Detecting Cognitive Impairment from Spoken Language

by

Tuka Alhanai

B.S., The Petroleum Institute (2011)
S.M., Massachusetts Institute of Technology (2014)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Author . . . .
Department of Electrical Engineering and Computer Science
May 22, 2019

Certified by . . . .
James R. Glass
Senior Research Scientist
Thesis Supervisor

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

Detecting Cognitive Impairment from Spoken Language

by

Tuka Alhanai

Submitted to the Department of Electrical Engineering and Computer Science
on May 22, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Dementia comes second only to spinal cord injuries in terms of its debilitating effects; from memory-loss to physical disability. The standard approach to evaluate cognitive conditions are *neuropsychological exams*, which are conducted via in-person interviews to measure memory, thinking, language, and motor skills. Work is on-going to determine biomarkers of cognitive impairment, yet one modality that has been relatively less explored is speech. Speech has the advantage of being easy to record, and contains the majority of information transmitted during neuropsychological exams.

To determine the viability of speech-based biomarkers, we utilize data from the Framingham Heart Study, that contains hour-long audio recordings of neuropsychological exams for over 5,000 individuals. The data is representative of a population and the real-world prevalence of cognitive conditions (3-4%).

We first explore modeling cognitive impairment from a relatively small set of 92 subjects with complete information on audio, transcripts, and speaker turns. We loosen these constraints by modeling with only a fraction of audio (∼2-3 minutes), of which the speaker segments are defined through text-based diarization. We next apply this diarization method to extract audio features from all 7,000+ recordings (most of which have no transcripts), to model cognitive impairment (AUC 0.83, spec. 78%, sens. 79%). Finally, we eliminate the need for feature-engineering by training a neural network to learn higher-order representations from filterbank features (AUC 0.85, spec. 81%, sens. 82%). Our speech models exhibit strong performance and are comparable to the baseline demographic model (AUC 0.85, spec. 93%, sens. 65%).

Further analysis shows that our neural network model automatically learns to detect specific speech activity which clusters according to: pause followed by onset of speech, short burst of speech, speech activity in high-frequency spectral energy bands, and silence.

Thesis Supervisor: James R. Glass
Title: Senior Research Scientist

Acknowledgments

I am very grateful for the mentorship and guidance I have received from my advisor Dr. James Glass. He allowed me to learn, grow, and pursue my interests. It was his sharp eye that recognized that this project could be interesting, and that there could be a lot of potential to this domain. I am also very grateful to Professor Rhoda Au who helped make this research happen through her open-mindedness and willingness to share data, resources, and experiences.

At The Framingham Heart Study I would like to thank the young and youthful team that helped process the largest dataset we ever had to deal with; Ida Xu, Brynna Wasserman, Maulika Kohli, Nancy Heard-Costa, Yulin Liu, Karen Mutalik, Mia Lavallee, Cody Karjadi, Alvin Ang, and Spencer Hardy.

I am also very grateful for the constructive feedback provided by my thesis committee members Professors Collin Stultz and Roger Mark, as well as Professor Victor Zue, Dr. Stephanie Seneff, Marcia Davidson, and my colleagues past and present at the Spoken Language Systems Group. I also appreciate the mentorship I received over the years from my academic advisor Professor Munther Dahleh.

The work in this thesis not only exists because of essential contributions made by individuals, but also by the momentum of non-profit organizations, and those individuals within who help make meaningful things happen, but are too often not there to experience first-hand the impact of their decisions. I am very grateful to the National Institute of Health and the Framingham Heart Study for allowing this research to exist through their 70+ year operation. I am also very grateful for the foresight the Abu Dhabi Department of Education and Knowledge in sponsoring myself and fellow students. I would like to acknowledge their talented team over the years; Fatema Alhosani, Reem Aljabri, Sultana Alketbi, Maha Basuwaid, and Mona Almansoori.

I would also like to acknowledge Dr. Nadia Alhasani for her guidance and belief in a young lady. I thank Dr. Raed Shubair for his energy and counsel.

Ultimately, to my family I remain eternally grateful.

Bibliographic Note

Portions of this thesis have appeared in peer-reviewed publications, which are:

- T. Alhanai, R. Au, and J. Glass. "*Spoken Language Biomarkers for Detecting Cognitive Impairment*", IEEE ASRU 2017, Okinawa JAPAN.

- T. Alhanai, R. Au, and J. Glass. "*Role-specific Language Models for Processing Neuropsychological Exams*", NAACL HLT 2018, New Orleans USA.

Source code related to this thesis is available at: https://github.com/talhanai

Contents

List of Figures

List of Tables

Glossary

APOE	Apolipoprotein
AUC	Area Under the Receiver Operating Characteristic Curve
CI	Cognitive Impairment
CNN	Convolution Neural Network
DER	Diarization Error Rate
HL-test	Hosmer-Lemeshow Test
HNR	Harmonic-to-noise Ratio
FHS	Framingham Heart Study
FPR	False Positive Rate
LSTM	Long Short-Term Memory
MCI	Mild Cognitive Impairment
MFCC	Mel-frequency Cepstral Coefficient
ROC	Receiver Operating Characteristic Curve
SSE	Sum of Squared Error
TPR	True Positive Rate
ZCR	Zero Crossing Rate
WER	Word Error Rate

Chapter 1

Setting the Scene

We meet Carol for the first time.

Interviewer: How old are you now?

Carol: 65? 65. I think, right?

Interviewer: Yeah.

Three years later ...

Interviewer: How old are you?

Carol: 80? ... No? I don't know.

Interviewer: You're actually 67.

Carol: 67?

Interviewer: 67.

Carol: Yeah?

Two years later ...

Interviewer: What's your husband's name?

Carol: My husband's? [laughter]

Interviewer: Your husband's name.

Carol: Yeah. [laughter]

Interviewer: The guy sitting to your left.

Carol: Yeah. [laughter]

Interviewer: That big guy who loves you.

Carol: Yeah. Yeah. Who loves me. [laughter]

Carol: [silence]

Carol: *shakes head*

Carol's condition continues to degenerate for the next decade. She loses her ability to speak and move. Her husband, the name of whom she struggles to recall, promises to support and care for her, 'until death do us part'. Many years later, her husband struggles with the difficult reality that the person he once knew is no more. A realization that leads him to suffer from severe depression and to contemplate suicide. As a former New York Police Department (NYPD) officer, he reports this was by far the most challenging situation he has ever had to cope with. With their life's savings dwindling, Carol is eventually moved to an elder care nursing home which can provide her with complete 24-hour support[1].

Carol's story illustrates the serious and slow toll dementia can take on the individual and their loved ones. Her story provides a snapshot into the necessity of having comprehensive access to mental, emotional, medical, and financial services. While providing such support is often complex, the ability to *detect* cognitive impairment can play a critical role in preparing individuals for the long journey ahead. With this thread of the challenge held firmly in our grasp, we will show in this thesis how cognitive impairment may be detected from audio recordings of speech alone, a modality that is rich in information and also easy to record.

[1] Featured on CBS 60 Minutes. *Alzheimer's Disease: Following a couple from diagnosis to the final stages of Alzheimer's*. April 22nd 2018. https://www.cbsnews.com/news/alzheimers-disease-following-a-couple-from-diagnosis-to-the-final-stages/

1.1 Contributions

This thesis provides the following contributions to the domains of speech, language, medical informatics, and machine learning.

- **Modeling at scale:** We evaluate the modeling and predictive performance of cognitive conditions at a population level of 5,000+ individuals.
- **Operating with real-world limits:** We present methods developed on, and evaluated for, real-world conditions where audio recordings are noisy, and ground truth transcriptions are often lacking. We are able to model cognitive impairment when only 1.4% of the data has manually generated transcriptions.
- **Defining spoken language biomarkers:** We present spoken language biomarkers that are associated with cognitive impairment. While not everyone may have the luxury to code up and input biomarkers into a machine learning algorithm, one may still utilize this information in their personal mental model of the world as they interact with individuals within their sphere.
- **Maintaining privacy:** We utilize methods in which personal information about speakers (content of what was said) is not necessary and therefore protected. In fact, our language modeling paradigm utilizes no personal information and allows us to diarize large amounts of audio while maintaining salient information for modeling cognitive impairment.
- **Generating interpretable results:** We apply techniques that map the association between spoken language biomarkers and cognitive conditions. When the associations are not easily drawn from standard models, as is the case of neural network models compared to logistic regression, we show how interpretable information can still be inferred.

1.2 Thesis Overview

This thesis is organized as follows, and builds on methods developed in each chapter. An overview of the studies presented in this thesis displayed in Figure 1-1.

In **Chapter 2** we define cognitive impairment, describe prior work on detecting biomarkers (speech and non-speech), and the dataset used in this thesis.

In **Chapter 3: Study 1** we model cognitive impairment on an initial cohort of 92 'gold' subjects utilizing fully annotated data (transcripts and speaker labels), and explore a set of audio and text based spoken language biomarkers.

Most of our data has no speaker turn labels, so in **Chapter 4: Study 2** we present our audio diarization pipeline that utilizes role-specific language modeling to extract the segments in which a subject was speaking to model cognitive impairment.

In **Chapter 5: Study 3** we utilize spoken language biomarkers from audio alone to model 5,000+ 'bronze' subjects.

In **Chapter 6: Study 4** we explore the use of neural network models that can learn to detect cognitive impairment straight from spectral energy of the audio waveform; eliminating the need for engineering hand-crafted features.

In **Chapter 7** we summarize our findings and describe future work.

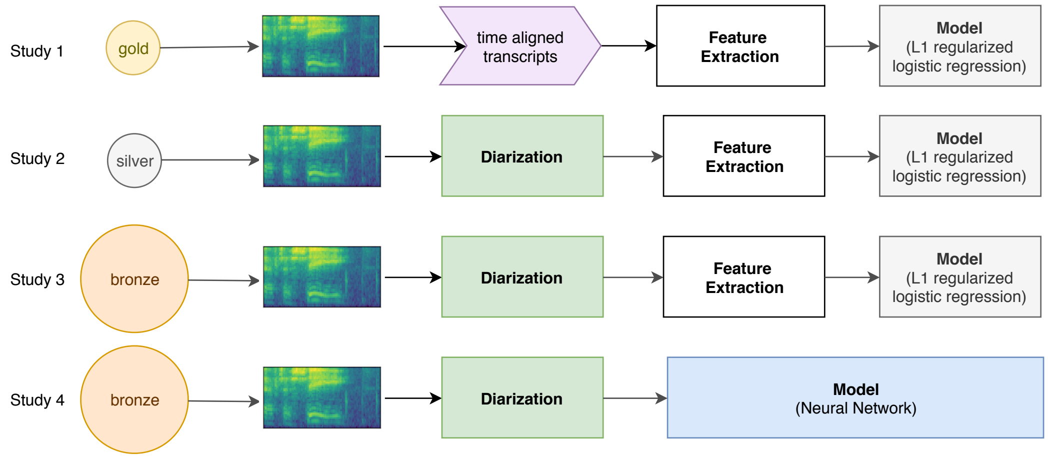


Figure 1-1: **Overview of studies presented in this thesis.** Gold, silver, and bronze refer to datesets with different levels of annotation quality.

.

Chapter 2

Background

Synopsis

Cognitive impairment is a noticeable decline in mental abilities which manifests in speech through memory loss, dis-connected thinking, and dis-fluent communication. The impairment may be caused by a variety of factors, such as Vascular dementia (stroke) and Alzheimer's disease, and affects about 4 million people above the age of 60 (1 in 20). The impact and prevalence of cognitive impairment creates a serious psychological and financial burden on patients, families, and greater society.

The medical community have developed structured tests conducted via in-person interviews to evaluate and score the presence for, and severity of cognitive impairment, often referred to as *neuropsychological exams*. Work is on-going to determine biomarkers of cognitive impairment, from brain imaging scans to laboratory tests, yet one modality for extracting bio-markers that has been relatively less explored, is speech. Speech has the advantage of containing the majority of information transmitted during neuropsychological exams, and may be recorded passively, using an off-the-shelf microphone, and at the convenience of the subject's desired schedule and location.

To determine the viability of speech-based biomarkers, we utilize data collected from the Framingham Heart Study, which contains 7,779 audio recordings of neuropsychological examinations for 5,180 individuals conducted over a 10 year period from 2005 to 2015. This data allows us to perform population-level modeling with samples representative of the real-word prevalence of cognitive conditions.

2.1 Cognitive Impairment and Dementia

The word *dementia* came into circulation in the late 1700s due to its usage by one of the founders of modern psychiatry, Philippe Pinel [Torack, 1983]. It comes from the French/Latin word *démence* or *demens* which means *out of one's mind* [Hippius and Neundörfer, 2003, Jellinger, 2010]. Such an expression alludes to how unexplainable the ailment was seen to be. As knowledge expanded and techniques improved, the times were ready for Alois Alzheimer to be one of the earliest individuals to formally record dementia. He first observed "a peculiar severe disease process of the cerebral cortex" in a 56 year-old woman, too young to have been suffering from memory loss due to aging [Hippius and Neundörfer, 2003]. Here is his description of one of his first patients, 56 year-old Johann F., and an example of an early neuropsychological exam in the history of this evaluation protocol [Moller and Graeber, 1998].

> The 56-year-old labourer Johann F. was admitted to the psychiatric clinic on 12 Nov. 1907. There was no history of excessive drinking. Two years before admission his wife died, since then he became quiet and dull. In the previous 6 months he had become forgetful, could not find his way, could not perform simple tasks or carried these out with difficulty. Pupillary reaction normal. Patellar-reflex a little brisk. No signs of nervous palsy. Language strikingly slow, but without articulatory disturbance. Dull, slightly euphoric, impaired understanding. Echoes questions put to him frequently and repeatedly instead of giving a reply. Can only solve very simple calculations after a long delay. ... he cannot name a matchbox.

In current usage, dementia is defined as a decline in mental abilities that is severe enough to interfere with daily life. Individuals suffering from dementia struggle to recall names of objects and loved ones, string a coherent sentence together, follow a cooking recipe, and even lose the ability to be mobile. Dementia is not a symptom of 'normal aging', but are the observable symptoms of specific pathologies that may develop for many years before they become manifest [Nestor et al., 2004, Braak and

Del Tredici, 2012]. Pathologies that underly dementia include Alzheimer's disease, Vascular dementia, and Lewy Body dementia.

The mechanisms behind these pathologies can be complex and varied, but there is evidence to suggest that some pathologies (such as Alzheimer's) develop in stages [Nestor et al., 2004, Perl, 2010]. This temporal element is more pronounced in the case of stroke patients who exhibit improved motor function along with associated cortical reorganization during recovery [Feydy et al., 2002]. Furthermore, different pathophysiological conditions may manifest with varying levels of impact on cognition [Iadecola, 2010]. Even though the mechanisms causing dementia remain undiscovered, there is some utility in a simplified model to capture the relationship between cognitive function and the underlying pathology; as neuro-degeneration becomes more aggressive (i.e. pathological load), then cognitive function declines as a result. Figure 2-1 illustrates cognitive function (blue line) and pathological load (red line, which may be viewed either as a function of time or function of severity), and highlights how normal cognitive function is maintained up to a critical threshold (yellow band), then evolving into mild cognitive impairment (MCI), after which dementia becomes manifest (green band). This simplified model not only serves as a description on the relationship between pathology and cognitive function, but it helps illustrate some of the opportunities available during the different 'stages' of cognitive function (expanded upon in 2.1.1), and which affects a significant number of people.

As of 2013, about 4 million people above the age of 60 (1 in 20) were suffering from dementia in the USA alone, and 35 million globally, with these numbers estimated to double every 20 years [Prince et al., 2013]. Dementia is both a physically and financially taxing condition. Individuals may live for over a decade with the condition, which comes second only to spinal cord injuries in terms of debilitating effects. In addition, individuals spend on average $80,000 a year in services and support resulting in $200-$600 billion a year spent in the USA [Hurd et al., 2013, Wimo et al., 2006]. For perspective, the amount spent is the same market value as some of the world's mightiest conglomerates; Coca Cola, Walmart, and Petrochina.

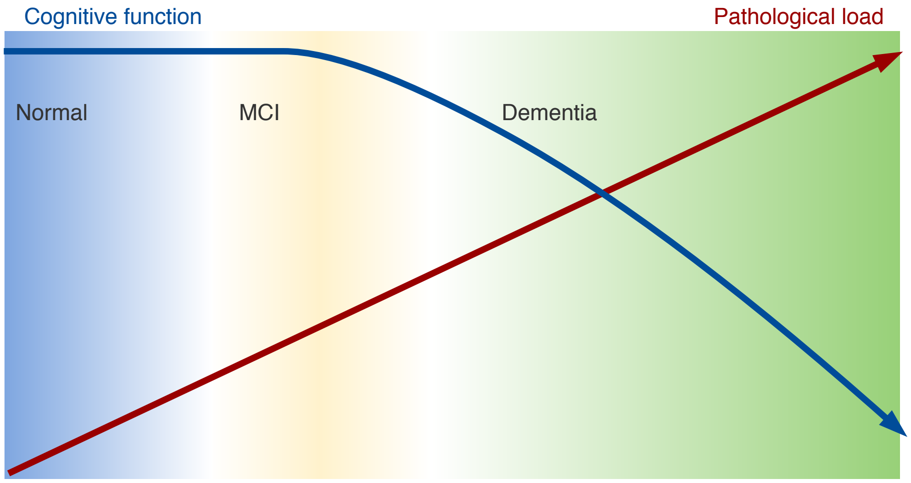


Figure 2-1: **Pathological load and cognitive function as a function of time** [Nestor et al., 2004]. (MCI = mild cognitive impairment)

2.1.1 The Value of Detecting Cognitive Conditions

The objective of this thesis may beg the question: why endeavour to detect cognitive impairment? Generally speaking, we would like to live longer and have healthier lives [Locke and Latham, 2002], so developing techniques to detect cogntive conditions will help extend our life-goal in the following three ways; as a pathway to treatment, as a motivation for lifestyle changes, and as an avenue for prevention.

Pathway to Treatment

Being notified of the presence of dementia can empower the patient and their friends/family to be strategic in acquiring support and services. Dementia is a costly condition with coverage often coming from personal savings or medical insurance (if it exists), so financial planning can help accommodate the condition and minimize its burden [Hirschman et al., 2008]. There are also different strategies patients and family can take when it comes to hospital care. There is evidence to suggest that hospital care in the home can reduce hospital visits and mortality rates since it reduces the stress and risk patients may be exposed to when visiting health-care facilities [Caplan et al., 2006]. There is also the issue of the hidden burden. Dementia is not only debilitating for those suffering from the condition, but it also takes its toll on caregivers such as friends and family, who often feel that it is their duty to provide their full support, and who over time have to internalize the reality that the person they loved is no more the same person they interact with. A person who may not remember them, or recall past memories and shared experiences. The result is that 80% of caregivers state that they frequently experience high levels of stress and up to 50% of caregivers suffer from depression, which can be severe enough that they contemplate suicide as an outlet from the degenerative situation [Etters et al., 2008].

Motivating Lifestyle Changes

Detecting dementia may also motivate lifestyle changes. Studies have shown that daily engagement in cognitive activities [Hall et al., 2009], a healthy body-mass index (BMI) [Chuang et al., 2016], and bilingualism [Bialystok et al., 2007], can delay onset by months and years. It has also been observed that physical activity and nutrition may reduce the risk of developing dementia. Just three days of exercise a week and eating a fish meal a week keeps the brain at its peak, reducing the risk of dementia by 40% and 70% respectively [Larson et al., 2006, Morris, 2009].

Detection for Prevention

Research in preventative strategies is still young, with several clinical trials having been conducted over the years, but no positive outcomes yet [Mangialasche et al., 2010, Schneider et al., 2014]. What research has found however, is that individuals who take non-steroidal anti-inflammatory drugs (NSAIDs) consistently, such as aspirin or ibuprofin, have a lower risk of developing dementia [Szekely et al., 2004]. Research on mice has shown that the SIRT3 protein protects brain cells against aging [Cheng et al., 2016], and that individuals suffering from mixed pathologies have a more damaging outcome than individuals suffering from Alzheimer's alone [Kawas et al., 2015]. This evidence highlights that early detection of the onset of dementia will be critical in encouraging research and development of preventative treatments that are also holistic and considers the interactions between the many different brain related pathologies.

Finally, it is important to consider that while detection has its utility to motivate treatment strategies, the continuous *monitoring* of dementia is also important as it provides a timely method for decision-making by stakeholders, such as: evaluating the progress of a clinical trial [Grundman et al., 2004], determining the efficacy of medication [Guo et al., 1999], assessing the impact of a lifestyle change [Fratiglioni et al., 2004], establishing a power of attorney [Gregory et al., 2007], and deciding when a patient should move to an advance care facility [Chang et al., 2009].

2.2 Biomarkers

Earlier we highlighted Alois Alzheimer's early experience with a patient suffering from dementia. He came to the conclusion that there was a particular pathology causing the symptoms via autopsy. Even with evidence to suggest that such a malady (cognitive impairment outside of normal aging) was observed by and recorded as early as the Greeks, from Plato to Hippocrates [Boller and Forbes, 1998], our sophistication in conclusively diagnosing the pathology still remains with an autopsy of the brain [Clark et al., 2003, Knopman et al., 2003, Walker et al., 2007]. However, there is a diverse range of work researching biomarkers for detecting dementia and the underlying pathology causing it, that aim to deduce an individual's cognitive condition while they are alive.

2.2.1 Laboratory Tests to Brain Imaging Scans

Studies of cognitive impairment and dementia have explored multiple modalities of information for assessment and diagnosis. This includes subjective measures of cognitive decline (e.g., patient's response to the question 'Has your memory become worse?') [Saykin et al., 2006, Jessen et al., 2010, Reisberg et al., 2010], medical profile (stroke, cardiovascular disease, blood pressure, etc.) [Newman et al., 2005, Satizabal et al., 2016], education level [Ott et al., 1995, Cobb et al., 1995], brain imaging exams [Au et al., 2006, Jack Jr et al., 2009, Mosconi et al., 2008], apolipoprotein E (APOE) genotype (from plasma samples) [Farrer et al., 1997, Kim et al., 2009, Myers et al., 1996], atherosclerosis (via ultrasonography) [Hofman et al., 1997], brain-derived neurotrophic factor (BDNF) [Weinstein et al., 2014], cerebro-spinal fluid [Shaw et al., 2009], and other laboratory measures (glucose homeostasis, markers of inflammation, blood homocysteine, folate, vitamin B-12, etc.) [Quadri et al., 2004, Van Himbergen et al., 2012].

Currently, none of these biomarkers are part of the diagnostic criteria for Alzheimer's disease (as recommended by the National Institute on Aging-Alzheimer's Association workgroups [NIA-AA]), with suggested usage contained to a research setting [McKhann et al., 2011]. The main reasons behind this constraint on the diagnostic criteria are due to the limited standardization of biomarkers, and because the core clinical criteria already provides good diagnostic accuracy.

2.2.2 Spoken Language

While the studies highlighted in Section 2.2.1 have explored associations between their measures and cognitive outcome, such information has a high barrier of acquisition due to the costly nature of laboratory tests and imaging scans, ranging anywhere from $100 for an APOE gene test [Keshava, 2017], to $4,000 for a brain imaging positron emission tomography (PET) scan (and which may not even be covered by health insurance) [Bahrampour, 2017]. This motivates the exploration of measures that are easier to record and are less invasive, specifically, speech.

Spoken Language as a Vessel for the Shadow of the Mind

Speech is commonly used as a proxy to assess an individual's cognitive state, by both medical professionals and untrained individuals (from friends to family), and is often the first indicator that something is amiss. The ability for spoken communication to relay cognitive health is exemplified through the design and usage of neuropsychological exams to quantify a subject's cognitive state [Kurlowicz et al., 1999]. These exams are conducted one-on-one with a subject, in a structured interview-like format. The examiner poses pre-defined questions to the subject to verbally recall details of a story, repeat a sequence of digits, define the similarity between two objects, read unfamiliar words, and answer questions on general knowledge. Depending on what the subject responded, a score is assigned to quantify their cognitive impairment. Since the standard screening process for cognitive impairment relies on neuropsychological exams, it follows that the speech relaying information from the mind, can in and of itself, be used to model an individual's cognitive state. Indeed, speech contains both linguistic information (i.e. words as a response to questions posed during the exam) as well as prosodic information (i.e. the intonation used to communicate words), of which higher-order information may be extracted (i.e. whether they answered a question correctly, how long they took to recall a detail, or if they exhibited hesitation in their response).

Spoken Language is Easy to Record

Speech may also be easily recorded, using an off-the-shelf microphone or one's phone. The accessibility of microphones means speech can be recorded at any place and at any time. Speech recordings are also easy to store, which makes observation of changes over longer periods of time (weeks, months, years) possible.

Spoken Language Features of Cognitive Impairment

Many different patterns and features may be extracted from speech, from intonation to language usage. We can categorize features into two categories: text-based features and audio-based features, i.e. *what* was said and *how* it was said. Studies utilizing text-based features have modeled a subject's language by capturing patterns of word usage (e.g. part-of-speech tag, and vocabulary-to-total-words ratio) [Ripich et al., 1991, Thomas et al., 2005a], and conversation acts (e.g. response to question, active listening) [Lopez-de Ipiña et al., 2015]. Studies utilizing audio-based features have extracted acoustic information from the speech waveform such as speech/silence segments, onset time of speech, time intervals between words in a sentence as aligned with a transcript [König et al., 2015], as well as prosodic information (e.g. pitch, shimmer, jitter) [Meilán et al., 2014, Lopez-de Ipiña et al., 2015].

Spoken Language Scenarios

Speech data utilized in studies have varied from being highly structured memory and reading tests [Meilán et al., 2014, König et al., 2015, Fraser et al., 2016], to organic, open-ended, conversation-like interactions (e.g. sharing stories) [Atay et al., 2015, Lopez-de Ipiña et al., 2015].

Scale of Speech-based Studies

In terms of scale, studies have varied from utilizing a dozen subjects, to 1,000 [Bot et al., 2016, Wan et al., 2018]. Generally, the ratio between subjects with a cognitive condition relative to healthy controls, does not represent the real-world prevalence of the condition. For example, the largest study yet, with 1,000 subjects [Bot et al., 2016, Wan et al., 2018], had 50% of subjects with the outcome of interest, Parkinson's disease, yet the condition actually has a real-word prevalence of 0.4% (for ages 60 to 69) [Pringsheim et al., 2014]. The advantage of balancing outcomes is that the study focuses on determining the salient features of a condition, in isolation to other conflicting factors introduced by imbalanced data. However, for a model to generalize, it will need be able to operate in scenarios which represent the real-world incidence of conditions, similar to approaches pursued for anomaly detection [Haixiang et al., 2017], where class outcome of the data are proportional to the prevalence of the anomaly being detected, such as e-mail spam, or malicious system attacks which are rare, sporadic, but critical to detect.

Terminology

For the purpose of this thesis we will be using the term *cognitive impairment* (rather than dementia) to refer to the state wherein an individual has recognizable lapses in recalling memories, deficiencies in motor ability, and speech dis-fluencies attributed to some hidden (abnormal) function in the brain. We use this terminology inline with the ground truth and vocabulary used at the Framingham Heart Study. We will revisit this definition in Section 2.3.2.

2.3 The Framingham Heart Study

In this thesis, we utilized audio recordings of neuropsychological examinations from the Framingham Heart Study (FHS) to model cognitive impairment. The FHS is an on-going longitudinal, multi-generational population study of 15,447 subjects from 1948 to the present [Mahmood et al., 2014]. It was established by the National Heart Act in a time when cardiovascular conditions became the main cause of death in the USA, and poor understanding surrounded it.

Data from this study provides snapshots into the lives and health of individuals in the town of Framingham located in the state of Massachusetts in the USA. Using evaluation criteria developed by the medical community, the data contains detailed information on the mental well-being, personal relationships, socio-economic status, medical history, physical condition, cognitive health, and autopsy reports. Breakthrough research coming out of the study was able to discern associations between lifestyle behaviors (such as smoking and physical activity) with cardiovascular outcomes [Kannel et al., 1961, Kannel and Sorlie, 1979, Kannel, 2000], and more recently, research was able to map how happiness spreads in an in-person social network [Fowler and Christakis, 2008].

2.3.1 Audio Recordings

Most relevant to this thesis, the data also contains 7,779 audio recordings of neuropsychological examinations for 5,180 individuals conducted over a 10 year period from 2005 to 2015. All this data combined allows for a holistic study, using established medical baselines, and both structured (medical records, exam format) and unstructured information (spoken language) to assess health and well-being.

2.3.2 Neuropyschological Examinations

Neuropsychological examinations are an important screening tool for the presence of cognitive conditions such as Alzheimer's disease, Parkinson's disease, and Vascular dementia [Elias et al., 2000]. The exam is composed of multiple components that measure a specific domain of cognition such as: thinking, immediate and delayed memory recall, speech, and physical movement. Each exam component is assigned a score by the tester according to the established rubric, from which an overall score for the severity of cognitive impairment is computed. The scores can be 0 = not demented, 0.5 = mild cognitive impairment, 1 = mildly demented, 2 = moderately demented, and 3 = severely demented. A complete copy of the exam may be viewed in [Satizabal et al., 2016], while some of the tests are highlighted below.

Anna Thompson - Logical Memory Immediate, Delayed, and Recall

An example of a test within the the neuropsychological examination is the logical memory test where subjects are asked to remember a story about a woman called Anna Thompson. An example of the tester delivering this test is provided below.

> **Prompt**: Well, the first test we'll be doing today I'll be reading you a little story of about four or five lines. When I'm through I'll ask you to tell me everything that I read to you, are you ready?
>
> **Story**: Anna Thompson, of South Boston, employed as a scrub woman in an office building, reported at the City Hall Station that she had been held up, on State Street, the night before, and robbed of fifteen dollars. She had four little children, the rent was due, and they had not eaten for two days. The officers, touched by the woman's story, made up a purse for her.
>
> **Question 1** (Immediate): Tell me everything that I read to you.
>
> **Question 2** (Delayed): Well at the beginning of this test I read a story to you I'd like you to tell me that story again.
>
> **Question 3a** (Recall): Okay well now I'm going to ask you some questions about that story and I'd like you to tell me which one of the three choices in each question is correct. Was the story about a woman a man or an animal?
>
> **Question 3b** (Recall): Was her name Annie Thomas, Anna Thompson or Annie Hall?

> **Question 3c** (Recall): Was she from the Southwest, South Boston or Cape Cod?

East-West - Verbal Paired Associates, Immediate, Delayed, and Recall

For the Verbal Paired Associates Test, the subject is asked to recall the word associated with a given word they've been shown before hand. Similar to the Anna Thompson test, the subject is asked to recall the pair immediately, later in the exam, and from a three-set option. Below as an example of the test prompt.

> **Prompt**: Okay, for this next text I'm going to read you a list of words two at a time when I'm through I will tell you one of the words and then I want you to tell me the word that went with it, so for example if the words were east-west gold-walk then when I say east what would you say ...?

Abstract Shapes, Digit Sequence, Animals, Clock-drawing and other tests

There are several other tests, briefly, these include a test to evaluate visual reproduction ability. The subject looks at a drawing composed of abstract shapes for 10 seconds, and then reproduce it by hand. Another test is the Digit Span where the subject is asked to recall a sequence of 4 to 9 digit numbers, forwards and backwards. The Animal Fluency test asks a subject to recall as many animals as possible. There is a clock-drawing test where the subject must reproduce a familiar object, a clock from memory, as well as fill in a pre-drawn clock. Yet another test asks a subject to define the similarity between two objects (How are an orange and banana alike? Saw and Axe? Eye and Ear? etc.). There is also a test to evaluate general knowledge (In what continent is Brazil? Who was Amelia Earhart?).

2.4 Challenges of Medical Speech Data

Speech is relatively easy to collect, requiring an off-the-shelf microphone and a data storage plan, however utilizing such a corpus in a medical context presents its own challenges.

2.4.1 Protecting Privacy

Privacy assurances comfort subjects and makes them more willing to share medical information, therefore raw audio recordings are hard to collect and and even more difficult to share because they (often inadvertently) contain identifiable information on the subject (e.g. their name, family details, work location).

2.4.2 High Quality Data Labels

Tools to automatically de-identify audio have yet to operate with the specificity demanded by medical institutions (i.e. human parity). It has been demonstrated that speech-to-text systems can be trained to transcribe with human precision on specific data ($<$ 6% word error rate [WER]) [Xiong et al., 2017], however such corpora have been available for research and development for over 20 years. The pursuit of generating high-quality transcriptions from audio recordings of individuals with speech dis-fluencies continues to be an on-going area of research [Christensen et al., 2012], with results fluctuating anywhere from 4% to 90% WER for attempts to automatically transcribe speech of individuals with varying severities of speech impediments.

The alternative to automatic speech transcriptions is to have humans manually transcribe and de-identify the audio. This is too costly to scale as a viable solution; at $100 per audio recording, it would cost on the order of $800K to transcribe all recordings of the Framingham Heart Study.

2.4.3 Real-world Constraints Define the Research Trajectory

These limitations established by the need to maintain privacy of subjects and utilizing data with very limited transcriptions and speaker labels, opens up interesting areas for research, and defines the trajectory of this thesis.

Thus our outcome of interest is to detect cognitive impairment with almost:

1. No transcription information (What was said?).

2. No speaker-audio alignments (Who spoke when?).

3. No test-audio alignment information (Which test? i.e. the component - memory, logic, etc. - of the exam that was being evaluated).

4. No engineered features.

2.5 Corpus

The methods presented in this thesis utilize three types of datasets: ‘gold’, ‘silver’, and ‘bronze’. The gold dataset is used for initial analysis on a subset of 92 subjects (Chapters 3). The silver dataset contains the same data as the gold, but without the human-generated text transcripts, and is utilized to evaluate our diarization method (Chapter 4). The bronze set contains the full set of 5,000+ subjects, which have no associated transcripts beyond the 92 gold subjects (utilized in Chapters 5 and 6). A summary of each set can be viewed in Table 2.1. Differences in bronze sets between Chapter 5 and 6 are due to differences in data pre-processing that filtered out some subjects.

	Gold Ch. 3	**Silver** Ch. 4	**Bronze** Ch. 5	**Bronze** Ch. 6
# subjects	92	92	4,836	5,063
# recordings	92	92	6,705	7,196
# impaired	21 (22.8%)	21 (22.8%)	224 (3.49%)	256 (3.56%)
Transcripts	Yes	No	No	No
Duration (mean)	65 mins	65 mins	64 mins	63 mins
Total duration	100 hrs	100 hrs	7200 hrs	7600 hrs
Age (mean)	68 yrs (+/17)	68 yrs (+/17)	64 yrs (+/- 15)	63 years (+/- 15)
Sex (female)	45 (49%)	45 (49%)	3,701 (55.2%)	3,985 (55.4%)

Table 2.1: **Corpus statistics.** Each study utilizes data with reduced annotations and increased number of subjects, and was composed of three sets defined as gold, silver, and bronze. Gold contains 92 subjects with complete transcripts and speaker turn labels. Silver contains the same 92 subjects as the gold, but no transcripts and speaker turn labels. Bronze contains the full set of subjects, with no transcripts or speaker turn labels.

2.5.1 Gold Dataset: Fully Annotated

This gold dataset is composed of 92 audio recordings of neuropsychological examinations with manually generated text transcripts. Recordings are on average, 65 minutes in duration, contain 2,496 words, with a vocabulary size of 527 words. 21 recordings (22.8%) contain subjects with cognitive impairment. Ten of these subjects had a severity rating less than mild, six were mild, five were moderate, and none were severe [Seshadri et al., 2006]. Fourteen subject were diagnosed as having Alzheimer's disease using the NINCDS-ADRDA[1] criteria [McKhann et al., 2011], and five were diagnosed with Vascular dementia based on the NINDS-AIREN[2] criteria [Román et al., 1993].

Transcripts for each audio file were generated manually using rules developed in [Glass et al., 2004]. Transcribers were instructed to include timestamps for each speaker turn (subject/tester), indicate who spoke when, transcribe speech phonetically (e.g. nineteen dollars instead of $19), include tags to highlight moments such as filled pauses (<um>), and to subjectively insert punctuation.

2.5.2 Silver Dataset: No Transcripts

The silver dataset is composed of the same data as the gold set, but without transcripts. By eliminating transcripts, we are able to mimic the conditions of the bronze dataset (i.e. audio with no transcripts), whereby audio has to be segmented by speaker turn first before further feature extraction and modeling can be performed. The silver dataset is utilized in Chapter 4 to perform experiments on diarization, while also allowing for fair comparisons with experiments that utilize the gold set.

[1] National Institute of Neurological and Communicative Disorders (NINCDS) and Stroke and the Alzheimer's Disease and Related Disorders Association

[2] National Institute of Neurological Disorders and Stroke (NINDS) and the Association Internationale pour la Recherche et l'Enseignement en Neurosciences (AIREN)

2.5.3 Bronze Dataset: All Subjects and No Transcripts

This bronze dataset in Chapter 5 is composed of 6,715 audio recordings belonging to 4,838 unique subjects. None of these recordings have manually generated text transcripts. The recordings are on average 64 minutes in duration. 224 recordings (3.49%) contain subjects with cognitive impairment.

This bronze dataset in Chapter 6 is composed of 7,196 audio recordings belonging to 5,063 unique subjects. None of these recordings have manually generated text transcripts. The recordings are on average 63 minutes in duration. 256 recordings (3.56%) contain subjects with cognitive impairment.

2.5.4 Largest Medical Speech Analysis to Date

The bronze set we utilize in our experiments represents the largest corpus of medical speech data yet, composed of 7,000+ recordings, 5,000+ subjects, and 7,600+ hours. As a reference, and to the best of our knowledge, the largest set of audio data is composed of 1,000 subjects to model Parksinon's disease [Wan et al., 2018].

2.6 Evaluating Model Performance

There are several evaluation metrics that are consistently used throughout this thesis. We use the True Positive Rate (TPR) and False Positive Rate (FPR) to measure the sensitivity and specificity of our models [Davis and Goadrich, 2006]. Our main metric for evaluating model performance is the Area Under the Receiver Operating Characteristic Curve (AUC) [Hanley and McNeil, 1982]. To evaluate the calibration of our models we use the Hosmer-Lemeshow test (HL-test) [Hosmer and Lemesbow, 1980].

2.6.1 Sensitivity and Specificity

A useful technique to quantify classification performance of a model is to use a confusion matrix. Figure 2-2 displays a confusion matrix between actual and predicted

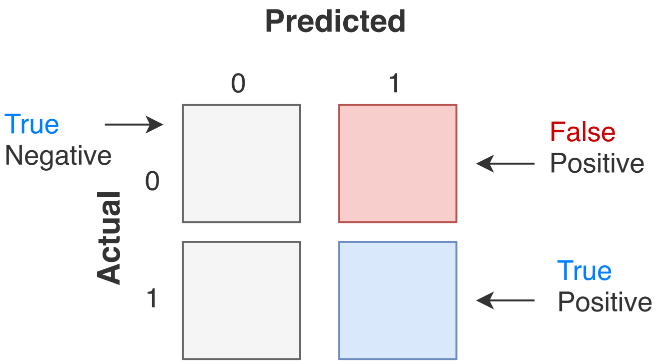


Figure 2-2: **Confusion Matrix**. The visualization displays the four quadrants in a binary classification task. Classification label '1' indicates cognitive impairment, while '0' indicates healthy. The y-axis corresponds to the *actual* class labels, while the x-axis corresponds to the *predicted* class labels. Each quadrant contains information on true negative, false positive, true positive, and false negative.

classes for a binary outcome, and displays the quadrants for true positive, false positive, true negative, and false negative. From this matrix we can compute an array of classification performance, whereby the outcome of '1' corresponds to cognitive impairment, and '0' is healthy. Most relevant to this thesis are the following:

- **Specificity** (TPR, or recall): is the number of subjects correctly classified with the condition (i.e. cognitive impairment), divided by the number of subjects that actually have the condition. Formally this is:

$$\text{Specificity (TPR)} = \frac{\sum \text{true predicted positives}}{\sum \text{actual positives}} \tag{2.1}$$

- **Sensitivity** (TNR, or 1 - FPR): is the number of subjects correctly classified as healthy, divided by the number of subjects that are actually healthy.

$$\text{Sensitiviy (TNR)} = \frac{\sum \text{true predicted negatives}}{\sum \text{actual negatives}} \tag{2.2}$$

Generally, we want a classifier to maximize sensitivity and specificity, although that criteria is ultimately determined by the decision-maker who would deploy the model [Hoffrage et al., 2000]. Sensitivity and Specificity are especially important in scenarios where conditions are relatively rare, but where population-level screening may result in more false positives than true positives. This can create a costly psychological burden in individuals (e.g. increased risk of suicide [Lu et al., 2013]), increases financial burden on individuals and the health-care system [Elmore et al., 1998], becomes a source of litigation [Wilson, 2000], and reduces trust in the screening process [Maxim et al., 2014, Vlahiotis et al., 2018].

2.6.2 Receiver Operating Characteristic Curve

Predictive models output values between 0 and 1 that convey the probability that an event has (or will) occur. These predicted probabilities are often used for *classification*, where a hard threshold (often set at 0.5) defines whether an input falls into one class (e.g. 0) or the other (e.g. 1) . This use case can be seen in work of the image, speech, and natural language processing communities where researchers determine the type of object (cat or not?), the word being uttered, or the sentiment of a sentence. Researchers often report their hard decision threshold in the form of accuracy, word error rate (WER), F1 score, precision, recall, or BLEU score [Papineni et al., 2002, Graves and Jaitly, 2014, Nakov et al., 2016].

However, there may be scenarios where there are a variety of costs of decisions and priors for different classification thresholds, so a metric that can convey these thresholds becomes more useful. One such metric that captures all possible classification thresholds of the predicted probabilities is the AUC score. For example, if we are trying to predict whether 5 subjects have cognitive impairment, our model may generate the following predicted probabilities $[.48, .50, .30, .75, .63]$. To calculate the AUC, we first sort the predicted probabilities $[.30, .48, .50, .63, .75]$, and then set a classification threshold at each predicted probability. Next we determine the TPR and FPR for that threshold. An example of this computation, as well as the values of the true class, predicted probabilities, thresholds, and (mis-)classification rates are displayed in Table 2.2.

We then plot the operating points (A-F) to map the Receiver Operating Characteristic Curve (ROC) [DeLong et al., 1988]. The area under the ROC becomes our metric to evaluate how well the model performs over a *range* of classification thresh-

Actual Class	0	1	0	1	1	
Predicted Probability	.30	.48	.50	.63	.75	
Operating Point	A	B	C	D	E	F
Threshold ≥	.30	.48	.50	.63	.75	0.75
TPR	100%	100%	67%	67%	33%	0%
FPR	100%	50%	50%	0%	0%	0%

Table 2.2: **Computing ROC**. The table displays an example of a model that output probabilities of cognitive impairment for 5 subjects. For each threshold, there is a corresponding TPR and FPR which is used to plot the ROC and calculate the AUC (displayed in Figure 2-3).

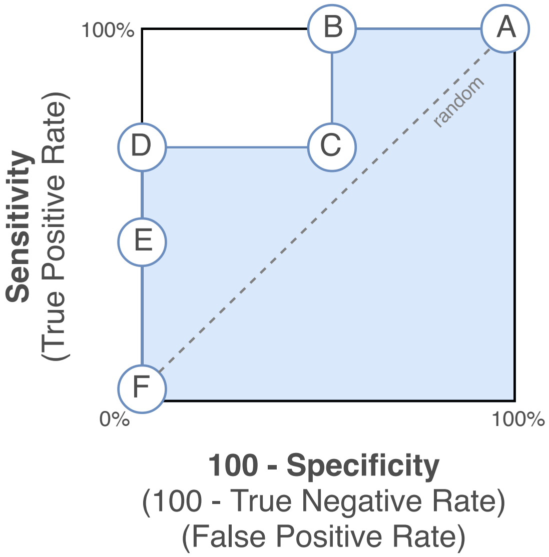


Figure 2-3: **Plotting ROC**. A-F are operating points for each classification threshold in Table 2.2. AUC (shaded in blue) is 0.83. ROC of random predictions is dotted diagonal line with an AUC of 0.5.

olds. Figure 2-3 displays the ROC, and AUC (shaded in blue). The dotted line across the diagonal is the performance of a model that outputs random predictions which is an AUC of 0.5. A perfect model would have an AUC of 1. Our 5 subject example has an AUC of 0.83. This approach to evaluating model performance is particularly useful in scenarios where the penalty and reward for incorrect and correct classification are not equal. Predicting that a subject has cognitive impairment when they indeed have the condition, is not the same as incorrectly predicting that a healthy subject is cognitively impaired. Indeed, not accounting for the different costs of true and false positives may yield some of the negative outcomes highlighted with poor sensitivity and specificity in the previous section.

2.6.3 Model Calibration

For a model to be deployed it must meet some reliability criteria, which we term as *model calibration*. One standard method to calculate model calibration is the Hosmer-Lemeshow test (HL-test) [Hosmer and Lemesbow, 1980]. The HL-test evaluates whether a model that provides a prediction is truly correct in proportion to the prediction value, that is, if a subject is predicted to have a 30% probability of cognitive impairment then the model should be correct 30% of the time. This allows decision-makers (such as clinicians) to scale their confidence and behavior accordingly. If a model satisfies the HL-test (or any other model calibration criteria) then we declare it to be well-calibrated. Figure 2-4 displays two plots of an imaginary model where one is well-calibrated (left) and the other is not (right). We can observe that the predicted and true probabilities scale in proportion to fit the 'ideal' line. The plot on the right may be output by a strong classifier, but the probabilties it generates are not reliable.

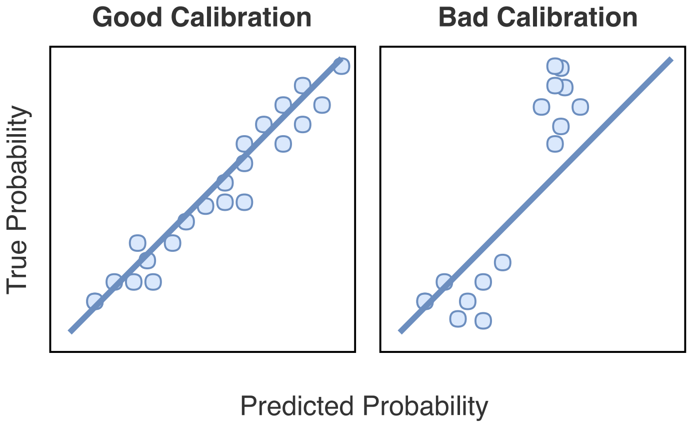


Figure 2-4: **Calibration**. Plot on the left displays predicted probabilities that tightly follow a linear trend, indicative of a model with good calibration. The plot on the right shows two clustering of probabilities which may result in strong classification performance, but is a model with poor calibration because the points do not evenly fit a linear trend.

Hosmer-Lemeshow Test

The HL-test takes the information on the true and predicted class outcomes to determine how well the model predictions match the true labels. Specifically, the HL-test is calculated by determining the sum of squared errors (SSE) between the number of observed and sum of predicted probabilities, equivalent to pearson's chi-squared test statistic:

$$\chi^2 = \sum_{i=1}^{N} \Big(\frac{X_i - \mu_i}{\sigma_i}\Big)^2 \tag{2.3}$$

Utilizing the pearson's chi-squared statistic (and by extension the HL-test) assumes that that the samples (predicted probabilities) are independent and normally distributed. With this assumption a chi-squared distribution arises from the SSE operation.

What differentiates the HL-test from the standard pearson's chi-squared statistic is the introduction of an initial step of sorting and allocating predictions to bins (typically 10), a transform pursued due to the nature of logistic regression (and final sigmoid function of a neural network); where the model outputs probabilities while the observations are annotated with hard class labels (i.e 1 or 0). Formally the HL-test is defined as:

$$H = \sum_{i=1}^{N} \frac{(O_{1i} - E_{1i})^2}{E_{1i}} + \frac{(O_{0i} - E_{0i})^2}{E_{0i}} \tag{2.4}$$

where N is the number of bins, '0' and '1' denote binary class outcomes, O_{1i} is the count of subjects observed to be in class '1' (i.e. cognitive impairment) and the i^{th} bin, and $E_{1i} = \sum p_i$ is the sum of the expected (i.e. predicted probability p) of subjects in the i^{th} bin. The summation in Equation 2.4 results in a single scalar which is used to determine the statistical significance of the differences between observed and expected values, from a chi-squared distribution with 8 degrees of freedom (10 bins - 2 class outcomes). If the resulting p-value is *greater* than 0.05, then the model predictions scale to the true class labels, which we define to be well-calibrated.

2.7 Tools of the Trade

A blacksmith is as good as the tools they choose to use, the same is true for Computer Science research. The work in this thesis utilizes a number of software tools, and in most cases open-source software, which are listed below according to their function. For feature extraction we use `OpenSmile` [Eyben et al., 2010], and `Kaldi` [Povey et al., 2011]. For data processing and statistical modeling we use `Matlab` [MATLAB, 2010], `glmnet` [Friedman et al., 2009], `Python`, and Python libraries `scikit-learn` [Pedregosa et al., 2011], `scipy` [Jones et al., 01], `numpy` [Oliphant, 06], `pandas` [pan, 2012], `kald-io`, `keras` [Chollet, 2015], `tensorflow` [Abadi et al., 2015], and `Pytorch` [Paszke et al., 2017]. For speech recognition training we use `Kaldi` [Povey et al., 2011], and for language modeling we use `SRILM` [Stolcke et al., 2002]. For visualization we use Python libraries `librosa` [McFee et al., 2015], and `matplotlib` [Hunter, 2007]. For virtual environments we use `virtualenv` and `conda` [Cohen-Boulakia et al., 2017]. For quick coding on a cloud environment we use `repl.it`. For distributed computing we use the `slurm` resource management utility [Yoo et al., 2003], and computer networks maintained by the `CSAIL Spoken Language Systems` group. For GPU training we used `Nvidia Titan X (Pascal)`, `GeForce GTX TITAN X`, and `GeForce GTX 1080`. And finally, for debugging we rely on information shared by the communities on `Stack Exchange`, `github.com`, and other online forums.

Chapter 3

A Motivating Study

Synopsis

In this chapter we analyze a subset of the Framingham Heart Study data: 92 subjects with audio and manually generated text transcriptions. We utilize 265 features: 230 audio, 21 text, and 14 demographic. The audio features measure vocal energy and prosody (e.g. pitch), the text features capture attributes of language (e.g. hesitations), and the demographic features (e.g. age), capture several well-known correlates of cognitive impairment. The features are used to train a logistic regression model, with cognitive impairment as the outcome. To reduce the number of features (and enhance generalizability), the model is regularized using Elastic-net. We compare performance of our model with a logistic regression that uses only the 14 demographic features, training using the data from 6,258 subjects in the greater study cohort. Our model performs better than the demographic baseline (0.92 AUC vs. 0.79 AUC) and is well-calibrated (HL-test > 0.05). The features selected by our model indicate that decreased pitch and jitter, shorter segments of speech, and responses phrased as questions are positively associated with cognitive impairment.[1]

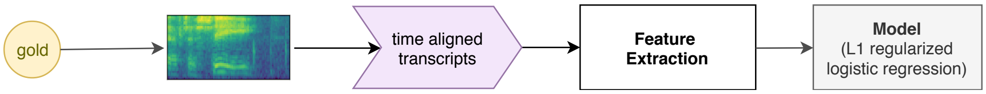


Figure 3-1: **Overview of first study**. The study utilizes gold subjects (*gold*) with time aligned transcripts (*time aligned transcripts*) to extract subjects' audio and text features (*Feature Extraction*) and model cognitive impairment (*Model*).

[1] Work in this chapter was previously published in [Alhanai et al., 2017].

3.1 Background

In our review of prior work (Section 2.2.2), we found that many studies use a large number of features and a relatively small number number of subjects (e.g. 450+ features from 284 subjects) [Yancheva et al., 2015]. The ratio of these two factors (number of features and number of subjects) have important implications for the degrees of freedom and fit of a model. To ensure generalizability, a high example-to-feature ratio needs to be maintained [Friedman, 1997, Hua et al., 2004].

There are several strategies to ensure the example-to-feature ratio remains high. The simplest strategy is to manually select a subset of the features. Another strategy is to select features based on their uni-variate correlation with the outcome [Yancheva et al., 2015]. Alternative feature selection strategies include: forward-feature selection (which iteratively evaluates one feature at a time, and then selects the most predictive set of features for modeling) [Abe, 2010], reducing the dimensionality of the data with mathematical transforms (e.g. principal component analysis [PCA] [Wold et al., 1987], or factor analysis [Child, 1990]) [Fraser et al., 2016], and model regularization [Ng, 2004].

In this chapter we use model regularization because it unites feature selection and model optimization under a single algorithmic framework. Furthermore, model regularization helps reduce over-fitting by eliminating co-linear features, and maintaining only the most informative ones.

3.2 Hypothesis

We hypothesize that it is possible to model cognitive impairment in 92 subjects with a high-dimensional set of audio and text features.

3.3 Objectives

We model cognitive impairment by (1) utilizing a set of audio and text features, (2) comparing model performance with a baseline demographic model from the larger cohort, and (3) evaluating the association between these features and the outcome.

3.4 Method

3.4.1 Data

We use 92 audio recordings of neuropsychological examinations that had available text transcripts. Recordings are on average, 65 minutes in duration, contain 2,496 words, with a vocabulary size of 527 words. We also include 6,258 subjects from the larger cohort (i.e. those that have no missing data) with the same set of demographic variables as the 92 subjects with audio and text transcripts. The audio recordings contain 21 subjects (22.8%) with cognitive impairment.

3.4.2 Features

We extract a total of 265 features of three types: 14 demographic, 230 acoustic, and 21 text. All continuous features are mean-variance normalized, and all categorical features are dummy coded.

Demographic Features

Demographic features contain the subject's age, sex, highest level of self-reported education (didn't graduate high-school, high-school graduate, attended but didn't graduate college, or college graduate or higher), and occupation (part-time, full-time, not working due to disability, retired, unemployed, never worked, volunteer, full-time student, or other). Age is modeled as a continuous feature, while all other features are dummy coded to represent categories, resulting in 14 features total.

Acoustic Features

Acoustic features are extracted using the openSMILE v2.1 toolkit over 20ms frames, shifted 10ms, on audio files that are downsampled to 8kHz [Eyben et al., 2010].

Features contain information on the subject's pitch, probability of voicing, root-mean-square (RMS) energy, Mel-frequency cepstral coefficients (MFCCs), harmonic-to-noise ratio (HNR), zero-crossing rate, shimmer, and jitter, as well as the difference between features in neighboring frames. This results in 46 frame-level features, the details of their calculation are described in [Eyben, 2015].

We are interested in capturing high-level statistics from frames that are most likely to be speech (as opposed to non-speech frames), and further process the features to generate the mean, maximum, minimum, median, and standard deviation. This results in 230 global features that describe each subject's speech characteristics over the entire exam. The specific steps to compute the global features from the frame-level features are as follows:

1. **Feature Normalization:** To remove information of the recording environment, all features except probability of voicing, pitch, shimmer, and jitter (where the absolute number matters) are mean-variance normalized within each subject.
2. **Speaker Turn:** Using the speaker turn labels and timestamps in the transcript, we process the frames that belong to the subject only, and not the tester.
3. **Normalization:** We mean-variance normalize the probability of voicing per subject. (This is only used to extract speech segments, the unnormalized probability of voicing is used to calculate the global feature.)
4. **Smoothing:** We calculate an envelope over the mean-variance normalized probability of voicing by a peak (upper) envelope using a spline over local maxima separated by at least 10 points (100ms).
5. **Speech Segments:** We label frames as speech if the threshold of the smoothed probability of voicing is greater than 0.1 standard deviations from the (zero) mean.
6. **Global Features:** We calculate the mean, maximum, minimum, median, and standard deviation of these frame-level features (from the speech segments) for each subject.

7. **Non-zero Pitch:** For the pitch, shimmer, and jitter, we calculate the same global features, but for non-zero values (i.e. presence of pitch activity) within the speech segments.

Text Features

Text features contain the subject's number of words, duration of speech, speaking rate, questions, hesitations, vocabulary, and language perplexity (how well can the subject's next word be predicted). A total of 21 features are generated, as follows:

- **Number of Words** (5 features): The number of words are calculated for each turn the subject spoke, then the mean, minimum, maximum, median, and sum are calculated across all segments for each subject.
- **Duration** (4 features): Duration is calculated for each turn the subject spoke, then the mean, minimum, maximum, and median are calculated across all segments for each subject.
- **Speaking Rate** (4 features): Words-per-minute (WPM) is calculated for each turn the subject spoke, then the mean, minimum, maximum, and median are calculated across all segments for each subject.
- **Questions** (2 features): A count of the question mark symbol '?' for each turn the subject spoke is calculated. Then the mean and cumulative sum across all segments are taken, for each subject.
- **Hesitation** (2 features): A count of the transcription tag <um> for each turn the subject spoke is calculated. Then the mean and cumulative sum across all segments are taken, for each subject.
- **Vocabulary** (1 feature): The number of unique words expressed per subject during the exam is calculated.
- **Out-of-Vocabulary (OOV) Rate** (1 feature): For each subject $s = i$, we identify the set of unique words $V_{s=i}$, and a list of unique words spoken by all

other subjects ($V_{s \neq i}$). The OOV is computed as:

$$\text{OOV} = \frac{|V_{s=i}^{C} \cap V_{s \neq i}|}{|V_{s=i}|}$$

where V^C denotes the complement of V.

- **Language Perplexity (PPL)**: (2 features) For each subject $s = i$, a tri-gram model with Kneser-Ney discounting is trained on all other subjects. Using the trained model, the language perplexity is evaluated on each subject and all other subjects $s \neq i$ [Jurafsky and Martin, 2014].

  $$\text{PPL} = 2^H \text{ and } H = -\frac{1}{M} \sum_{m=1}^{M} \log p(w_m)$$

  where M is the size of the training vocabulary ($M = |V_{s \neq i}|$) of all subjects s that aren't the i^{th} subject, and w_m is the m^{th} word in the vocabulary. The SRILM Toolkit is used for this calculation [Stolcke et al., 2002].

3.4.3 Model Choice and Evaluation Metrics

Given the limited number of subjects with audio and text transcripts, and the importance of model interpretability, logistic regression is chosen as our modeling framework. We will describe each of the generated models in detail, below.

To assess the performance of our models, the evaluation metrics we report are the AUC, as well as the performance of the models at various points on the ROC: Accuracy, FPR, and TPR. We also perform the HL-test to evaluate model calibration [Hosmer and Lemesbow, 1980]. To evaluate the generalizability and robustness of our modeling techniques, we perform leave-one-out cross-validation. We also report the values of the model coefficients (β), and where applicable: odds ratio, confidence interval at 95%, and the statistical significance of the features via the Wald Test.

Baseline Models

A baseline model using the 14 demographic features is trained on 6,258 subjects who had undergone neuropsychological examinations. Many subjects underwent multiple examinations over the years, which resulted in 12,258 training examples. We refer to

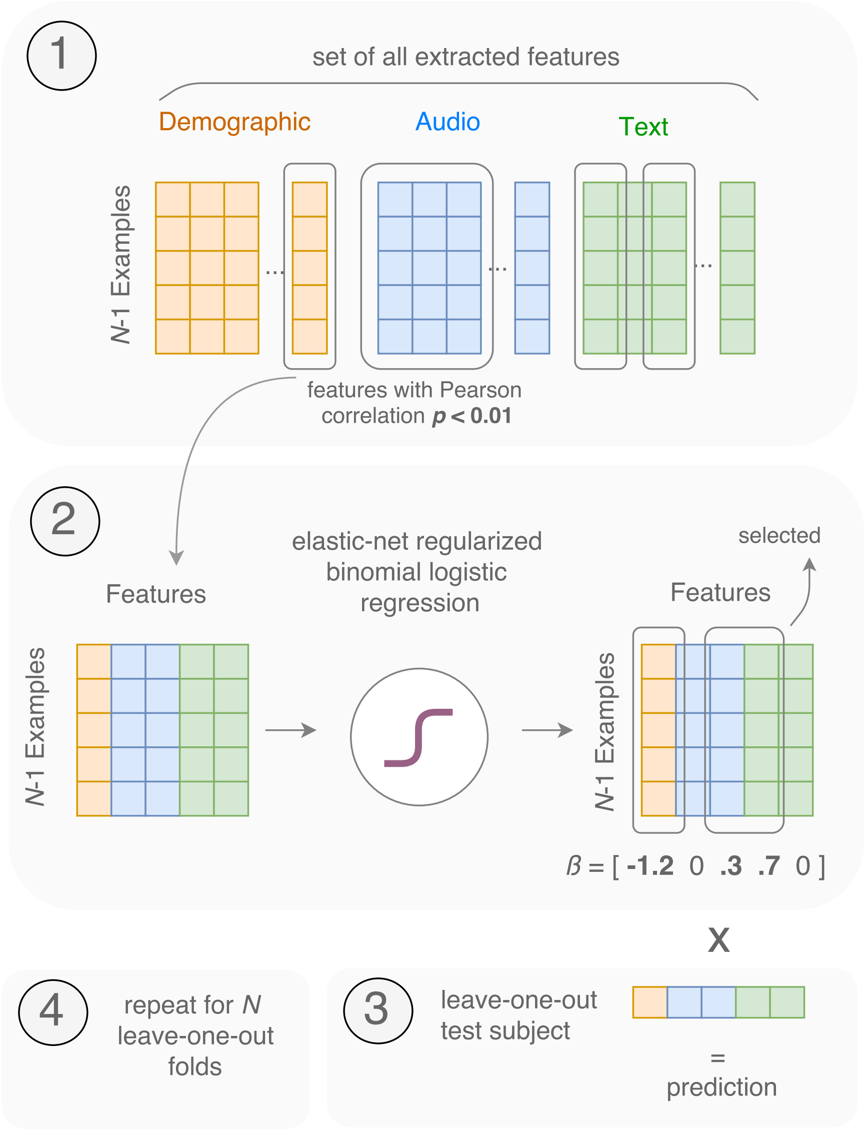

1
set of all extracted features
Demographic
Audio
Text
N-1 Examples
...
features with Pearson correlation p < 0.01
2
elastic-net regularized binomial logistic regression
Features
N-1 Examples
selected
Features
N-1 Examples
β = [-1.2 0 .3 .7 0]
X
4
repeat for N leave-one-out folds
3
leave-one-out test subject
=
prediction

Figure 3-2: **Modeling Technique and Feature Selection**. (1) An initial set of features are selected if they had a statistically significant ($p < 0.01$) uni-variate Pearson correlation with the (training set, i.e. $N-1$ subjects) outcome. (2) An Elastic-net regularized binomial logistic regression is trained with these features, resulting in further feature selection. Features with non-zero model coefficients (β) are selected. (3) This model is the evaluated on the held out test subject. (4) This training method is repeated for all leave-one-out (N) folds.

this as the *global* demographic model.

A second baseline model using only demographic features is trained using data from the 92 subjects with audio and text transcriptions. This model uses 9 of our 14 features (5 features are excluded due to zero variance). We refer to this as the *local* demographic model.

Proposed Model

In Figure 3-2 we show a visual representation of the proposed modeling pipeline, which we also describe here. While there are enough degrees of freedom to fit the local demographic model (9 features and 92 examples), our dataset of 92 subjects is too small to accommodate the addition of the 251 text and acoustic features. We address the large feature to sample ratio in two steps: (1) an initial subset of features are selected that have statistically significant ($p < 0.01$) uni-variate Pearson correlations with the training set outcomes in each fold, (2) we then provide the subset of correlated features to an Elastic-net regularized binomial logistic regression model. In addition to the standard approach of minimizing the difference between predicted and true outcomes (i.e. deviance for binomial logistic regression), Elastic-net also minimizes the linear combination of $L1$ and $L2$ penalties of the estimated coefficients, which has the effect of producing sparse model coefficients thus implicitly selecting features via the resulting non-zero model coefficients. The objective function for an Elastic-net regularized binomial logistic regression model is defined as:

$$\min_{\beta} \left[DEV(\hat{\beta}) + \lambda \Big((1-\alpha)||\beta||_2^2 + \alpha||\beta||_1 \Big) \right] \tag{3.1}$$

where β are the model coefficients being estimated, $DEV(\hat{\beta})$ is the standard binomial logistic regression model objective function, λ scales the influence of the regularization term, while α ranges between 0 and 1, and allows for a combination of $L1$ and $L2$ penalties. Thus when $\alpha = 0$, only the $L2$ penalty forms the regularization term, and when $\alpha = 1$, only the $L1$ penalty term appears. A more complete discussion of Elastic-net is described in [Zou and Hastie, 2005].

We use the MATLAB implementation of the Elastic-net method, GLMNET [Hastie and Qian, 2014], cross-validated over the training set, with the following parameters, $\alpha \in [0, 1]$ with a step-size of 0.01, the cross-validated training *loss* set to 'deviance', and the *number-of-folds* over the training set equal to the number of subjects (N - 1 = 91 examples). Up to 100 λ values are automatically explored by GLMNET until a minimum deviance error threshold is reached. For our evaluation, λ that is 1 standard error from the mean minimum error value across the cross-validated (training) folds is used. All test performance values reported are according to the top performing model on the training set.

3.5 Results

3.5.1 Demographic Model Coefficients

Table 3.1 displays the demographic model coefficients (β), odds ratio (e^{β}), 95% confidence interval and p-values for the *global* and *local* demographic models. For the *global* model, as expected, age is positively associated with cognitive impairment ($p <$ 0.001). Similarly, as is also-well documented [Elias et al., 2000], an increasing level of education with at least some years of college (relative to some years of high school) is negatively associated with cognitive impairment ($p < 0.001$). Employment status of retired, unemployed, and unemployed due to disability (relative to full-time employment) are positively associated with cognitive impairment ($p < 0.05$). High school degree, other categories of employment, and sex do not have a statistically significant association with outcome.

In the *local* model, age is again positively associated with cognitive impairment ($p < 0.05$), while a high school degree and some college (relative to some years of high school) are negatively associated with cognitive impairment ($p < 0.05$). A college degree, employment status, and sex do not have a statistically significant association with the outcome.

Global Demographic Model *(N = 6,258)*

Features	β	**Odds Ratio** (e^{β})	**95% CI**	p**-val**
Age	2.02	7.55	[6.47, 8.81]	< 0.001
Education (w.r.t some high school)				
high school	-0.16	0.85	[0.68, 1.07]	0.16
some college	-0.48	0.62	[0.48, 0.80]	< 0.001
college	-0.59	0.55	[0.43, 0.72]	< 0.001
Employment (w.r.t full-time)				
part-time	0.07	1.08	[0.43, 2.68]	0.87
retired	1.42	4.16	[2.00, 8.64]	< 0.001
unemployed	1.66	5.27	[2.54, 10.94]	< 0.001
disability	1.68	5.34	[1.02, 28.13]	< 0.05
never	0.88	2.41	[0.63, 9.25]	0.20
volunteer	-0.21	0.81	[0.23, 2.82]	0.74
student	-94.47	0.00	[0.00, -]	1.00
homemaker	-94.47	0.00	[0.00, -]	1.00
other	0.64	1.90	[0.21, 17.65]	0.57
Sex (w.r.t male)				
female	-0.05	0.95	[0.81, 1.12]	0.57

Local Demographic Model *(N = 92)*

Features	β	**Odds Ratio** (e^{β})	**95% CI**	p**-val**
Age	1.43	4.20	[1.18, 14.87]	< 0.05
Education (w.r.t some high school)				
high school	-3.44	0.03	[0.00, 0.45]	< 0.05
some college	-4.30	0.01	[0.00, 0.28]	< 0.01
college	-2.20	0.11	[0.01, 1.24]	0.08
Employment (w.r.t full-time)				
part-time	-100.4	0.00	[0.00, -]	1.00
retired	0.40	1.49	[0.11, 20.13]	0.76
unemployed	-100.6	0.00	[0.00, -]	1.00
volunteer	2.31	10.05	[0.19, 531.93]	0.25
Sex (w.r.t male)				
female	-0.01	0.99	[0.25, 4.03]	0.99

Table 3.1: **Model Coefficients.** Logistic regression model coefficients using *global* and *local* demographic features. Note that age is mean-variance normalized, hence, a 1 unit increase in age is not a one year increase but a 1 standard deviation increase (15 years).

Logistic Regression (baseline)

Features	**AUC**	**Acc.** (%)	**TPR @ FPR 10%**	**TPR @ FPR 5%**	**TPR @ FPR 0%**
Trivial - No Impairment	0.50	77	0	0	0
Demographic - Local	0.74	72	0.56	0.44	0
Demographic - Global	0.79	83	0.14	0.14	0.14

Elastic-net†

Features	**AUC**	**Acc.** (%)	**TPR @ FPR 10%**	**TPR @ FPR 5%**	**TPR @ FPR 0%**
Text	0.69	67	0.38	0.24	0.14
Dem. + Text	0.73	74	0.33	0.24	0.10
Audio	0.90	84	0.71	0.48	0.14
Dem. + Audio	0.90	84	0.67	0.48	0.14
Audio + Text	0.92	89	0.76	0.62	0.29
Dem. + Text + Audio	0.92	89	0.76	0.62	0.38

Table 3.2: **Results**. The table displays the performance of all models and feature sets. Top table displays results of logistic regression model utilizing demographic features sets, while bottom table displays results of Elastic-Net regularized models utilizing various combinations of audio, text, and demographic features sets. Combining audio, text, and demographic features sets yields the overall best performing model.†All Elastic-net models are well-calibrated (HL-test > 0.05).

3.5.2 Speech and Language Features

Table 3.2 shows the performance of the baseline models, as well as the performance of the Elastic-net regularized binomial logistic regression models for different combinations of speech and language features.

For reference, a model that consistently guesses cognitive impairment (*trivial* model) would have an AUC of 0.5 (random) and an accuracy of 77%. Using demographic features from the 92 subject subset (i.e. *local* model), the AUC is 0.74. Exposing the demographic-based model to a larger group of subjects (i.e. *global* model), and evaluating on the 92 subset, increases performance to an AUC of 0.79. This model also has a higher accuracy (83%), and an improved TPR and FPR. Neither of the baseline demographic models are well calibrated according to the HL-test.

For the Elastic-net regularized models, an approach using text features alone does not yield a higher performing system (0.69 AUC) than the baselines. However, using audio based features results in higher performance (0.90 AUC) than the baseline models. Combining audio and text features results in the best performing model (0.92 AUC), while introducing demographic features along with audio, and audio-text combined do not improve performance with respect to AUC. Combining all three feature sets results in the best TPR at FPR of 0%, but also increases FPR at TPR of 95%. All Elastic-net regularized models are well calibrated according to the HL-test (> 0.05).

Further results comparing several modeling frameworks (support vector machine, disciminant analysis, decision tree, and k-NN) are displayed in Figure 3.3, utilizing the best performing features set (audio, text, and demographic).

	AUC	**Acc.** (%)	**TPR @ FPR 10%**	**TPR @ FPR 5%**	**TPR @ FPR 0%**	**HL-test**
SVM						
Linear	0.91	84	0.62	0.52	0.14	> 0.05
Cubic	0.90	85	0.57	0.57	0.19	> 0.05
Quadratic	0.91	85	0.62	0.33	0.10	> 0.05
Medium Gaussian	0.88	86	0.57	0.52	0.14	> 0.05
Coarse Gaussian	0.88	78	0.62	0.52	0.10	> 0.05
Discriminant Analysis						
Linear	0.87	83	0.67	0.67	0.14	< 0.05
Quadratic	0.87	82	0.57	0.52	0.29	< 0.05
Tree						
Simple	0.66	74	0.48	0.00	0.00	< 0.05
Medium	0.69	73	0.29	0.00	0.00	< 0.05
Complex	0.69	73	0.29	0.00	0.00	< 0.05
RUS Boosted	0.77	78	0.38	0.05	0.00	< 0.05
Ensemble Boosted	0.53	77	0.10	0.10	0.05	< 0.05
Ensemble Bagged	0.85	84	0.57	0.43	0.05	< 0.05
k-NN						
Fine	0.71	80	0.00	0.00	0.00	< 0.05
Medium	0.88	86	0.67	0.62	0.10	< 0.05
Cosine	0.86	82	0.52	0.19	0.00	< 0.05
Cubic	0.85	87	0.57	0.52	0.05	< 0.05
Weighted	0.89	89	0.67	0.62	0.19	< 0.05
Elastic-net	0.92	89	0.76	0.76	0.62	> 0.05

Table 3.3: **Full results**. Results of all modeling frameworks using all features that were found to be significant during regularization (demographic, text, and audio). Analysis is leave-on-one-out cross-validation.

3.5.3 Selected Features

In Table 3.4 we show the demographic, audio and text features selected by Elastic-net. The values of the model coefficients shown in the table result from training the Elastic-net on all 92 subjects. We do not report confidence intervals and hypothesis testing because sparse estimators such as Elastic-net are difficult to interpret in the same way as a standard logistic regression model [Dezeure et al., 2015].

Two text features are selected by the model: the mean duration of each subject's turn (segment duration), as well as the number of '?' symbols (question marks) transcribed in the text. The segment duration has a negative association with the outcome (cognitive impairment), while the number of question marks have a positive association with the outcome. From the audio features, pitch based measures; minimum pitch and standard deviation of jitter are selected. Both features have negative associations with the outcome. The rest of the features selected are energy based (MFCCs 2 [sd], 3 [max], 6 [median], 10 [sd], 13 [min]) and include the difference between frames (MFCC 3 [mean, median], 8 [median]). All energy based features have a mix of negative and positive associations with the outcome.

Features		**Type**	β	% **Selected**
MFCC 13	(min)	Audio	0.0043	60
Segment Duration	(mean)	Text	-0.0083	82
MFCC 10	(sd)	Audio	-0.1346	89
MFCC 2	(sd)	Audio	-0.2514	98
MFCC 3 diff.	(mean)	Audio	-0.1526	99
Question Mark	(sum)	Text	0.0171	100
MFCC 6	(median)	Audio	0.1741	100
Pitch	(min)	Audio	-0.2430	100
MFCC 8 diff.	(median)	Audio	0.4187	100
Jitter	(sd)	Audio	-0.5337	100
MFCC 3	(max)	Audio	-0.6168	100
MFCC 3 diff.	(median)	Audio	-0.7620	100

Table 3.4: **Selected Features.** Features selected from demographic, audio, and text feature set using Elastic-net regularization (0.92 AUC, $\alpha = 0.99$, $\lambda =$ 'lambda_1se'). The % selected indicates the proportion of leave-one-out folds the feature was selected.

3.6 Discussion

The superior performance of audio features indicates that there is low level information about the speaker's pitch and its variance (jitter) that is predictive of cognitive impairment. These findings are in line with the literature [Meilán et al., 2014, Horley et al., 2010], where decreasing pitch and variance may indicate less expressive speech. We note that different forms of cognitive impairment (e.g. Alzheimer's and Vascular dementia) and co-occurring conditions (e.g. Parkinson's Disease) may exhibit differing speech pathologies, as observed by [Illes, 1989]. Therefore we interpret our results to be capturing a broad spectrum of cognitive disorders.

Although text features alone are not found to be predictive of the outcome, some are found to be meaningful when combined with audio features. Features that capture hesitation (via counts of the question mark symbol '?'), and shorter time taken to respond to the question (via mean segment duration) are indicative of the subject's struggles in responding fully and/or recalling words, agreeing with observations in the literature [Lopez-de Ipiña et al., 2015]. However, features capturing syntax (vocabulary size, OOV rate) and coherence in speech (perplexity) were not selected, although prior research suggests that early predictors of cognitive impairment (let alone onset) may be observed at the syntactic and semantic level of speech, more so than with acoustics [Taler and Phillips, 2008].

While demographic features perform well (0.79 AUC), they are not necessary when combined with audio and text features as evidenced by their absence from the features selected by the Elastic-net regularization model. This indicates that text and audio features are capturing information that is at least equivalent to information contained in the demographic features: such as age and gender of the speaker (via pitch and energy), and/or are capturing information that is even more predictive of cognitive impairment, overshadowing the information content of demographic features. These results suggest that cognitive impairment can be screened for *without having any information on the subject's demographic profile*, using audio recordings alone, and without the constraint of medical visit logistics, missing medical history, or sparse medical evaluations.

The results of modeling with audio and text features indicate that this source of information allows for models that are not only high-performing, but also well calibrated. Importantly, real-world diagnostic systems require that predictions made by models are well calibrated. That is, if a model assigns a 30% probability of cognitive impairment, it is important that cognitive impairment occur 30% of the time. Well calibrated systems allow clinicians and families to make informed judgments about risk.

3.7 Conclusion

In this study we utilize audio recordings of 92 subjects undergoing neuropsychological examinations at the FHS to model cognitive impairment. We found that combining audio and text features provides the best performance in detecting cognitive impairment (0.92 AUC), and is superior to the baseline approach that uses demographic features from the 6,258 subject cohort (0.79 AUC). Given the high-dimensionality of the feature set (265 total), we use an Elastic-net binomial logistic regression model which consistently selects for 12 features. We found that decreasing pitch, decreasing jitter, shorter speech segment lengths, and an increasing number of questions by the subject are positively associated with cognitive impairment. Our methodology does not explicitly model the structure and components of the neuropsychological exams subjects underwent, which allows for it to generalize to other scenarios such as informal spoken interactions.

Chapter 4

Role Specific Language Modeling for Diarization

Synopsis

While audio is relatively easy to record, it remains a challenge to automatically diarize (*who spoke when?*), decode (*what did they say?*), and assess a subject's cognitive health. This chapter demonstrates a method to determine the cognitive health (impaired or not) of 92 subjects, from audio that is diarized using an automatic speech recognition system trained on TED talks and on the structured language used by testers and subjects. Using leave-one-out cross-validation and logistic regression modeling we show that even with noisily decoded data (81% WER) we can still perform accurate enough diarization (0.02 % confusion rate) to determine the cognitive state of a subject (0.76 AUC)[1].

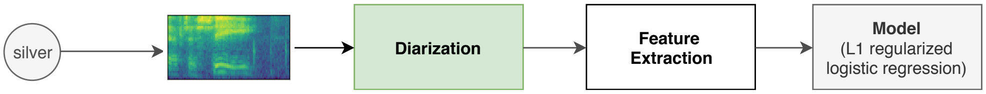


Figure 4-1: **Overview of second study**. This study utilizes silver subjects (*silver*) to automatically diarize audio (*Diarization*), then extract acoustic features of subjects' speech (*Feature Extraction*) to model cognitive impairment (*Model*).

[1] Work in this chapter was published in [Al Hanai et al., 2018a]

4.1 Background

The application of automatic speech processing technologies to medical domains requires a pipeline with multiple stages. Such a system requires audio pre-processing to locate speech and speaker segments (i.e. diarization) [Anguera et al., 2012], the transcription of spoken utterances [Besacier et al., 2014], and a feature representation of the speaker's latent condition to determine disease biomarkers for classification purposes [Cummins et al., 2015].

Research in this domain can be categorized into two areas. First is the utilization of acoustic and linguistic information to perform speaker diarization and verification using standard corpora (e.g. Switchboard, NIST) [Stolcke et al., 2006, Reynolds et al., 2003]. The second category of work seeks to evaluate speech and language biomarkers for the detection of cognitive impairment utilizing measures such as speaking rate, pauses, *n*-grams, and Word Error Rates (WERs) [Pakhomov et al., 2010, Lehr et al., 2012, Fraser et al., 2014, Pakhomov and Hemmy, 2014, Vincze et al., 2016], as well as Automatic Speech Recognition (ASR) for phonetic alignment and acoustic feature extraction [Tóth et al., 2015]. However, systems for diarization and transcription are typically developed using noise-free data from healthy speakers. Such systems do not easily translate to the context of medical speech because subjects may exhibit speech disfluencies and audio recordings are often noisy. Hence existing off-the-shelf speech tools need to be adapted to diarize and transcribe medical speech data [Tóth et al., 2015, Weiner et al., 2016].

Our study differentiates itself from prior work by combining speaker-specific language modeling and ASR for speaker diarization, with the ultimate goal of assessing the cognitive condition of the subjects using the acoustic information contained in the hypothesized (and less than ideal) segments. This is an extension of work presented in the previous chapter that used gold standard speaker segmentations and transcriptions to evaluate cognitive outcomes. Further details on feature selection, modeling, and the relation to previous work in that domain are described in chapter 3.

The approach pursued in this chapter captures real-world scenarios where automatically diarized and transcribed data may not be at human parity, but the usage of less than ideal annotation is necessary for deploying screening technologies at scale. Moreover, audio recordings are often sub-optimal (e.g. placing digital recorders on a desk), which is the case of the data used in this study. Therefore the ability to detect cognitive conditions must accommodate the presence of noisy data.

Before delving into our methodological approach, we provide an overview of ASR and its core components: acoustic model, lexicon, and language model. The ASR system was core to our diarization pipeline.

4.1.1 An Automatic Speech Recognition (ASR) System

An ASR system transcribes speech into text, by taking as input a set of acoustic observations O and translating them to the sequence of underlying words W associated with the speech. Formally, an ASR system is trying to maximize the most likely set of words W given an observation O, from the set of words in the lexicon L.

$$W^* = \arg\max_{W \in L} P(W|A) = \arg\max_{W \in L} P(A, W) \tag{4.1}$$

$P(A, W)$ can be further decomposed into three components of an ASR system.

$$P(A, W) = \sum_{U} P(A|U)P(U|W)P(W) \tag{4.2}$$

The three components are (1) an acoustic model $P(A|U)$ that captures statistics on phonemes (e.g. a neural network that outputs probabilities for a set of phonemes that generated O), (2) a lexicon $P(U|W)$ that contains mappings of words to their underlying phonetic units U, and (3) a language model $P(W)$ that defines the likelihood of a sequence of words. A pictorial representation of these components is shown in Figure 4-2.

The acoustic, lexicon, and language models capture probability distributions on acoustic observations, acoustic units, and words that are represented using a Fi-

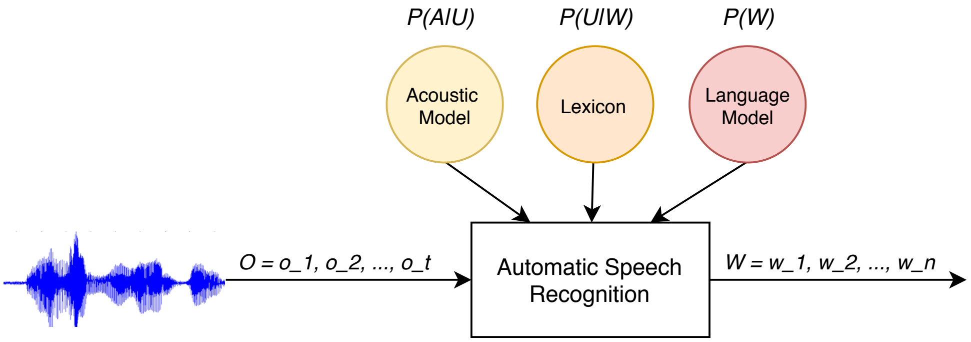


Figure 4-2: **Automatic speech recognition system**. An ASR system takes as input a sequence of observations and outputs a sequence of words. An ASR system is composed of three components: an acoustic model, lexicon, and language model.

nite State Transducer (FST) [Mohri et al., 2002]. The FST is a graph the contains (weighted) mappings from states, to acoustic units, to words. Once all these components have been combined together into a single representation (e.g. FST), the acoustic observations are (historically) decoded using Hidden Markov Models (HMM) and Viterbi Search [Jurafsky and Martin, 2014].

Acoustic Model

The acoustic model $P(A|U)$ generates the probability of a given acoustic observation conditioned on the phonetic unit. Acoustic observations are usually frame-level representations of speech (20 ms window, 10 ms shift) in the form of Mel-frequency cepstral coefficients (MFCCs) or filterbanks. Current methods utilize neural networks to model acoustics, which attempt to map (through a non-linear transform) the acoustic observation to the most likely phonetic unit [Hinton et al., 2012].

Lexicon

The lexicon $P(U|W)$ contains the set of words W that map to phonetic units U (i.e. a text file with two column entries, words and phonetic sequence), and may contain multiple entries for the same word mapping to different sequences of phonetic units (e.g. *cat* maps to $/k/$ $/ae/$ $/t/$). A lexicon can be generated by hand, be composed of each letter in a word (i.e. grapheme, *cat* maps to $/c/$ $/a/$ $/t/$), or can be generated automatically using some set of rules or probabilistic model that maps from word to phonetic units [Bisani and Ney, 2008].

Language Model

The language model $P(W)$ models the probability over word sequences K-long, and is usually generated using some form of N-gram modeling [Dunning, 1994], which assumes that the probability of any given word w_i is a function on the past $N-1$ words, and is formalized as:

$$P(W) = P(w_1, ..., w_K) = \prod_{i=1}^{K} P(w_i | w_{i-(N-1)}, ..., w_{i-1}) \tag{4.3}$$

A common order for N-grams is tri-gram (used in this thesis). Because most possible combination of words in a language do not appear in a text transcript during training (making language effectively infinite [Savitch, 1993]), we need to find a method to avoid over-fitting (i.e. $P(w_i|w_{i-2}w_{i-1}) = 0$ for an unseen tri-gram $w_{i-2}w_{i-1}w_i$). To improve generalizability, a language model needs to account for these unseen word sequences, an operation termed *smoothing*. There are several methods to smooth existing probabilities, an example of which is Kneser-Ney smoothing (an approach utilized in this thesis) [Chen and Goodman, 1999]. A strategy for smoothing is to discount from the probaility mass of the higher-order model, and assign it to the lower order model. Kneser-Ney smoothing extends this strategy by also taking into account how diverse a given lower-order word sequence when it appears in the higher-order model. That is, a word such as `Fransisco` may appear relatively often, but only in the context of the bi-gram `San Fransisco`. Therefore, when we discount from `San Fransisco`, the probability of the uni-gram `Fransisco` should remain relatively rare when following any word other than `San`. This kind of language modeling behavior is what Kneser-Ney's smoothing attempts to account for; words constrained to particular contexts should not be more likely to appear in an absolute sense, but words that make diverse appearances (such as the function word `the`) should maintain a common and diverse appearance.

An important aspect of speech recognition is that the accuracy of audio decoding can more strongly impacted by a language model that does not match the domain of the data being transcribed (e.g. broadcast news data used to transcribe a cooking show). Domain mismatch of the acoustic model also has a negative impact [Bellegarda, 2004].

Evaluating Language Models

Because domain mismatch plays a strong role in speech recognition performance, and language modeling in general, metrics to evaluate how well a language model can predict the next word in a sequence have been developed and are commonly reported. A simple metric is the *out-of-vocabulary (OOV) rate*, which measures the percentage of words W in the target data vocabulary $V(W_{target})$ that does not exist in the source data vocabulary $V(W_{source})$.

$$oov = \frac{V(W_{target}) \notin V(W_{source})}{|V(W_{target})|} \tag{4.4}$$

Another metric is the *perplexity* score, which captures how well the language model predicts the word statistics in the target data, and is equivalent to estimating the effective vocabulary size, or how many times the model can be sampled before correctly finding the next word in the word sequence (sometimes termed as the branching factor). For text data composed of m sentences $S = s_1, s_2, ..., s_n$ the probability of all sentences in the target set is $p(T) = \prod_{k=1}^{m} p(s_k)$, so the higher the value the better the language model is modeling the target data. However, having an estimate at the world level may be more useful; especially for speech recognition which is evaluated at the word-level, rather than sentence-level. We can define the target data to have a total of L words where $N = \sum_{k=1}^{m} n_k$ where n_k is the length of the k^{th} sentence. The average log probability (or entropy) of a given word is defined as:

$$l = \frac{1}{N}\log_2 \prod_{k=1}^{m} p(s_k) = \frac{1}{N}\sum_{k=1}^{m} \log_2 p(s_k) \tag{4.5}$$

In order to report a more intuitive (i.e. linear) score, we calculate perplexity as

2^{-l}. The smaller the perplexity value, the better our model is at predicting words in a sequence in the target data.

4.2 Hypothesis

We hypothesize that it is possible to automate data curation for clinical use by conditioning on speaker roles, because speakers (subject/tester) during neuropsychological exams have different word usage and speaking patterns due to the question and answer nature of the evaluation. We also hypothesize that not all segments of the exam will be equally valuable in evaluating for cognitive conditions, due to potential confusion between speakers when automatically annotating speaker segments, polluting the features used for modeling cognitive conditions.

4.3 Objectives

Given our observations of prior work, we are motivated to (1) develop a method to automatically extract and identify segments of speech that were most likely to belong to the subject, and (2) to evaluate the type of segments that were most predictive of a subject's cognitive condition.

4.4 Method

4.4.1 Data

The data used in this chapter is the same as that of Chapter 3. Briefly, the data is composed of 92 audio recordings of neuropsychological examinations that have text transcripts. Recordings are on average, 65 minutes in duration, contain 2,496 words, with a vocabulary size of 527 words.

4.4.2 Outcome of Interest

Our overarching goal is to determine whether the subject being evaluated is cognitively impaired, but we also need to determine who spoke when (subject or tester). To this end, we model two levels of outcomes. Our first outcome of interest is a binary indicator of the speaker type (subject or tester), with the subject coded as 1. Our second outcome of interest is a binary indicator of cognitive impairment, with impairment coded as 1.

4.4.3 Model Choice and Evaluation Metrics

To evaluate speaker diarization we use the Diarization Error Rate (DER) metric, which combines statistics on the percentage of speech classified as non-speech (Miss - E_{miss}), the percentage of non-speech classified as speech (False Alarm - E_{FA}), and the percentage of speech mis-classified as belonging to the other speaker (Confusion Rate - E_{conf}) [Tranter and Reynolds, 2006]. We use a time-based diarization approach, ignoring segments less than 250ms in duration. Formally,

$$DER = E_{miss} + E_{FA} + E_{conf} \tag{4.6}$$

To evaluate the performance of the ASR system we use the Word Error Rate (WER) metric [Wang et al., 2003], which combines information on the rate of deletion (Del), insertion (Ins), and confusion (Sub) between the reference and hypothesized transcription. Formally,

$$WER = \frac{N_{ref} - Sub - Del - Ins}{N_{ref}} \tag{4.7}$$

where N_{ref} the total number of words in the reference transcript.

Given the importance of model interpretability for detecting spoken language biomarkers, logistic regression is chosen as our modeling framework. The evaluation metric we use for detecting cognitive impairment is the AUC which has the advantage of evaluating model performance across the whole range of probability cutoffs, rather than a single point estimate such as accuracy or F1 score [Huang and Ling, 2005]. To assess the generalizability and robustness of our modeling techniques, we perform leave-one-out cross-validation.

4.4.4 Experimental Pipeline

Figure 4-3 displays an overview of the experimental pipeline. We perform three sets of experiments: (1) determining speaker role from the text transcripts, (2a) diarizing audio with an ASR system that utilizes a role-specific language model and oracle segmentation of audio, (2b) diarizing audio with automatic segmentation, and (3) determining cognitive impairment based on the subject's segments of the diarized audio.

4.5 Experiment 1: Speaker ID from Text

4.5.1 Language Model

We first investigate the language patterns of speakers to determine whether a subject or tester was speaking (i.e. a binary class problem). We start with the segmentation from the speaker turns labeled in the transcripts. Next, we train a tri-gram language model with Kneser-Ney smoothing for each speaker type. The two language models are then used to generate the language perplexity of the spoken (text) segment. The training and testing is performed with leave-one-out cross-validation (i.e. 92 folds, one fold for each of the 92 subject-tester interactions). Six features are used in the logistic regression model to generate the predicted probability that a subject is speaking (code as 1):

- **OOV-rate** (2 features): The out-of-vocabulary rate of the subjects' and testers' vocabulary (from their respective training sets).

- **Perplexity** (2 features): The language model perplexity for the subjects and testers.

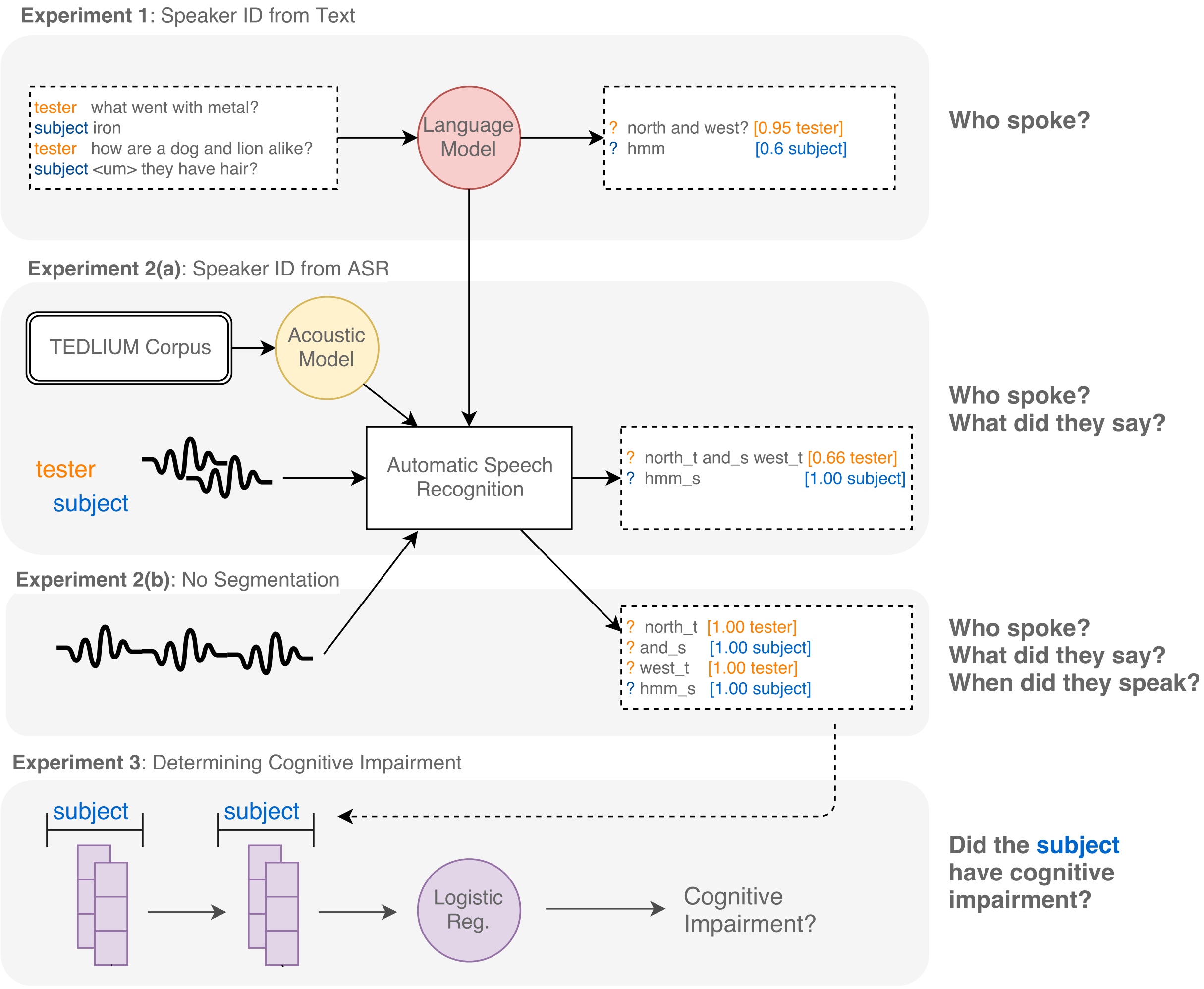

Experiment 1: Speaker ID from Text
tester what went with metal?
subject iron
tester how are a dog and lion alike?
subject <um> they have hair?
Language Model
? north and west? [0.95 tester]
? hmm [0.6 subject]
Who spoke?
Experiment 2(a): Speaker ID from ASR
TEDLIUM Corpus
Acoustic Model
tester
subject
Automatic Speech Recognition
? north_t and_s west_t [0.66 tester]
? hmm_s [1.00 subject]
Who spoke?
What did they say?
Experiment 2(b): No Segmentation
? north_t [1.00 tester]
? and_s [1.00 subject]
? west_t [1.00 tester]
? hmm_s [1.00 subject]
Who spoke?
What did they say?
When did they speak?
Experiment 3: Determining Cognitive Impairment
subject
subject
Logistic Reg.
Cognitive Impairment?
Did the subject have cognitive impairment?

Figure 4-3: **Experimental Pipeline**. We perform three sets of experiments: (1) determining speaker role from the text transcripts, (2a) diarizing audio with an ASR system that utilizes a role-specific language model (with oracle segmentation of audio), (2b) diarization with automatic segmentation of audio , and (3) determining cognitive impairment based on the subject's segments of the diarized audio.

- **Perplexity sans** <s> (2 features): The language model perplexity for the subjects and testers, excluding the start and end of sentence tags (<s>,</s>).

4.5.2 Results

This results in a classification accuracy of 84% (±0.06), and an AUC of 0.93 (±0.07).

4.5.3 Discussion

Our results show that language usage between the subject and tester differ significantly, and that each speaker's language style is consistent across recordings (i.e. subjects consistently spoke like other subjects, and testers consistently spoke like other testers). Therefore, with the availability of highly accurate transcriptions of the same structure (neuropsychological exams), a highly accurate text-based speaker diarization can be conducted.

4.6 Experiment 2: Speaker ID from ASR

For this experiment, we decode the audio using an ASR system, with a language model trained on each speaker (subject/tester), and an acoustic model trained on the TEDLIUM corpus. Each component of the ASR system is developed as follows:

4.6.1 Acoustic Model

The TEDLIUM corpus contains over 1,400 audio recordings and text transcription of TED talks, for a total of 120 hours of data and 1.7M words [Rousseau et al., 2012]. Using this corpus, we train the acoustic model as a feedforward Neural Network (6 layers x 2048 hidden units) with the Minimum Bayes Risk (MBR) criterion using 40 Mel filterbank features, via the Kaldi speech recognition toolkit using the 's5' TEDLIUM recipe [Povey et al., 2011, Rousseau et al., 2012].

4.6.2 Language Model

A tri-gram language model is trained for the speaker and tester using the SRILM toolkit [Stolcke et al., 2002]. In practice, each word in the transcripts is denoted by a special character $\_s$ or $\_t$ for each of the subject and tester. A single language model is then trained, which learns language patterns across two sets of speaker roles. This approach allows for a single FST to be built for ASR decoding, and which represents

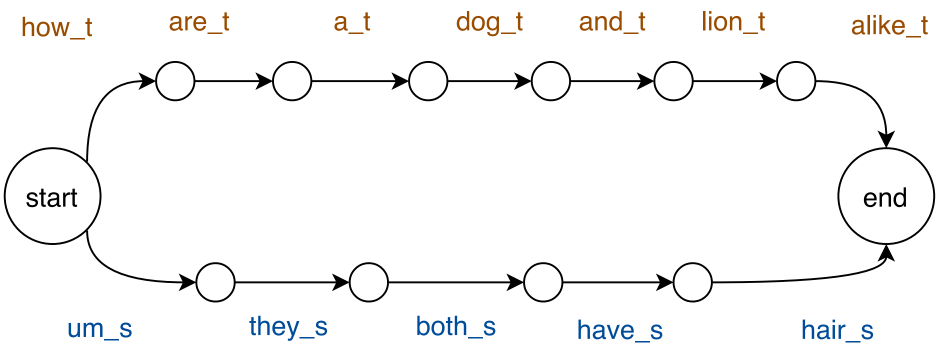


Figure 4-4: **Example Role-based FST**. Top path marks the tester's _*t* language path. Bottom path marks the subject's _*s* language path. Language patterns for each speaker role can be represented within a single search space.

the search space for both speaker roles. A visual example of the language model and FST is displayed in Figure 4-4.

4.6.3 Lexicon

We generate the word pronunciations using the LOGIOS lexical tool[2]. This results in a lexicon where each word maps to a phonetic sequence.

4.6.4 Decoding Audio

Once we've built the ASR components, we decode the audio in three ways:

- **Oracle**: A language model is trained across all 92 transcripts, with utterances segmented according to the reference transcript.

- Exp. 2(a) **Leave-one-out**: A language model is trained on all transcripts *excluding* the transcript of the audio being decoded. Utterances are segmented according to to the reference transcripts. This approach is represented in Figure 4-3 part 2(a).

- Exp. 2(b) **Leave-one-out + automatic segmentation**: A language model is trained on all transcripts *excluding* the transcript of the audio being decoded.

[2] http://www.speech.cs.cmu.edu/tools/lextool.html

Utterances are *not* segmented by speaker turn, rather, the audio is decoded in full. This approach is represented in Figure 4-3 part 2(b).

4.6.5 Results

The results of our decoding are displayed in Table 4.1. Our Oracle system performs with a WER of 66.7%, while decoding without language modeling information (of the test audio being decoded) results in a WER of 68.6%. This relatively small difference in performance (68.6% vs. 66.7%) indicates that the language usage across the audio recordings was consistent.

We also compare the DER across the different setups (Table 4.1). This helps us evaluate how well a speaker can be identified given various levels of information about the underlying segments being decoded.

All Segments

	Oracle	**Loocv**	**Loocv auto seg.**
WER (%)	66.7	68.6	81.3
DER (%)	35.8 (± 5.9)	37.2 (± 5.5)	40.5 (± 05.4)
Miss	00.2 (± 0.4)	00.2 (± 0.4)	00.2 (± 15.3)
False Alarm	03.9 (± 1.2)	04.1 (± 1.3)	03.7 (± 01.3)
Confusion	31.7 (± 5.9)	32.9 (± 5.5)	36.7 (± 05.3)
Cognitive ID			
AUC	0.72	0.70	0.68

Optimum Segments

		95% subj. & 10+ words	**Top 9 longest**
DER (%)	-	98.2 (± 1.6)	99.9 (± 0.2)
Miss	-	97.9 (± 2.1)	99.9 (± 0.4)
False Alarm	-	00.0 (± 0.0)	00.0 (± 0.0)
Confusion	-	00.3 (± 0.6)	0.02 (± 0.2)
Cognitive ID			
AUC	-	0.75	0.76

Table 4.1: **Results of ASR (Exp. 2(a)-(b)), Speaker ID (Exp. 2(a)-(b)), and Cognitive ID (Exp. 3).** The top table displays ASR, diarization, and cognitive modeling when utilizing the full audio. The bottom table displays diarization and cognitive modeling when utilizing a subset of the segments.

4.6.6 Discussion

This set of experiments validates the observation from the previous experiment on language usage patterns across speaker roles (i.e. subjects consistently spoke like other subjects, testers consistently spoke like other testers, and subjects and testers did not speak like each other). Also, transcriptions with high WERs (between 66.7% and 81.3%) still contain information that is robust enough for further usage in diarization and modeling of cognitive impairment.

4.7 Experiment 3: Cognitive ID

Using the classified speaker segments, we are interested in determining the subject's cognitive condition (impaired or not). Our modeling approach (feature extraction, model, and evaluation metrics) is the same as Chapter 3.

4.7.1 Transcript-based Speaker Turn Segmentations

For the experimental setup that uses segmentations by speaker turn from the reference transcript (Exp. 2(a) in Section 4.6.4), we are interested in evaluating how well we can model cognitive impairment using segmentations of various quality (e.g. different lengths and speaker purity). To this end, we explore the hyperparameter space for segmentation quality and perform a grid-search of segmentation types along two dimensions: (1) by length (equivalent to the total number of words decoded), and (2) by purity (equivalent to the percentage of words decoded that were hypothesized to belong to the subject). The results of evaluating cognitive impairment in this search space can be viewed in Figure 4-5. The highest AUC (of 0.75) is found when modeling with segments that have been decoded with at least 10 words, and 95% of which are hypothesized to belong to the subject (Table 4.1).

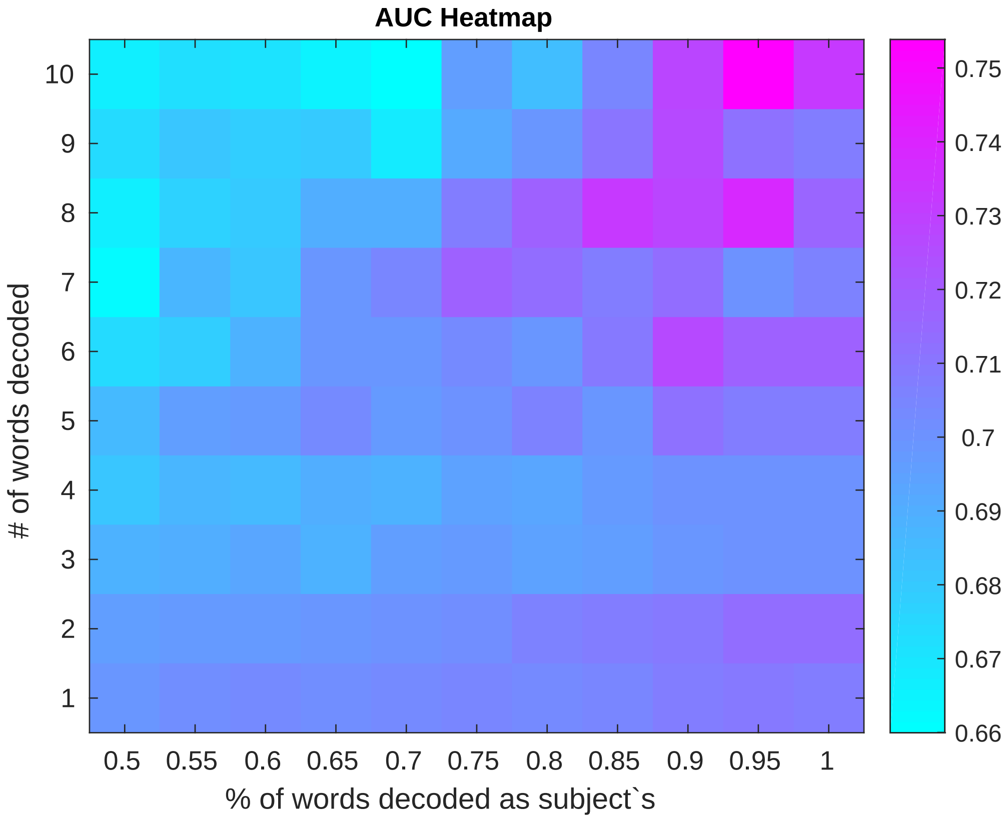


Figure 4-5: **Heatmap for Modeling Cognitive Impairment for Different Segment Length and Speaker Purity**. Heatmap of AUC for cognitive impairment model across two thresholds: (y-axis) segment *length* (i.e. number of words decoded), and (x-axis) segment *purity* (i.e. percentage of words in a segment classified as the subject's).

4.7.2 Discarding Transcript-based Speaker Segmentations

For the experimental setup that is decoded without oracle speaker turn segmentation (Exp. 2(b) in Section 4.6.4), we first segment the decoded hypothesis along silences that are longer than 1.5 seconds. This is followed by further segmentation according to the hypothesized speaker. For example the following hypothesized transcript:

> Hypothesis Transcript: `how_t <sil-1.6s> are_t a_t dog_t and_t cat_t alike_t um_s they_s both_s have_s hair_s`

We first split according to silences (which are defined to be longer than 1.5 seconds). The result of this processing yields the following segmentation which eliminates:

Segment 1: `how_t`

Segment 2: `are_t a_t dog_t and_t cat_t alike_t um_s they_s both_s have_s hair_s`

Finally, we segment along transitions between speakers, which results in:

Segment 1: `how_t`

Segment 2: `are_t a_t dog_t and_t cat_t alike_t`

Segment 3: `um_s they_s both_s have_s hair_s`

Now we have segmented audio according to speaker, and can label each segment as belonging to either the tester or subject. So segments 1 and 2 belong to the tester, and segment 3 belongs to the subject. For each word and silence token in our hypothesis transcript there exists a timestamp to mark the start and end of the word/token(s). We clip the audio with this information to perform feature extraction and modeling with the audio segment of interest.

Since segments may be of different lengths, we explore utilizing the longest segments. This is informed by our previous experiment (Section 4.7.1) where longer segments (i.e. segments with more words) are found to be the most robust for modeling cognitive impairment. For modeling, we select the N longest segments that are hypothesized to be the subject's. We evaluate N from 1 to 15. 99% of hypothesized segments are under 25 seconds in duration, and as a pre-processing step we discard the longest 1% of hypothesized segments, which are many minutes long and several standard deviations beyond the mean (i.e. spurious decodings). The highest AUC (of 0.76) is found when modeling the 9 longest segments hypothesized as the subject's (Table 4.1). The length of these segments are on average 150 seconds (±20 sec) of audio per subject, or 7% of a subject's total audio duration.

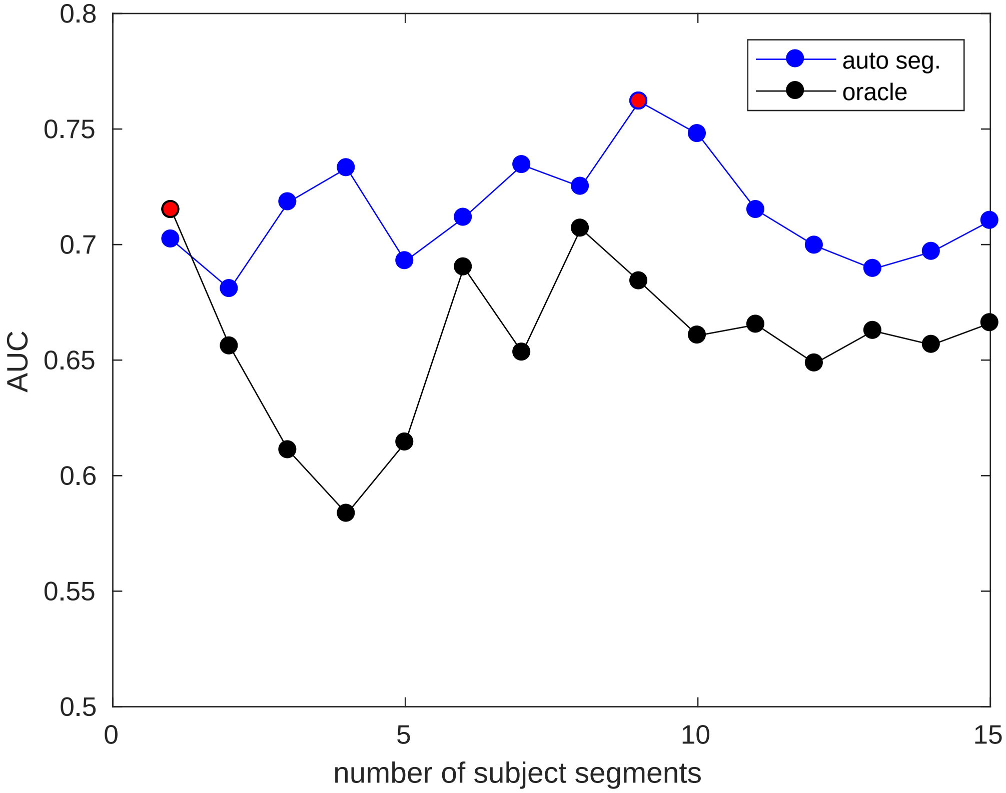


Figure 4-6: **Cognitive ID by Number of Segments**. Plot of AUC (y-axis) with respect to the number of segments per-subject (by descending order of length) used for modeling their cognitive impairment (x-axis). Red points indicate best performance (AUC 0.72 and 0.75 for oracle and automatic segmentation systems respectively).

4.7.3 Discussion

This experiment shows that it is possible to perform modeling of cognitive impairment utilizing automatically segmented subject speaker turns, with performance that is on par with the oracle speaker segmentation. Furthermore, our experiments find that 9 segments is sufficient for robust evaluation. As shown in Figure 4-6, we also find that not all diarization is equal, nor are all segment lengths equally powerful at modeling subjects' cognitive state. In the case where no oracle segmentation is available, and automatic segmentation is utilized, longer segments contain information that is more discriminative (AUC 0.68 vs. 0.76). For the oracle system, the longest segment is the both the most and equally predictive of cognitive impairment, when compared to all segments taken together. This highlights that tests that elicit longer responses allow for more robust diarization. Moreover, longer segments evaluate cognitive performance that is (via speech) most strongly associated with the outcome. Finally, longer spoken segments provide more opportunity to capture patterns associated with cognitive impairment.

Furthermore, the modeling paradigm we have explored is robust enough that neither the underlying neuropsychological test need be explicitly modeled [Lehr et al., 2012], nor do the features utilized require word or phone alignments (alignments which require accurate transcriptions in order to generate) [Tóth et al., 2015].

4.8 Conclusion

In this chapter we showed how structured spoken interactions such as a neuropsychological exams can be diarized using the word patterns of speakers, and how the diarized segments contain enough information to model cognitive impairment. We found that even with noisily transcribed data (81% WER) we can diarize audio with minimal confusion between speakers (4% confusion rate). Furthermore, we are able to determine a subject's cognitive condition using only a fraction of their recorded spoken interaction (about 150 seconds from an hour long recording).

Chapter 5

Population-level Modeling

Synopsis

The efficacy of spoken language biomarkers has yet to be evaluated at the level of a representative population. In this chapter we perform population-level modeling of cognitive impairment (4,500+ subjects, 3.49% impairment). Subjects' cognitive impairment is modeled with an L1 regularized logistic regression model that utilizes 676 prosodic and energy based speech statistics. Age, gender, education level, exam battery type, and presence of APOE e4 are accounted for as confounders in the model. The model has an AUC of 0.91, and selects for 252 features. In general, increased and monotonous speech activity in higher frequencies is associated with cognitive impairment. Further analyses shows that specific audio features implicitly model speaking rate, and that from all tests, the 'verbal fluency' test is most predictive of cognitive impairment.

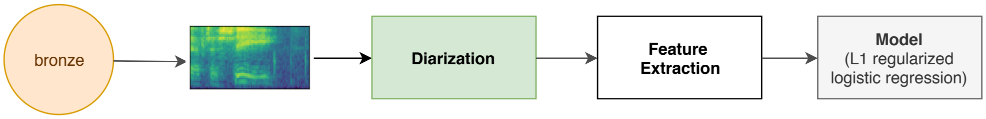


Figure 5-1: **Overview of third study**. This study utilizes the bronze dataset (*bronze*) automatically diarized to extract subjects' speech segments (*Diarization*), followed by acoustic feature extraction (*Feature Extraction*) for modeling cognitive impairment (*Model*).

5.1 Background

As we expand upon below, prior work has evaluated spoken language biomarkers of cognitive impairment but have been limited by (1) small and skewed samples, (2) content-dependant feature extraction methods, and (3) context-dependant modeling.

5.1.1 Small and Skewed Samples

The samples of many studies are limited in size (range of N = 22 to 242 [Ripich et al., 1991, Orimaye et al., 2014]) and are rarely representative of the patient population in question [Fraser et al., 2016]. To facilitate statistical power, small datasets necessarily balance the proportions of cognitively impaired and healthy controls. Such data do not reflect the real-world prevalence of disease (e.g. 3% to 32% prevalence for Alzheimer's Disease [Hebert et al., 2001]), and may therefore impact the generalizability of models that use them.

5.1.2 Content-dependent features

Prior work is further limited by its dependence on accurate audio transcriptions [Orimaye et al., 2014]. With few exceptions [König et al., 2015, Al-Hameed et al., 2016], researchers have used language *content* to capture differences in syntax [Fraser et al., 2016], density of ideas [Roark et al., 2011], word category [Shibata et al., 2016], and dialogue across cognitive conditions [Atay et al., 2015]. When audio is incorporated into studies, it is usually *aligned* with the language content to extract features including duration [Slegers et al., 2018], speaking rate [Fraser et al., 2013], and phonetic-based statistics [Tóth et al., 2015]. While content-dependent models have shown great promise for the detection of cognitive impairment, they require meticulous labeling of: speaker turns, word alignments, and phonetic units. These labels are typically produced manually because automated systems do not yet perform at human parity [Xiong et al., 2017], and only a few systems have been developed to process speakers with speech disfluencies [Liu et al., 2006]. The requirement of human audio transcription makes content-dependent models difficult to scale.

5.1.3 Context-dependent Modeling

Finally, prior work use data collected from a limited number of highly specific tests (e.g. cookie-theft [König et al., 2015], animal fluency [Becker et al., 1994]), allowing for explicit modeling of speech context (e.g. the word 'cookie' was recalled within the test [Fraser et al., 2016]) [Hernandez-Dominguez et al., 2018]. However, it is not clear how models with such strict contextual assumptions can be used for discrete monitoring of cognitive impairment "in the wild" [Snowdon et al., 1996, Le et al., 2011].

5.2 Hypothesis

In Chapter 3 we were able to model cognitive impairment in a small subset of subjects (92 in total) which had been diarized to extract speaker segments. We hypothesize that diarization and modeling is possible on a much larger set of subjects.

5.3 Objectives

Given the limitations of prior work, we are motivated to perform modeling of cognitive outcomes with the following three objectives: (1) to analyze a large number of subjects that represent the real-world prevalence of healthy/impaired individuals (4,836 subjects of which 3.49% are cognitively impaired), (2) to evaluate spoken language features extracted from audio recordings independent of the language content of speech, and (3) to utilize recordings of neuropsychological exams that contain a comprehensive number of tests (37 tests across all recordings), and without explicitly modeling the underlying context. Hence we are able to perform population-level modeling, with purely audio-based features, and conduct analysis that clarifies the association between speech, content, and context.

5.4 Methods

5.4.1 Data

Our study uses 6,705 audio recordings from 4,836 unique subjects who were undergoing neuropsychological evaluations. An audio recording has an average length of 64 minutes. The average age of a subject is 64 years, 55.2% are female. 224 recordings (3.49%) contain subjects that have cognitive impairment.

5.4.2 Spoken Language Features

Speaker Segmentation

Audio recordings are processed to extract segments belonging to the subject (i.e. tester audio was removed). Full details are described in Chapter 4, with a visual representation of the process displayed in Figure 5-2. In brief, an ASR system (i.e. speech-to-text) is used to timestamp when each speaker (tester or subject) is most likely to be speaking. Specifically, a tester tends to utter specific prompts, while subjects are likely to respond with certain phrases that are different to what testers are likely to utter. The ASR system learns to differentiate between speaker 'roles' with a language model trained on a manually transcribed subset of the data (of 92 subjects, 1.9% of the total data), along with an acoustic model trained on an external corpus of over 750 TED talks (Figure 5-2(a)) [Rousseau et al., 2012].

Feature Extraction

Following audio segmentation by speaker turn, 51 spoken language frame-level features are extracted at 10 ms intervals over a 20 ms window, utilizing the opensmile and kaldi toolkits [Eyben et al., 2010, Povey et al., 2011]. The features are composed of prosodic and energy information. The prosodic features are: probability of voicing, harmonics-to-noise ratio (HNR), pitch (F0), shimmer, jitter, zero-crossing rate (ZCR), length of the segment, and number of frames with speech activity and pitch activity. The energy features are: spectral energy (40 Mel filterbanks), and root-mean-square

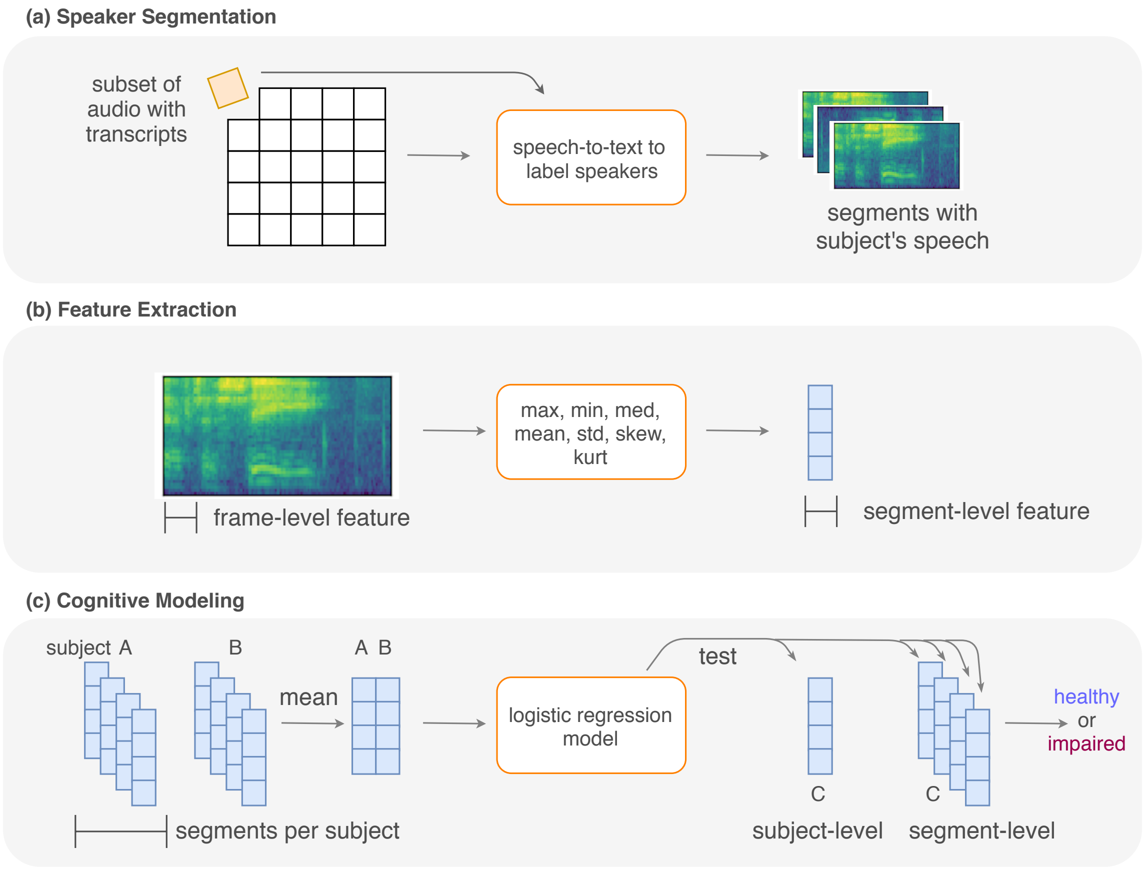


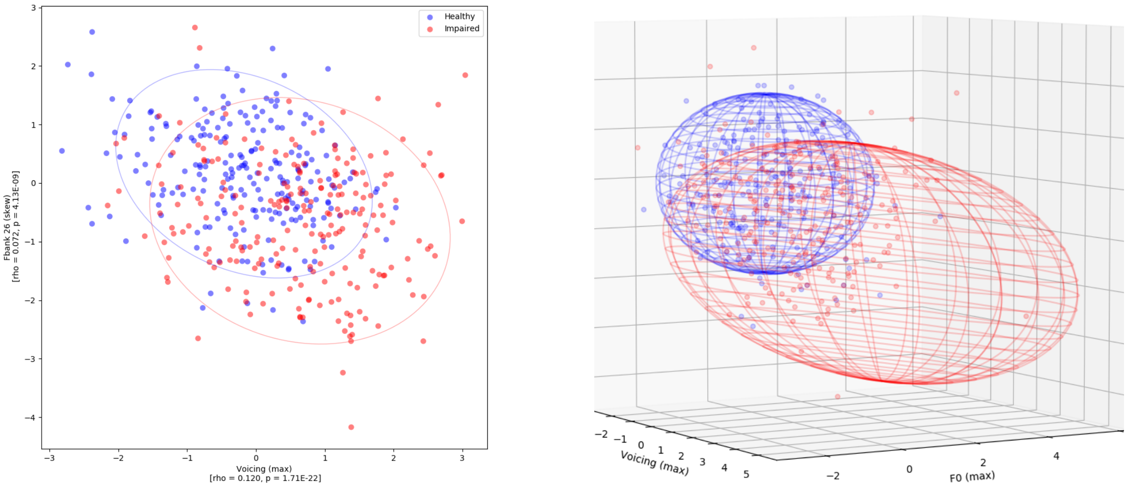
Figure 5-2: **Method**. The figure illustrates the three steps required for modeling cognitive impairment. (a) We first perform diarization to extract segments belonging to the subject. (b) Next we generate higher-order acoustic features to represent the subject's speech patterns. (c) We model cognitive impairment (coded as 1) with leave-one-subject-out cross-validation.

energy (RMS). The frame-level difference for all features are also generated (i.e. delta; jitter[x=1] - jitter[x=0]). Next, high-order statistics are calculated (mean, median, maximum, minimum, standard deviation, skewness, and kurtosis) across all 20 ms frames in a segment (Figure 5-2(b)).

Cognitive Modeling

For each subject, 11 segments of 25 second duration or less are selected after performing a grid-search (on the leave-one-subject-out training set) for the optimal number of segments (1 to 15) and duration (1 to 30 seconds) to utilize for detecting cognitive impairment. Finally, the average of all these segment-level statistics are used to represent the subject in an audio recording, yielding 676 features. All features are zero-mean and variance normalized when input to the model (Figure 5-2(c)).

In Figure 5-3 we plot visually how features may separate between class outcomes in 2D, and how an additional dimension (i.e. 3D) improves separability. This serves as an illustrative motivator for utilizing high-dimensional representations of speech activity.



(a) 2D plot

(b) 3D plot

Figure 5-3: **Class Separation**: The figure displays a 2D and 3D plot of subject predictor values colored according to the outcome (blue is healthy, red is impaired). (Left - 2D plot) The features on the x-axis and y-axis are the maximum voicing probability and skew of the 26th filterbank, respectively. (Right - 3D plot) The features on the three axes are the maximum voicing probability, skew of the 26th filterbank, and maximum pitch (F0). The 2D plot displays some separation between the two classes, while the 3D displays more pronounced separation between the two classes. This motivates the use of speaker representations of higher-dimensions to maximize separability.

5.4.3 Confounders

To account for the influence of an individual's physiology, their predisposition to developing cognitive impairment, and examination factors, we included age, sex, highest education level, presence of the genetic marker apolipoprotein E (APOE) e4, and exam battery type as 'demographic' features in our model. These factors have been shown to have associations with cognitive impairment (age, sex, education) [Satizabal et al., 2016], as well as specific pathologies such as Alzheimer's (APOE) [Haan et al., 1999]. Age is coded as a continuous variable and zero-mean variance normalized. Sex is coded as a binary variable with 1 representing female. Education is dummy coded into 4 categories (some high-school, high-school, some college, college) with some high-school set as the reference category. APOE is binary coded with the presence of e4 represented as 1. The exam battery is of two types, long and short, with the long one composed of additional tests (approximately 10 more) relative to the short [Downer et al., 2015]. We code the long exam as 1.

5.4.4 Model

Given the relatively high feature to example ratio (682 to 6,705) and the propensity for highly co-linear features, an L1 regularized logistic regression model is used to perform implicit feature selection [Tibshirani, 1996]. The model coefficients of a regularized model may be used to interpret the strength and polarity of the association between features and outcome, although not as odds ratios due to the penalty imposed by regularization. The L1 model is selected based on prior evaluations of several modeling paradigms (e.g. support vector machine, decision trees, etc.) and regularization penalties (e.g. elastic-net, L2 regularization, etc.) performed on a subset of the data reported in Chapter 4, and displayed in Tables 3.2 and 3.3. We use a standard logistic regression model when feature selection is not the goal.

To assess the performance of our models, the evaluation metrics we use are the AUC, sensitivity and specificity at Youden's Index (the probability threshold at which sensitivity and specificity are jointly maximized) [Youden, 1950]. To evaluate the generalizability and robustness of our modeling techniques, we perform leave-one-subject-out cross-validation [Kearns and Ron, 1999]. The results we report are at the subject-level.

5.5 Results

Table 5.1 displays the results of modeling cognitive impairment utilizing three types of feature sets: (1) demographic only, (2) audio features only, and (3) demographic and audio features combined. The demographics-only model uses a standard logistic regression model, while the other two feature sets use an L1 regularized logistic regression model. The best performing model utilizes both demographic and audio features (AUC 0.91, sensitivity 89%, specificity 81%).

Features	**Model**	**AUC**	**Sens.**	**Spec.**	**HL-test**
demographic	standard	0.84	99%	15%	< 0.05
audio	L1	$0.87^\dagger$	83%	79%	< 0.05
audio + demographic	L1	$0.91^\dagger$	89%	81%	< 0.05

Table 5.1: **Results**. Audio feature set performs better than demographic feature set. Best model combines both audio and demographic feature sets. †Results have statistically significant improvement ($p < 0.001$) via t-test in detecting cognitive impairment.

For the best performing model, 252 features are selected by the (L1 regularized) model across all folds; 202 energy, 44 prosodic, and all 6 demographic features. The 10 features with the largest absolute model coefficient values are shown in Table 5.2, and are composed of 2 demographics, and 8 audio energy features. Features that are positively associated with outcome, in descending order, are battery type, filterbanks 30 (median), 26 (maximum), and 34 (kurtosis). Features that are negatively associated with outcome, in descending order, are filterbank 24 (kurtosis), filterbank 21 difference (standard deviation), education (some college), filterbank 21 (standard deviation), filterbank 9 (maximum), and filterbank 35 (skewness).

5.6 Discussion

Our goal in this study is to investigate the utility of audio for the passive detection of cognitive impairment. Specifically, we seek to understand the utility of audio data when the linguistic content, and speaker context, are unknown.

Feature	**Statistic**	**Type**	**Model Coefficient** (e^β)
Battery Type	/	dem.	3.2
Filterbank 24 1337 - 1535 Hz	kurtosis	energy	-1.8
Filterbank 21 Diff. 1072 - 1245 Hz	standard deviation	energy	-1.6
Filterbank 31 2120 - 2394 Hz	median	energy	1.6
Education (some college)	/	dem.	-1.4
Filterbank 21 1072 - 1245 Hz	standard deviation	energy	-1.4
Filterbank 9 315 - 415 Hz	maximum	energy	-1.4
Filterbank 35 2695 - 3026 Hz	skewness	energy	-1.2
Filterbank 27 1642 - 1870 Hz	maximum	energy	1.2
Filterbank 35 2695 - 3026 Hz	kurtosis	energy	1.2

Table 5.2: **Model coefficients**. Model coefficients of features with 10 largest absolute values in audio + demographic L1 regularized logistic regression model. Features are composed of 2 demographics, and 8 audio energy features.

5.6.1 Including Demographics Improves Model Performance

We observe that the audio-only feature set performs relatively stronger than the demographics-only feature set (from AUC 0.84 to 0.87, $p < 0.001$) which indicates that audio alone contains information that can discriminate between subjects that are healthy and cognitively impaired. We also observe that including demographics with the audio feature set improves model performance (from AUC 0.87 to 0.91, $p < 0.001$). This improvement may be because the introduction of demographics allows the model to formally capture these dimensions of the subject's profile without trying to infer it from the audio, hence the model is free to find other patterns in the audio while conditioning on the sex and age of a subject. Furthermore, the information content in the demographic features may be complementary to the audio features. It should be noted that the battery type demographic feature has a strong and positive association with outcome because subjects who are more likely to develop cognitive impairment (due to age, family history, or existing cognitive condition) undergo the long version of the neuropsychological exam. Furthermore, it has been shown that higher education levels are negatively associated with cognitive impairment [Satizabal et al., 2016].

5.6.2 Interpreting Audio Features

Energy in the different frequency bands of speech (i.e. filterbank features 1 through 40) are modeled with 7 segment-level statistics (minimum, maximum, median, mean, standard deviation, skewness, kurtosis) which capture the following information:

1. maximum captures upper peak speech activity (i.e. loudness).
2. minimum captures lower peak speech activity (i.e. silence).
3. mean captures the overall speech activity.
4. median captures the overall speech activity excluding outliers.
5. standard deviation captures the range of the speech activity.
6. skewness captures whether the speech activity is increasing (low skewness) or decreasing (high skewness) (i.e. symmetry of activity).
7. kurtosis captures how concentrated (high kurtosis) or varied (low kurtosis) the speech activity is (i.e. correlation of activity).

As an aid in understanding the behavior of these statistics, an illustration of the mean, standard deviation, skewness, and kurtosis are displayed in Figure 5-4.

We observe that the most predictive features were capturing a combination of the maximum, median, standard deviation, skewness, and kurtosis of spectral energy in speech. Specifically, when a speaker has increased speech activity that is consistent and monotonous (filterbank 31 [med], filterbank 27 [max], filterbank 35 [skew]) in higher frequencies (1600 - 3000 Hz), then they are more likely to be cognitively impaired. While a speaker with high and varied speech activity (filterbank 21 diff [sd], filterbank 9 [max], filterbank 24 [kurt]) in lower frequencies (300 - 1500 Hz) is more

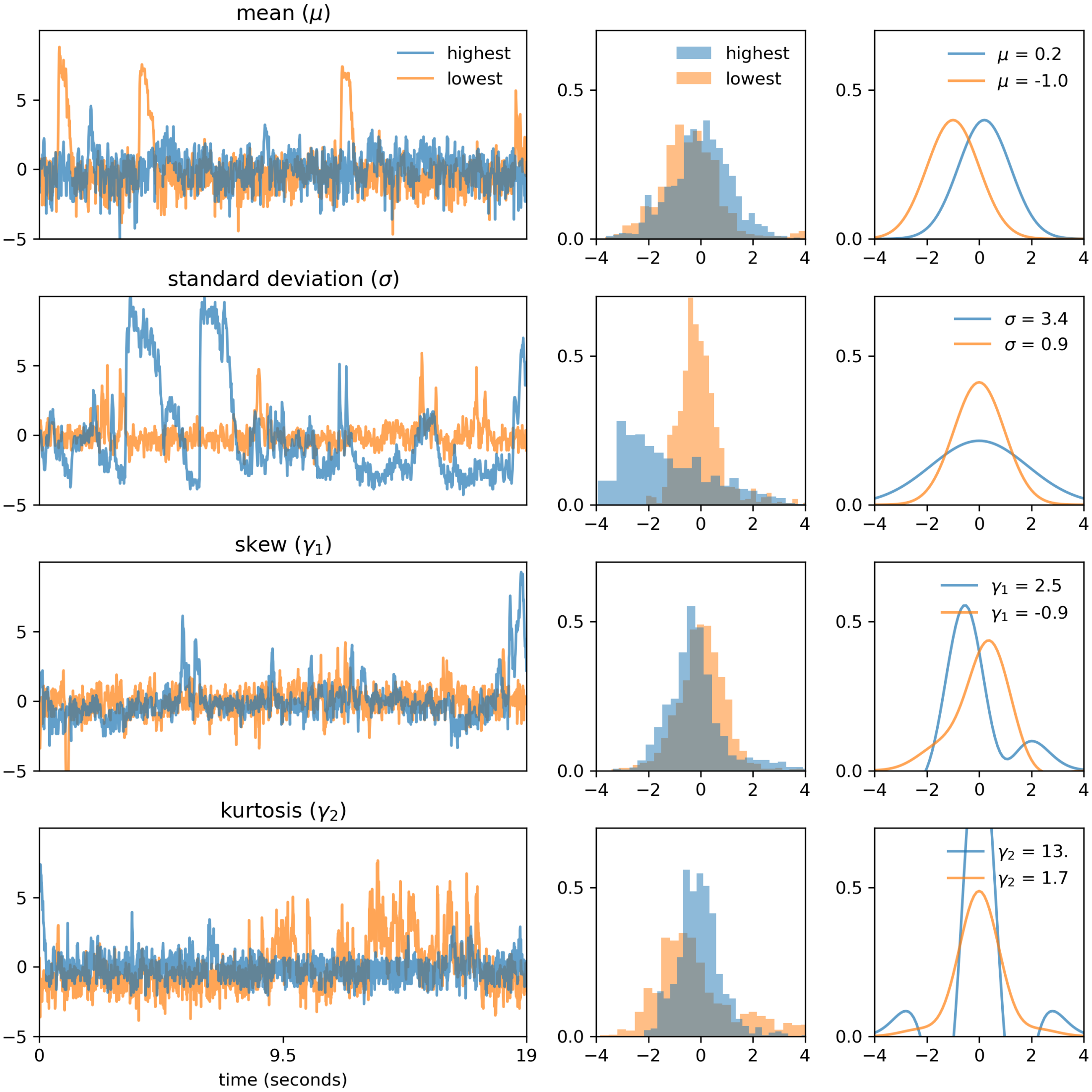


Figure 5-4: **Signal Statistics**. The plots shows raw filterbank signals (column 1) and their distributions (columns 2-3). Each row highlights signals with differing means, standard deviations, skewness, and kurtosis. (Column 1) blue corresponds to the signal with the higher value statistics, and the orange with the lower value. (Column 2) displays the distribution of the signals in column 1. (Column 3) displays the distributions normalized with respect to all statistics except the one under focus (i.e. first row: μ={0.2, 1.0}, ρ=1. second row: μ=0, ρ={0.9, 3.4}. third row: μ=0, ρ=1, γ_1={-0.9, 2.5}. fourth row: μ=0, ρ=1, γ_1=0, γ_2={1.7, 13}).

likely to be healthy. These observations correspond to prior findings reported in the literature of prosodic qualities of speech being impacted by cognitive impairment, and that changes in these qualities are independent of language function (i.e. their grammar may be sound, but quality of voice is impacted). In addition to prosodic information, spectral energy has been found to differentiate between cognitive states, and is uncorrelated with semantic and syntactic impairment but correlated with speaking rate [Al-Hameed et al., 2016]. This indicates that acoustic information may indeed be used for classification, without the need for transcripts and language-based features.

5.6.3 Inferring Speaking Rate

Performing further analysis on the the spoken language features that are most predictive of cognitive impairment (Table 5.2), we find that three are significantly ($p < 0.001$) and positively correlated with speaking rate. These three features corresponded to filterbanks 21 (sd), filterbanks 9 (max) and 27 (max) with correlations of $r = 0.22$, 0.18, and 0.24 respectively. These correlation results indicate that certain statistics extracted from the audio may be used to indirectly model speaking rate. Furthermore, these correlation results indicate that the rest of the highly predictive audio features are capturing other types of speech patterns, possibly differences in articulation (since they had no statistically significant correlation with speaking rate) [Gómez-Vilda et al., 2017]. Our work confirms observations described in the literature, that speaking rate is associated with cognitive impairment [Hoffmann et al., 2010], and may be capturing speech dis-fluencies [Singh et al., 2001, Hoffmann et al., 2010].

5.6.4 Differences in Speech Activity and Articulation

The features we use for modeling are capturing segment-level statistics. In order to infer how speech activity over time is predictive of cognitive impairment, we plot in Figure 5-5 (top) the distribution over time of the frame within the segment that is most representative of the segment-level feature representation (via euclidean dis-

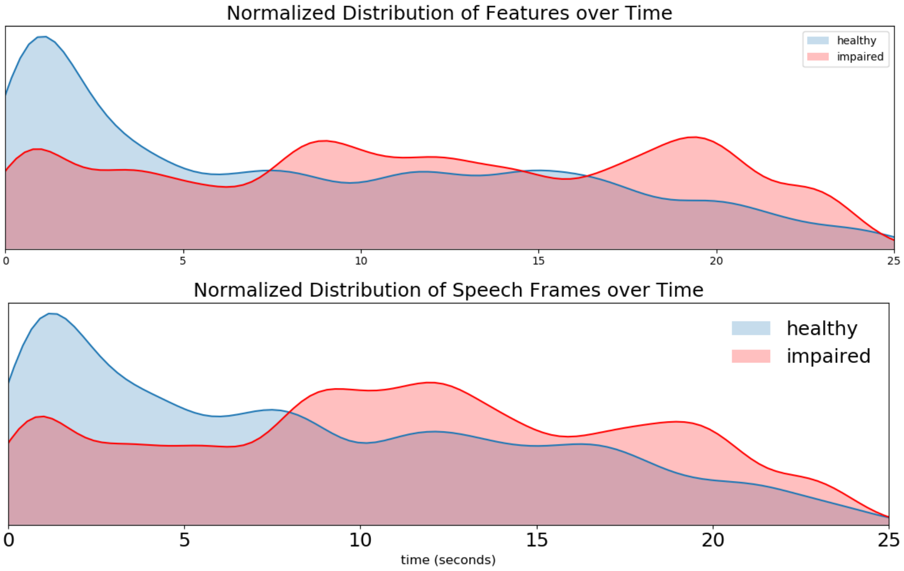


Figure 5-5: **Distribution Over Time of Features and Speech Activity**. (Top) Plot displays the distribution of frame-level *features* most closely matching segment-level feature representation. (Bottom) Plot displays the distribution of *speech activity* over time. Colors according to subject class outcomes, healthy (blue) and cognitive impairment (red).

tance). That is, for each segment, we first calculate the time-point which corresponds to the most similar frame to the segment-level representation via euclidean distance. Next we plot a smoothed histogram with kernal density estimation[1] (where the x-axis corresponds to time).

It can be noticed that the most representative feature frames for healthy subjects occurs at the beginning of the segment, while for impaired subjects it is more likely to occur later in the segment. These differences may be due to the style of speech, where impaired subjects are more likely to exhibit hesitations and delayed responses [Gómez-Vilda et al., 2017], as indicated by the distribution of speech over time in Figure 5-5 (bottom).

5.6.5 'Verbal fluency' Test Most Predictive of Outcome

Our modeling paradigm does not require the formal knowledge of the context that elicits speech, however we present a context-based analysis, which is an established

[1] scalar factor `bw` = 100 in the python seaborn library

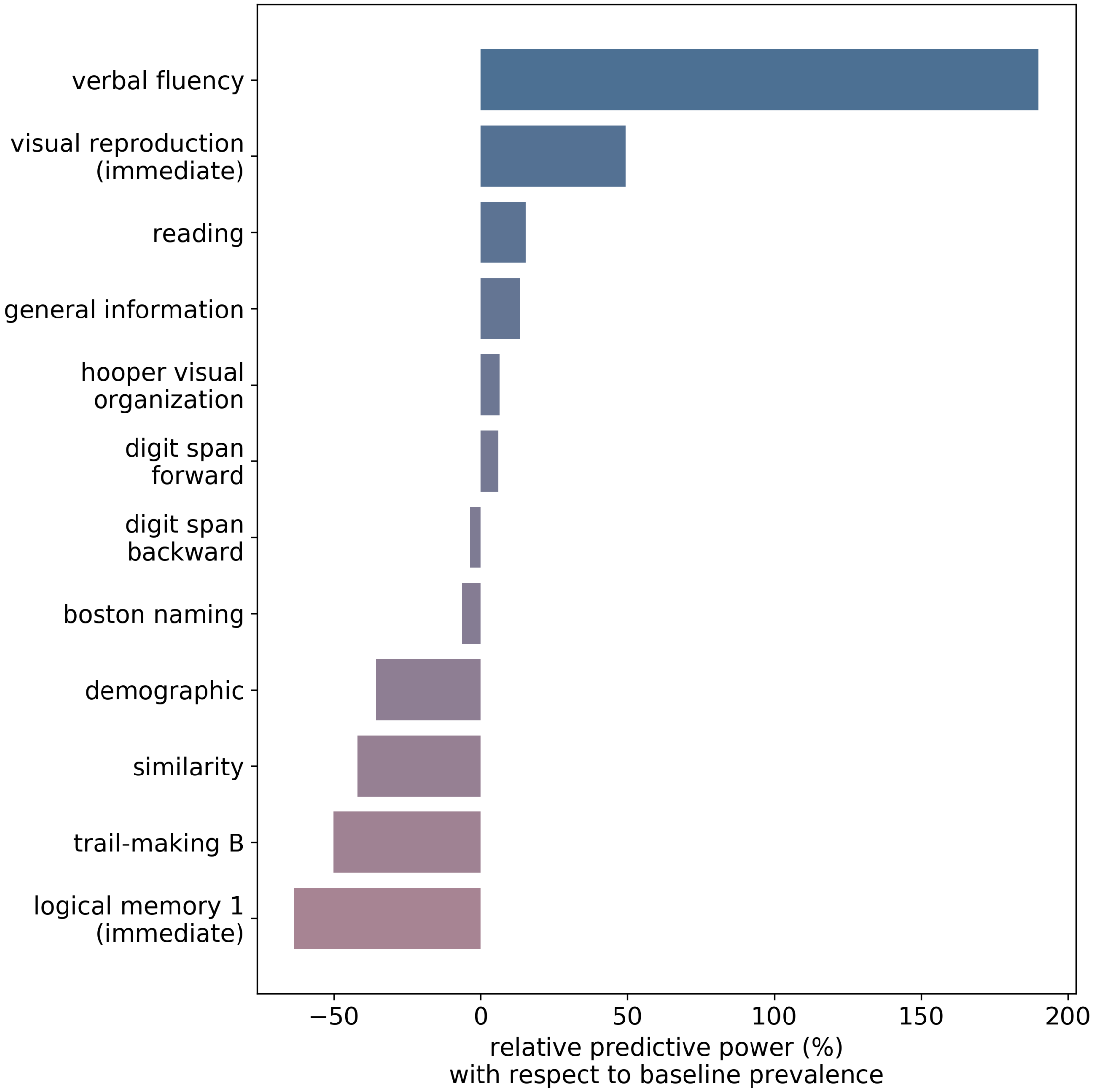


Figure 5-6: **Most Predictive Tests**. The figure displays the relative predictive power (%) of a test with respect to its prevalence in the data. At least 20% of the subjects were modeled with the represented tests. Descriptions of the tests are available in Table 5.3.

approach, and useful for comparison with prior work. We analyze segment-level information, specifically, the type of test that was conducted, and how (the audio patterns expressed within) those segments are most predictive of cognitive impairment. Figure 5-6 displays how predictive each test is (relative to its prevalence in the dataset) in determining the correct outcome. We plot the 12 tests (out of a total of 33 in our data) that exists in at least 20% of the subjects[2]. We observe that the most predictive test in determining the correct outcome of a subject is the 'verbal fluency' test, which asks subjects to recall in 60 seconds as many words as they can that starts with a specific letter 'F', 'A', or 'S'. The predictive power of this test may be because it is evaluating general recall (generate words from the universe of memories without any prior conditioning or exposure to stimulus), but with a simple constraint (words have to start with a particular letter). Other tests that evaluate recall are relatively more constrained (visual reproduction and logical memory; recall an image or detail from a previously shown set of images or narrated story). Our results are supported by prior findings in the literature: the efficacy of the 'verbal fluency' test at differentiating between cognitive states has been explored by [Crossley et al., 1997, Cerhan et al., 2002, Laws et al., 2009], while [Ewers et al., 2012] performed a comprehensive evaluation of neuropsychological tests, and found that verbal fluency had a statistically significant correlation with Alzheimer's disease.

5.7 Conclusion

In this chapter we modeled cognitive impairment from audio recordings of 4,800+ subjects, utilizing audio features alone, and without the need to model linguistic context (i.e. text transcripts) or test structure. Subjects' cognitive impairment was modeled with an L1 regularized logistic regression model that utilized 676 prosodic and energy based speech statistics. Age, gender, education level, exam battery type, and presence of APOE e4 were accounted for as confounders in the model. In general, increased and monotonous speech activity in higher frequencies was associated with cognitive impairment. Further analyses showed that some of the selected audio features correlated with speaking rate, and that from all tests, the 'verbal fluency' test was most predictive of cognitive impairment.

[2] 37 tests were administered, but not all are modeled due to the segment selection criteria (11 segments per subject 25 seconds long).

Test	**Description**
Verbal fluency	Repeat in 60 seconds words that being with a specific letter (‘F’, ‘A’, or ‘S’).
Visual reproduction (immediate)	Reproduce images shown from memory.
Reading	Read words such as ‘efficacious’, ‘precocious’ to measure phone production accuracy.
General information	Answer general knowledge questions: How many weeks in a year? In which continent is Brazil?
Hooper visual organization	Mentally re-organize a jigsaw of an image and recognize the name of the object.
Digit span - forward	Repeat sequence of digits forward (e.g. four-five-seven-three).
Digit span - backward	Repeat sequence of digits forward
Boston naming	Recognize the name of the object in each image.
Demographic	Questions about the subject (not formally part of the exam): How old are you? What is your occupation? Right- or left-handed?
Similarity	Identify the similarity between two words: Orange and apple. Axe and saw. etc.
Trail making B	Connect the dots in the order of numbers and letters, 1 - A - 2 - B ... - 5 - E, etc.

Table 5.3: **Description of Tests in the Neuropsychological Exam**. The tests listed correspond to Figure 5-6.

Chapter 6

Representation Learning with Neural Networks

Synopsis

> In this chapter we demonstrate how a convolution neural network can model cognitive impairment from spoken interactions of 5,000+ subjects undergoing neuropsychological exams. We do away with hand-crafted features, and instead utilize spectral energy information (i.e. filterbank features) to learn higher-order summary representations of speech. Our model exhibits high performance (AUC 0.85), and can be used to infer moments in an utterance that are indicative of cognitive impairment, such as hesitation and speech dis-fluencies in the form of: pause followed by onset of speech, short burst of speech, high frequency speech activity, and silence.

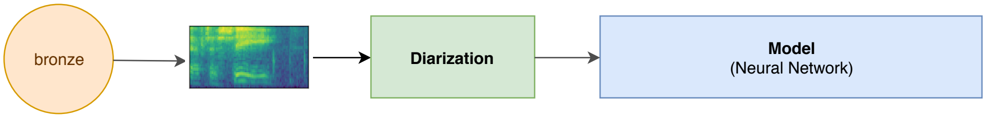


Figure 6-1: **Overview of fourth study**. The study in this chapter utilizes the bronze dataset (*bronze*) which is automatically diarized to extract subjects' segments (*Diarization*). Spectral energy features are extracted and then used to train a neural network to model cognitive impairment (*Model*).

6.1 Background

We showed in Chapter 5 that cognitive impairment can be modeled utilizing hand-crafted acoustic features from 4,800+ subjects. This result and evidence from prior work indicates that it is possible to model cognitive impairment, however, what has been relatively less explored are modeling paradigms that learn feature representations from the data itself [Fraser et al., 2013, Alhanai et al., 2017, Vasquez-Correa et al., 2017, Zhang et al., 2018].

6.1.1 Neural Network-based Modeling

Recent approaches to modeling cognitive impairment have looked to automatically learn features from data through the use of neural networks [Orimaye et al., 2016, Yancheva and Rudzicz, 2016, Karlekar et al., 2018]. The motivation behind the application of artificial neural networks (ANN) is the empirical observation that they are powerful at observing patterns and learning higher representations from the raw data; eliminating the tedious process of generating hand-crafted features [Schmidhuber, 2015, Jansen et al., 2017]. Prior work has focused on text transcripts to model language similarities with word vector representations [Mirheidari et al., 2018]. A more sophisticated approach has been to utilize the sequential nature of language to model cognitive impairment with a feedforward model [Orimaye et al., 2016], or a combination of Convolution Neural Networks (CNN) and Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) models [Karlekar et al., 2018]. Work on audio recordings of speech has shown the effectiveness of using 2D CNNs to model Parkinson's disease [Vasquez-Correa et al., 2017, Zhang et al., 2018].

6.1.2 Inference of Neural Networks

The neural network's ability to learn representations is due to the mathematics that define it, through a stochastic learning algorithm a model is able to iteratively learn complex non-linear transforms from the raw data [Goodfellow et al., 2016]. However these transforms are difficult to interpret [Olden and Jackson, 2002]. This has posed a challenge for individuals interested in applying these networks in scenarios where classification is not the ultimate goal (i.e. beyond image/speech recognition) [Ribeiro et al., 2016], but being able to learn from the model is important (i.e. human health and behavioral modeling) [Conroy et al., 2003, Dynan, 2000], or where sanity checking its logic is critical for deployment (i.e. spoof-proof speaker verification systems) [Wu et al., 2015, Moosavi-Dezfooli et al., 2016, Carlini and Wagner, 2017]. To this end, researchers have looked at interpreting models by either inserting mechanisms to force it to report what it's learning (i.e. attention models) [Chan et al., 2016], or to peel away the layers and map the activations according to different hypothesized class labels (i.e. heatmaps) [Zhou et al., 2016], or to optimize objective functions that learn alignments across modalities (i.e. segmentation of images-to-captions) [Karpathy and Fei-Fei, 2015, Harwath et al., 2018].

In the domain of modeling health from speech, multiple approaches have been pursued to develop interpretable models. In the context of modeling Alzheimer's disease, certain Part-of-Speech (POS) tag clusters have shown to be associated with Alzheimer's disease when inferred from a CNN-LSTM model [Karlekar et al., 2018], while in the context of Parkinson's disease, each layer of a CNN was probed to ascertain its significance in separating between classification outcomes [Vasquez-Correa et al., 2017], and in the context of speech dysarthria, bottleneck features in a feed-forward network may be explicitly trained to learn specific speech characteristics [Tu et al., 2017].

We highlight these approaches as a guide for our own work, since there is an expectation from medical professionals that they should be able to probe and interpret what the model has learned [Cabitza et al., 2017], be it a neural network or other modeling paradigm, and which is exemplified by the standard approach of utilizing regression for modeling outcomes and risk factors [Tu, 1996, Bellazzi and Zupan, 2008].

6.2 Hypothesis

We hypothesize that it is possible to model cognitive impairment from audio recordings without the need for feature engineering.

6.3 Objectives

Given the current state of machine learning for speech-based health modeling as described above, we are motivated to perform modeling of cognitive outcomes with the following two objectives: (1) to automatically learn speech characteristics that are predictive of cognitive impairment using neural networks, and (2) to evaluate what patterns these neural networks have learned. These two objectives would allow us to obviate the need for creating hand-crafted features, and would aid in the development of approaches to probe what a neural network has learned.

6.4 Methods

6.4.1 Data

Our study uses 7,196 audio recordings from 5,063 unique subjects who were undergoing neuropsychological evaluations. An audio recording has an average length of 63 minutes. The average age of a subject is 63 years, and 55.4% are female. 256 recordings (3.56%) contain a subject that has some level of cognitive impairment.

6.4.2 Data Processing and Feature Extraction

Audio recordings were processed to extract segments belonging to the subject (i.e. examiner audio was removed), according to the procedure described in Chapter 4. Following audio segmentation by speaker turn, 40 Mel filterbank (frame-level) features were extracted at 10 ms intervals over a 20 ms window, utilizing the kaldi toolkit [Povey et al., 2011]. For each subject, 10 segments of 25 second duration or less were utilized for detecting cognitive impairment, a threshold selected based on experiments in Chapters 4 and 5.

The data is randomly split into 'training', 'validation', and 'testing' sets with a ratio of 69% (49,668 examples, 4,982 subjects), 14% (10,164 examples, 1,017 subjects), and 17% (11,917 examples, 1,197 subjects) respectively. A speaker with multiple recordings is assigned to the same split, thus yielding an experimental setup that is speaker-independent across sets. To aid in analysis, all subjects in the gold set are assigned to the test set.

6.4.3 Performance Metrics

To assess the performance of our models, the evaluation metrics used are the AUC [Hanley and McNeil, 1982], sensitivity and specificity at Youden's Index (the probability threshold at which sensitivity and specificity are jointly maximized) [Youden, 1950], as well as model calibration using the HL-test [Hosmer Jr et al., 2013]. We report results at the subject-level by calculating the mean probability across all segments per subject.

6.4.4 Model

We are interested in modeling the temporal component of cognitive impairment in speech, and therefore utilize a CNN model. We compare this model with two baseline models: (1) a logistic regression model with demographic information (age, gender, education, APOE e4) as the input features, and (2) the model developed in Chapter 5. We conduct experiments according to the following configurations:

- **Baseline (demographic)**:

  Model: Logistic regression.

  Input: Demographic information (age, gender, education-level, APOE e4).

- **Baseline (feature-based L1 model)**:

  Model: L1 regularized logistic regression.

  Input: 676 prosodic and energy based speech statistics.

- **Global Mean/Max-pool CNN**:

  Model: 4-layer CNN with global mean/max-pooling.

  Input: 40 Mel filterbank.

- **Local Mean/Max-pool CNN**:

  Model: 4-layer CNN with global max-pooling and local mean/max-pooling.

  Input: 40 Mel filterbank.

- **10/20 second input to CNN**:

  Model: 4-layer CNN with global mean/max-pooling and local mean/max-pooling.

  Input: 40 Mel filterbank with input features either 10 or 20 seconds long.

Convolution Neural Network (CNN)

CNNs have been effective at modeling categorical outcomes of data with spatio-temporal properties such as images and videos [Russakovsky et al., 2015, Karpathy et al., 2014]. They are popular not just due to their accuracy but also because CNNs can be trained relatively faster than other equally powerful models (e.g. LSTMs), when maintaining an equivalent number of trainable parameters [Lei et al., 2018]. Furthermore, the ability for CNNs to capture salient spatio-temporal patterns of the input features have provided opportunities to 'peel away' layers of the model and project back onto the input features to discern the influence of each data-point/pixel/movie-frame in determining the outcome [Zhou et al., 2016, Bulat and Tzimiropoulos, 2016]. It is for these two reasons (historically strong/speedy performance and spatio-temporal modeling abilities) that we decide to explore the application of CNNs for modeling cognitive impairment from audio recordings.

Hyperparamter Optimization

To find the optimum topology of the CNN model we explore a set of hyper-parameter settings and select the best according to the smallest loss observed on the validation set. It can be challenging to find the optimum topology because a single training cycle can take several hours and sometimes days. Moreover there are a high number of hyperparameters than can be explored. Performing hyperparameter optimization as a grid-search is effectively intractable, and oftentimes an intuition is developed as to what works without a clear systematic understanding of why a particular setup works. Evidence suggests that a random search converges to a close-to-optimum topology [Bergstra and Bengio, 2012], so we adopt that approach, and manually introduce and remove sets of hyperparameters as they show gains, no gains, or even deterioration. Further evidence suggests that allowing for an expansive algorithmically driven search of optimal neural network topologies would not converge to some of the most successful topologies in use (e.g. searching for RNNs that out-perform an LSTM and/or Gated Recurrent Unit [GRU] [Jozefowicz et al., 2015]). So in the end, the evolution of neural networks is dependant on human-based insights (e.g. attention modeling as an interpretable constraint) and wisdom (e.g. 'change learning rate')

With the reality of hyperparameter optimization in mind, we next describe the hyperparameters in a CNN that we explored and what their function is. To begin, a CNN is generally composed of one or multiple convolution layers, non-linearity layers, pooling layers, and fully connected layers [Goodfellow et al., 2016].

Convolution layer: A convolution layer is composed of filters; each filter outputs a weighted sum of the input (i.e. $\sum w_i x_i + b$ for a set of weights w and inputs x and bias term b). A specific number of filters are defined over a subset of dimensions of the input data, with a width, height, and shift in the case of a 2D tensor, and where each filter focuses to learn different patterns in the input. In our own work we explore networks that have 1 to 5 convolution layers, and an increasing number of filters (by powers of 2) in each layer ranging from 32 to 4096 filters.

Activation function: The convolution layer results in a map of activation values that can then be further transformed to ensure that values are scaled within a specific range, which mitigates the problem of exploding gradients and ultimately non-convergence of the model during training. Some common re-scaling transforms are the sigmoid ($\frac{1}{1+e^{-x}}$), hyperbolic tangent ($\frac{2}{1+e^{-2x}} - 1$), and rectified linear unit (RELU, $max(0, x)$). We explored networks using these 3 functions.

Pooling layer: Once a convolution has been performed and activation function has been applied, it may be useful to learn some higher-order and generalizable information, an operation called 'pooling'. Pooling layers commonly calculate the mean or maximum values within a specific region of the activation value map (e.g. m_1-element-wide shifted by m_2 elements for an M-long vector where $M > m_1$). Work exploring pooling layer operations have shown how networks trained on images learn shapes of objects with mean-pooling, and textures in images with max-pooling [Gatys et al., 2015]. We term the output of this layer as generating a *feature map*. Additionally, the pooling layer also results in down-sampling the input activation map, which in our case is by a power of 2, and explore both mean- and max-pooling with a size between 2 and 4, and shifts between 2 and 30.

Global pooling layer: In addition to pooling over a subset of inputs, it may be useful to calculate values over the whole activation map and collapse a dimension. In our case, we were interested in generating an utterance-level feature map to use for the final classification, that is collapsed across time (i.e. m_1-element-width $=$ M-long vector). We explore both mean- and max-pooling over the entire temporal dimension which yields a feature vector the same length as the number of filters in the final convolution layer.

Classification layer: For a model trained as a classifier, the final layer is designed to generate a class output. In our case we utilize a sigmoid activation function since it produces a binary output; a probability scaled between 0 and 1.

Fully connected layer: Before the final classification output, it may be useful to have a fully connected layer learn a transform to better discriminate between classes. A fully connected layer provides a similar operation to the convolution layer, but is connected to each input, rather than a subset of inputs. We performed an initial hyperparameter search that included a fully connected layer, ranging from 1 to 3 layers, 4 to 256 hidden units, and activation functions sigmoid, hyperbolic tangent, and RELU. We decide not to include a fully connected layer in our final models because we wanted to be able to directly map the inputs of the final sigmoid classifier, to the feature map inputs.

Batch normalization: In addition to bounding convolved values with an activation function, we might want to rescale the data to perform mean and variance normalization. The strength of a neural network is that it can learn these normalization parameters during training. We explored utilizing batch normalization in every layer, or not at all. In previous chapters we performed mean and variance normalization across all input features, for our neural network training we attempt to do this on every batch, but find that it makes training unstable, therefore, we decide to perform batch normalization on the input features [Harwath et al., 2018].

Dropout: To avoid overfitting, the method of blocking a percentage of hidden units in the neural network was developed, and has shown to be powerful at yielding improvements in model performance [Srivastava et al., 2014]. We explore utilizing dropout, with rates ranging from 0 to 0.9.

Training algorithm: There are several algorithms that can be used to train a neural network, in our case we utilized stochastic gradient descent. There are several hyperparameters for stochastic gradient descent; we explored various values for the learning rate (1e-01 to 1e-09), momentum (0 to 1), and decay (0 to 1e-09). There are also various stopping criteria, in our case we train networks for a maximum of 150 epochs, or if an early stopping criterion is met; the decrease in validation loss is less then 1e-04 for 5 consecutive epochs.

Batch size: We select a size that would fit into 80-90% of an 8GB GPU memory (from 32 to 4096 depending on the size of the network).

Loss function: Our model is solving a binary classification problem, we therefore explore a set of binary loss functions; mean absolute error (MAE, Eq. 6.1), mean-squared-error (MSE, Eq. 6.2), and binary cross-entropy (BCE, Eq. 6.3). We define y_i to be the actual value (0 or 1) for the i^{th} example. $\hat{y}_i$ is the predicted value for the i^{th} example, and N is the total number of examples in the data.

$$\text{MAE} = -\frac{1}{N}\sum_{i=1}^{N} ||\hat{y}_i - y_i|| \tag{6.1}$$

$$\text{MSE} = -\frac{1}{N}\sum_{i=1}^{N}||\hat{y}_i - y_i||^2 \tag{6.2}$$

$$\text{BCE} \ = -\frac{1}{N}\sum_{i=1}^{N} y_i \text{log}(\hat{y}_i) + (1-y_i)\text{log}(1-\hat{y}_i) \tag{6.3}$$

Data input: The input features are convolved across time, with each filterbank convolved separately (i.e. treated as a 1D signal). Filters in subsequent layers were also convolved separately. This decision is made because speech in spectrograms do not have the same translation properties across the x- and y-axes (i.e. temporal based spectral energy activity) as images do (i.e. objects rotate and scale in both the x and y axes as a function of movement in the physical world), and since filterbanks are highly co-linear, we want to learn representations that are maximally de-correlated. This approach of convolving across time has been previously utilized for neural network modeling of speech with CNNs [Harwath et al., 2018, Abdel-Hamid et al., 2014]. However, we note that prior work has also found it effective to utilize 2D CNNs with speech filterbank features [Zhang et al., 2018, Vasquez-Correa et al., 2017, Zhang et al., 2017]. More broadly in the research community, work is on-going to learn representations straight from the raw waveform [Aytar et al., 2016, Sainath et al., 2015].

All utterances in our data are less than 25 seconds long, and vary in length, so we explore utilizing segments that are 2048 (∼20 secs), and 1024 (∼10 secs) frames long, with longer segments truncated, and shorter segments zero-padded at the end. It should be noted that segment lengths of powers of 2 are used since it simplifies the process of downsampling (by 2) and padding. Approximately 50% of the segments are longer than 20 seconds, while approximately 90% are longer than 10 seconds. Figure 6-2 displays a cumulative distribution of the segment lengths in our data.

The Optimum Hyperparamters

Finally, the optimum hyper-parameter setting was found to be the following, of which a visual display is shown in Figure 6-3 . F corresponds to the number of filters, W is

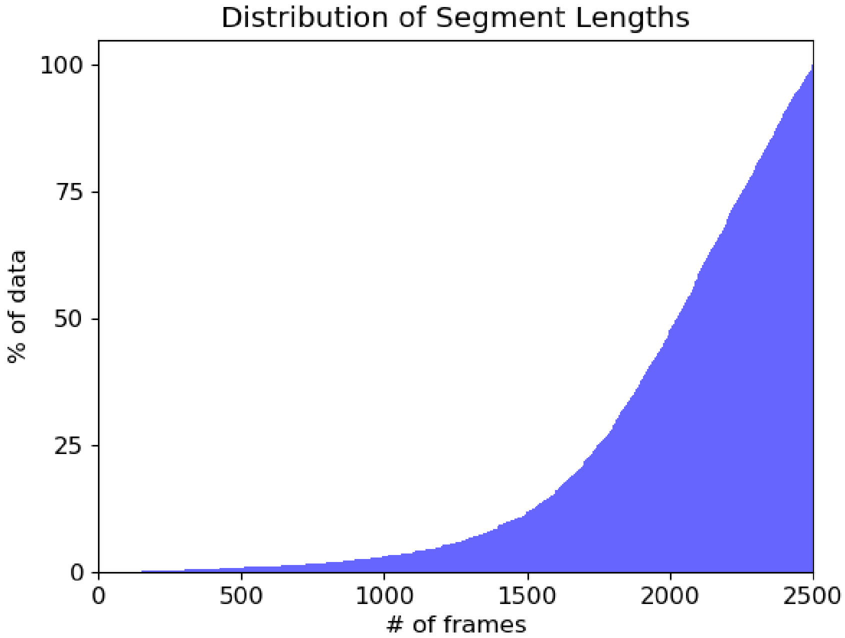


Figure 6-2: **Cumulative Distribution of Data Segment Lengths**. Approximately 50% of the segments are longer than 20 seconds, while approximately 90% are longer than 10 seconds.

for window width, H for window height, and S for stride.

Topology: Batch-normalized input, with 4-layer convolution network, ReLU activations, and global max-pooling of the final output as a 2048 dimensional (collapsed across time) as input to the sigmoid classifier.

- *Conv 1*: $F = 256$, $W = 1$, $H = 40$, $S = 1$.
- *Conv 2*: $F = 512$, $W = 18$, $H = 1$, $S = 1$.
- *Max-pool*: $W = 3$, $S = 2$.
- *Conv 3*: $F = 1024$, $W = 18$, $H = 1$, $S = 1$.
- *Max-pool*: $W = 3$, $S = 2$.
- *Conv 4*: $F = 2048$, $W = 18$, $H = 1$, $S = 1$.
- *Global max-pool*: $W = 512$, $S = 0$.
- *Sigmoid*: $F = 2048$, $W = 1$, $H = 1$.

Optimizer: Stochastic gradient descent, with learning rate = 1e-01, momentum = 0.8, and decay = 0. Loss function is binary cross-entropy.

Data: Input $H = 40$ (Mel filterbanks), $W = 2048$ frames (approximately 20 seconds), batch size = 160, epochs = 150 and an early stopping criterion of validation loss < 1e-04 for 5 consecutive epochs.

Baseline: Demographic Model

Prior work has looked at modeling associations between an individual's physiology, and cognitive outcomes. Therefore we were interested in comparing our speech model with prior models for two reasons: (1) to evaluate the relative gain in accuracy with established approaches, and (2) to determine, in an interpretable way, what components our speech model is capturing (i.e. is it learning a speaker's age or sex, or complementary information from their speech?).

To this end we use a logistic regression model with demographic features as the input to serve as the baseline. The demographic features contain the subject's age, sex, highest level of self-reported education (didn't graduate high-school, high-school graduate, attended but didn't graduate college, or college graduate or higher), and the presence of the apolipoprotein E (APOE) e4 gene. These factors have been shown to have associations with cognitive impairment (age, sex, education), as well as specific pathologies such as Alzheimer's (APOE) [Ward et al., 2012]. Age is coded as a continuous variable and zero-mean variance normalized. Sex is coded as a binart variable with 1 representing female. Education is dummy coded into 4 categories (some high school, high school, some college, college) with some high school set as the reference category. APOE is binary coded with the presence of e4 represented as 1.

Baseline: Feature Engineering + L1 Model

We also train our model from Chapter 5; hand-crafted features input to an L1 regularized logistic regression model. The pre-processing method results in some feature values that are *Null* which therefore eliminates some examples from the data. Hence, the training set has slightly less subjects (46,117 examples, 4,977 subjects), but with

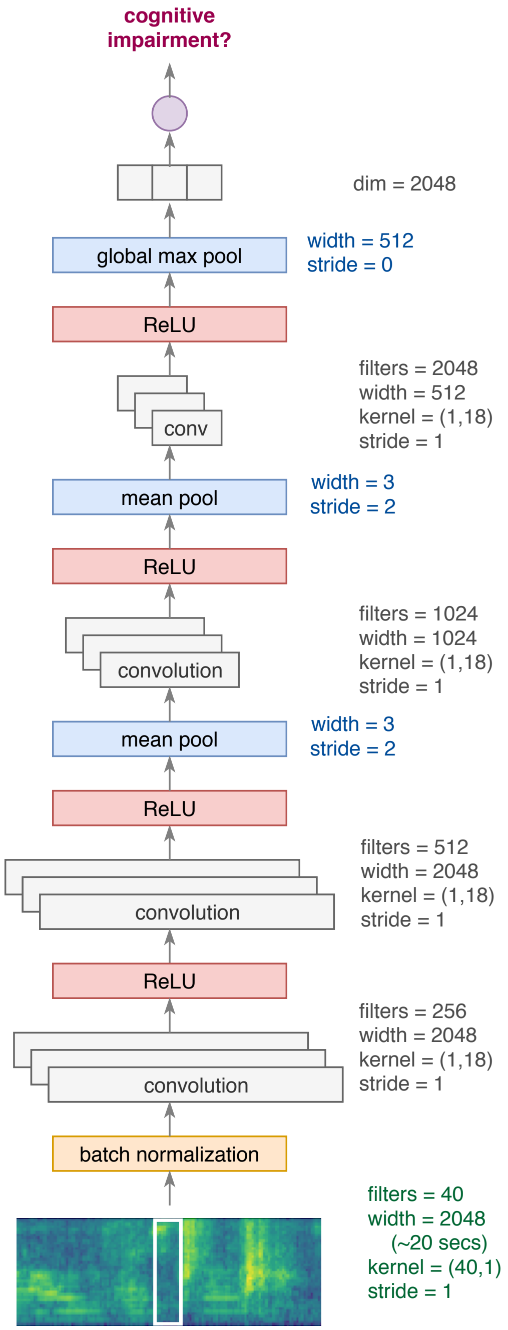


Figure 6-3: **CNN Model**. The figure displays the optimum topology found after hyper-parameter optimization. The CNN has a batch-normalized input, with 4-layers of convolution, ReLU activations, and global max-pooling of the final output as a 2048 dimensional embedding (collapsed across time) input to the sigmoid classifier. The optimizer is stochastic gradient descent, with learning rate = 1e-01, momentum = 0.8, and decay = 0. Loss function is binary cross-entropy. The data input is H = 40 (Mel filterbanks), W = 2048 frames (approximately 20 seconds), batch size = 160, epochs = 150 and an early stopping criterion of validation loss < 1e-04 for 5 consecutive epochs.

the exact same number of test subjects (11,094 examples, 1,197 subjects0), relative the training examples available for the CNN model.

6.5 Results

The results of our experiments are summarized in Table 6.1. The best performing speech based model is the CNN max-pool model (0.85 AUC), which has comparable performance to the baseline demographic model (0.85 AUC), and higher performance than the baseline feature-based L1 model (AUC 0.83). The mean-pool CNN has the best sensitivity (82%) and is well-calibrated (HL-test > 0.05).

Model	**AUC**	**Spec.**	**Sens.**	**HL-test**
Demographic model	0.85	93%	65%	1e-16
Feature-based L1 model	0.83	78%	79%	0.002
CNN global mean-pool	0.81	67%	84%	0.069‡
CNN global max-pool	0.85	81%	82%	0.013

Table 6.1: **Results**. CNN max-pool model exhibits equivalent classification performance to the demographic model (AUC 0.85). The CNN mean-pool has the best sensitivity (82%) and is well-calibrated‡ (HL-test > 0.05).

6.6 Discussion

6.6.1 Speech and Demographics

Our results show that the speech-model performs as well as the demographic (baseline) model, therefore we are interested in discerning what, if any, information the speech model is capturing that relates to a person's demographic profile such as age, sex, and education level. We perform this analysis by combining the output of the speech model (predicted probabilities from the sigmoid) as an additional feature with the demographic features as input to a logisitic regression model (visualization displayed in Figure 6-4).

Table 6.2 displays the logistic regression model coefficients for the model with demographic features only, as well as the model that includes the max-pool CNN

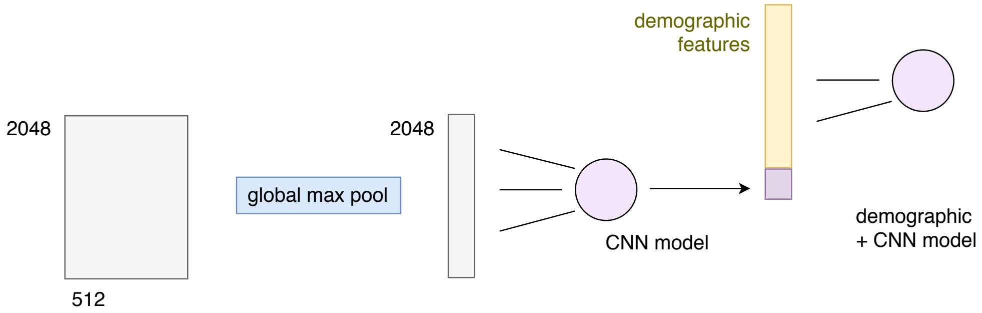


Figure 6-4: **Combined CNN and Demographic Model**. CNN output probability is an input feature with demographic features into logistic regression model.

output. For the demographic model, increasing age is positively and significantly ($p < 0.001$) associated with cognitive impairment, while being female, and having higher levels of education (relative to no high-school degree) is negatively and significantly ($p < 0.001$) associated with cognitive impairment. The presence of APOE e4 does not have a significant association with outcome.

	Demographic model		Demographic model + CNN model prob.	
Feature	**coeff** (e^{β})	p**-val**	**coeff** (e^{β})	p**-val**
Age	3.10	< 0.001	1.99	< 0.001
Sex	0.49	< 0.001	0.43	< 0.001
Education - high-school	0.06	< 0.001	0.05	< 0.001
Education - some college	0.03	< 0.001	0.02	< 0.001
Education - college	0.01	< 0.001	0.01	< 0.001
APOE e4	1.16	0.014	1.30	< 0.001
max-pool CNN prob.	/	/	1.82	< 0.001

Table 6.2: **Logistic Regression Model Coefficients**. (Demographic model) Increasing age, male sex, and lower education levels are positively associated with cognitive impairment. (Demographic model + CNN) Including output prediction of max-pool CNN model yields a positive and significant association with the outcome, reduces model coefficients of other features, except for the APOE e4 biomarker, which now has a significant (and positive) association with cognitive impairment.

Including the output prediction of the max-pool CNN model in the logistic regression model yields a positive and significant association with the outcome. The values of the rest of the model coefficients either stay the same (e.g. education - college), decrease slightly (1% to 6% absolute, e.g sex and education - high school/some college), or in the case of age, decrease by a large amount (112% absolute). This changes in model coefficients indicate that the speech model contains information that is complimentary to age, sex, or education level. Most interestingly, the APOE e4 biomarker, now has a significant (and positive) association with cognitive impairment. The APOE e4 biomarker marks the presence of the e4 variant of the APOE gene, which has been found to be a risk factor for the late-onset of Alzheimer's disease [Ward et al., 2012]. Its significance in this model indicates that the speech model may be learning patterns unique to a single or subset of cognitive conditions independent of what the APOE biomarker captures.

There remains a number of variations that can be applied to our modeling approach which may help in understanding what attributes of a speaker a neural network is learning from speech. For example, including demographic information along with filterbank features for neural network training is one approach that may be further explored, as well as introducing demographic features into different layers in the network [Alhanai and Ghassemi, 2017].

6.6.2 Model Calibration

While analysis of model sensitivity and specificity at Youden's Index is illuminating, our models output probabilities, therefore the threshold for classification can be adjusted, which means that model calibration becomes an important attribute. A well-calibrated model (HL-test > 0.05) means that if a model predicts a 30% chance of cognitive impairment, then 30% of the time, a subject will be found to truly have cognitive impairment. This aspect of a model's performance makes it reliable for use as a screening tool, since a decision-maker (e.g. medical professional) can adjust their behaviour according to the probabilities that the model outputs, such as being conservative in demanding follow-up screening for lower probabilities, and aggressive with higher probabilities. Model calibration has important implications when deployed, but it has been relatively less explored for neural network models [Guo et al., 2017]. In our own experiments we observe that the CNN model utilizing global mean-pooling is well calibrated, albeit with a slightly lower performance (AUC $= 0.81$) than the max-pool CNN (AUC $= 0.85$, and not well-calibrated). Utilizing the mean-pool

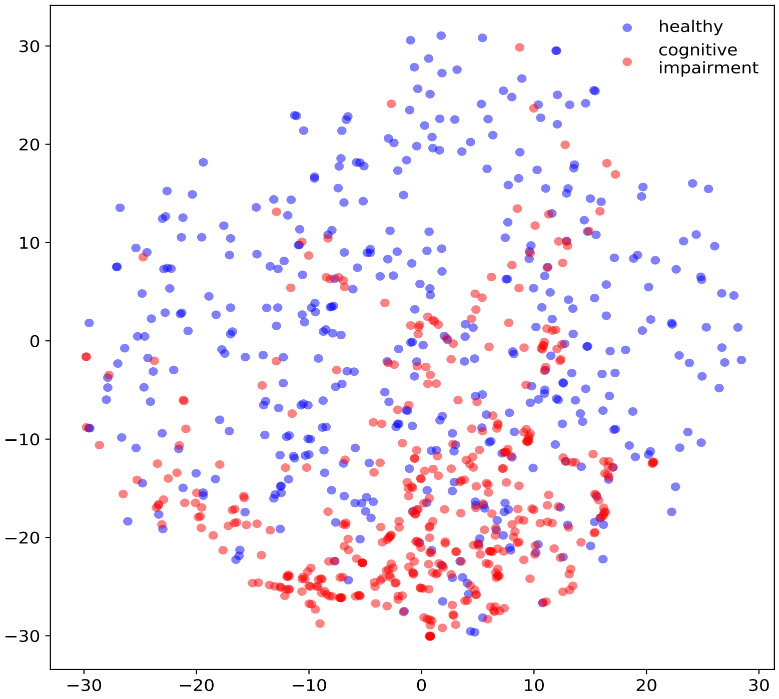


Figure 6-5: **t-SNE of Utterance-level Embeddings**. t-SNE of 2048-dim embeddings of a random selection of utterances from test set subjects, colored according to ground truth outcomes. Axes are dimensionless.

model would be a classification performance trade-off, in return for a more robust model for deployment. We hypothesize that the global mean-pool operation may be focusing more on overall differences in speech, rather than extrema, but is an avenue that would be informative to pursue in future work.

6.6.3 Utterance-level Representations

We are interested in utilizing what the network has learned to perform frame-level analysis. As a step towards this analysis, we perform a qualitative assessment of the utterance-level representations learned by the model utilizing the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm, a method that can learn a non-linear transform of data, and has been empirically shown to be powerful at revealing the underlying structure of high-dimensional data [Maaten and Hinton, 2008]. The algorithm allows us to reduce the dimension of the 2048 utterance-level embeddings down to 2, so that we can visualize the distribution in 2D. We utilize the Multicore-t-SNE implementation with hyperparameters set at perplexity of 5, and 500 training iterations [Ulyanov, 2016]. The results are displayed in Figure 6-5, which shows a separability between the two classes. Given this qualitative assessment of class separability, and the high performance of the model (AUC 0.85), we are motivated to perform further analysis on what these representations are learning.

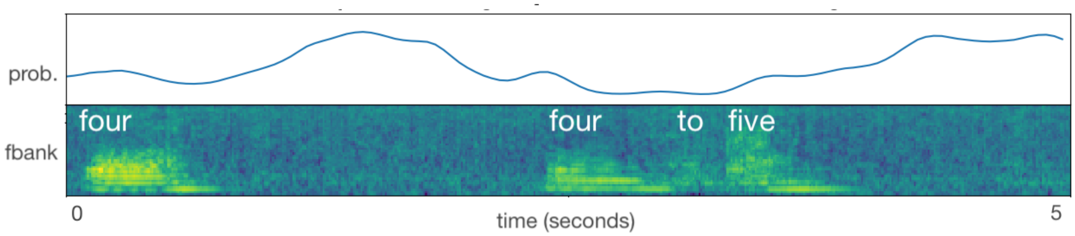


Figure 6-6: **Probability Profile of Frame-level Embeddings**. (Top) Probability profile of feature-map embeddings, (bottom) along with the corresponding input features (i.e. spectrogram). Increasing probability corresponds to a higher likelihood of cognitive impairment as hypothesized by the sigmoid classifier. Each point in the probability profile corresponds to 1260ms in the spectrogram. The classifier outputs a higher probability of cognitive impairment with increased silence. During this specific test (called trail-making) of the neuropsychological exam the subject is attempting a 'dot-to-dot' like activity to connect consecutive numbers on a sheet and is speaking out loud during the process.

6.6.4 Frame-level Representations

Before the global max-pooling layer, the final convolution layer outputs a *feature map* that is composed of 2048 filters and 512 elements. The 512 elements correspond to the time-axis, with each element capturing information spanning 1260ms[1] and shifting every 40ms[2]. Since the feature map is capturing temporal information before global max-pooling (which collapses across the time axis) we would like to see what speech activity at each timestep the model is detecting, so we remove the global max-pooling

[1] (10ms x 18 kernel width) + (10ms x 18 kernel width x 2 downsampling) + (10ms x 18 kernel width x 4 downsampling) = 180ms + 360ms + 720ms

[2] 10ms shift of input feature frame x 4 downsampling

operation and record the predicted probability of the sigmoid output when fed the 2048-dim feature map embedding at each timestep. Figure 6-6 shows an example of the probability profile and the corresponding feature input (i.e. spectrogram that is ultimately transformed into the feature map embedding) to aid in visualization. We observed in the figure that the embeddings during each timestep along with the sigmoid classifier are operating as a function of speech activity, where increased duration of silence increases the predicted probability that the speaker has cognitive impairment, and increased speech activity is predictive of a healthier outcome.

Next, we accumulate the values of the probability profile and plot it as a distribution, shown in Figure 6-7. We observe that subjects with an outcome of cognitive impairment have higher frame-level probability values than subjects with a healthy outcome. This separation between the frame-level distributions motivates deeper analysis of the probability profile.

6.6.5 Peaks and Dips

We perform further analysis to determine if there were particular patterns of speech that the model had learned to capture. We accumulate the feature map embeddings that resulted in peaks and dips of the probability profile (an example is displayed in Figure 6-8). This results in 8,398 peaks and 8,429 dips calculated every 2.5 seconds (for a total of 16,827 samples). Many of these peaks and dips correspond to periods of silences. We therefore focus specifically on spoken components of the utterance, since they may contain information on *how* a person spoke that the model finds to be predictive of cognitive impairment. We therefore eliminate peaks and dips of complete silences, which results in 2,095 peaks and 1,633 dips (3,728 samples total). We reduce the dimensionality of the embeddings via t-SNE (from 2048 to 2)[3], and plot the results in Figure 6-9. We observe that a pattern emerges; embeddings that generate similar probabilities for cognitive impairment cluster together, so peaks (large, red points) and dips (small, blue points) cluster together. This pattern indicates the model learns to respond to specific patterns, with similar representations and potentially similar

[3] Trained with perplexity of 5, and 500 iterations.

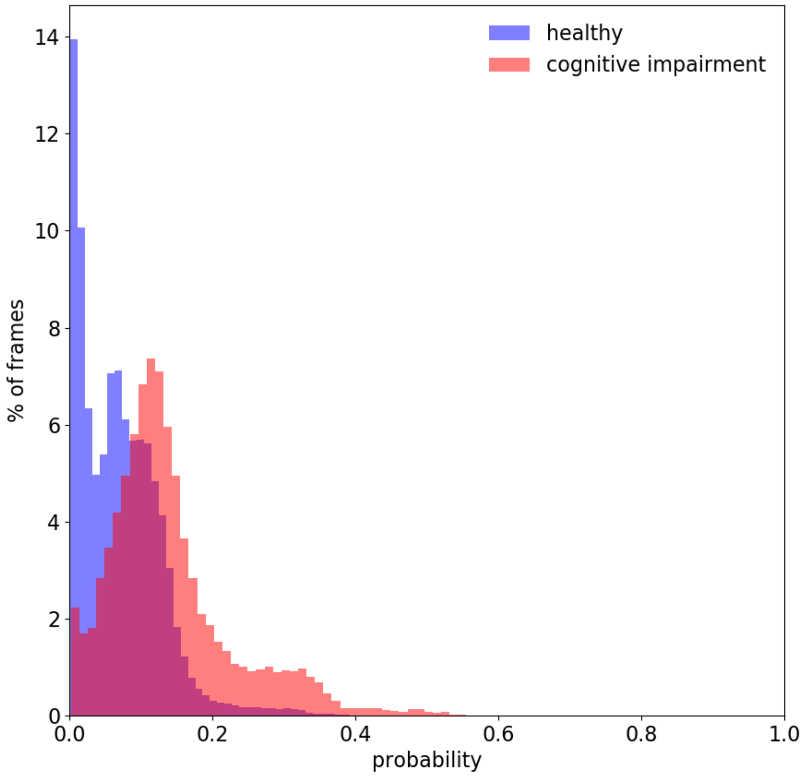


Figure 6-7: **Distribution of Frame-level Probabilities.** The probability values are from before the global max-pooling operation, which are 2048-dim representations at each time-step fed into the sigmoid binary classifier. We can observe some separation between the two classes.

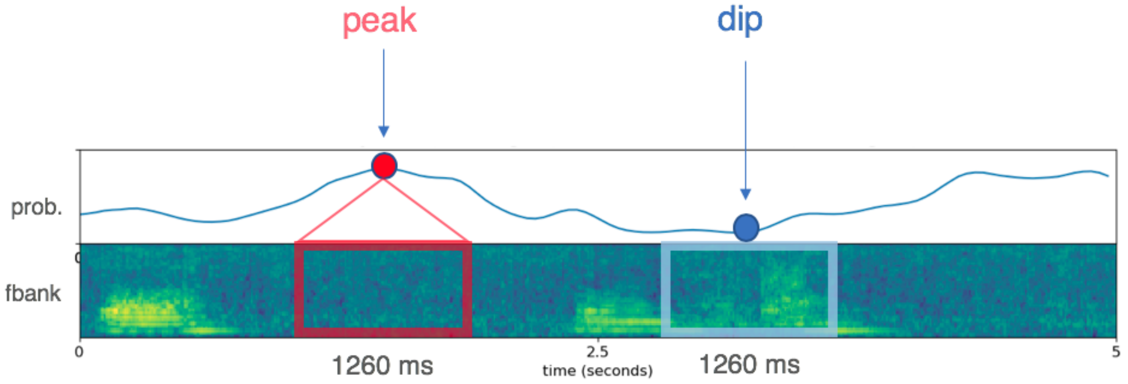


Figure 6-8: **Peak and Dip**. A peak and dip are calculated every 2.5 seconds. Each peak/dip captures information that spans 1260ms of the input filterbank feature.

underlying speech activity. To further explore this idea, we (1) plot the underlying 1300ms spectrograms of these corresponding points in t-SNE space, and (2) perform k-means clustering to see what types of speech patterns appeared in our data.

6.6.6 Spectrogram Clusters in t-SNE Space

We generate t-SNE spectrogram plot by replacing each point in Figure 6-9 with the underlying 1300ms spectrogram that it represents. This t-SNE spectrogram plot is viewable in Figure 6-10. We can observe that speech is clustered into several types of patterns; a sudden onset of speech, a short burst of speech activity, spectral energy in high-frequency bands (e.g. strong emphasis on */s/* for the word *senior*), and short gaps of silence (e.g. `[uh] <sil> powersaw`). This serves as a visual illustration

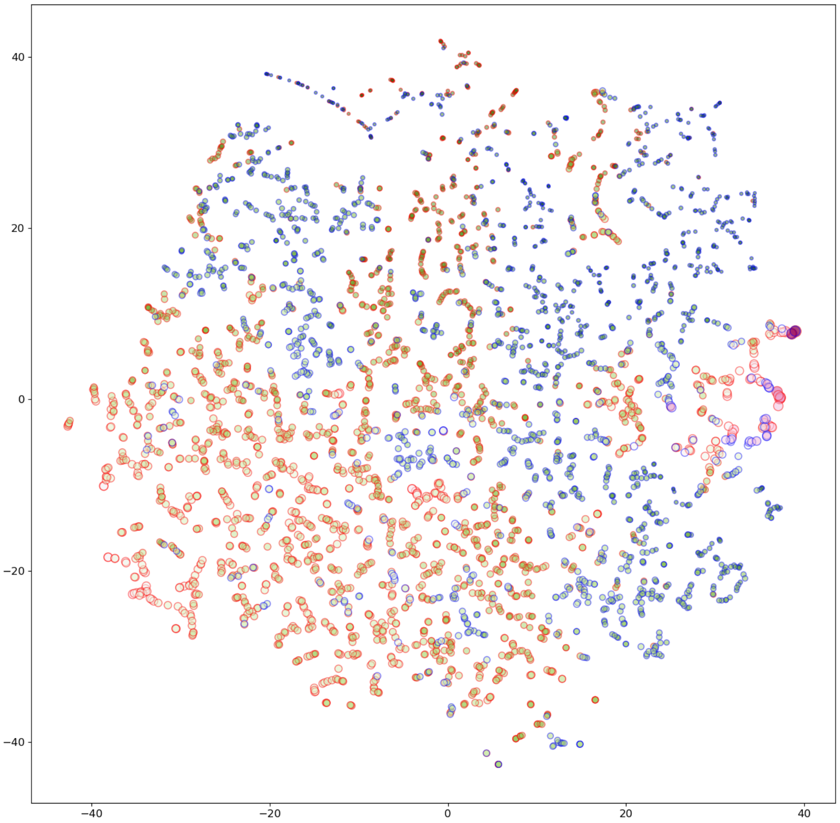


Figure 6-9: **t-SNE of Frame-level Embeddings**. t-SNE of 2048-dim frame-level feature map embeddings colored according to peak (red) or dip (blue) while size is according to probability value. Axes are dimensionless. We can observe that similar probability values cluster together which motivates understanding what the underlying speech patterns may be.

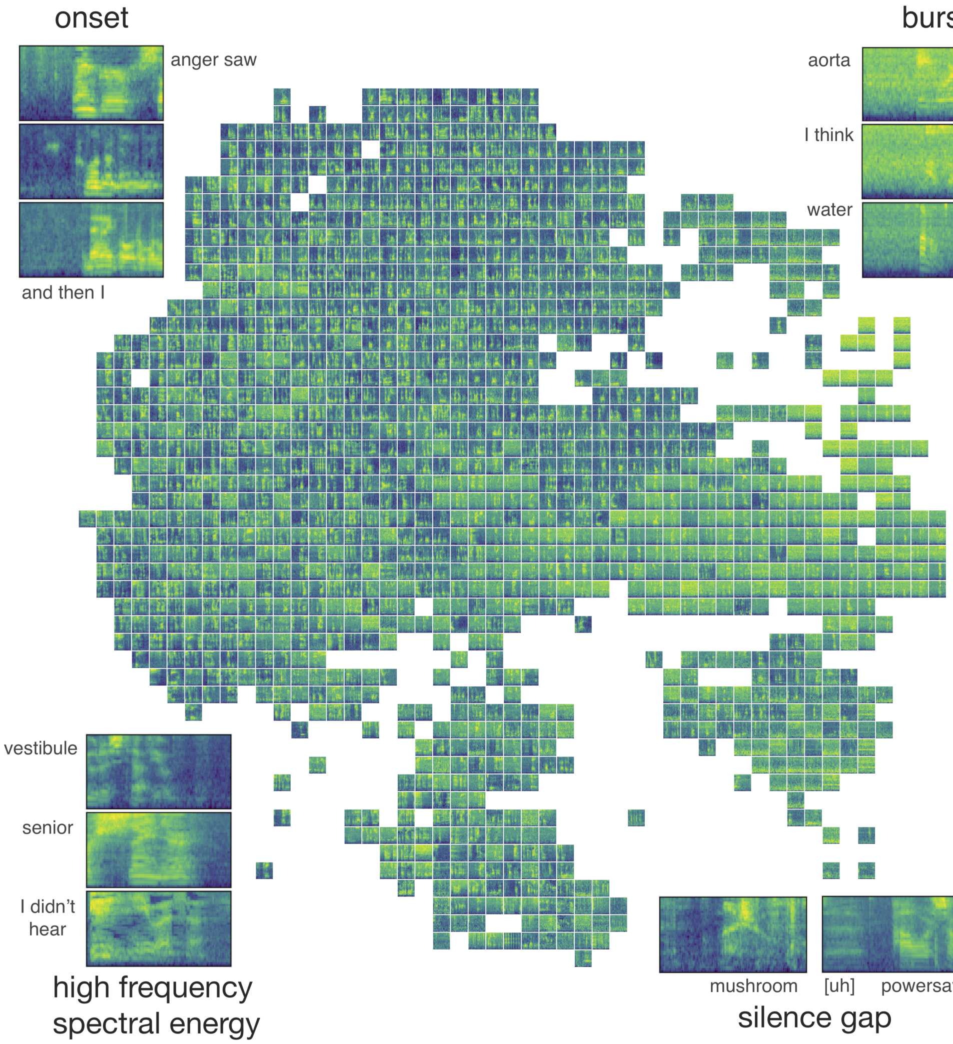


Figure 6-10: **t-SNE Plot of Spectrograms.** Plot of spectrograms (1300ms duration) in the t-SNE 2D co-ordinate space of feature map embeddings that elicit peaks when input to the sigmoid classifier. Clustering of speech activities can be seen in the plot, with some examples of this activity in the form of: speech onset, short burst of activity, spectral energy in high-frequency bands, and gaps of silence between words.

of the kind of speech activity the model is learning, and that it does indeed learn representations that cluster according to similar underlying speech activity.

Speech Profile Peaks are Independent of Speaker, Test, and Vocabulary

We would like to verify that none of the observed clustering are artifacts due to speakers and tests. Therefore, we plot the t-SNE points colored according to speaker and test type. We also train word embeddings from the transcripts[4], perform t-SNE, and plot the word embeddings colored according to class outcome. We note that because the word embeddings showed clustering by words utilizing the existing text transcripts, we do not attempt to use a large word corpus of text. Furthermore this may avenue of work will need to establish whether grammar from a Wikipedia-like corpus translates to the specific context of neuropsychological exams.

Plotting the results in Figure 6-11, we find that the clustering of speech activity is independent of the speaker, neuropsychological test, and vocabulary, which indicates that the model learns to focus on particular properties of speech articulation. The speech patterns our model learns agrees with prior findings. Indeed, speech differences across different cognitive states may appear independent of syntactic and semantic patterns of language [Fraser et al., 2016]. Furthermore, speech variances across spectral energy bands were useful in modeling cognitive impairment [Al-Hameed et al., 2017]. Finally, dis-fluencies manifesting as hesitations such as short responses, reduced speech activity, and pauses (silence or filled paused: 'sigh' or 'um') have been used as features in models to good effect [Hoffmann et al., 2010, Corley and Stewart, 2008, López-de Ipiña et al., 2013, Satt et al., 2013, Pistono et al., 2016].

6.6.7 Frame-level Speech Clustering via Euclidean Distance

Our speech model is not trained to formally learn relationships between embeddings through a distance metric (e.g. objective function which measures the distance between inputs and updates the gradients of the neural network accordingly [Chen and Salman, 2011]). Rather, our model is trained to classify a variety of speech patterns into two classes once a final representation has been input to a final sigmoid classifier. So there is no guarantee that the frame-level embeddings will capture relationships

[4] Wor2Vec trained with perplexity of 5, 500 iterations, word embedding dimension of 50.

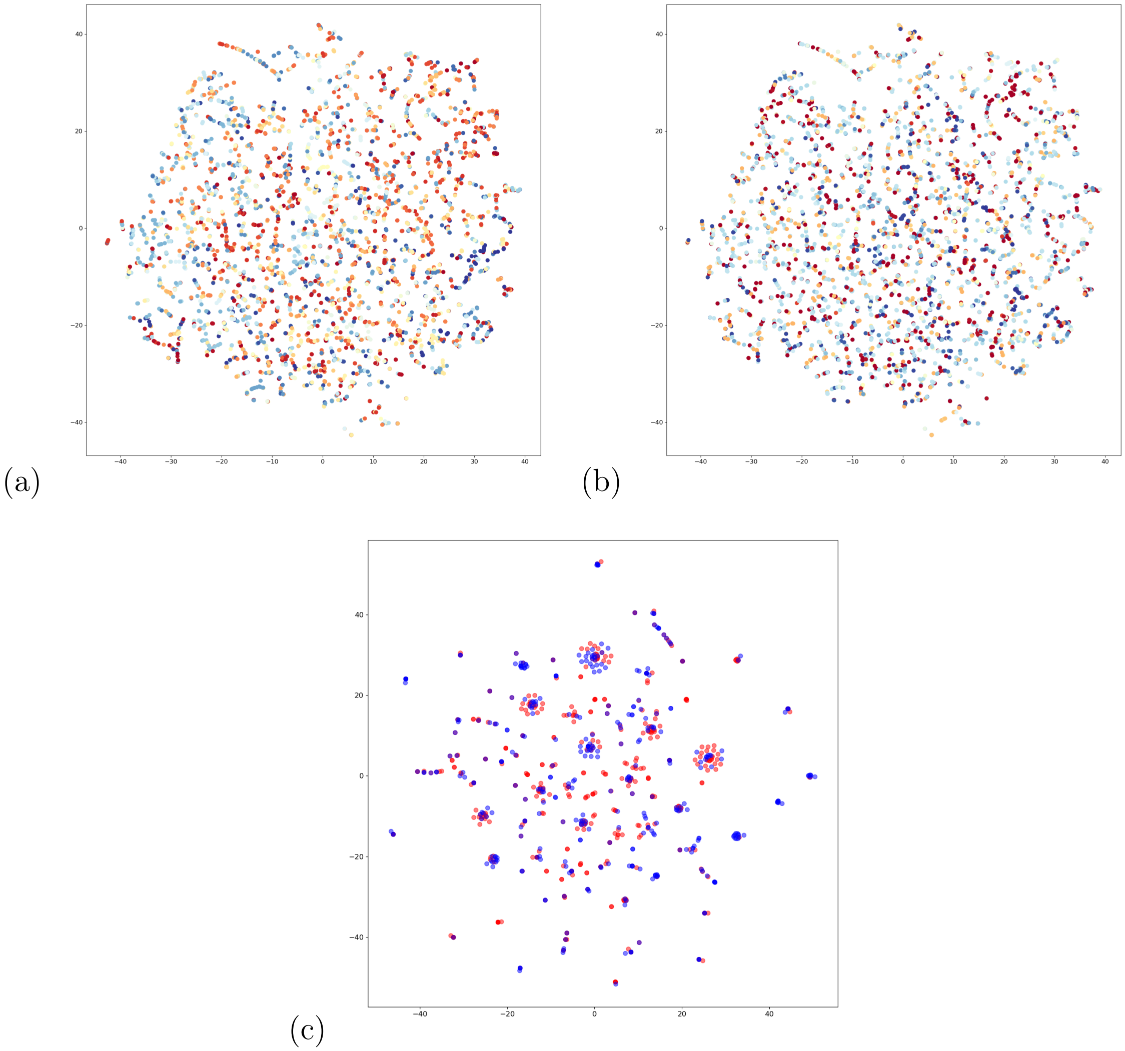


Figure 6-11: **t-SNE of Speakers, Tests, and Vocabulary**. Plots of t-SNE for feature map vectors colored by (a) speaker and (b) neuropsychological test, as well as (c) t-SNE of word2vec word embeddings (colored by peaks and dips). None of the plots exhibit clustering by speaker, test, or words indicating that the features learned by the speech model are independent of speech elicitation source and language usage.

that can be clustered with a distance metric (e.g. euclidean distance). To find out if such a relationship exists, we perform k-means clustering of the frame-level embeddings; an algorithm that assigns samples to a single cluster by minimizing the intra-cluster variability via the sum of squared error (SSE) function. We define K to be the number of clusters, $\mathbf{x}$ as the vector of values representing the sample (in our case 2048-dim embedding), and c_i as the i-th centroid out of a possible k clusters.

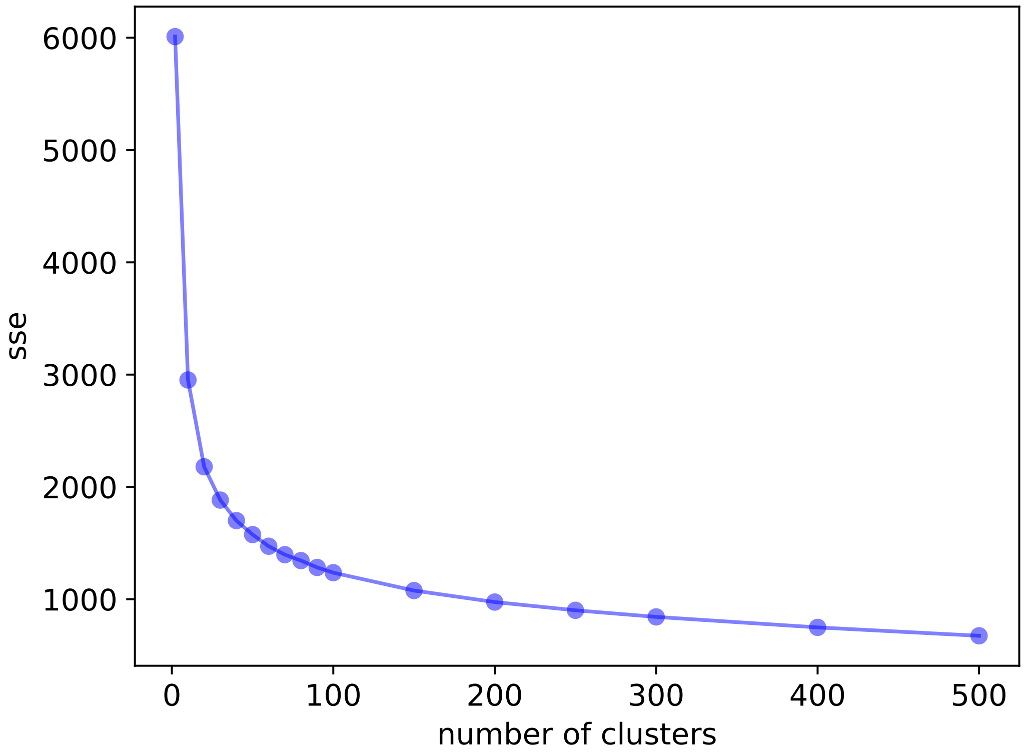


Figure 6-12: ***k*-means SSE elbow plot.** We cluster with k ranging from 2 to 50, and focus our analysis for sum of squared error (SSE) less than 1000, at $k = 300$.

$$\text{SSE} = \arg\min_{K} \sum_{i=1}^{k} \sum_{\mathbf{x}\in K_i} ||\mathbf{x} - c_i||^2 \tag{6.4}$$

In our work we cluster with k ranging from 2 to 500 and calculated the number of samples (N) in each cluster, the variance (σ^2) of each cluster, the class purity (C, the ratio of samples belonging to the ground truth for cognitive impairment), and peak purity (P, the ratio of samples belonging to peaks as opposed to dips). In Figure 6-12 we plot an elbow plot of the SSE as a function of k, and focus our analysis for SSE less than 1000, at $k = 300$.

Next we plot representative examples of spectrograms from clusters that have the lowest intra-cluster variance, which assumes a homogeneity in the speech activity they capture. We also notice that smaller cluster sizes have a larger range of intra-cluster variance (displayed in Figure 6-13). We therefore focus our analysis on small

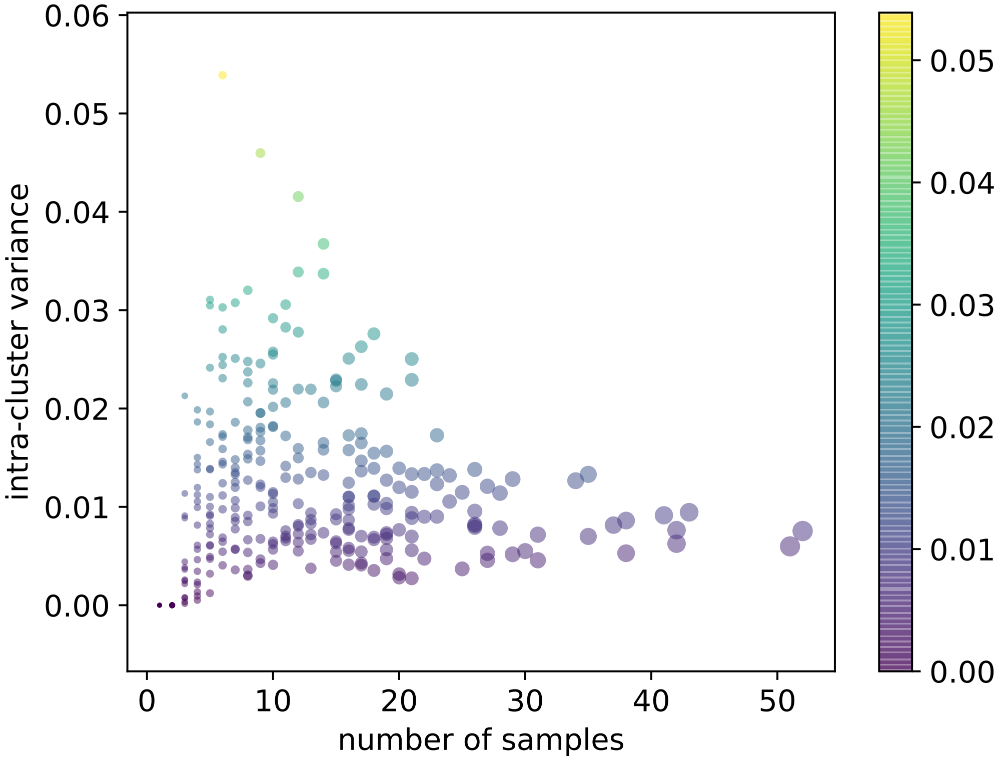


Figure 6-13: **Scatter Plot of Cluster Size vs. Variance** Plot indicates that clusters with smaller number of samples have a large range of variances, and vice versa. Bubble size maps to number of samples in cluster, and color maps to intra-cluster variance. We focus our analysis on clusters with the smallest variance to ensure that samples within a cluster are homogeneous.

cluster sizes with low intra-cluster variance to ensure homogeneity among samples in the cluster. Figure 6-14 displays 5 columns of spectrograms, with each column corresponding to a cluster. The first two columns belong to clusters that mostly capture dips (i.e. peak purity P is low) and contains rich speech activity. The next two columns (from mostly peaks) are noisy and contain short bursts of speech. The last column (also mostly peaks) displays speech activity with relatively more energy in high-frequency bands and with a gap for silence. This clustering of speech activity matches clustering found in the previous section (that was conducted via t-SNE). The ability to find such patterns via k-means indicates that the frame-level embeddings are encoding relationships that preserve the similarity and differences between frame-

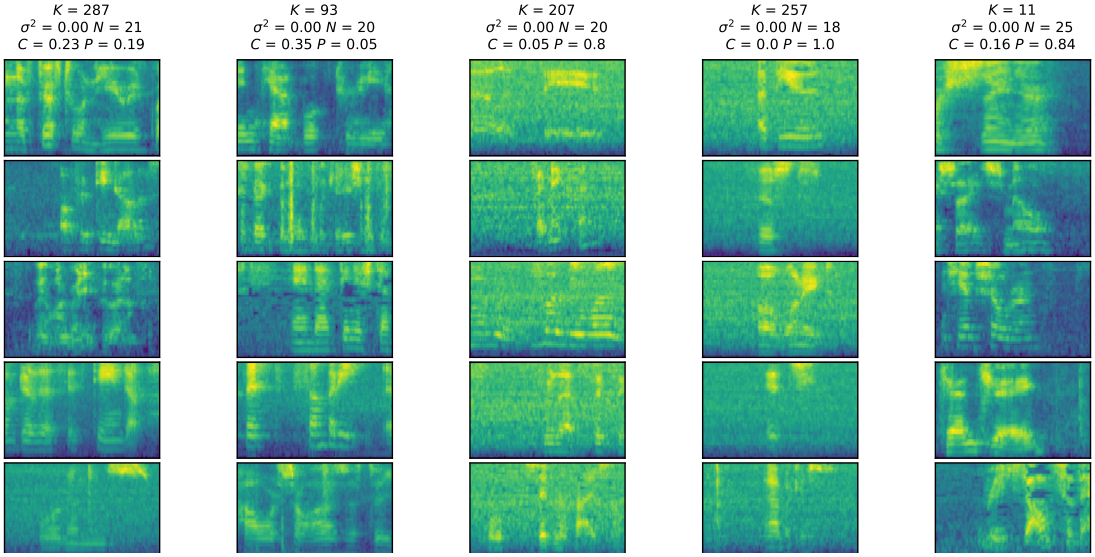


Figure 6-14: **Representative Spectrograms from k-means Clustering.** Spectrograms from clusters by k-means clustering for $k = 300$. Each column displays spectrograms from a cluster where the variance (σ^2) is the lowest, and where the number of samples in a cluster (N) is greater than 10. Spectrograms displayed are the closest distance to the cluster centroid and are speaker independent. **Predictive of healthy**: The two left columns are mostly from dips (i.e. P is small) and display a rich amount of speech activity. **Predictive of impairment**: The next two columns (from mostly peaks) are noisy and contain short bursts of speech. The last column (also mostly peaks) displays relatively more speech activity in high-frequency energy bands, with a gap for silence.

level embeddings. Our k-means analysis implies that methods that directly learn a similarity function between frames of an utterance, or even whole utterances, may be able to do so with strong effect (e.g. training a siamese network with contrastive error loss to learn similarities between speakers who have the same severity of cognitive impairment or the same pathology).

6.6.8 Neural Network Hyperparameters

Determining the optimum hyperparameters for a neural network is challenging due to the complex interactions between hyperparameters and outcome. Indeed, working rules like Occam's Razor (i.e. a simple model is best) are counter-intuitive for operating in this domain [Novak et al., 2018], and searching for hyperparameter optimization trends methodically (i.e. grid-search) are considered inefficient [Bergstra and Bengio, 2012]. Nevertheless, we present analysis that attempts to summarize patterns we find during our hyperparameter search. We generate boxplots (Figure 6-15) to capture distributions across 6 hyperparameters; number of convolution layers, learning rate, regularization type, dropout rate, global pooling operation, and local pooling operation (of the interim layers). The boxplots display on the y-axis the network AUC performance evaluated on the test set. Each row corresponds to separate sets of experiments.

If we consider the optimum number of layers (Figure 6-15a) in a network we find that AUC values were generally the same (hovering above AUC 0.5), a similar pattern emerges across different learning rates (Figure 6-15b). However, our optimized neural network appears in the outliers, within high number of layers and low learning rates.

Our next set of experiments explored introducing regularization and dropout into our network. If we consider the type of regularization (Figure 6-15c), we find a larger distribution of high AUC values when no regularization is used, compared to L1 and L2 regularization. For the dropout rate (Figure 6-15d), we observe that lower rates (e.g. 0.1 and 0.2) yields higher performing networks, compared to higher dropout rates.

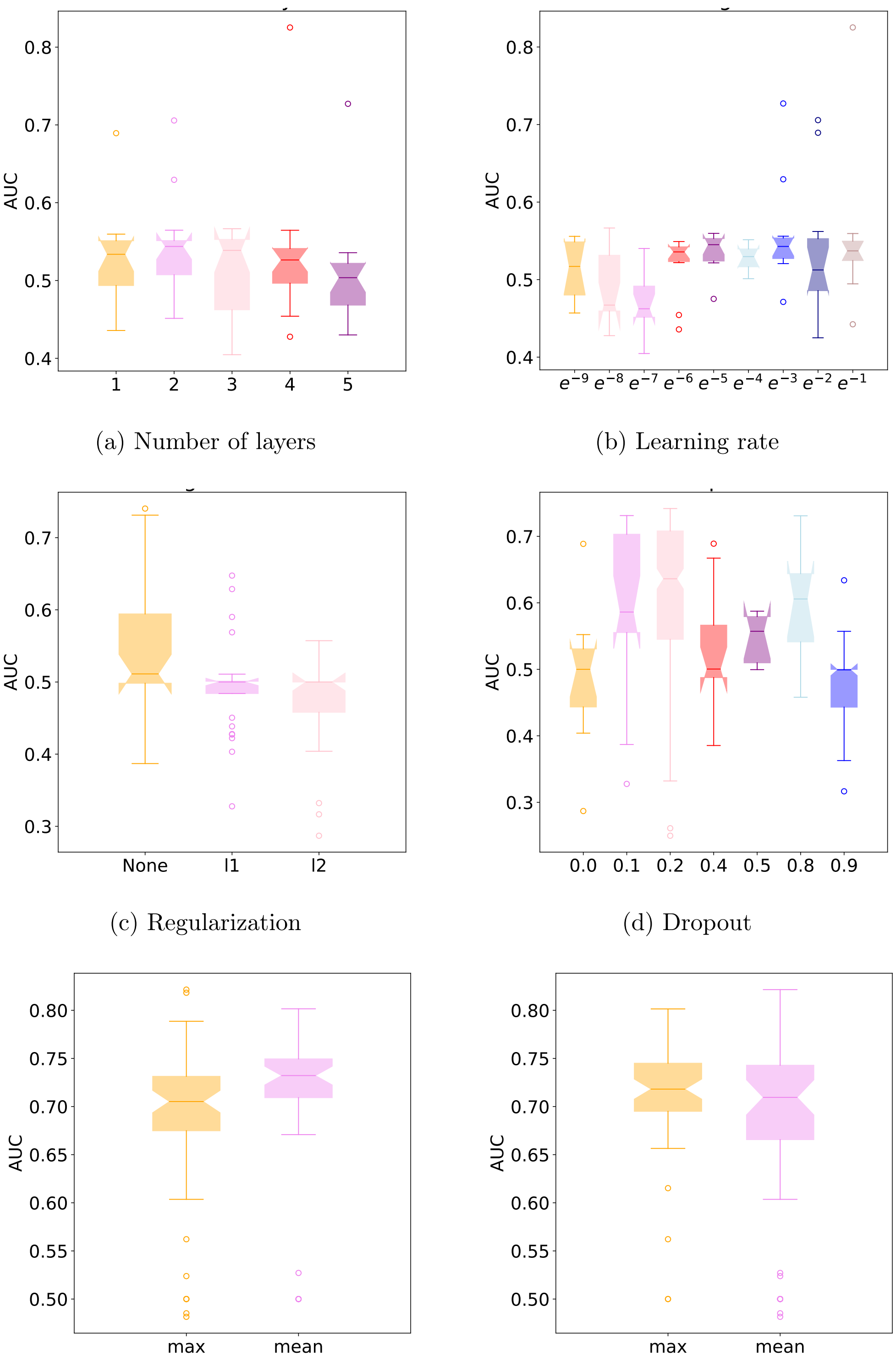
Once we established the number of layers and learning rate of our final network (defined in Section 6.4.4), we explored the effect of global pooling operations (Figure 6-15e), and observed that mean-pooling yielded on average higher AUC performance than max-pooling (which yielded more extreme performance values). This behavior was opposite in the case of local pooling (Figure 6-15f), where max-pooling had slightly higher mean performance, but mean-pooling yielded more extreme values.

These results highlight how neural network training tends to focus on finding *specific* parameters, that may not generalize when training slightly different topologies. Furthermore, these results motivate future work that explores how knowledge learned in one network may be transferred to a slightly different context; different set of audio recordings or multi-class modeling of cognitive impairment [Hoo-Chang et al., 2016].

6.7 Conclusion

In this chapter we utilized a CNN to learn salient representations of higher-order patterns in speech, and without the need for hand-crafted features. Our modeling approach captured information in speech to model cognitive impairment exhibiting high performance (AUC 0.85), on population-level data (5,000+ subjects) that was representative of the real-world prevalence of cognitive impairment (3.56%).

We also presented analysis that probed a neural network to understand the representations it had learned, and the types of speech the model considered indicative of cognitive impairment, which clustered according to: pause followed by onset of speech, short burst of speech, high frequency speech activity, and silence. Our results encourage further work in understanding the complex interactions that occur in speech, from patterns over a short window of speech (e.g. less than 1.5 seconds for phonetic and word-level resolution) to longer utterance level activity (e.g. 10 to 20 seconds for sentence-level resolution), or even longer-term changes in patterns over months and years [Ahmed et al., 2013].



(a) Number of layers

(b) Learning rate

(c) Regularization

(d) Dropout

(e) Global pooling

(f) Local pooling

Figure 6-15: **Statistics of CNN Hyperparameters**. Boxplots capture distributions of model performance across 6 hyperparameters; (a) number of layers, (b) learning rate, (c) regularization type, (d) dropout rate, (e) global pooling operation, and (f) local pooling operation (of the interim layers). y-axis is AUC performance evaluated on the test set. Each row is a separate set of experiments.

Chapter 7

Conclusions

Synopsis

In this chapter we provide a summary of our findings and discuss avenues that would be worth pursuing in future work, such as developing personalized models, evaluating changes over time, distinguishing between universal features and culturally dependant information, teaching a model to understand context, incorporating physician intuition, modeling with a combined set of modalities, and synthesizing speech. We remain reservedly optimistic that with increased data, we can perform more nuanced modeling of cognitive outcomes and underlying pathologies.

7.1 Summary of Findings

In the following two sub sections we highlight the findings presented in this thesis. The first set of findings contribute to an engineering understanding of data modeling paradigms and feature extraction approaches. The second set of findings contribute towards a scientific understanding of cognitive impairment and how it manifests in speech and language.

7.1.1 Reducing Data Labeling Constraints

Work in each chapter of this thesis was built on methods from previous chapters. We first explored in Chapter 3 the potential of modeling cognitive impairment from a relatively small set of 92 subjects with complete information on audio, transcripts, speaker turn, and utilizing hand-crafted features.

We loosened these constraints in Chapter 4, by performing modeling on a fraction of the audio recording, of which the speaker segments were generated through text-based diarization.

We next applied this diarization and segmentation method to the bronze set in Chapter 5 to extract audio features alone and model cognitive impairment on data representative of a population (5,000+ subjects) and the real-world prevalence of cognitive conditions (3-4%).

Finally, in Chapter 6 we did away with engineering of audio features (which had required voice activity detection, scaling, thresholding, calculating higher-order statistics, and segment-level representations) by training a CNN to learn higher-order representations from filterbank features to model cognitive impairment. For each chapter we found some interesting findings, beyond strong model performance.

7.1.2 Findings Beyond Performance

In Chapter 3 we found that combining audio and text features provided the best performance in detecting cognitive impairment (0.92 AUC). From these features we found that decreasing pitch, decreasing jitter, shorter speech segment lengths, and

and an increasing number of questions by the subject were positively associated with cognitive impairment.

In Chapter 4 we found that even with noisily transcribed data (81% WER) we could diarize audio with minimal confusion between speakers (4% confusion rate). This was because examiners and subjects operated in different roles during the neuropsychological exam, therefore they tended to generate different sets of vocabulary and sentence structure. Furthermore, we were able to determine a subject's cognitive condition using only a fraction of their recorded spoken interaction (about 150 seconds from an hour long recording).

In Chapter 5 we found that increased and monotonous speech activity in high-frequency spectral energy bands was associated with cognitive impairment. Further analyses showed that certain audio features correlated with speaking rate, and that from all tests, the 'verbal fluency' test was most predictive of cognitive impairment.

In Chapter 6 we utilized a CNN to learn salient representations of higher-order patterns in speech as utterance-level representations. We then probed the neural network model to understand the frame-level representations it had learned, and the types of speech the model considered indicative of cognitive impairment, which clustered according to: pause followed by onset of speech, short burst of speech, high frequency speech activity, and silence.

7.2 Cost Analysis of Model Performance

A missed detection of cognitive impairment may have costly effects in the long term. Neuropsychological exams are conducted every few years at best (if at all), thus a missed detection may prohibit an individual from adopting treatments (improving cardiovascular health [Langa and Levine, 2014, alz, 2018], increasing social engagement [Dubois et al., 2015]) to delay the onset of more severe cognitive impairment and plan for the future (defining caregiver responsibilities and structuring finances [Hirschman et al., 2008]).

In monetary terms, delayed detection can cost $64,000 per individual, which is an 18% increase in overall healthcare costs [alz, 2018]. There is also a cost to false detection of cognitive impairment, where an individual may be exposed to: further testing (e.g. imaging scans, spinal taps), misdiagnosis of the underlying pathology (e.g falsely detecting Alzheimer's disease), and even prescription of unnecessary medication [Rabinovici et al., 2017]. Financially, the cost of a false positive ranges from $3,000 to $10,000 to cover one year of testing and medication [Beck, 2014].

To balance the cost of missed detection and false positives, a model should perform with high sensitivity (correct detection) while maintaining high specificity (true negative). In Chapter 5, our top performing model (audio and demographic feature set) had the best overall sensitivity (89%) and specificity (81%) at Youden's Index, which would result in $11.3 million in savings[1] relative to the demographics-only model (99% sensitivity, 15% specificity). Our model performance is comparable to reported performance of physicians (30-35% sensitivity, 92% specificity) [Bouwmans and Weber, 2012], and screening tests (50-100% sensitivity, 33-100% specificity) [Cullen et al., 2007, Pinto and Peters, 2009, Aslam et al., 2018], albeit these studies were performed with a different set of subjects with a range of cognitive impairment severities, underlying pathologies, and comorbidities.

In Chapter 7, our speech-based models had higher sensitivity but lower specificity at Youden's Index relative to the demographic model (93% specificity, 65% sensitivity). Our best speech model (global max-pool CNN; 81% specificity, 82% sensitivity) relative to the demographic model may result in $415,440 of unnecessary screening[2], but may yield $467,840 in savings from positive detection[3]; an overall savings of approximately $50,000 for 1200 subjects.

While not all costs of true and missed detection may be captured in monetary terms, the above analysis helps anchor discussions of model performance in a quantifiable manner.

[1] (0.81-0.15) * 6,471 negative subjects * $3,000 per-false-positive + (0.89 - 0.99) * 234 CI subjects * $64,000 per-missed-detection

[2] (0.81-0.93) x 1154 negative subjects x $3,000 per-false-positive

[3] (0.82 - 0.65) x 43 positive subjects x $64,000 per-missed-detection

7.3 Future Work

We found a number of interesting results throughout our work in this thesis, but there remains even more interesting avenues to explore which we expand upon below, and which may provide a good starting point for the intrepid researcher.

7.3.1 Personalized Modeling

Our evaluation metrics were performed on data splits that were speaker independent, so no speaker in the test set had examples in the training set. This splitting of data shows that there are population-level patterns for cognitive impairment that generalize to other, unseen speakers. While this is very encouraging, speakers may have different levels of expression that are unique to their own style of communication. Therefore it may be more powerful to model a person's cognitive state with *respect* to their own baseline. Indeed, it has already been shown that models customized to a person's context improves classification performance, such as: incorporating personal information into speech-to-text systems (such as the names in a person's contacts list) improves speech recognition on mobile devices [McGraw et al., 2016], adapting the weights of a neural network to a speaker's patterns for Xbox voice-based search [Yu et al., 2013], incrementally adapting a model for personalized emotion detection [Kim et al., 2011], and calibrating medication dosages to an individual's health profile (e.g. age, weight, sex) [Ghassemi et al., 2018a]. Furthermore, a personalized model may incorporate information from the past, to monitor the progression of cognitive impairment over time.

7.3.2 Modeling over Time

One approach towards developing robust screening tools, is to not just rely on a single snapshot of spoken language interactions, but to aggregate over a longer period of time, to capture the average cognitive state and trajectory of an individual, to avoid spurious false positives and false negatives. Indeed, work that has looked at cognitive health of individuals over time has been able evaluate changes as observed in the diaries of the Nun study [Snowdon et al., 1996], through the writings of renowned fiction writers such as Agatha Christie [Le et al., 2011], and from the interviews of the boxer Cassius Clay [Berisha et al., 2017]. On a larger cohort, analysis has been performed to monitor changes in Parkinson's [Tsanas et al., 2010], Alzheimer's [Yancheva et al., 2015], and Vascular dementia [Walker et al., 2004].

7.3.3 Modeling Across Cultures and Languages

We presented methods that utilized data from speakers of English in the town of Framingham in Massachusetts in the USA. While studies to model cognitive impairment from speech have contained speakers of different languages and from different cultures (e.g. Colombia [Orozco-Arroyave et al., 2014], Canada [Thomas et al., 2005b], South Carolina USA [Rudzicz et al., 2014]), it is yet to be determined how a spoken language based model from one culture translates to another. Currently, it has been shown that many neuropyschological tests translate across cultures (such as the USA, UK, China, Japan, India, and Nigeria) [Guruje et al., 1995, Chan et al., 2003, Tsoi et al., 2015], but further analysis across different cultures and languages would help determine what spoken language features of cognitive impairment are universal, and which features may be contained to specific populations. Indeed, the discussion surrounding emotion across cultures (an outcome with a tangential relationship to cognitive impairment [Rock et al., 2014]) has been rich in determining what aspects of emotions are cultural and which are universal, indicating that there are some universally recognizable signals (such as facial expressions of Americans) [Elfenbein and Ambady, 2002]. In a similar spirit, studies on the semantics of language have found that some relationships between words (e.g. emotions, kinship) cut across cultures, while others are unique to the in-group (e.g. concept of 'shame' in Japan vs. USA) [Romney et al., 1996, Romney et al., 1997]; a cultural/linguistic mapping that may be relevant for evaluating responses to some of the neuropsychological tests that ask subjects to define relationships between objects.

7.3.4 Incorporating Physician Intuition

The ground truth in our work is generated by a panel of medical professionals, and is the oracle our model seeks to reach. The medical professionals had access to far more information and context than our model did (e.g. MRI scans, scores for each test), which is very encouraging for our own line of work. However, it has been shown that humans can be very unforgiving of algorithms that make errors, even if they perform better than humans [Dietvorst et al., 2015]. Such evaluations on ability and expertise seems to be deep-rooted, with children as young as 3 years old already drawing boundaries between the abilities of medical professionals and car mechanics [Koenig and Jaswal, 2011].

One opportunity to resolve this aversion to algorithms is by incorporating the intuition of medical professionals to both improve model performance, and translate the trust we already have for medical judgment into the algorithm. Indeed, there is evidence from the literature that medical professionals are utilizing more information than is contained in structured medical records (e.g. age, sex, diabetes, severity of illness) to arrive at their decisions. A feature as simple as sentiment contained in their medical notes has been shown to be associated with decisions that aid in prognostication (e.g. the number of medical imaging scans ordered), in addition to the structured information available to them [Ghassemi et al., 2018b]. This indicates that medical professionals are making judgments on observations that have yet to be formalized [Croskerry, 2009]. A simple approach to incorporating medical intuition would be to include their opinion as a feature in the classifier, or by mapping their thought process in the form of a decision tree and incorporating that classification outcome into a second comprehensive model.

7.3.5 Modeling Context

Our ability to model cognitive impairment relied heavily on acoustic patterns of speech, and was almost completely independent of the linguistic content of speech, and the test being administered. Our model performance under these constraints are very encouraging, yet to truly be able to model cognitive impairment at a deeper level requires determining moments when a subject had made an error (e.g. incorrect answer to a general knowledge question), the seriousness of their error (e.g. poor general knowledge or degenerating cognitive functions), and to what condition their error may be attributed to (e.g. tiredness, poor recall due to Alzheimer's, or poor hand-eye co-ordintation due to Parkinson's disease). To concretely illustrate this point, pigeons have been trained using differential food reinforcement learning to distinguish between benign and malignant human breast histopathology and to detect calcification in mammograms [Levenson et al., 2015], but do they truly understand what cancer is, where the image comes from, and what the underlying cause of a tumor is? Our speech-based model is similar to this pigeon. Although it may be powerful at detecting patterns in speech, it would need to be provided with far more background knowledge [Erhan et al., 2010], and imbued with hard-set rules [Sluckin, 2017], in order to perform higher order reasoning.

7.3.6 Multi-modal Modeling

Our work focused on speech and language to model cognitive impairment, but there remains a plethora of signals that may and do indeed contain information on a subject's cognitive state. Other researchers have looked at utilizing video to assess cognitive impairment from how stroke patients dressed [Walker et al., 2004], as well as observing digital pen recordings of the clock-drawing test to evaluate Parkinson's disease [Piers et al., 2017]. It stands to reason that a person's engagement with a specific task, as captured by gaze, movement, posture may contain further information on their cognitive state [Lotfi et al., 2012, Pampouchidou et al., 2017, Alhanai and Ghassemi, 2017]. With the ever increasing accessibility to sensors (e.g. video, motion, smart-watch) and the affordability of storing data, the development of pre-processing pipelines to filter signals for salient information and fuse modalities will become a riper domain to explore [Karpathy and Fei-Fei, 2015, Ngiam et al., 2011, Poria et al., 2015, Aytar et al., 2018, He et al., 2018]. For example, audio and text may contain useful information on an individual's mental state (e.g. depression), but not at an equivalent resolution of time (e.g. using the word 'sad' may flag for depression whereas montony requires a longer duration of speech to detect). Being able to capture this dual information in a complementary manner requires developing models that can accommodate the differing styles in which cognitive states may manifest in prosody and language [Al Hanai et al., 2018b].

7.3.7 Generating Synthetic Data

There are some limitations to how much data may be collected for modeling, for example generating ground-truth labels is a challenge because it requires specialized knowledge, and collecting health-related data requires advanced approval and informed consent. One strategy to circumvent these limitations is to develop data augmentation methods to improve generalizability. Standard approaches in speech recognition add various noise profiles to audio, increase and decrease the speed of recordings, and alter a speaker's pitch [Ko et al., 2015, Cui et al., 2015, Ko et al., 2017]. While such data augmentation strategies have shown to be effective, they may not be applicable to our own modeling objectives because these variances (acoustic profile, speaker speed and pitch) may be important for our model, since differences in pitch and speaking rate may be associated with cognitive impairment (as we observed in Chapter 4 and 5). A promising development that may better serve our modeling objectives has been the rise of neural network based generative models that learn latent representations to compactly model statistics on speech such as phones, gender, and noise profiles [Blaauw and Bonada, 2016, Hsu et al., 2016, Hsu et al., 2017]. Moreover, there has been efforts in synthesizing speech with autoregressive-based convolution neural networks [Van Den Oord et al., 2016], and generative adversarial networks [Pascual et al., 2017], that are sounding ever more natural and are already being deployed in voice-based assistive technologies [Martin, 2018]. These developments may be applicable to modeling cognitive impairment, as a method to learn potentially interpretable statistics on speech patterns, and generate examples that may be useful for training algorithms and humans in recognizing signs of cognitive impairment, its severity, and its progression over time. Such generative-based examples would maintain a speaker's privacy since any given example would be based on aggregate statistics rather than a single person's voice recording.

7.4 Reserved Optimism

When Alois Alzheimer started describing his first few patients with a condition that was to be named after him, he had only observed a small number of subjects from the many millions of humans that existed in his time. Indeed, evidence suggests that Johann F. (Alois' patient described in Chapter 2.1) was thought to have probable vascular dementia, rather than a different (undiscovered) pathology [Graeber et al., 1997]. It follows that in our own work, utilizing 5,000 subjects, remains relatively small to the many billions of people that have lived since the FHS was established. With such perspective, there still remains many samples of the population to explore before we can robustly differentiate between different pathologies of the brain.

Bibliography

[pan, 2012] (2012). pandas: Python Data Analysis Library. Online.

[alz, 2018] (2018). 2018 alzheimer's disease facts and figures. *Alzheimer's Dementia*, 14(3):367 – 429.

[Abadi et al., 2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Abdel-Hamid et al., 2014] Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545.

[Abe, 2010] Abe, S. (2010). Feature selection and extraction. In *Support Vector Machines for Pattern Classification*, pages 331–341. Springer.

[Ahmed et al., 2013] Ahmed, S., Haigh, A.-M. F., de Jager, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, 136(12):3727–3737.

[Al-Hameed et al., 2016] Al-Hameed, S., Benaissa, M., and Christensen, H. (2016). Simple and robust audio-based detection of biomarkers for Alzheimer's disease. In *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 32–36.

[Al-Hameed et al., 2017] Al-Hameed, S., Benaissa, M., and Christensen, H. (2017). Detecting and Predicting Alzheimer's Disease Severity in Longitudinal Acoustic Data. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, pages 57–61. ACM.

[Al Hanai et al., 2018a] Al Hanai, T., Au, R., and Glass, J. (2018a). Role-specific language models for processing recorded neuropsychological exams. In *Proceedings* *of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 746–752.

[Al Hanai et al., 2018b] Al Hanai, T., Ghassemi, M., and Glass, J. (2018b). Detecting depression with audio/text sequence modeling of interviews. In *Proc. Interspeech*, pages 1716–1720.

[Alhanai et al., 2017] Alhanai, T., Au, R., and Glass, J. (2017). Spoken language biomarkers for detecting cognitive impairment.

[Alhanai and Ghassemi, 2017] Alhanai, T. W. and Ghassemi, M. M. (2017). Predicting latent narrative mood using audio and physiologic data. In *AAAI*, pages 948–954.

[Anguera et al., 2012] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370.

[Aslam et al., 2018] Aslam, R. W., Bates, V., Dundar, Y., Hounsome, J., Richardson, M., Krishan, A., Dickson, R., Boland, A., Fisher, J., Robinson, L., and Sikdar, S. (2018). A systematic review of the diagnostic accuracy of automated tests for cognitive impairment. *Int J Geriatr Psychiatry*, 33(4):561–575. 29356098[pmid].

[Atay et al., 2015] Atay, C., Conway, E. R., Angus, D., Wiles, J., Baker, R., and Chenery, H. J. (2015). An automated approach to examining conversational dynamics between people with dementia and their carers. *PloS one*, 10(12):e0144327.

[Au et al., 2006] Au, R., Massaro, J. M., Wolf, P. A., Young, M. E., Beiser, A., Seshadri, S., DâĂŹAgostino, R. B., and DeCarli, C. (2006). Association of white matter hyperintensity volume with decreased cognitive functioning: the framingham heart study. *Archives of neurology*, 63(2):246–250.

[Aytar et al., 2018] Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., and Torralba, A. (2018). Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2303–2314.

[Aytar et al., 2016] Aytar, Y., Vondrick, C., and Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900.

[Bahrampour, 2017] Bahrampour, T. (2017). PET scans show many Alzheimer's patients may not actually have the disease. *The Washington Post.*

[Beck, 2014] Beck, M. (2014). How to bring the price of health care into the open. *Wall Street Journal.*

[Becker et al., 1994] Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

[Bellazzi and Zupan, 2008] Bellazzi, R. and Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97.

[Bellegarda, 2004] Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. *Speech communication*, 42(1):93–108.

[Bergstra and Bengio, 2012] Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

[Berisha et al., 2017] Berisha, V., Liss, J., Huston, T., Wisler, A., Jiao, Y., and Eig, J. (2017). Float like a butterfly sting like a bee: Changes in speech preceded parkinsonism diagnosis for Muhammad Ali. *Proc. Interspeech 2017*, pages 1809–1813.

[Besacier et al., 2014] Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

[Bialystok et al., 2007] Bialystok, E., Craik, F. I., and Freedman, M. (2007). Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia*, 45(2):459–464.

[Bisani and Ney, 2008] Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

[Blaauw and Bonada, 2016] Blaauw, M. and Bonada, J. (2016). Modeling and transforming speech using variational autoencoders. In *INTERSPEECH*, pages 1770–1774.

[Boller and Forbes, 1998] Boller, F. and Forbes, M. M. (1998). History of dementia and dementia in history: an overview. *Journal of the neurological sciences*, 158(2):125–133.

[Bot et al., 2016] Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E. R., et al. (2016). The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific data*, 3:160011.

[Bouwmans and Weber, 2012] Bouwmans, A. E. and Weber, W. E. (2012). Neurologists' diagnostic accuracy of depression and cognitive problems in patients with parkinsonism. *BMC Neurology*, 12(1):37.

[Braak and Del Tredici, 2012] Braak, H. and Del Tredici, K. (2012). Where, when, and in what form does sporadic Alzheimer's disease begin? *Current opinion in neurology*, 25(6):708–714.

[Bulat and Tzimiropoulos, 2016] Bulat, A. and Tzimiropoulos, G. (2016). Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer.

[Cabitza et al., 2017] Cabitza, F., Rasoini, R., and Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518.

[Caplan et al., 2006] Caplan, G. A., Meller, A., Squires, B., Chan, S., and Willett, W. (2006). Advance care planning and hospital in the nursing home. *Age and ageing*, 35(6):581–585.

[Carlini and Wagner, 2017] Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE.

[Cerhan et al., 2002] Cerhan, J., Ivnik, R., Smith, G., Tangalos, E., Petersen, R., and Boeve, B. (2002). Diagnostic utility of letter fluency, category fluency, and fluency difference scores in alzheimer's disease. *The Clinical neuropsychologist*, 16(1):35–42.

[Chan et al., 2003] Chan, A. S., Shum, D., and Cheung, R. W. (2003). Recent development of cognitive and neuropsychological assessment in asian countries. *Psychological Assessment*, 15(3):257.

[Chan et al., 2016] Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4960–4964. IEEE.

[Chang et al., 2009] Chang, E., Daly, J., Johnson, A., Harrison, K., Easterbrook, S., Bidewell, J., Stewart, H., Noel, M., and Hancock, K. (2009). Challenges for professional care of advanced dementia. *International Journal of Nursing Practice*, 15(1):41–47.

[Chen and Salman, 2011] Chen, K. and Salman, A. (2011). Extracting speaker-specific information with a regularized siamese deep network. In *Advances in Neural Information Processing Systems*, pages 298–306.

[Chen and Goodman, 1999] Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

[Cheng et al., 2016] Cheng, A., Yang, Y., Zhou, Y., Maharana, C., Lu, D., Peng, W., Liu, Y., Wan, R., Marosi, K., Misiak, M., et al. (2016). Mitochondrial sirt3 mediates adaptive responses of neurons to exercise and metabolic and excitatory challenges. *Cell metabolism*, 23(1):128–142.

[Child, 1990] Child, D. (1990). *The essentials of factor analysis.* Cassell Educational.

[Chollet, 2015] Chollet, F. (2015). keras. https://github.com/fchollet/keras.

[Christensen et al., 2012] Christensen, H., Cunningham, S., Fox, C., Green, P., and Hain, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. In *Thirteenth Annual Conference of the International Speech Communication Association.*

[Chuang et al., 2016] Chuang, Y.-F., An, Y., Bilgel, M., Wong, D. F., Troncoso, J. C., O'Brien, R. J., Breitner, J., Ferruci, L., Resnick, S. M., and Thambisetty, M. (2016). Midlife adiposity predicts earlier onset of Alzheimer's dementia, neuropathology and presymptomatic cerebral amyloid accumulation. *Molecular psychiatry*, 21(7):910.

[Clark et al., 2003] Clark, C. M., Xie, S., Chittams, J., Ewbank, D., Peskind, E., Galasko, D., Morris, J. C., McKeel, D. W., Farlow, M., Weitlauf, S. L., et al. (2003). Cerebrospinal fluid tau and β-amyloid: how well do these biomarkers reflect autopsy-confirmed dementia diagnoses? *Archives of neurology*, 60(12):1696–1702.

[Cobb et al., 1995] Cobb, J., Wolf, P. A., Au, R., White, R., and D'agostino, R. (1995). The effect of education on the incidence of dementia and Alzheimer's disease in the Framingham Study. *Neurology*, 45(9):1707–1712.

[Cohen-Boulakia et al., 2017] Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Le Bras, Y., Lemoine, F., et al. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, 75:284–298.

[Conroy et al., 2003] Conroy, R., Pyörälä, K., Fitzgerald, A. e., Sans, S., Menotti, A., De Backer, G., De Bacquer, D., Ducimetiere, P., Jousilahti, P., Keil, U., et al. (2003). Estimation of ten-year risk of fatal cardiovascular disease in europe: the score project. *European heart journal*, 24(11):987–1003.

[Corley and Stewart, 2008] Corley, M. and Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602.

[Croskerry, 2009] Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic medicine*, 84(8):1022–1028.

[Crossley et al., 1997] Crossley, M., D'arcy, C., and Rawson, N. S. (1997). Letter and category fluency in community-dwelling canadian seniors: A comparison of normal participants to those with dementia of the Alzheimer or vascular type. *Journal of Clinical and Experimental Neuropsychology*, 19(1):52–62. PMID: 9071641.

[Cui et al., 2015] Cui, X., Goel, V., and Kingsbury, B. (2015). Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(9):1469–1477.

[Cullen et al., 2007] Cullen, B., O'Neill, B., Evans, J. J., Coen, R. F., and Lawlor, B. A. (2007). A review of screening tests for cognitive impairment. *J Neurol Neurosurg Psychiatry*, 78(8):790–799. 17178826[pmid].

[Cummins et al., 2015] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

[Davis and Goadrich, 2006] Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

[DeLong et al., 1988] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

[Dezeure et al., 2015] Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N., et al. (2015). High-dimensional inference: Confidence intervals, *p*-values and r-software hdi. *Statistical science*, 30(4):533–558.

[Dietvorst et al., 2015] Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

[Downer et al., 2015] Downer, B., Fardo, D. W., and Schmitt, F. A. (2015). A summary score for the framingham heart study neuropsychological battery. *Journal of aging and health*, 27(7):1199–1222.

[Dubois et al., 2015] Dubois, B., Padovani, A., Scheltens, P., Rossi, A., and Dell'Agnello, G. (2015). Timely Diagnosis for Alzheimer's Disease: A Literature Review on Benefits and Challenges. *J Alzheimers Dis*, 49(3):617–631. 26484931[pmid].

[Dunning, 1994] Dunning, T. (1994). *Statistical identification of language*. Computing Research Laboratory, New Mexico State University.

[Dynan, 2000] Dynan, K. E. (2000). Habit formation in consumer preferences: Evidence from panel data. *American Economic Review*, 90(3):391–406.

[Elfenbein and Ambady, 2002] Elfenbein, H. A. and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203.

[Elias et al., 2000] Elias, M. F., Beiser, A., Wolf, P. A., Au, R., White, R. F., and D'agostino, R. B. (2000). The preclinical phase of Alzheimer disease: a 22-year prospective study of the Framingham Cohort. *Archives of neurology*, 57(6):808–813.

[Elmore et al., 1998] Elmore, J. G., Barton, M. B., Moceri, V. M., Polk, S., Arena, P. J., and Fletcher, S. W. (1998). Ten-year risk of false positive screening mammograms and clinical breast examinations. *New England Journal of Medicine*, 338(16):1089–1096.

[Erhan et al., 2010] Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660.

[Etters et al., 2008] Etters, L., Goodall, D., and Harrison, B. E. (2008). Caregiver burden among dementia patient caregivers: a review of the literature. *Journal of the American Association of Nurse Practitioners*, 20(8):423–428.

[Ewers et al., 2012] Ewers, M., Walsh, C., Trojanowski, J. Q., Shaw, L. M., Petersen, R. C., Jack, C. R. J., Feldman, H. H., Bokde, A. L. W., Alexander, G. E., Scheltens, P., Vellas, B., Dubois, B., Weiner, M., Hampel, H., and (ADNI), N. A. A. D. N. I. (2012). Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiol Aging*, 33(7):1203–1214. 21159408[pmid].

[Eyben, 2015] Eyben, F. (2015). *Real-time speech and music classification by large audio feature space extraction*. Springer.

[Eyben et al., 2010] Eyben, F., Wollmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

[Farrer et al., 1997] Farrer, L. A., Cupples, L. A., Haines, J. L., Hyman, B., Kukull, W. A., Mayeux, R., Myers, R. H., Pericak-Vance, M. A., Risch, N., and Van Duijn, C. M. (1997). Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis. *Jama*, 278(16):1349–1356.

[Feydy et al., 2002] Feydy, A., Carlier, R., Roby-Brami, A., Bussel, B., Cazalis, F., Pierot, L., Burnod, Y., and Maier, M. (2002). Longitudinal study of motor recovery after stroke: recruitment and focusing of brain activation. *Stroke*, 33(6):1610–1617.

[Fowler and Christakis, 2008] Fowler, J. H. and Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *Bmj*, 337:a2338.

[Fraser et al., 2014] Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., and Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60.

[Fraser et al., 2016] Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

[Fraser et al., 2013] Fraser, K. C., Rudzicz, F., and Rochon, E. (2013). Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *INTERSPEECH*, pages 2177–2181.

[Fratiglioni et al., 2004] Fratiglioni, L., Paillard-Borg, S., and Winblad, B. (2004). An active and socially integrated lifestyle in late life might protect against dementia. *The Lancet Neurology*, 3(6):343–353.

[Friedman et al., 2009] Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4).

[Friedman, 1997] Friedman, J. H. (1997). On bias, variance, 0/1âĂŤloss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77.

[Gatys et al., 2015] Gatys, L., Ecker, A., and Bethge, M. (2015). A neural algorithm of artistic style. *Nature Communications*.

[Ghassemi et al., 2018a] Ghassemi, M., AlHanai, T., Westover, M., Mark, R., and Nemati, S. (2018a). Personalized medication dosing using volatile data streams.

[Ghassemi et al., 2018b] Ghassemi, M. M., Al-Hanai, T., Raffa, J. D., Mark, R. G., Nemati, S., and Chokshi, F. H. (2018b). How is the doctor feeling? icu provider sentiment is associated with diagnostic imaging utilization. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4058–4064. IEEE.

[Glass et al., 2004] Glass, J., Hazen, T. J., Hetherington, L., and Wang, C. (2004). Analysis and processing of lecture audio data: Preliminary investigations. In *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*, pages 9–12. Association for Computational Linguistics.

[Gómez-Vilda et al., 2017] Gómez-Vilda, P., de Ipiña, M. L., Rodellar-Biarge, V., Palacios-Alonso, D., and Ecay-Torres, M. (2017). Articulation characterization in ad speech production. In *Converging Clinical and Engineering Research on Neurorehabilitation II*, pages 861–866. Springer.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

[Graeber et al., 1997] Graeber, M., Kösel, S., Egensperger, R., Banati, R., Müller, U., Bise, K., Hoff, P., Möller, H., Fujisawa, K., and Mehraein, P. (1997). Rediscovery of the case described by Alois Alzheimer in 1911: historical, histological and molecular genetic analysis. *Neurogenetics*, 1(1):73–80.

[Graves and Jaitly, 2014] Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pages 1764–1772.

[Gregory et al., 2007] Gregory, R., Roked, F., Jones, L., and Patel, A. (2007). Is the degree of cognitive impairment in patients with Alzheimer's disease related to their capacity to appoint an enduring power of attorney? *Age and ageing*, 36(5):527–531.

[Grundman et al., 2004] Grundman, M., Petersen, R. C., Ferris, S. H., Thomas, R. G., Aisen, P. S., Bennett, D. A., Foster, N. L., Jack Jr, C. R., Galasko, D. R., Doody, R., et al. (2004). Mild cognitive impairment can be distinguished from Alzheimer disease and normal aging for clinical trials. *Archives of neurology*, 61(1):59–66.

[Guo et al., 2017] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330.

[Guo et al., 1999] Guo, Z., Fratiglioni, L., Zhu, L., Fastbom, J., Winblad, B., and Viitanen, M. (1999). Occurrence and progression of dementia in a community population aged 75 years and older: relationship of antihypertensive medication use. *Archives of neurology*, 56(8):991–996.

[Guruje et al., 1995] Guruje, O., Unverzargt, F., Osuntokun, B., Hendrie, H., Baiyewu, O., Ogunniyi, A., and Hali, K. (1995). The cerad neuropsychological test battery: norms from a yoruba-speaking nigerian sample. *West African Journal of Medicine*, 14(1):29–33.

[Haan et al., 1999] Haan, M. N., Shemanski, L., Jagust, W. J., Manolio, T. A., and Kuller, L. (1999). The role of apoeâĹŁ 4 in modulating effects of other risk factors for cognitive decline in elderly persons. *Jama*, 282(1):40–46.

[Haixiang et al., 2017] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239.

[Hall et al., 2009] Hall, C., Lipton, R., Sliwinski, M., Katz, M., Derby, C., and Verghese, J. (2009). Cognitive activities delay onset of memory decline in persons who develop dementia. *Neurology*, 73(5):356–361.

[Hanley and McNeil, 1982] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

[Harwath et al., 2018] Harwath, D., Recasens, A., D. Suris, G. C., Torralba, A., and Glass, J. (2018). Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European Conference on Computer Vision*, pages 659–677.

[Hastie and Qian, 2014] Hastie, T. and Qian, J. (2014). Glmnet vignette.

[He et al., 2018] He, W., Motlicek, P., and Odobez, J.-M. (2018). Joint localization and classification of multiple sound sources using a multi-task neural network. *Proc. Interspeech 2018*, pages 312–316.

[Hebert et al., 2001] Hebert, L. E., Beckett, L. A., Scherr, P. A., and Evans, D. A. (2001). Annual incidence of Alzheimer disease in the United States projected to the years 2000 through 2050. *Alzheimer Disease & Associated Disorders*, 15(4):169–173.

[Hernandez-Dominguez et al., 2018] Hernandez-Dominguez, L., Ratte, S., Sierra-Martinez, G., and Roche-Bergua, A. (2018). Computer-based evaluation of AlzheimerâĂŹs disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:260–268.

[Hinton et al., 2012] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

[Hippius and Neundörfer, 2003] Hippius, H. and Neundörfer, G. (2003). The discovery of Alzheimer's disease. *Dialogues in clinical neuroscience*, 5(1):101.

[Hirschman et al., 2008] Hirschman, K. B., Kapo, J. M., and Karlawish, J. H. (2008). Identifying the factors that facilitate or hinder advance planning by persons with dementia. *Alzheimer disease and associated disorders*, 22(3):293.

[Hoffmann et al., 2010] Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., and Kálmán, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *International journal of speech-language pathology*, 12(1):29–34.

[Hoffrage et al., 2000] Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information.

[Hofman et al., 1997] Hofman, A., Ott, A., Breteler, M. M., Bots, M. L., Slooter, A. J., van Harskamp, F., van Duijn, C. N., Van Broeckhoven, C., and Grobbee, D. E. (1997). Atherosclerosis, apolipoprotein E, and prevalence of dementia and Alzheimer's disease in the Rotterdam Study. *The Lancet*, 349(9046):151–154.

[Hoo-Chang et al., 2016] Hoo-Chang, S., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285.

[Horley et al., 2010] Horley, K., Reid, A., and Burnham, D. (2010). Emotional prosody perception and production in dementia of the Alzheimer's type. *Journal of Speech, Language, and Hearing Research*, 53(5):1132–1146.

[Hosmer and Lemesbow, 1980] Hosmer, D. W. and Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10):1043–1069.

[Hosmer Jr et al., 2013] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

[Hsu et al., 2016] Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., and Wang, H.-M. (2016). Voice conversion from non-parallel corpora using variational auto-encoder. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*, pages 1–6. IEEE.

[Hsu et al., 2017] Hsu, W.-N., Zhang, Y., and Glass, J. (2017). Learning latent representations for speech generation and transformation. In *Interspeech*, pages 1273–1277.

[Hua et al., 2004] Hua, J., Xiong, Z., Lowey, J., Suh, E., and Dougherty, E. R. (2004). Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*, 21(8):1509–1515.

[Huang and Ling, 2005] Huang, J. and Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310.

[Hunter, 2007] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

[Hurd et al., 2013] Hurd, M. D., Martorell, P., Delavande, A., Mullen, K. J., and Langa, K. M. (2013). Monetary costs of dementia in the United States. *New England Journal of Medicine*, 368(14):1326–1334.

[Iadecola, 2010] Iadecola, C. (2010). The overlap between neurodegenerative and vascular factors in the pathogenesis of dementia. *Acta neuropathologica*, 120(3):287–296.

[Illes, 1989] Illes, J. (1989). Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, Huntington's, and Parkinson's. *Brain and language*, 37(4):628–642.

[Jack Jr et al., 2009] Jack Jr, C. R., Lowe, V. J., Weigand, S. D., Wiste, H. J., Senjem, M. L., Knopman, D. S., Shiung, M. M., Gunter, J. L., Boeve, B. F., Kemp, B. J., et al. (2009). Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. *Brain*, 132(5):1355–1365.

[Jansen et al., 2017] Jansen, A., Plakal, M., Pandya, R., Ellis, D. P., Hershey, S., Liu, J., Moore, R. C., and Saurous, R. A. (2017). Towards learning semantic audio representations from unlabeled data. *signal*, 2(3):7–11.

[Jellinger, 2010] Jellinger, K. A. (2010). Should the word ‘dementia’ be forgotten? *Journal of cellular and molecular medicine*, 14(10):2415–2416.

[Jessen et al., 2010] Jessen, F., Wiese, B., Bachmann, C., Eifflaender-Gorfer, S., Haller, F., Kölsch, H., Luck, T., Mösch, E., van den Bussche, H., Wagner, M., et al. (2010). Prediction of dementia by subjective memory impairment: effects of severity and temporal association with cognitive impairment. *Archives of general psychiatry*, 67(4):414–422.

[Jones et al., 01] Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python.

[Jozefowicz et al., 2015] Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning*, pages 2342–2350.

[Jurafsky and Martin, 2014] Jurafsky, D. and Martin, J. H. (2014). *Speech and language processing*, volume 3. Pearson London.

[Kannel, 2000] Kannel, W. B. (2000). Fifty years of Framingham Study contributions to understanding hypertension. *Journal of human hypertension*, 14(2):83.

[Kannel et al., 1961] Kannel, W. B., Dawber, T. R., Kagan, A., Revotskie, N., and Stokes, J. (1961). Factors of risk in the development of coronary heart disease - six-year follow-up experience: the Framingham Study. *Annals of internal medicine*, 55(1):33–50.

[Kannel and Sorlie, 1979] Kannel, W. B. and Sorlie, P. (1979). Some health benefits of physical activity: the Framingham Study. *Archives of internal medicine*, 139(8):857–861.

[Karlekar et al., 2018] Karlekar, S., Niu, T., and Bansal, M. (2018). Detecting Linguistic Characteristics of Alzheimer’s Dementia by Interpreting Neural Models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 701–707.

[Karpathy and Fei-Fei, 2015] Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

[Karpathy et al., 2014] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

[Kawas et al., 2015] Kawas, C. H., Kim, R. C., Sonnen, J. A., Bullain, S. S., Trieu, T., and Corrada, M. M. (2015). Multiple pathologies are common and related to dementia in the oldest-old the 90+ study. *Neurology*, 85(6):535–542.

[Kearns and Ron, 1999] Kearns, M. and Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, 11(6):1427–1453.

[Keshava, 2017] Keshava, M. (2017). FDA approves sale of genetic tests for risk of Alzheimer's and other diseases. *Stat*.

[Kim et al., 2009] Kim, J., Basak, J. M., and Holtzman, D. M. (2009). The role of apolipoprotein E in Alzheimer's disease. *Neuron*, 63(3):287–303.

[Kim et al., 2011] Kim, J.-B., Park, J.-S., and Oh, Y.-H. (2011). On-line speaker adaptation based emotion recognition using incremental emotional information. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4948–4951. IEEE.

[Knopman et al., 2003] Knopman, D. S., Parisi, J. E., Boeve, B. F., Cha, R. H., Apaydin, H., Salviati, A., Edland, S. D., and Rocca, W. A. (2003). Vascular dementia in a population-based autopsy study. *Archives of Neurology*, 60(4):569–575.

[Ko et al., 2015] Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[Ko et al., 2017] Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE*, pages 5220–5224.

[Koenig and Jaswal, 2011] Koenig, M. A. and Jaswal, V. K. (2011). Characterizing childrenâĂŹs expectations about expertise and incompetence: Halo or pitchfork effects? *Child Development*, 82(5):1634–1647.

[König et al., 2015] König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P. H., et al. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124.

[Kurlowicz et al., 1999] Kurlowicz, L., Wallace, M., et al. (1999). The mini-mental state examination (mmse). *Journal of gerontological nursing*, 25(5):8–9.

[Langa and Levine, 2014] Langa, K. M. and Levine, D. A. (2014). The diagnosis and management of mild cognitive impairment: a clinical review. *JAMA*, 312(23):2551–2561. 25514304[pmid].

[Larson et al., 2006] Larson, E. B., Wang, L., Bowen, J. D., McCormick, W. C., Teri, L., Crane, P., and Kukull, W. (2006). Exercise is associated with reduced risk for incident dementia among persons 65 years of age and older. *Annals of internal medicine*, 144(2):73–81.

[Laws et al., 2009] Laws, K., Duncan, A., and Gale, T. (2009). ‘Normal’ semanticâĂŞphonemic fluency discrepancy in Alzheimer’s disease? A meta-analytic study. *Cortex; a journal devoted to the study of the nervous system and behavior*, 46:595–601.

[Le et al., 2011] Le, X., Lancashire, I., Hirst, G., and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Literary and Linguistic Computing*, 26(4):435–461.

[Lehr et al., 2012] Lehr, M., Prud’hommeaux, E., Shafran, I., and Roark, B. (2012). Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *Thirteenth Annual Conference of the International Speech Communication Association*.

[Lei et al., 2018] Lei, T., Zhang, Y., Wang, S. I., Dai, H., and Artzi, Y. (2018). Simple recurrent units for highly parallelizable recurrence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4470–4481.

[Levenson et al., 2015] Levenson, R. M., Krupinski, E. A., Navarro, V. M., and Wasserman, E. A. (2015). Pigeons (columba livia) as trainable observers of pathology and radiology breast cancer images. *PLoS One*, 10(11):e0141357.

[Liu et al., 2006] Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on audio, speech, and language processing*, 14(5):1526–1540.

[Locke and Latham, 2002] Locke, E. A. and Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist*, 57(9):705.

[Lopez-de Ipiña et al., 2015] Lopez-de Ipiña, K., Alonso, J. B., Solé-Casals, J., Barroso, N., Henriquez, P., Faundez-Zanuy, M., Travieso, C. M., Ecay-Torres, M., Martinez-Lage, P., and Eguiraun, H. (2015). On automatic diagnosis of AlzheimerâĂŹs disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, 7(1):44–55.

[López-de Ipiña et al., 2013] López-de Ipiña, K., Alonso, J.-B., Travieso, C. M., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., et al. (2013). On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors*, 13(5):6730–6745.

[Lotfi et al., 2012] Lotfi, A., Langensiepen, C., Mahmoud, S. M., and Akhlaghinia, M. J. (2012). Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour. *Journal of ambient intelligence and humanized computing*, 3(3):205–218.

[Lu et al., 2013] Lu, D., Fall, K., Sparen, P., Ye, W., Adami, H.-O., Valdimarsdottir, U., and Fang, F. (2013). Suicide and suicide attempt after a cancer diagnosis among young individuals. *Annals of oncology*, 24(12):3112–3117.

[Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[Mahmood et al., 2014] Mahmood, S. S., Levy, D., Vasan, R. S., and Wang, T. J. (2014). The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, 383(9921):999–1008.

[Mangialasche et al., 2010] Mangialasche, F., Solomon, A., Winblad, B., Mecocci, P., and Kivipelto, M. (2010). Alzheimer's disease: clinical trials and drug development. *The Lancet Neurology*, 9(7):702–716.

[Martin, 2018] Martin, T. (2018). How to get all of google assistantâĂŹs new voices right now.

[MATLAB, 2010] MATLAB (2010). The MathWorks Inc., Natick, Massachusetts.

[Maxim et al., 2014] Maxim, L. D., Niebo, R., and Utell, M. J. (2014). Screening tests: a review with examples. *Inhalation toxicology*, 26(13):811–828.

[McFee et al., 2015] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25.

[McGraw et al., 2016] McGraw, I., Prabhavalkar, R., Alvarez, R., Arenas, M. G., Rao, K., Rybach, D., Alsharif, O., Sak, H., Gruenstein, A., Beaufays, F., et al. (2016). Personalized speech recognition on mobile devices. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5955–5959. IEEE.

[McKhann et al., 2011] McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., et al. (2011). The diagnosis of dementia due to AlzheimerâĂŹs disease: Recommendations from the National Institute on Aging-AlzheimerâĂŹs Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, 7(3):263–269.

[Meilán et al., 2014] Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., and Arana, J. M. (2014). Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.

[Mirheidari et al., 2018] Mirheidari, B., Blackburn, D., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2018). Detecting signs of dementia using word vector representations. *Proc. Interspeech 2018*, pages 1893–1897.

[Mohri et al., 2002] Mohri, M., Pereira, F., and Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

[Moller and Graeber, 1998] Moller, H.-J. and Graeber, M. B. (1998). The case described by alois alzheimer in 1911. *European Archives of Psychiatry and Clinical Neuroscience*, 248(3):111–122.

[Moosavi-Dezfooli et al., 2016] Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582.

[Morris, 2009] Morris, M. (2009). The role of nutrition in Alzheimer's disease: epidemiological evidence. *European Journal of Neurology*, 16(s1):1–7.

[Mosconi et al., 2008] Mosconi, L., Tsui, W. H., Herholz, K., Pupi, A., Drzezga, A., Lucignani, G., Reiman, E. M., Holthoff, V., Kalbe, E., Sorbi, S., et al. (2008). Multicenter standardized 18F-FDG PET diagnosis of mild cognitive impairment, Alzheimer's disease, and other dementias. *Journal of Nuclear Medicine*, 49(3):390–398.

[Myers et al., 1996] Myers, R., Schaefer, E., Wilson, P., d'Agostino, R., Ordovas, J., Espino, A., Au, R., White, R., Knoefel, J., Cobb, J., et al. (1996). Apolipoprotein e element 4 association with dementia in a population-based study the framingham study. *Neurology*, 46(3):673–677.

[Nakov et al., 2016] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., and Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18.

[Nestor et al., 2004] Nestor, P. J., Scheltens, P., and Hodges, J. R. (2004). Advances in the early detection of Alzheimer's disease. *Nature medicine*, 10(7):S34.

[Newman et al., 2005] Newman, A. B., Fitzpatrick, A. L., Lopez, O., Jackson, S., Lyketsos, C., Jagust, W., Ives, D., DeKosky, S. T., and Kuller, L. H. (2005). Dementia and Alzheimer's disease incidence in relationship to cardiovascular disease in the Cardiovascular Health Study cohort. *Journal of the American Geriatrics Society*, 53(7):1101–1107.

[Ng, 2004] Ng, A. Y. (2004). Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM.

[Ngiam et al., 2011] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

[Novak et al., 2018] Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. (2018). Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*.

[Olden and Jackson, 2002] Olden, J. D. and Jackson, D. A. (2002). Illuminating the âĂIJblack boxâĂİ: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, 154(1-2):135–150.

[Oliphant, 06] Oliphant, T. (2006–). NumPy: A guide to NumPy. USA: Trelgol Publishing.

[Orimaye et al., 2016] Orimaye, S. O., Wong, J. S.-M., and Fernandez, J. S. G. (2016). Deep-deep neural network language models for predicting mild cognitive impairment. In *BAI@ IJCAI*, pages 14–20.

[Orimaye et al., 2014] Orimaye, S. O., Wong, J. S.-M., and Golden, K. J. (2014). Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 78–87.

[Orozco-Arroyave et al., 2014] Orozco-Arroyave, J. R., Arias-Londoño, J. D., Bonilla, J. F. V., Gonzalez-Rátiva, M. C., and Nöth, E. (2014). New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *LREC*, pages 342–347.

[Ott et al., 1995] Ott, A., Breteler, M. M., Van Harskamp, F., Claus, J. J., Van Der Cammen, T. J., Grobbee, D. E., and Hofman, A. (1995). Prevalence of Alzheimer's disease and vascular dementia: association with education. The Rotterdam study. *Bmj*, 310(6985):970–973.

[Pakhomov and Hemmy, 2014] Pakhomov, S. V. and Hemmy, L. S. (2014). A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex*, 55:97–106.

[Pakhomov et al., 2010] Pakhomov, S. V., Smith, G. E., Marino, S., Birnbaum, A., Graff-Radford, N., Caselli, R., Boeve, B., and Knopman, D. S. (2010). A computerized technique to assess language use patterns in patients with frontotemporal dementia. *Journal of neurolinguistics*, 23(2):127–144.

[Pampouchidou et al., 2017] Pampouchidou, A., Simos, P., Marias, K., Meriaudeau, F., Yang, F., Pediaditis, M., and Tsiknakis, M. (2017). Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing*.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Pascual et al., 2017] Pascual, S., Bonafonte, A., and Serrà, J. (2017). Segan: Speech enhancement generative adversarial network. *Proc. Interspeech 2017*, pages 3642–3646.

[Paszke et al., 2017] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *NIPS-W*.

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

[Perl, 2010] Perl, D. P. (2010). Neuropathology of Alzheimer's disease. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine*, 77(1):32–42.

[Piers et al., 2017] Piers, R. J., Devlin, K. N., Ning, B., Liu, Y., Wasserman, B., Massaro, J. M., Lamar, M., Price, C. C., Swenson, R., Davis, R., et al. (2017). Age and Graphomotor Decision Making Assessed with the Digital Clock Drawing Test: The Framingham Heart Study. *Journal of Alzheimer's Disease*, 60(4):1611–1620.

[Pinto and Peters, 2009] Pinto, E. and Peters, R. (2009). Literature review of the clock drawing test as a tool for cognitive screening. *Dementia and Geriatric Cognitive Disorders*, 27(3):201–213.

[Pistono et al., 2016] Pistono, A., Jucla, M., Barbeau, E. J., Saint-Aubert, L., Lemesle, B., Calvet, B., Kopke, B., Puel, M., and Pariente, J. (2016). Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. *Journal of Alzheimer's Disease*, 50(3):687–698.

[Poria et al., 2015] Poria, S., Cambria, E., and Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544.

[Povey et al., 2011] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society.

[Prince et al., 2013] Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., and Ferri, C. P. (2013). The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 9(1):63–75.

[Pringsheim et al., 2014] Pringsheim, T., Jette, N., Frolkis, A., and Steeves, T. D. (2014). The prevalence of Parkinson's disease: A systematic review and meta-analysis. *Movement disorders*, 29(13):1583–1590.

[Quadri et al., 2004] Quadri, P., Fragiacomo, C., Pezzati, R., Zanda, E., Forloni, G., Tettamanti, M., and Lucca, U. (2004). Homocysteine, folate, and vitamin B-12 in mild cognitive impairment, Alzheimer disease, and vascular dementia. *The American journal of clinical nutrition*, 80(1):114–122.

[Rabinovici et al., 2017] Rabinovici, G. D., Gatsonis, C., Apgar, C., Gareen, I. F., Hanna, L., Hendrix, J., Hillner, B. E., Olson, C., Romanoff, J., Siegel, B. A., Whitmer, R. A., and Carrillo, M. C. (2017). Impact of amyloid pet on patient management: Early results from the ideas study. *Alzheimer's Dementia*, 13(7, Supplement):P1474. 2017 Abstract Supplement.

[Reisberg et al., 2010] Reisberg, B., Shulman, M. B., Torossian, C., Leng, L., and Zhu, W. (2010). Outcome over seven years of healthy adults with and without subjective cognitive impairment. *Alzheimer's & Dementia*, 6(1):11–24.

[Reynolds et al., 2003] Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., et al. (2003). The supersid project: Exploiting high-level information for high-accuracy speaker recognition. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 4, pages IV–784. IEEE.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

[Ripich et al., 1991] Ripich, D. N., Vertes, D., Whitehouse, P., Fulton, S., and Ekelman, B. (1991). Turn-taking and speech act patterns in the discourse of senile dementia of the Alzheimer's type patients. *Brain and Language*, 40(3):330–343.

[Roark et al., 2011] Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., and Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090.

[Rock et al., 2014] Rock, P., Roiser, J., Riedel, W., and Blackwell, A. (2014). Cognitive impairment in depression: a systematic review and meta-analysis. *Psychological medicine*, 44(10):2029–2040.

[Román et al., 1993] Román, G. C., Tatemichi, T. K., Erkinjuntti, T., Cummings, J., Masdeu, J., Garcia, J. a., Amaducci, L., Orgogozo, J.-M., Brun, A., Hofman, A., et al. (1993). Vascular dementia diagnostic criteria for research studies: Report of the ninds-airen international workshop. *Neurology*, 43(2):250–250.

[Romney et al., 1996] Romney, A. K., Boyd, J. P., Moore, C. C., Batchelder, W. H., and Brazill, T. J. (1996). Culture as shared cognitive representations. *Proceedings of the National Academy of Sciences*, 93(10):4699–4705.

[Romney et al., 1997] Romney, A. K., Moore, C. C., and Rusch, C. D. (1997). Cultural universals: Measuring the semantic structure of emotion terms in english and japanese. *Proceedings of the National Academy of Sciences*, 94(10):5489–5494.

[Rousseau et al., 2012] Rousseau, A., Deléglise, P., and Esteve, Y. (2012). Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.

[Rudzicz et al., 2014] Rudzicz, F., Chan Currie, L., Danks, A., Mehta, T., and Zhao, S. (2014). Automatically identifying trouble-indicating speech behaviors in Alzheimer's disease. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pages 241–242. ACM.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

[Sainath et al., 2015] Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W., and Vinyals, O. (2015). Learning the speech front-end with raw waveform cldnns. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[Satizabal et al., 2016] Satizabal, C. L., Beiser, A. S., Chouraki, V., Chêne, G., Dufouil, C., and Seshadri, S. (2016). Incidence of dementia over three decades in the framingham heart study. *New England Journal of Medicine*, 374(6):523–532.

[Satt et al., 2013] Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A., and Tsolaki, M. (2013). Evaluation of speech-based protocol for detection of early-stage dementia. In *INTERSPEECH*, pages 1692–1696.

[Savitch, 1993] Savitch, W. J. (1993). Why it might pay to assume that languages are infinite. *Annals of Mathematics and Artificial Intelligence*, 8(1-2):17–25.

[Saykin et al., 2006] Saykin, A., Wishart, H., Rabin, L., Santulli, R., Flashman, L., West, J., McHugh, T., and Mamourian, A. (2006). Older adults with cognitive complaints show brain atrophy similar to that of amnestic mci. *Neurology*, 67(5):834–842.

[Schmidhuber, 2015] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

[Schneider et al., 2014] Schneider, L. S., Mangialasche, F., Andreasen, N., Feldman, H., Giacobini, E., Jones, R., Mantua, V., Mecocci, P., Pani, L., Winblad, B., et al. (2014). Clinical trials and late-stage drug development for Alzheimer's disease: an appraisal from 1984 to 2014. *Journal of internal medicine*, 275(3):251–283.

[Seshadri et al., 2006] Seshadri, S., Beiser, A., Kelly-Hayes, M., Kase, C. S., Au, R., Kannel, W. B., and Wolf, P. A. (2006). The lifetime risk of stroke. *Stroke*, 37(2):345–350.

[Shaw et al., 2009] Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., Blennow, K., Soares, H., Simon, A., Lewczuk, P., et al. (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Annals of neurology*, 65(4):403–413.

[Shibata et al., 2016] Shibata, D., Wakamiya, S., Kinoshita, A., and Aramaki, E. (2016). Detecting Japanese patients with AlzheimerâĂŹs disease based on word category frequencies. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 78–85.

[Singh et al., 2001] Singh, S., Bucks, R. S., and Cuerden, J. M. (2001). Evaluation of an objective technique for analysing temporal variables in dat spontaneous speech. *Aphasiology*, 15(6):571–583.

[Slegers et al., 2018] Slegers, A., Filiou, R.-P., Montembeault, M., and Brambati, S. M. (2018). Connected Speech Features from Picture Description in Alzheimer's Disease: A Systematic Review. *Journal of Alzheimer's Disease*, (Preprint):1–26.

[Sluckin, 2017] Sluckin, W. (2017). *Imprinting and early learning.* Routledge.

[Snowdon et al., 1996] Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Jama*, 275(7):528–532.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

[Stolcke et al., 2002] Stolcke, A. et al. (2002). Srilm-an extensible language modeling toolkit. In *Interspeech*, volume 2002, page 2002.

[Stolcke et al., 2006] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M. (2006). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Dialogue*, 26(3).

[Szekely et al., 2004] Szekely, C. A., Thorne, J. E., Zandi, P. P., Ek, M., Messias, E., Breitner, J. C., and Goodman, S. N. (2004). Nonsteroidal anti-inflammatory drugs for the prevention of Alzheimer's disease: a systematic review. *Neuroepidemiology*, 23(4):159–169.

[Taler and Phillips, 2008] Taler, V. and Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of clinical and experimental neuropsychology*, 30(5):501–556.

[Thomas et al., 2005a] Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E. (2005a). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *IEEE International Conference Mechatronics and Automation, 2005*, volume 3, pages 1569–1574 Vol. 3.

[Thomas et al., 2005b] Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E. (2005b). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *Mechatronics and Automation, 2005 IEEE International Conference*, volume 3, pages 1569–1574. IEEE.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[Torack, 1983] Torack, R. M. (1983). The early history of senile dementia.

[Tóth et al., 2015] Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., Zsura, F., Pákáski, M., and Kálmán, J. (2015). Automatic detection of mild cognitive impairment from spontaneous speech using asr. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[Tranter and Reynolds, 2006] Tranter, S. E. and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*, 14(5):1557–1565.

[Tsanas et al., 2010] Tsanas, A., Little, M. A., McSharry, P. E., and Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE transactions on Biomedical Engineering*, 57(4):884–893.

[Tsoi et al., 2015] Tsoi, K. K., Chan, J. Y., Hirai, H. W., Wong, S. Y., and Kwok, T. C. (2015). Cognitive tests to detect dementia: a systematic review and meta-analysis. *JAMA internal medicine*, 175(9):1450–1458.

[Tu, 1996] Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231.

[Tu et al., 2017] Tu, M., Berisha, V., and Liss, J. (2017). Interpretable objective assessment of dysarthric speech based on deep neural networks. In *Proc. Interspeech*, pages 1849–1853.

[Ulyanov, 2016] Ulyanov, D. (2016). Multicore-tsne. https://github.com/DmitryUlyanov/Multicore-TSNE.

[Van Den Oord et al., 2016] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*.

[Van Himbergen et al., 2012] Van Himbergen, T. M., Beiser, A. S., Ai, M., Seshadri, S., Otokozawa, S., Au, R., Thongtang, N., Wolf, P. A., and Schaefer, E. J. (2012). Biomarkers for insulin resistance and inflammation and the risk for all-cause dementia and Alzheimer disease: results from the Framingham Heart Study. *Archives of neurology*, 69(5):594–600.

[Vasquez-Correa et al., 2017] Vasquez-Correa, J., Orozco-Arroyave, J. R., and Noth, E. (2017). Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease. *Proc. Interspeech 2017*, pages 314–318.

[Vincze et al., 2016] Vincze, V., Gosztolya, G., Toth, L., Hoffmann, I., Szatloczki, G., Banreti, Z., Pakaski, M., and Kalman, J. (2016). Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 181–187.

[Vlahiotis et al., 2018] Vlahiotis, A., Griffin, B., Stavros, A. T., and Margolis, J. (2018). Analysis of utilization patterns and associated costs of the breast imaging and diagnostic procedures after screening mammography. *ClinicoEconomics and outcomes research: CEOR*, 10:157.

[Walker et al., 2004] Walker, C., Sunderland, A., Sharma, J., and Walker, M. (2004). The impact of cognitive impairment on upper body dressing difficulties after stroke: a video analysis of patterns of recovery. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(1):43–48.

[Walker et al., 2007] Walker, Z., Jaros, E., Walker, R. W., Lee, L., Costa, D. C., Livingston, G., Ince, P., Perry, R., McKeith, I., and Katona, C. L. (2007). Dementia with Lewy bodies: a comparison of clinical diagnosis, FP-CIT SPECT imaging and autopsy. *Journal of Neurology, Neurosurgery & Psychiatry*.

[Wan et al., 2018] Wan, S., Liang, Y., Zhang, Y., and Guizani, M. (2018). Deep multi-layer perceptron classifier for behavior analysis to estimate parkinsonâĂŹs disease severity using smartphones. *IEEE Access*, 6:36825–36833.

[Wang et al., 2003] Wang, Y.-Y., Acero, A., and Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 577–582. IEEE.

[Ward et al., 2012] Ward, A., Crean, S., Mercaldi, C. J., Collins, J. M., Boyd, D., Cook, M. N., and Arrighi, H. M. (2012). Prevalence of apolipoprotein E4 genotype and homozygotes (APOE e4/4) among patients diagnosed with Alzheimer's disease: a systematic review and meta-analysis. *Neuroepidemiology*, 38(1):1–17.

[Weiner et al., 2016] Weiner, J., Herff, C., and Schultz, T. (2016). Speech-Based Detection of Alzheimer's Disease in Conversational German. In *INTERSPEECH*, pages 1938–1942.

[Weinstein et al., 2014] Weinstein, G., Beiser, A. S., Choi, S. H., Preis, S. R., Chen, T. C., Vorgas, D., Au, R., Pikula, A., Wolf, P. A., DeStefano, A. L., et al. (2014). Serum brain-derived neurotrophic factor and the risk for dementia: the Framingham Heart Study. *JAMA neurology*, 71(1):55–61.

[Wilson, 2000] Wilson, R. M. (2000). Screening for breast and cervical cancer as a common cause for litigation: A false negative result may be one of an irreducible minimum of errors. *BMJ: British Medical Journal*, 320(7246):1352.

[Wimo et al., 2006] Wimo, A., Jonsson, L., and Winblad, B. (2006). An estimate of the worldwide prevalence and direct costs of dementia in 2003. *Dementia and geriatric cognitive disorders*, 21(3):175–181.

[Wold et al., 1987] Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

[Wu et al., 2015] Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., and Li, H. (2015). Spoofing and countermeasures for speaker verification. *Speech Communication*, 66(C):130–153.

[Xiong et al., 2017] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., Yu, D., and Zweig, G. (2017). Toward Human Parity in Conversational Speech Recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(12):2410–2423.

[Yancheva et al., 2015] Yancheva, M., Fraser, K., and Rudzicz, F. (2015). Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 134–139.

[Yancheva and Rudzicz, 2016] Yancheva, M. and Rudzicz, F. (2016). Vector-space topic models for detecting Alzheimer's disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2337–2346.

[Yoo et al., 2003] Yoo, A. B., Jette, M. A., and Grondona, M. (2003). Slurm: Simple linux utility for resource management. In *Workshop on Job Scheduling Strategies for Parallel Processing*, pages 44–60. Springer.

[Youden, 1950] Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

[Yu et al., 2013] Yu, D., Yao, K., Su, H., Li, G., and Seide, F. (2013). Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7893–7897. IEEE.

[Zhang et al., 2018] Zhang, H., Wang, A., Li, D., and Xu, W. (2018). DeepVoice: A voiceprint-based mobile health framework for Parkinson's disease identification. In *Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on*, pages 214–217. IEEE.

[Zhang et al., 2017] Zhang, Y., Chan, W., and Jaitly, N. (2017). Very deep convolutional networks for end-to-end speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 4845–4849. IEEE.

[Zhou et al., 2016] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.