T. E. Llorente, F. Martínez-Sánchez, and J. J. G. Meilán, "Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: A systematic review article," *Frontiers in Psychology*, vol. 12, 2021, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.620251. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.620251.

[39] A. Meghanani, C. S. Anoop, and A. G. Ramakrishnan, "Recognition of alzheimer's dementia from the transcriptions of spontaneous speech using fasttext and cnn models," *Frontiers in Computer Science*, vol. 3, 2021, ISSN: 2624-9898. DOI: 10.3389/fcomp.2021.624558. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.624558.

[40] G. A. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.

[41] N. B. Mota, N. A. P. Vasconcelos, N. Lemos, *et al.*, "Speech graphs provide a quantitative measure of thought disorder in psychosis," en, *PLoS One*, vol. 7, no. 4, e34928, Apr. 2012.

[42] S. Nasreen, M. Rohanian, J. Hough, and M. Purver, "Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features," *Frontiers in Computer Science*, vol. 3, 2021, ISSN: 2624-9898. DOI: 10.3389/fcomp.2021.640669. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.640669.

[43] K. North, M. Zampieri, and M. Shardlow, "Lexical complexity prediction: An overview," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–42, 2023.

[44] W. H. Organization. "Mental disorders." (2022), [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders (visited on 06/12/2023).

[45] W. H. Organization. "Dementia." (2023), [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia (visited on 06/12/2023).

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[47] M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: A review," *Expert systems with applications*, vol. 150, p. 113 213, 2020.

[48] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[49] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/2284239 (visited on 07/17/2023).

[50] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2081–2090, 2011.

[51] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, "Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity," *arXiv preprint arXiv:2009.00700*, 2020.

[52] B. Schuller, S. Steidl, A. Batliner, *et al.*, "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[53] B. Schuller, S. Steidl, A. Batliner, *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[54] T. Searle, Z. Ibrahim, and R. Dobson, "Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech," *arXiv preprint arXiv:2006.07358*, 2020.

[55] Z. Shah, S.-A. Qi, F. Wang, *et al.*, "Exploring language-agnostic speech representations using domain knowledge for detecting alzheimer's dementia," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–2. DOI: 10.1109/ICASSP49357.2023.10095593.

[56] Z. Shah, J. Sawalha, M. Tasnim, S.-a. Qi, E. Stroulia, and R. Greiner, "Learning language and acoustic models for identifying alzheimer's dementia from speech," *Frontiers in Computer Science*, vol. 3, 2021, ISSN: 2624-9898. DOI: 10.3389/fcomp.2021.624659. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.624659.

[57] M. Soltan and J. Girguis, "How to approach the mental state examination," en, *BMJ*, vol. 357, j1821, May 2017.

[58] M. Valstar, B. Schuller, K. Smith, *et al.*, "Avec 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, ACM, 2013, pp. 3–10.

[59] E. Vuorinen, M. Laine, and J. Rinne, "Common pattern of language impairment in vascular dementia and in alzheimer disease," *Alzheimer Disease & Associated Disorders*, vol. 14, no. 2, pp. 81–86, 2000.

[60] J. R. Williamson, D. Young, A. A. Nierenberg, J. Niemi, B. S. Helfer, and T. F. Quatieri, "Tracking depression severity from audio and video based on speech articulatory coordination," *Computer Speech & Language*, vol. 55, pp. 40–56, 2019, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2018.08.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230817303510.

[61] J. Yuan, X. Cai, Y. Bian, Z. Ye, and K. Church, "Pauses for detection of alzheimer's disease," *Frontiers in Computer Science*, vol. 2, 2021, ISSN: 2624-9898. DOI: 10.3389/fcomp.2020.624488. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2020.624488.

[62] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, "Exploring deep transfer learning techniques for alzheimer's dementia detection," *Frontiers in Computer Science*, vol. 3, 2021, ISSN: 2624-9898. DOI: 10.3389/fcomp.2021.624683. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.624683.

Appendix A

Machine learning for detection of post-traumatic stress disorder using speech

A.1 Introduction

In this appendix, we present another mental health related speech analysis project we conducted. This project was done in collaboration with the Computational Psychiatry unit at IBM T.J. Watson Research Center, the Psychiatry and Computing Science departments at the University of Alberta, and Alberta Machine Intelligence Institute (Amii). We faced significant challenges during the data collection part of this project (including recruitment problems possibly related to the pandemic), due to which the scope of this research had to be limited to pilot study status. For this reason, I have chosen to present this project as an appendix of the main thesis, instead of a chapter in the main body.

The diagnosis of psychiatric disorders, such as depression and psychosis, is a challenging task. There are oftentimes no objective, readily measurable biomarkers to characterize the disorder – i.e., there are no associated biophysical symptoms (e.g. measurable by blood tests). Brain imaging technologies, such as fMRI and CT scans, do provide some means to measure brain function. However, these modalities are plagued by the need for highly specialized, costly equipment, and sometimes invasive procedures. Could we make use of an easier-to-measure, more available and less costly proxy, that can aid

in objectively and reliably characterizing brain health? Variations in speech patterns, both acoustic and semantic, are frequently used by experienced clinicians for assessing a subject's emotional and psychological state. Hence, this project was focused on studying the use of speech as a convenient and reliable window into the human brain and its health. Specifically, our plan was to develop and demonstrate a machine learning system based on speech analysis for the diagnosis of Post Traumatic Stress Disorder (PTSD) in Alberta's military veterans.

Several fundamental research challenges need to be addressed in building a robust performance system for PTSD diagnosis using speech analysis. We planned to pursue the following two complementary directions:

(1) **Acoustic and prosodic analysis:** In general, this can be categorized as a signal processing task. It pertains to analysis of lower level voice features such as pitch, intonation, and pause duration. Our hypothesis is that subjects with PTSD exhibit acoustic and prosodic features that are distinct from those without PTSD.

(2) **Semantic and syntactic analysis:** This is characterized as a natural language processing task. The idea here is to analyze the actual semantic content of the speech as well as measures of speech complexity. Again, we hypothesize that speech samples from PTSD subjects would have semantic and syntactic characteristics that can reliably distinguish them from non-PTSD subjects.

A.2 Data collection

Recruitment and data collection efforts for this project spanned the period of March 2020 to April 2021. Two participant groups were recruited during this time: a target group consisting of individuals having a positive current diagnosis of PTSD, and a control group having no diagnosis of PTSD (See Table A.1).

Table A.1: Recruitment criteria for PTSD speech study

Target group	**Control group**
Military personnel and veterans	Military personnel, veterans, and general population
Meet DSM-5 diagnostic criteria for PTSD (with comorbid disorders)	No mental health diagnosis
18+ years old	18+ years old
Male/Female (4:1)	Male/Female (4:1)

The data collection process involved (after obtaining the participant's signed consent) a set of two recorded interview sessions via Zoom. During the first session, each participant filled a demographic questionnaire online, and completed two tasks: (a) a baseline reading task, and (b) a picture description task. Between the first and second Zoom sessions, participants completed a mental health questionnaire, which consisted of many scales such as PCL-5, AUDIT, and others. During the second Zoom session, the interviewer asked each participant a predefined set of questions (See Table A.2). These questions were designed to elicit a meaningful response from the participant, potentially enabling the capture of such data that would support the development of machine learning algorithms to distinguish between participants with versus without PTSD.

Additionally, for each participant a binary outcome label was assigned based on their score on the PCL-C questionnaire (with a score greater than 22 indicating a participant with PTSD, and a score less than 22 indicating one without PTSD). The final data set consisted of demographic data, self-report questionnaires and speech data from (only) 14 participants, out of which 8 participants met the DSM-5 criteria for PTSD (based on PCL-C scores).

Table A.2: Interview questions asked of participants to collect conversational speech samples

1. Where are you from originally?
2. How would your best friend describe you?
3. How would you describe yourself?
4. What is your dream job?
5. What color best describes your personality?
6. What are you most passionate about?
7. What do you consider your best attributes?
8. If you have friends coming for supper what would you cook?
9. What makes you happy/sad/angry?
10. Are you a religious person? Why or why not?
11. Do you have any pets? Why or why not? Are you an animal lover?
12. What is one thing on your bucket list?
13. If you could visit any place in the world, where would it be and why?
14. If you could have dinner with anyone you wanted, dead or alive, who would you choose? And why?
15. Is there anything you'd change about yourself? What would that be?
16. What book/movie are you reading/watching (recently), what is it about?

A.3 Data analysis

For a preliminary data analysis, we used audio recordings from the first of the two Zoom interview sessions. We hypothesized that audio-based features computed from the baseline reading task and picture description task will be adequate for distinguishing participants with and without PTSD. Since we also did not have good quality transcripts available for the second Zoom interview at this time, we decided to conduct this analysis with data from the first interview session only.

We used a set of 130 acoustic low-level descriptor features computed from the audio recordings of each of the 14 participants. These proprietary features were computed using IBM's internal feature generation pipeline, and were then provided to us by our collaborators at IBM. We systematically trained a set of machine learning models on these input features (using the scikit-learn package [46]). We used a nested leave-one-out cross validation procedure to search over various model configurations for the best candidate in terms of test accuracy. With 8 out of 14 positive samples, the baseline model accuracy was 57% (from guessing majority class). See Table A.3 for a list of models and hyperparameters that were explored.

Table A.3: Models and hyperparameters

Model	**Hyperparameter values**
Decision Tree (DT) - config A	'min_samples_split': [2, 3, 4, 5] 'max_features': [0.25, 0.5, 0.75, 1.0]
Decision Tree (DT) - config B	'max_depth': [2, 5, 10]
Logistic Regression (LR) - config A	'penalty': ['L1','L2','Elasticnet'] 'C': [100, 10, 1.0, 0.1, 0.01]
Logistic Regression (LR) - config B	'penalty': ['L1'] 'C': [0.01, 0.001, 1e-4, 1e-5]
Naive Bayes (NB)	'var_smoothing': [1e-9, 1e-6, 1e-3]
Linear SVM (SVM_linear) - config A	'penalty': ['L1','L2'] 'C': [100, 10, 1.0, 0.1, 0.01]
Linear SVM (SVM_linear) - config B	'penalty': ['L1','L2'] 'C': [0.01, 0.001, 1e-4, 1e-5]
PCA + Decision Tree (PCA_DT)	'PCA -n_components': [0.3, 0.5, 0.7, 'MLE'] 'Decision tree -max_depth': [2, 5, 10]

A.4 Results and discussion

Figure A.1 shows the resubstitution accuracies and test accuracies obtained from the machine learning model training procedure described in the previous section. These results indicate that, given the input features, almost all of the trained models had a hard time beating the baseline (with only the decision tree configuration A demonstrating a test accuracy increase of about 7% over the baseline). This is likely due to the small size of the data set, and possibly also because we used only one interview session per participant in the current analysis. Another contributing factor may be that only audio features were used, since transcripts for all participants were not available at this time.

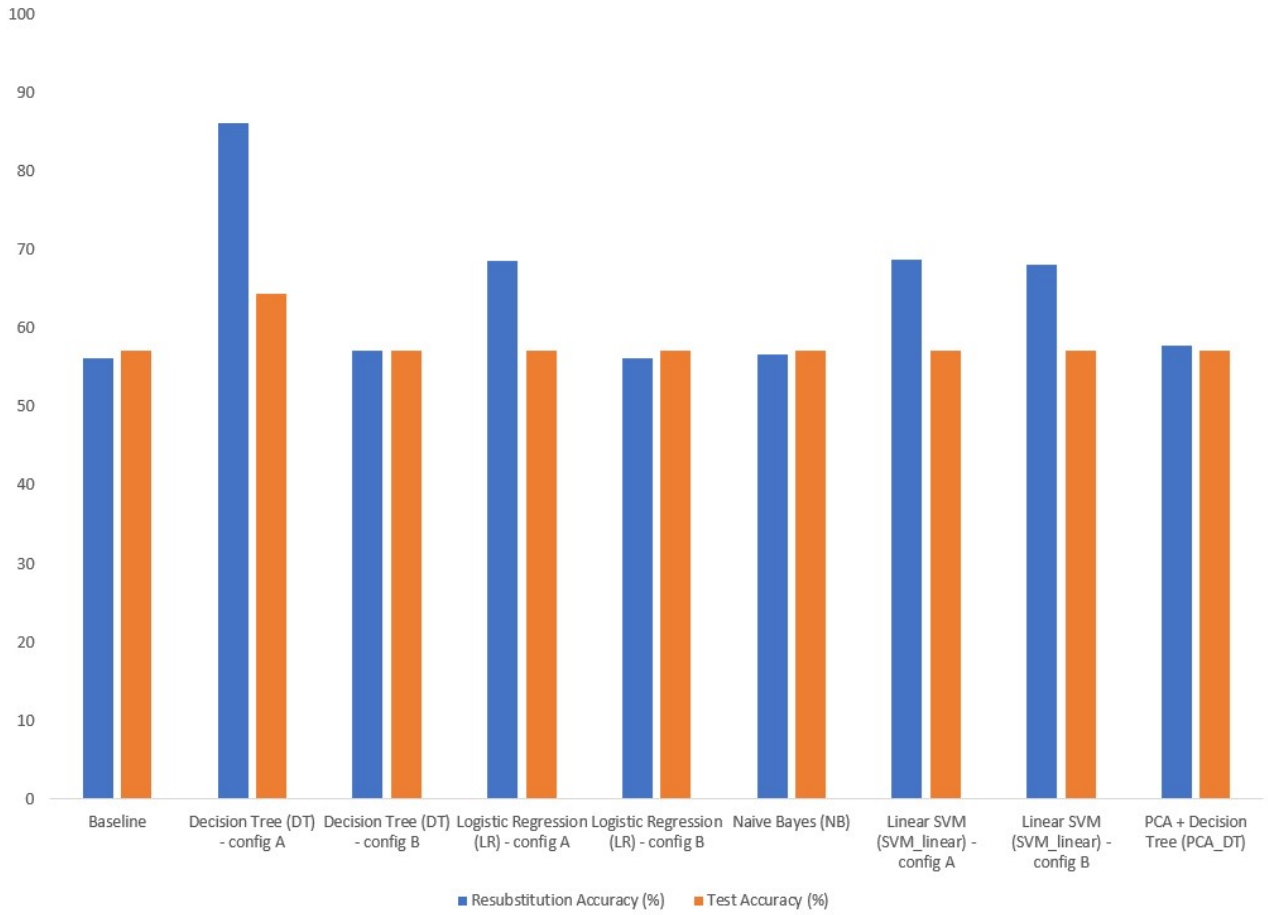

Figure A.1: Accuracy using leave-one-out cross validation

Results from this pilot study show that the machine learned models built using IBM's proprietary acoustic feature extraction pipeline were unable to dis-

tinguish between study participants with versus without PTSD. However, for several reasons we cannot generalize these results beyond this study. The most obvious reason is that the data set we are working with is extremely small from a machine learning perspective, with a total of only 14 data points across both experimental groups. This is not enough to characterize the high-dimensional feature space we are working with. Even with aggressive regularization, the predictive performance we achieved was not better than baseline.

Additionally, due to limited resources and shifting priorities, we were unable to obtain good quality manual transcriptions and could not pursue experiments using linguistic features. In our other experiments with Alzheimer's dementia, we did find that linguistic features had greater predictive power compared to acoustic features, so it is possible that for PTSD as well, we may discover specific linguistic features to be important. This will however need to be validated using larger data sets.

To conclude, there is immense potential for training clinically relevant predictive models based on speech analysis to aid in screening for psychiatric disorders such as PTSD and dementia. However, this opportunity can only truly be explored if the data bottleneck is addressed, and researchers are given access to large well-curated speech data sets covering a variety of psychiatric disorders collected from multiple study sites in a collaborative manner.