Detecting late-life depression in Alzheimer's disease through analysis of speech and language

Kathleen C. Fraser[1] and **Frank Rudzicz**[2,1] and **Graeme Hirst**[1]
[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Toronto Rehabilitation Institute-UHN, Toronto, Canada
{kfraser,frank,gh}@cs.toronto.edu

Abstract

Alzheimer's disease (AD) and depression share a number of symptoms, and commonly occur together. Being able to differentiate between these two conditions is critical, as depression is generally treatable. We use linguistic analysis and machine learning to determine whether automated screening algorithms for AD are affected by depression, and to detect when individuals diagnosed with AD are also showing signs of depression. In the first case, we find that our automated AD screening procedure does not show false positives for individuals who have depression but are otherwise healthy. In the second case, we have moderate success in detecting signs of depression in AD (accuracy = 0.658), but we are not able to draw a strong conclusion about the features that are most informative to the classification.

1 Introduction

Depression and dementia are both medical conditions that can have a strong negative impact on the quality of life of the elderly, and they are often comorbid. However, depression is often treatable with medication and therapy, whereas dementia usually occurs as the result of an irreversible process of neurodegeneration. It is therefore critical to be able to distinguish between these two conditions.

However, distinguishing between depression and dementia can be extremely difficult because of overlapping symptoms, including apathy, crying spells, changes in weight and sleeping patterns, and problems with concentration and attention.

It is also important to detect when someone has both AD and depression, as this serious situation can lead to more rapid cognitive decline, earlier placement in a nursing home, increased risk of depression in the patient's caregivers, and increased mortality (Thorpe, 2009; Lee and Lyketsos, 2003).

Separate bodies of work have reported the utility of spontaneous speech analysis in distinguishing participants with depression from healthy controls, and in distinguishing participants with dementia from healthy controls. Here we consider whether such analyses can be applied to the problem of detecting depression in Alzheimer's disease (AD). In particular, we explore two questions: (1) In previous work on detecting AD from speech (elicited through a picture description task), are cognitively healthy people with depression being misclassified as having AD? (2) If we consider only participants with AD, can we distinguish between those with depression and those without, using the same picture description task and analysis?

2 Background

There has been considerable work on detecting depression from speech and on detecting dementia from speech, but very little which combines the two. We will first review the two tasks separately, and then discuss some of the complexity that arises when depression and AD co-occur.

2.1 Detecting depression from speech

Depression affects a number of cognitive and physical systems related to the production of speech, including working memory, the phonological loop, articulatory planning, and muscle tension and control (Cummins et al., 2015). These changes can result in word-finding difficulties, articulatory errors, decreased prosody, and lower verbal productivity.

Over the past decade or so, there has been growing interest in measuring properties of the speech signal that correlate with the changes observed in depression, and using these measured variables to train machine learning classifiers to automatically detect depression from speech.

Ozdas et al. (2004) found that mean jitter and the slope of the glottal flow spectrum could distinguish between 10 non-depressed controls, 10 participants with clinical depression, and 10 high-risk suicidal participants.

Moore et al. (2008) considered prosodic features as well as vocal tract and glottal features. They performed sex-dependent classification and found that glottal features were more discriminative than vocal tract features, but that the best results were achieved using all three types of features.

Cohn et al. (2009) examined the utility of facial movements and vocal prosody in discriminating participants with moderate or severe depression from those with no depression. They achieved 79% accuracy using only two prosodic features: variation in fundamental frequency, and latency of response to interviewer questions. They used a within-subjects design, in which they predicted which participants had responded to treatment in a clinical trial.

Low et al. (2011) analyzed speech from adolescents engaged in normal conversation with their parents (68 diagnosed with depression, 71 controls). They grouped their acoustic features into 5 groups: spectral, cepstral, prosodic, glottal, and those based on the Teager energy operator (TEO, a nonlinear energy operator). They achieved higher accuracies using sex-dependent models than sex-independent models, and found that the best results were achieved using the TEO-based features (up to 87% for males and 79% for females).

Cummins et al. (2011) distinguished 23 depressed participants from 24 controls with a best accuracy of 80% in a speaker-dependent configuration and 79% in a speaker-independent configuration. Spectral features, particularly mel-frequency cepstral coefficients (MFCCs), were found to be useful.

Alghowinem et al. (2012) analyzed speech from 30 participants with depression and 30 healthy controls. The speech was elicited through interview questions about situations that had aroused significant emotions. Higher accuracy was achieved on detecting depression in women than in men. Energy, intensity, shimmer, and MFCC features were all informative, and positive emotional speech was more discriminatory than negative emotional speech.

Scherer et al. (2013) differentiated 18 depressed participants from 18 controls with 75% accuracy, using interviews captured with a simulated virtual human. They found that glottal features such as the normalized amplitude quotient (NAQ) and quasi-open quotient (QOQ) differed significantly between the groups.

Alghowinem et al. (2013) compared four classifiers and a number of different feature sets on the task of detecting depression from spontaneous speech. They found loudness and intensity features to be the most discriminatory, and suggested pitch and formant features may be more useful for longitudinal comparisons within individuals.

While most of the literature concerning the detection of depression from speech has focused solely on the speech signal, there is an associated body of work on detecting depression from writing that focuses on linguistic cues. Rude et al. (2004) found that college students with depression were significantly more likely to use the first-person pronoun *I* in personal essays than college students without depression, and also used more words with negative emotional valence. Other work has found differences in the frequency of different parts-of-speech (POS) (De Choudhury et al., 2013) and in the general topics chosen for discussion (Resnik et al., 2015). Other work has accurately identified depression (and differentiated PTSD and depression) in Twitter social media texts with high accuracies using *n*-gram language models (Coppersmith et al., 2015). Similarly, Nguyen et al. (2014) showed that specialized lexical norms and Linguistic Inquiry and Word Count[1] features significantly differentiate clinical and control groups in blog post texts. Howes et al. (2014) showed that lexical features (in style and dialogue) could also be used to predict the severity of depression and anxiety during Cognitive Be-

[1] http://liwc.wpengine.com.

havioural Therapy treatment. It is not obvious that these results generalize to the case where the topic and structure of the narrative is constrained to a picture description.

2.2 Detecting Alzheimer's disease from speech

A growing number of researchers have tackled the problem of detecting dementia from speech and language. Most of this work has focused on Alzheimer's disease (AD), which is the most common cause of dementia. Although the primary diagnostic symptom of AD is memory impairment, this and other cognitive deficits often manifest in spontaneous language through word-finding difficulties, a decrease in information content, and changes in fluency, syntactic complexity, and prosody. Other work, including that of Roark et al. (2007), focuses on mild cognitive impairment, which is also broadly applicable.

Thomas et al. (2005) classified spontaneous speech samples from 95 AD patients and an unspecified number of controls by treating the problem as an authorship attribution task, and employing a "common N-grams" approach. They were able to distinguish between patients with severe AD and controls with a best accuracy of 94.5%, and between patients with mild AD and controls with an 75.3% accuracy.

Habash and Guinn (2012) built classifiers to distinguish between AD and non-AD language samples using 80 conversations between 31 AD patients and 57 cognitively normal conversation partners. They found that features such as POS tags and measures of lexical diversity were less useful than measuring filled pauses, repetitions, and incomplete words, and achieved a best accuracy of 79.5%.

Meilán et al. (2012) distinguished between 30 AD patients and 36 healthy controls with temporal and acoustic features alone, obtaining an accuracy of 84.8%. For each participant, their speech sample consisted of two sentences read from a screen. The discriminating features were percentage of voice breaks, number of voice breaks, number of periods of voice, shimmer, and noise-to-harmonics ratio.

Jarrold et al. (2014) used acoustic features, POS features, and psychologically-motivated word lists to distinguish between semi-structured interview responses from 9 AD participants and 9 controls with an accuracy of 88%. They also confirmed their hypothesis that AD patients would use more pronouns, verbs, and adjectives and fewer nouns than controls.

Rentoumi et al. (2014) considered a slightly different problem: they used computational techniques to differentiate between picture descriptions from AD participants with and without additional vascular pathology ($n = 18$ for each group). They achieved an accuracy of 75% when they included frequency unigrams and excluded binary unigrams, syntactic complexity features, measures of vocabulary richness, and information theoretic features.

Orimaye et al. (2014) obtained F-measure scores up to 0.74 on transcripts from DementiaBank, combining participants with different etiologies rather than focusing on AD. In previous work, we also studied data from DementiaBank (Fraser et al., 2015). We computed acoustic and linguistic features from the "Cookie Theft" picture descriptions and distinguished 240 AD narratives from 233 control narratives with 81% accuracy using logistic regression.

2.3 Relationship between dementia and depression

The relationship between dementia and depression is complicated, as the two conditions are not independent of each other and in fact frequently cooccur. When someone is diagnosed with dementia, feelings of depression are common. At the same time, depression is a risk factor for developing Alzheimer's disease (Korczyn and Halperin, 2009). The diagnosis of a third medical condition (e.g., heart disease) can trigger depression and also independently increase the risk of dementia. Similarly, some risk factors for depression and dementia are the same, such as alcohol use and cigarette smoking (Thorpe, 2009). Furthermore, changes in white matter connectivity have been linked to both depression (Alexopoulos et al., 2008) and dementia (Prins et al., 2004).

The prevalence of depression in AD has been estimated to be 30–50% (Lee and Lyketsos, 2003), although these figures have been shown to vary widely depending on the diagnostic method used (Müller-Thomsen et al., 2005). In contrast, the prevalence of depression in the general population older than 75 is estimated to be 7.2% (major depression) and 17.1% (depressive disorders) (Luppa et al., 2012).

The prevalence of Alzheimer's disease is 11% for people aged 65 and older, increasing to 33% for people ages 85 and older (Alzheimer's Association, 2015).

Symptoms which are common in both depression and dementia include: poor concentration, impaired attention (Korczyn and Halperin, 2009), apathy (Lee and Lyketsos, 2003), changes to eating and sleeping patterns, and reactive mood symptoms, e.g., tearfulness (Thorpe, 2009). However, both dementia and depression are heterogeneous in presentation, which can lead to many possible combinations of symptoms when they co-occur.

Studies examining spontaneous speech tasks to discriminate between dementia and depression are rare. Murray (2010) investigated whether clinical depression could be distinguished from AD by analyzing narrative speech. She found that there were significant differences in the amount of information that was conveyed in a picture description task, with depressed participants communicating the same amount of information as healthy controls, and AD patients showing a reduction in information content. Other discourse measures relating to the quantity of speech produced and the syntactic complexity of the narrative did not differ between the groups. In contrast to the current work, the study described in Murray (2010) did not include participants with *both* dementia and depression, involved a much smaller data set (49 participants across 3 groups), and did not seek to make predictions from the data.

3 Methods

3.1 Data

We use narrative speech data from the Pitt corpus in the DementiaBank database[2]. These data were collected between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh. Detailed information about the study cohort is available from Becker et al. (1994), and demographic information is presented for each experiment below in Tables 1 and 3. Diagnoses were made on the basis of a personal history and a neuropsychological battery; a subset of these diagnoses were confirmed post-mortem. The language samples were elicited using the "Cookie Theft" picture description task from the Boston Diagnostic Aphasia Examination (BDAE) (Goodglass and Kaplan, 1983), in which participants are asked to describe everything they see going on in a picture. We extract features from both the acoustic files (converted from MP3 to 16-bit mono WAV format with a sampling rate of 16 kHz) and the associated transcripts. All examiner speech is excluded from the sample.

A subset of the participants also have Hamilton Depression Rating Scale (HAM-D) scores (Hamilton, 1960). The HAM-D is still one of the gold standards for depression rating (although it has also received criticism; see Bagby et al. (2014) for an example). It consists of 17 questions, for which the patient's responses are rated from 0–4 or 0–2 by the examiner. A total score between 0–7 is considered normal, 8–16 indicates mild depression, 17–23 indicates moderate depression, and greater than 24 indicates severe depression (Zimmerman et al., 2013).

3.2 Features

We extract a large number of textual features (including part-of-speech tags, parse constituents, psycholinguistic measures, and measures of complexity, vocabulary richness, and informativeness), and acoustic features (including fluency measures, MFCCs, voice quality features, and measures of periodicity and symmetry). A complete list of features is given in the Supplementary Material, and additional details are reported by Fraser et al. (2015).

3.3 Classification

We select a subset of the extracted features using a correlation-based filter. Features are ranked by their correlation with diagnosis and only the top *N* features are selected, where we vary *N* from 5 to 400. The selected features are fed to a machine learning classifier; in this study we compare logistic regression (LR) with support vector machines (SVM) (Hall et al., 2009). We use a cross-validation framework and report the average accuracy across folds. The data is partitioned across folds such that samples from a single speaker occur in either the training set or test set, but never both. Error bars are computed using the standard deviation of the accuracy across folds. In some cases we also report *sensitivity* and *specificity*, where sensitivity indicates the proportion of people with AD (or depression) who were

[2] `https://talkbank.org/DementiaBank/`

	AD	Controls	Sig.
	$n = 196$	$n = 128$	
Age	71.7 (8.7)	63.7 (7.6)	**
Education	12.4 (2.9)	13.9 (2.4)	**
Sex (M/F)	66/130	49/79	

Table 1: Mean and standard deviation of demographic information for participants in Experiment I. ** indicates $p < 0.01$.

correctly identified as such, and specificity indicates the proportion of controls who were correctly identified as such.

4 Experiment I: Does depression affect classification accuracy?

Our first experiment examines whether depression is a confounding factor in our current diagnostic pipeline. To answer this question, we consider the subset of narratives for which associated HAM-D scores are available. This leaves a set of 196 AD narratives and 128 control narratives from 150 AD participants and 80 control participants. Since participants may have different scores on different visits, we consider data per narrative, rather than per speaker. Demographic information is given in Table 1. The groups are not matched for age or education, which is a limitation of the complete data set as well; the AD participants tend to be both older and less educated.

We then perform the classification procedure using the analysis pipeline described above, with 10-fold cross-validation. The results for a logistic regression and SVM classifier are shown in Figure 1. This is a necessary first step to examine if depression is a confounding factor.

Choosing the best result of 0.799 (SVM classifier, 70 features), we then perform a more detailed analysis. Accuracy, sensitivity, and specificity for the full data set are reported in the first row of Table 2. We first break down the data into two separate groups: those with a Hamilton score greater than 7 (i.e., “depressed”) and those with a Hamilton score less than or equal to 7 (“non-depressed”) (Zimmerman et al., 2013). The accuracy, sensitivity, and specificity for these sub-groups are also reported in Table 2. Because there are far more AD participants with depression ($n = 65$) than controls with depression ($n = 9$), we also report the accuracy of a majority class classifier as a baseline with which to compare the reported accuracies. Alternatives to this approach, including synthetically balancing the classes, e.g., with synthetic minority oversampling (Chawla et al., 2002), is to be the subject of future work.

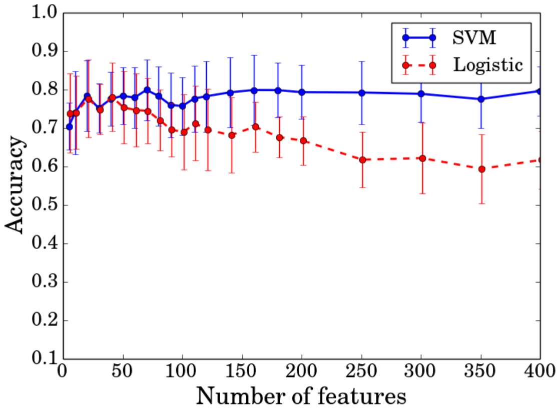


Figure 1: Classification accuracy on the task of distinguishing AD from control narratives for varying feature set sizes.

Data set	Baseline	Accuracy	Sens.	Spec.
All	0.605	0.799	0.826	0.758
Depressed	0.743	0.864	0.846	1.000
Non-depressed	0.552	0.780	0.816	0.739

Table 2: Accuracy, sensitivity, and specificity for all participants, depressed participants, and non-depressed participants.

A key result from this experiment is that although there are only a few control participants who are depressed, none of those are misclassified as AD (specificity = 1.0 in this case).

Furthermore, if we partition the participants by accuracy (those who were classified correctly versus incorrectly), we find no significant difference on HAM-D scores ($p > 0.05$). This suggests that the accuracy of the classifier is not affected by the presence or absence of depression.

5 Experiment II: Can we detect depression in Alzheimer’s disease?

In our second experiment, we tackle the problem of detecting depression when it is comorbid with

	Depressed	Non-dep.	Sig.
	$n = 65$	$n = 65$	
Age	71.4 (8.6)	71.6 (8.6)	
Education	11.7 (2.6)	12.9 (3.0)	*
Sex (M/F)	21/44	19/46	
MMSE	18.1 (5.5)	17.9 (5.4)	

Table 3: Mean and standard deviation of demographic information for AD participants in Experiment II. * indicates $p < 0.05$.

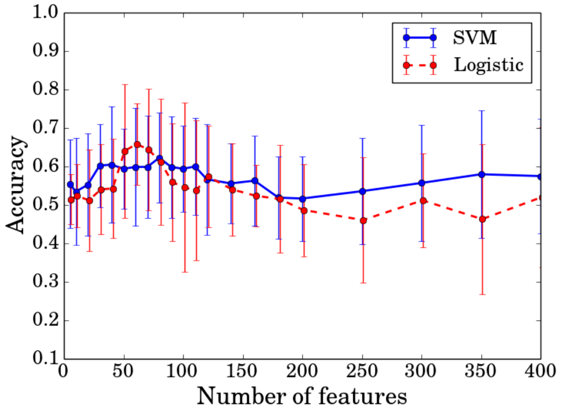


Figure 2: Classification accuracy on the task of distinguishing depressed from non-depressed AD narratives for varying feature set sizes.

Alzheimer's disease. From the previous section, we have 65 narratives from participants with both AD and depression (HAM-D > 7). We select an additional 65 narratives from participants with AD but no depression. These additional data are selected randomly but such that participants are matched for dementia severity, age, and sex. Demographic information is given in Table 3.

5.1 Standard processing pipeline

We begin by using our standard processing pipeline to assess whether it is capable of detecting depression. The classification accuracies are given in Figure 2. In this case, since the groups are the same size, the baseline accuracy is 0.5. The best accuracy of 0.658 is achieved with the LR classifier using 60 features (sensitivity: 0.707, specificity: 0.610). This represents a significant increase (paired t-test, $p < 0.05$) of 15 percentage points over the random baseline, but there is clearly room for improvement.

Rank	Feature	r	Trend
1	Skewness MFCC 1	0.270	↑
2	Info unit: *boy*	−0.265	↓
3	Mean $\Delta\Delta$MFCC 8	0.229	↑
4	VP → VB NP	−0.223	↓
5	Kurtosis MFCC 4	0.223	↑
6	Kurtosis MFCC 3	0.217	↑
7	Kurtosis ΔMFCC 2	0.213	↑
8	Skewness $\Delta\Delta$MFCC 2	0.211	↑
9	Kurtosis MFCC 10	0.209	↑
10	Determiners	−0.206	↓

Table 4: Highly ranked features for distinguishing people with AD and depression from people with only AD. The third column shows the correlation with diagnosis, and the fourth column shows the direction of the trend (increasing or decreasing) with depression.

Data set	Baseline	Accuracy	Sens.	Spec.
All	0.500	0.658	0.707	0.610
Females	0.511	0.588	0.519	0.653
Males	0.525	0.650	0.580	0.717

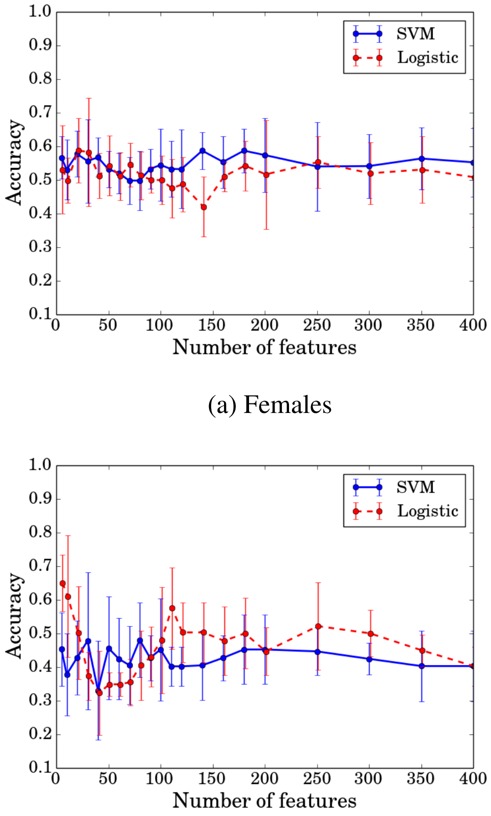
Table 5: Accuracy, sensitivity, and specificity for all participants, just females, and just males.

Table 4 shows the features which are most highly correlated with diagnosis (over all data). Even for the top-ranked features, the correlation is weak, and the difference between groups is not significant after correcting for multiple comparisons. We therefore cannot conclusively draw conclusions about the selected features, although we do note the apparent importance of the MFCC features here.

5.2 Sex-dependent classification

Given that acoustic features naturally vary across the sexes, and that previous work achieved better results using sex-dependent classifiers, we also consider a sex-dependent configuration. The drawback to this approach is the reduction in data, particularly for males. In these experiments we attempt to classify 21 males with depression+AD versus 19 males with AD only, and 44 females with depression+AD versus 46 females with AD only. The results for these experiments are shown in Figure 3, and the best accuracies are given in Table 5.

The features which are most correlated with di-



(a) Females

(b) Males

Figure 3: Sex-dependent classification accuracy on the task of distinguishing depressed from non-depressed AD narratives for varying feature set sizes.

Rank	Feature	r	Trend
1	Info unit: *boy*	−0.323	↓
2	Mean ΔMFCC 9	0.284	↑
3	VP → VB NP	−0.274	↓
4	Kurtosis MFCC 3	0.266	↑
5	Kurtosis Δ energy	0.261	↑
6	Skewness MFCC 1	0.260	↑
7	Kurtosis MFCC 4	0.256	↑
8	NP → PRP$ NNS	0.251	↑
9	Skewness ΔΔMFCC 2	0.249	↑
10	NP → NP NP .	0.243	↑

(a) Females

Rank	Feature	r	Trend
1	Mean ΔΔMFCC 9	0.447	↑
2	Skewness ΔΔMFCC 12	−0.406	↓
3	VP → VB S	0.405	↑
4	Mean ΔΔMFCC 2	−0.381	↓
5	Info unit: *stool*	0.352	↑
6	VP → VBG NP	−0.351	↓
7	Key word: *chair*	0.346	↑
8	Mean ΔMFCC 11	−0.325	↓
9	Key word: *girl*	−0.318	↓
10	Mean ΔΔMFCC 8	0.316	↑

(b) Males

Table 6: Highly ranked features for distinguishing individuals with AD and depression from individuals with only AD, in the sex-dependent case. (No differences are significant after correcting for multiple comparisons.)

agnosis for females are listed in Table 6a, and those which are most correlated with diagnosis for males are listed in Table 6b. Again, the selected features tend to be either informational, grammatical, or cepstral in nature, although none of the differences are significant after correcting for multiple comparisons.

5.3 Additional features

To help our classifiers better distinguish between people with and without depression, we implement a number of additional features which have been reported to be valuable in detecting depression. Many of the acoustic features from the literature were already present in our feature set, but we now consider a number of glottal features, including the mean and standard deviations of the maximum voiced frequency, glottal closure instants, linear prediction residuals, peak slope, glottal flow (and derivative), normalized amplitude quotient (NAQ), quasi-open quotient (QOQ), harmonic richness factor, parabolic spectral parameter, and cepstral peak prominence. These features are implemented in the COVAREP toolkit (version 1.4.1) (Degottex et al., 2014).

We also include three additional psycholinguistic variables relating to the affective qualities of words: valence, arousal, and dominance. Valence describes the degree of positive or negative emotion associated with a word, arousal describes the intensity of the emotion associated with a word, and dominance describes the degree of control associated with a word. We use the crowd-sourced norms presented by Warriner et al. (2013) for their broad coverage, and measure the mean and maximum value of each variable.

Finally, we count the frequency of occurrence of first-person words (*I*, *me*, *my*, *mine*). In general, the picture description task is completed in the third person, but first-person words do occur.

However, including these new features actually had a slightly negative effect on the sex-independent classification, reducing the maximum accuracy from 0.658 to 0.650, as well as on the males-only case, reducing maximum accuracy from 0.650 to 0.585. This suggests that some of the new features are being selected in individual training folds, but not generalizing to the test folds. In contrast, the new features did make a small, incremental improvement in the females-only case, from 0.588 to 0.609 for females. The new features that were most highly ranked for females were the standard deviation of the peak slope (rank 12, $r = -0.237$) and the standard deviation of NAQ (rank 35, $r = -0.186$), both showing a weak negative correlation with diagnosis. The most useful new feature for males was the mean QOQ (rank 16, $r = 0.273$), with a weak positive correlation with diagnosis.

6 Conclusion

In this paper, we considered two questions. The first is related to previous work in the field showing that speech analysis and machine learning can lead to good, but not perfect, differentiation between participants with AD and healthy controls. We wondered whether some control participants were being misclassified as having AD when in fact they were depressed. However, in our experiment we found that none of the 9 depressed controls were misclassified as having AD. This is a small sample, but it is consistent with the findings of Murray (2010), who found that although AD participants and controls could be distinguished through analysis of their picture descriptions, there were no differences between depressed participants and controls.

We then considered only participants with AD, and tried to distinguish between those with comorbid depression and those without. Our best accuracy for this task was 0.658, which is considerably lower than reported accuracies for detecting depression in the absence of AD, but reflects the difficulty of the task given the wide overlap of symptoms in the two conditions. In fact, previous work on detecting depression from speech has focused overwhelmingly on young and otherwise healthy participants, and much work is needed on detecting depression in other populations and with other comorbidities.

One limitation of this work is the type of speech data available; previous work suggests that emotional speech is more informative for detecting depression. Another limitation is that we are assigning our participants to the depressed and non-depressed groups on the basis of a single test score, rather than a confirmed clinical diagnosis. A related factor to consider is the relatively mild depression that is observed in this data set, which was developed for the study of AD rather than depression – only 8 participants met the criteria for "moderate" depression, and none met the criteria for severe depression. Furthermore, while the controls in Experiment 2 all had scores below the threshold for mild depression, in most cases the scores were still non-zero, and so the classification task is not as clearly binary as we have framed it here. Finally, limitations of the dataset introduced issues of confounding variables (namely age and education), and prohibited us from contrasting speech from participants with only depression versus those with only AD. We are currently undertaking our own data collection to overcome the various challenges of this dataset.

Depression and Alzheimer's disease both present in different syndromes, and so it is probably unrealistic to clearly delineate between the many potential combinations of depression, AD, and other possible medical conditions through the analysis of a single language task. On the other hand, previous work suggests that this type of analysis can be very fine-grained and sensitive to subtle cognitive impairments. Ideally, future work will focus directly on the task of distinguishing AD from depression, using clinically validated data with a stronger emotional component.

Acknowledgments

The authors would like to thank Dr. John Strauss for his comments on an early draft of this paper. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), grant number RGPIN-2014-06020.

References

George S. Alexopoulos, Christopher F. Murphy, Faith M. Gunning-Dixon, Vassilios Latoussakis, Dora Kanellopoulos, Sibel Klimstra, Kelvin O. Lim, and Matthew J. Hoptman. 2008. Microstructural white matter abnormalities and remission of geriatric depression. *The American Journal of Psychiatry*, 165(2):238–244.

Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Michael Breakspear, Gordon Parker, et al. 2012. From joyous to clinically depressed: Mood detection using spontaneous speech. In *Proceedings of the Florida Artificial Intelligence Research Society (FLAIRS) Conference*, pages 141–146.

Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Tom Gedeon, Michael Breakspear, and Gordon Parker. 2013. A comparative study of different classifiers for detecting depression from spontaneous speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8022–8026.

Paavo Alku, Helmer Strik, and Erkki Vilkman. 1997. Parabolic spectral parameter – a new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79.

Alzheimer's Association. 2015. 2015 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 11(3):332.

R. Michael Bagby, Andrew G. Ryder, Deborah R. Schuller, and Margarita B. Marshall. 2014. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *American Journal of Psychiatry*, 161(12):2163–2177.

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

Sarah D. Breedin, Eleanor M. Saffran, and Myrna F. Schwartz. 1998. Semantic factors in verb retrieval: An effect of complexity. *Brain and Language*, 63:1–31.

Étienne Brunet. 1978. *Le vocabulaire de Jean Giraudoux structure et évolution*. Éditions Slatkine.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Romola S. Bucks, Sameer Singh, Joanne M. Cuerden, and Gordon K. Wilcock. 2000. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Jieun Chae and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–147.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Jeffrey F. Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De La Torre. 2009. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the Shared Task for the NAACL Workshop on Computational Linguistics and Clinical Psychology*.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Bernard Croisile, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. 1996. Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and Language*, 53(1):1–19.

Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. 2011. An investigation of depressed speech detection: Features and normalization. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pages 2997–3000.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP – a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964.

Thomas Drugman and Yannis Stylianou. 2014. Maximum voiced frequency estimation: Exploiting amplitude and phase spectra. *Signal Processing Letters, IEEE*, 21(10):1230–1234.

Thomas Drugman, Baris Bozkurt, and Thierry Dutoit. 2012. A comparative study of glottal source estimation techniques. *Computer Speech & Language*, 26(1):20–34.

Thomas Drugman. 2014. Maximum phase modeling for sparse linear prediction of speech. *Signal Processing Letters, IEEE*, 21(2):185–189.

Rubén Fraile and Juan Ignacio Godino-Llorente. 2014. Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control*, 14:42–54.

Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2015. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Ken J. Gilhooly and Robert H. Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods*, 12:395–427.

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

Anthony Habash and Curry Guinn. 2012. Language analysis of speakers with dementia of the Alzheimer's type. In *Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium*, pages 8–13.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Max Hamilton. 1960. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, 23(1):56–62.

Antony Honoré. 1979. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172–177.

Christine Howes, Matthew Purver, and Rose McCabe. 2014. Linguistic indicators of severity and progress in online text-based therapy for depression. In *In Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, pages 27–36.

John Kane and Christer Gobl. 2011. Identifying regions of non-modal phonation using features of the wavelet transform. In *12th Annual Conference of the International Speech Communication Association 2011 (INTERSPEECH 2011)*, pages 177–180.

Amos D. Korczyn and Ilan Halperin. 2009. Depression and dementia. *Journal of the Neurological Sciences*, 283(1):139–142.

Hochang B. Lee and Constantine G. Lyketsos. 2003. Depression in Alzheimer's disease: heterogeneity and related issues. *Biological Psychiatry*, 54(3):353–362.

Lu-Shih Alex Low, Namunu C. Maddage, Margaret Lech, Lisa B. Sheeber, and Nicholas B. Allen. 2011. Detection of clinical depression in adolescents' speech during family interactions. *Biomedical Engineering, IEEE Transactions on*, 58(3):574–586.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Melanie Luppa, Claudia Sikorski, Tobias Luck, L. Ehreke, Alexander Konnopka, Birgitt Wiese, Siegfried Weyerer, H-H. König, and Steffi G. Riedel-Heller. 2012. Age-and gender-specific prevalence of depression in latest-life–systematic review and meta-analysis. *Journal of Affective Disorders*, 136(3):212–221.

Juan J.G. Meilán, Francisco Martínez-Sánchez, Juan Carro, José A. Sánchez, and Enrique Pérez. 2012. Acoustic markers associated with impairment in language processing in Alzheimer's disease. *The Spanish Journal of Psychology*, 15(02):487–494.

Elliot Moore, Mark Clements, John W. Peifer, and Lydia Weisser. 2008. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *Biomedical Engineering, IEEE Transactions on*, 55(1):96–107.

Tomas Müller-Thomsen, Sönke Arlt, Ulrike Mann, Reinhard Maß, and Stefanie Ganzer. 2005. Detecting depression in Alzheimer's disease: evaluation of four different scales. *Archives of Clinical Neuropsychology*, 20(2):271–276.

Laura L. Murray. 2010. Distinguishing clinical depression from early Alzheimer's disease in elderly people: Can narrative analysis help? *Aphasiology*, 24(6-8):928–939.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the 1st Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 78–87.

Asli Ozdas, Richard G. Shiavi, Stephen E. Silverman, Marilyn K. Silverman, and D. Mitchell Wilkes. 2004. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *Biomedical Engineering, IEEE Transactions on*, 51(9):1530–1540.

Niels D. Prins, Ewoud J. van Dijk, Tom den Heijer, Sarah E. Vermeer, Peter J. Koudstaal, Matthijs Oudkerk, Albert Hofman, and Monique M.B. Breteler. 2004. Cerebral white matter lesions and the risk of dementia. *Archives of Neurology*, 61(10):1531–1534.

Vassiliki Rentoumi, Ladan Raoufian, Samrah Ahmed, Celeste A. de Jager, and Peter Garrard. 2014. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease*, 42:S3–S17.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015. The University of Maryland CLPsych 2015 shared task system. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 54–60, Denver, Colorado, June 5.

Brian Roark, John-Paul Hosom, Margaret Mitchell, and Jeffrey A. Kaye. 2007. Automatically derived spoken language markers for detecting mild cognitive impairment. In *Proceedings of the 2nd International Conference on Technology and Aging (ICTA)*, Toronto, ON, June.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. 2013. Investigating voice quality as a speaker-independent indicator of depression and PTSD. In *Proceedings of Interspeech*, pages 847–851.

Hans Stadthagen-Gonzalez and Colin J. Davis. 2006. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38(4):598–605.

Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. 2005. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *Proceedings of the IEEE International Conference on Mechatronics and Automation*, pages 1569–1574.

Lilian Thorpe. 2009. Depression vs. dementia: How do we assess? *The Canadian Review of Alzheimer's Disease and Other Dementias*, pages 17–21.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Victor Yngve. 1960. A model and hypothesis for language structure. *Proceedings of the American Physical Society*, 104:444–466.

Mark Zimmerman, Jennifer H. Martinez, Diane Young, Iwona Chelminski, and Kristy Dalrymple. 2013. Severity classification on the Hamilton depression rating scale. *Journal of Affective Disorders*, 150(2):384–388.

Supplementary Materials

POS tags Counts of POS tags, normalized by the total number of word tokens in the sample. Includes: nouns, verbs, inflected verbs, determiners, demonstratives, adjectives, adverbs, function words, interjections, subordinate conjunctions, and coordinate conjunctions.

POS tag ratios Noun: verb ratio (ratio of nouns to verbs), noun ratio (ratio of nouns to nouns and verbs), pronoun ratio (ratio of pronouns to nouns), and subordinate:coordinate ratio (ratio of subordinate conjunctions to coordinate conjunctions).

Yngve depth A measure which quantifies to what extent a sentence is left-branching rather than right-branching (Yngve, 1960). We compute the maximum, mean, and total Yngve depth for each sentence, then average over all sentences in each narrative sample.

Parse tree height The average height of each parse tree in the sample.

Mean length of sentence (MLS) Total number of words divided by number of sentences.

Mean length of clause (MLC) Total number of words divided by number of clauses (as computed by Lu's syntactic complexity analyzer (Lu, 2010)).

Mean length of T-unit (MLT) Total number of words divided by number of T-units(as computed by Lu's syntactic complexity analyzer (Lu, 2010)). A T-unit is a minimally terminable syntactic unit consisting of a main clause and its dependent clauses.

Mean word length Mean number of letters in the words in the sample.

Disfluency frequencies Frequency of occurrence of the token *um* and *uh*, normalized by the total number of word tokens.

"Not in dictionary" (NID) words Frequency of occurrence of word tokens of length greater than two which do not occur in the English dictionary.

Total words Total number of words produced, excluding filled pauses and NID words.

Type:token ratio (TTR) V/N where V is the number of word types and N is the number of word tokens.

Moving-average type:token ratio (MATTR) An adaptation of TTR which reduces the effect of narrative sample length (Covington and McFall, 2010). MATTR$_w$ is the TTR calculated over a moving window of size w, and averaged over all windows.

Brunéts index $NV0.165$ where V is the number of word types and N is the number of word tokens (from Bucks et al. (2000) citing Brunet (1978)).

Honorés statistic $100logN/(1V_1/V)$ where V_1 is the number of words used only once, V is the total number of word types, and N is the number of word tokens (from Bucks et al. (2000) citing Honoré (1979)).

CFG production rules The frequency of occurrence of different grammatical constituents in the data, normalized by the total number of constituents in the sample. Dependency parsing is the subject of future work.

Phrase type proportion Length of each phrase type (noun phrase NP, verb phrase VP, or prepositional phrase PP), divided by total narrative length (see Chae and Nenkova (2009)).

Average phrase type length Total number of words in a phrase type (noun phrase NP, verb phrase VP, or prepositional phrase PP), divided by the number of phrases of that type (see Chae and Nenkova (2009)).

Phrase type rate Number of phrases of a given type (noun phrase NP, verb phrase VP, or prepositional phrase PP), divided by total narrative length (see Chae and Nenkova (2009)).

Frequency Frequency with which a word occurs in some corpus of natural language, here Brysbaert and New (2009).

Familiarity Subjective rating of how familiar a word seems. (Gilhooly and Logie, 1980; Stadthagen-Gonzalez and Davis, 2006).

Imageability Subjective rating of how easily a word generates an image in the mind (Gilhooly and Logie, 1980; Stadthagen-Gonzalez and Davis, 2006).

Age of acquisition (AOA) Subjective rating of how old a person is when they first learn that word (Gilhooly and Logie, 1980; Stadthagen-Gonzalez and Davis, 2006) .

Light verbs Number of occurrences of *be*, *have*, *come*, *go*, *give*, *take*, *make*, *do*, *get*, *move*, and *put*, normalized by total number of verbs (Breedin et al., 1998).

Information units Binary feature that measures whether or not any of the words relating to a given information unit were mentioned (from the list of relevant information units in Croisile et al. (1996)) . For example, in the sentence *The boy is getting a cookie and the boy is falling off the stool*, the feature Info unit: boy would have a value of 1.

Key words Integer count of how often specific relevant words are mentioned. For example, in the sentence *The boy is getting a cookie and the boy is falling off the stool*, the key word feature for boy would have a value of 2.

Cosine distance The cosine distance measures the similarity between two utterances; if they are identical, then their cosine distance is zero. The feature ave_cos_dist measures the average cosine distance between every pair of utterances in the transcript. The feature min_cos_dist measures the minimum cosine distance between pairs of utterances. We also measure the proportion of sentence pairs whose cosine distance is less than or equal to a threshold, for threshold = 0.0, 0.3, and 0.5.

Table 7: Text features.

Total duration of speech Total length of all non-silent segments, in milliseconds.
Phonation rate Total duration of active speech divided by the total duration of the sample (including pauses).
Mean pause duration Mean length of pauses > 150 ms.
Short pause count Number of pauses > 150 ms and < 400 ms.
Long pause count Number of pauses ≥ 400 ms.
Pause:word ratio Ratio of silent segments longer than 150 ms to non-silent segments.
Mean/var. F0:3 Mean and variance of the fundamental frequency and first three formant frequencies.
Jitter Measure of the short-term variation in the pitch (frequency) of a voice.
Shimmer Measure of the short-term variation in the loudness (amplitude) of a voice.
Zero-crossing rate (ZCR) An approximation for average pitch of an utterance, defined as the number of sign changes along a signal, per second.
Mean instantaneous power Measure related to the loudness of the voice.
First autocorrelation function Mean and maximum of the first autocorrelation function.
Skewness Measure of lack of symmetry in the distribution of the amplitude of a signal, associated with a tense or "creaky" voice.
Kurtosis Measure of the "peakedness" of a signals amplitude, or specifically the 4th moment of its distribution.
Mean recurrence period density entropy (MRPDE) Measure of periodicity of a signal. Specifically, it measures the extent to which a time series repeats in the phase space. It is similar to linear autocorrelation.
Mel-frequency cepstral coefficient (MFCC) features We measure six features relating to the MFCCs: the mean, variance, skewness, and kurtosis of the energy and the first 13 MFCCs (plus their individual velocities, indicated by Δ, and accelerations, indicated by $\Delta\Delta$), as well as the skewness and kurtosis of their individual means.

Table 8: Acoustic features.

Valence Degree of positive or negative emotion associated with a word.
Arousal Intensity of the emotion associated with a word.
Dominance Degree of control associated with a word.
First person words Normalized occurrence count of *I*, *me*, *my*, *mine*
Maximum voiced frequency Frequency boundary separating periodic and aperiodic components of the speech signal. (Drugman and Stylianou, 2014)
Glottal closure instants Instances of significant excitation of the vocal tract.
Linear prediction residuals Difference between a source-filter speech model and the observed signal (Drugman, 2014).
Peak slope Parameter which differentiates breathy versus tense voice quality (Kane and Gobl, 2011).
Glottal flow (and derivative) Estimate of the air flow through the vocal folds.
Normalized amplitude quotient (NAQ) Parametrization of glottal closing phase (Drugman et al., 2012).
Quasi-open quotient (QOQ) Parameter describing the relative open time of the glottis (Drugman et al., 2012).
Harmonic richness factor Measure of the amount of harmonics in the glottal source (Drugman et al., 2012).
Parabolic spectral parameter Frequency domain measure of the glottal flow (Alku et al., 1997).
Cepstral peak prominence Measure of voice quality based on the cepstrum (Fraile and Godino-Llorente, 2014).

Table 9: New features.