Speech recognition in Alzheimer’s disease and in its assessment

Luke Zhou1, Kathleen C. Fraser3, Frank Rudzicz2,1

1Department of Computer Science, University of Toronto; 2Toronto Rehabilitation Institute, University Health Network; luke.zhou@mail.utoronto.ca, {kfraser,frank}@cs.toronto.edu

Abstract

Narrative, spontaneous speech can provide a valuable source of information about an individual’s cognitive state. Unfortunately, clinical transcription of this type of data is typically done by hand, which is prohibitively time-consuming. In order to automate the entire process, we optimize automatic speech recognition (ASR) for participants with Alzheimer’s disease (AD) in a relatively large clinical database. We extract text features from the resulting transcripts and use these features to identify AD with an SVM classifier. While the accuracy of automatic assessment decreases with increased WER, this is weakly correlated (−0.31). This relative robustness to ASR error is aided by selecting features that are resilient to ASR error.

Index Terms: speech recognition; older voices; Alzheimer’s disease; assessment

1. Introduction

Diagnosing and screening for Alzheimer’s disease (AD) is an expensive and laborious process. Current approaches to the detection of AD and other dementias are unsustainable, given that the incidence of dementia is expected to rise significantly as the global population ages [1]. To ease this burden, modern approaches are being proposed to automate relevant aspects of assessment. Since AD can diminish vocabulary, syntactic complexity, and speech fluency, even in the earliest stages, various systems have been proposed to automatically detect signs of cognitive impairment from speech [2, 3, 4, 5, 6, 7, 8].

Currently, to extract clinically useful lexicosyntactic measures, spontaneous speech must be transcribed by professionals, which is impossible on a very large scale. Unfortunately, no available automatic speech recognition (ASR) system exists for people with AD or cognitive disorders. Moreover, even an optimized system would undoubtedly produce errorful transcripts, which may deteriorate automated assessment performance. For example, estimating the number of nouns in a narrative sample depends on those nouns being correctly recognized.

In this work, we first produce alternative ASR systems to optimize performance in older adults with and without dementia. We then use the resulting transcripts in our existing automatic assessment [8] to classify speakers by diagnostic group (healthy vs. AD), and discuss the features which are both relevant to binary classification and robust to the noisy recognition.

2. Previous work

In this study, we use speech recognition as the input to a system that can analyze a spoken narrative and predict whether the speaker is cognitively healthy or has AD. Other studies in this area have used manually transcribed transcripts [2, 4, 7, 8]. One strategy which combines ASR technology with manual transcripts is to use forced-alignment to measure features such as rate of speech [3, 9]. However, for a speech analysis system to be available online or as part of an in-home continuous monitoring system, there must be no reliance on manual transcriptions at the word-level, which forces alignment requires.

In general, the accuracy of ASR systems on elderly voices tends to decrease with the age of the speaker [10]. Elderly voices typically have increased breathiness, jitter, shimmer, and a decreased rate of speech [10]. Older speakers may also exhibit articulation difficulties, changes in fundamental frequency, and decreased voice intensity [11]. These factors can result in speech that is less intelligible to both human listeners and ASR systems. For example, Hakkani-Tur et al. [12] found that in automatic scoring of a speech-based cognitive test, their ASR system had a higher WER for healthy speakers over the age of 70 than for those under the age of 70, with WERs between 26.3% and 34.1% for the elderly speakers, depending on the task and the gender of the speaker, while the error rates ranged between 21.1% and 28.2% for the younger speakers.

Effective speech recognition can be further challenged by the presence of linguistic impairments such as those occurring in dementia; however, there have been relatively few results reported in this area. Peitner et al. [13] analyzed speech from patients with frontotemporal lobar degeneration, two variants of which affect language: progressive nonfluent aphasia (PNFA) and semantic dementia (SD). They achieved a WER of 61% for speakers with PNFA and 37% for those with SD. They also tested a control group, who had an average WER of 20%.

In previous work, we used commercial ASR to generate transcripts of narrative speech from patients with primary progressive aphasia, which is a neurodegenerative language impairment [14]. We used Nuance Dragon NaturallySpeaking 12.5 Premium with the ‘older voices’ model, and tested it both with the default vocabulary and with a reduced vocabulary of words relevant to our task. Counter-intuitively, Dragon gave higher WER with the reduced vocabulary (97.5%) than with the default vocabulary (67.5%). Because that system is proprietary, we were unable to investigate the changes to the underlying models that led to this result. To overcome this issue, we now use the open-source Kaldi ASR toolkit and a relatively large corpus of speech in AD.

3. Data

Our data are derived from the DementiaBank (DB) corpus1 [15]. These data were collected between 1983 and 1988 at the University of Pittsburgh. Information about the study cohort is

1Downloaded from https://talkbank.org/DementiaBank/ on November 22, 2013

http://dx.doi.org/10.21437/Interspeech.2016-1228
available from Becker et al. [16]. From the ‘AD’ group, we include participants with a diagnosis of ‘possible’ or ‘probable’ AD, resulting in 240 samples from 167 participants over 298.34 minutes. We also include ‘CTRL’ (control) participants, resulting in 233 additional files from 97 speakers over 220.09 minutes. Narrative speech was elicited using the standard ‘Cookie Theft’ picture description task from the Boston Diagnostic Aphasia Examination [17]. Each speech sample was recorded and manually transcribed at the word level following the CHAT protocol [18]. The AD participants produced an average of 104.3 (σ = 59.0) words per narrative, while the control participants produced an average of 114.4 (σ = 59.5) words per narrative. We convert the audio associated with each transcript from MP3 to 16-bit, 16 kHz mono WAV.

To train a baseline system, we also use the Wall Street Journal (phase 2) corpus (WSJ) [19], which consists of 78,000 training utterances (~73 hours of speech), 4,000 of which are the result of spontaneous dictation by journalists.

4. Experiments

Two experiments are conducted. In the first, we optimize ASR for individuals with AD performing the Cookie Theft task. In the second, we perform binary classification of AD using both manual transcripts, as in our previous work, and transcripts derived from the optimized ASR.

4.1. Experiment I: ASR for Alzheimer’s disease

Four experimental systems are produced with Kaldi, using various subgroups of the WSJ and DB data. The mono model uses monophones, tri1 is a triphone model with δ and δδ features, tri2 adds linear discriminant analysis and maximum likelihood linear regression (MLLR) transforms, and tri3 also adds speaker-adaptive training using full feature space MLLR [20].

All experiments use 10-fold cross-validation. In all cases, speakers are partitioned across folds so that individuals never appear in both the training and test sets, and so that each fold is balanced by the diagnosis of the speaker where relevant. In all cases, lexicons and finite-state transducer grammars are obtained on training sets. Insertion penalties (i.e., the cost of inserting a new word into the transcript) are empirically swept from 0.0 to 1.0. The language model (LM) weights, which regulate the importance of the language model relative to the acoustic model, vary along the default range from 9 to 20.

We first verify our methodology by training and testing on WSJ data, using the cross-validation strategy outlined above. Our best WER is 3.72%, using the tri1 model, which is comparable to state-of-the-art for WSJ [21], indicating that our models work well on traditional speech data.

4.1.1. Train on WSJ, test on DB

We test how well the models trained on the WSJ data generalize to older voices and the Cookie Theft task. Here, the models are trained on WSJ data but tested on DB data. We again use the 10-fold cross validation strategy, to allow for appropriate comparisons across experiments. We test with both CTRL and AD data separately, to observe any differences in WER between groups, as shown in Table 1. In both cases, the best WER is achieved with the tri3 model and an insertion penalty of 1.0, but even the best WER is extremely high: 91.74% among CTRLs and 93.54% among those with AD.

![Table 1: WER for 4 models and 3 insertion penalties, averaged over 10 folds and 13 LM weights. Results in bold are optimal, within each configuration of training (Tr) and testing (Te.) data.](image-url)

4.1.2. Train on DB, test on DB

We now train models on DB data alone. While we do not expect these models to generalize beyond the Cookie Theft task, we do expect that constraining the vocabulary to the specific task, and the acoustic models to the specific recording conditions, will improve accuracy. This is confirmed by the greatly improved WER (Table 1). The lowest WERs are again achieved using the tri3 models, but with an insertion penalty of 0. The best WER for controls (36.28%) is significantly lower (t(22) = 10.79, p < 0.001) than for people with AD (44.97%), in keeping with previous work [22].

Although we separated CTRL and AD data to observe the difference in WER between groups, in real-life screening or diagnosis, such as those described in Section 1, which ASR model to use will not be known a priori. Therefore, we also train a model which combines data from both the CTRL and AD data, and test on a combination of these data. In this case, we obtain a best average WER of 38.24%. This result is closer to the CTRL result than the AD result when trained separately, possibly a benefit of increased training set size.

Ultimately, our results suggest that for this particular task and population, training on a relatively small amount of in-domain data achieves better results than training on a large amount of out-of-domain data.

4.2. Experiment II: Diagnosis with noisy transcripts

With the ASR transcripts from the previous section, we can now classify according to diagnosis. We use transcripts generated by models trained on a combination of AD and control data, to mimic expected real-life scenarios. We use the same 10-fold cross-validation framework for assessing binary classification accuracy, using the same folds as in the ASR experiments. This ensures that no information from any test set is present in the training data. Each datum represents one session.

4.2.1. Diagnostic accuracy

To test the effect of WER on classification, we consider each combination of LM weight and insertion penalty separately, and we report the results for each fold. We use language-based fea-
Figure 1: Classification accuracy by WER. Individual points are colored according to the folds to which they belong. Round points are generated from the final experiment in Section 4.1. Square points are generated from gold transcripts and a subset of the other experiments in Section 4.1, and included to show how accuracy changes over the extremes of the WER range.

The accuracy of each fold is based on the points generated from the final experiment in Section 4.1, and included to show how accuracy changes over the extremes of the WER range.

4.2.2. Selected features

It is important to know which language features are still relevant to diagnosis, given erroneous ASR transcripts. Since it gives near-optimum accuracy, we restrict our analysis to the $N = 10$ case, and consider only the data trained and tested on DB using the tri3 model, as well as the gold-standard data. Since different features can be selected in each fold, we report the proportion of folds in which each feature is selected. We subdivide the training folds by WER, to determine if the selected features vary with WER, as shown in Figure 2.

Three features are selected across all folds and WERs, including gold transcripts: frequency (i.e., the frequency with which a word occurs in the SUBTL speech corpus [24]), word length in characters, and the frequency of the construction VP → VBG PP (verb phrases as gerund verb and prepositional phrase). The number of verbs and the informational feature concept: window are also selected more than 50% of the time across WERs. These features appear to be robust to ASR error. Other features are selected very frequently in the gold transcripts but rarely in ASR transcripts, including NP → PRP (noun phrases as prepositional phrases), not-in-dictionary (NID, aka out-of-vocabulary) words, and pronoun ratio (i.e., the ratio of pronouns to pronouns + nouns). Other features are selected only in the ASR transcripts, e.g., keyword: window and concept: girl. That these features distinguish CTRL from AD only in those transcripts suggests an asymmetry in how related words are recognized in the two groups.

Interestingly, the concept: window feature (which includes mentions of words relating to windows, such as “frame” or “glass”) is selected across the range of WERs, but the keyword: window feature (which includes only mentions of the word “window”) is only selected in ASR transcripts. In fact, both of these features are significantly different between the groups in the gold transcripts, but the difference is greater for the concept, and the keyword is not generally selected until we allow $N = 20$. In the ASR transcripts, there are fewer significant differences, and so keyword: window is selected earlier.
5. Discussion and future work

Our ASR results improve on previous work, and provide useful information for diagnosis. However, we are limited by the somewhat poor quality of the audio in DementiaBank. Some files are dominated by noise, and are difficult to decipher even by a human listener. Preliminary attempts to reduce noise using least squares amplitude estimation did not significantly improve the WER, and this remains the subject of ongoing work.

This paper is the first to detail the impact of ASR WER on the usefulness of text-based features in clinical diagnosis. While the broad trends in our results are expected (diagnostic accuracy decreases as WER increases), we observe a wide variance in accuracy associated with narrow bands of WER. Considering only the transcripts derived from our most successful ASR experiments, there is no significant effect on accuracy of any of the independent variables we examined. This underscores the fact that two ASR transcripts can have the same WER without containing the same diagnostically relevant information. Future work will seek to determine common properties of the transcripts which are associated with higher diagnostic accuracies, potentially including demographic information.

We have identified several features which appear to be robust to ASR error, but further work is required to better understand why these features remain significantly different in the ASR transcripts (and conversely, why some features are fragile to the recognition process). Further work will also explore the features which are significant only in the ASR transcripts. While these features cannot inform our knowledge of pathological speech patterns in dementia, if the differences are systematic, then they could still have utility in a diagnostic framework.

6. Acknowledgements

This research is funded by the Natural Sciences and Engineering Research Council of Canada (including grant RGPIN 435874) and Dr. Rudzicz’s Young Investigator award from the Alzheimer Society of Canada.
7. References


