Leveraging Syntactic, Semantic Patterns and Acoustic signals of pathological speech to Predict Clinical Scores for Alzheimer's Dementia: The Adress Challenge

Shahla Farzana

University of Illinois Chicago

sfarza3@@uic.edu

Abstract

We use a set of 636 lexicosyntactic, non-verbal features extracted from 108 speech samples in ADRESS Challenge dataset, which is a subcorpus of DementiaBank to predict clinical MMSE scores, an indicator of the severity of cognitive decline associated with dementia. As the speech audio sample is available, we used segment based approach to predict MMSE scores over each audio segment , and obtain a root mean squared error (RMSE) of 6.59 in predicting MMSE, whereas using text based features gave us better result in predicting MMSE (RMSE 4.67).

Index Terms: speech recognition, human-computer interaction, MMSE

1. Introduction

Recent advancements in artificial intelligence have opened new pathways for improving patient and clinician healthcare experiences, with technologies including but not limited to predictive disease modeling, ambient healthcare monitoring, and clinical record-keeping assistance. At the same time, shifting population demographics have instigated new health care concerns especially for older adults. Dementia is one such increasingly critical concern as median population ages around the globe continue to rise. There is no single laboratory test that can identify dementia with absolute certainty. Typically, probable dementia is diagnosed using a cognitive test called the Mini Mental State Examination (MMSE). It provides a score on a scale of 0 (greatest cognitive decline) to 30 (no cognitive decline), based on a series of questions in five areas: orientation, registration, attention, memory, and language [1]. This cognitive test can be time-consuming and relatively costly, often requiring a trained neuropsychologist or physician to administer the test in a clinical setting.

Changes in cognitive ability due to neurodegeneration associated with Alzheimer's disease (AD) lead to a progressive decline in memory and language quality. Patients experience deterioration in sensory, working, declarative, and non-declarative memory. It may lead to a decrease in the grammatical complexity and lexical content of their speech [2]. However, these changes can be subtle, particularly in the early stages of dementia. Automated dementia detection models may be able to learn patterns of Alzheimer's disease and other related dementia that are not readily apparent to clinicians. Consequently, early diagnosis of Alzheimer's disease and related dementia (ADRD) may lead to improved outcomes including mitigation or slowing of harmful symptoms.

~~1.1. Goals and Hypothesis~~

Since dynamic changes in linguistic ability in patients with AD differ from those in typical healthy older adults [2], previous research shows that considering speech samples over time would aid in estimating underlying cognitive status [3]. In this project, we plan to leverage the longitudinal speech samples of ADRD and healthy control subjects to estimate MMSE scores.

Another aspect of this problem is that collecting speech samples of older adults and administering cognitive tests to get MMSE scores is time consuming and resource intensive. Although limited publicly-available data exists for these tasks, the data that is available often contains common underlying syntactic and semantic structures. We hypothesize that these structures can be used to learn fine-grained measures of cognitive status, like MMSE score. Due to the scarcity of labelled data, we also hypothesize that semi-supervised learning approaches can offer promise when few labels are available by allowing models to supplement their training with unlabeled data [4]. In this project, we would like to experiment with models that can leverage speech samples with missing labels to predict MMSE scores for new samples in our test set.

2. Related Works

Research into automatic dementia detection has recently gained increased attention. Some of this growth may be attributed to heightened public interest as shifting age demographics become apparent [1]. Progress has also been accelerated by the release of large-scale language-based datasets to facilitate dementia detection [2], and advancements in machine learning techniques [3] and neural models [4], allowing automated techniques to more productively make use of this data.

While these studies have obtained promising results in classifying patients with dementia based on linguistic features, there is limited work modelling the progression of such features over time. A similar analysis performed by Yancheva [5] examined the progression of a small set of lexico-syntactic features in 6 patients with AD or mild cognitive impairment (MCI), with a minimum of 3 longitudinal samples in DementiaBank [9]. Analysis of the features over time did not reveal conclusive patterns; however, neither study involved neural models. Few previous works have pondered much on the regression problem of predicting MMSE scores from longitudinal speech samples or leveraging unlabeled data. We will explore these two approaches for text regression tasks using neural models that combine implicitly-learned linguistic features (e.g., word embeddings) with handcrafted acoustic features.

We should try to think of a shorter name

Start the abstract with a one-line "hook"

Were these selected from a list?

We should tweak the wording here so it doesn't sound like a proposal

Add a list of our key contributions

Is this the same citation as the next sentence?

We should update this section so it talks about general MMSE prediction, not longitudinal modeling.

	AD		non-AD	
Age Interval	Male	Female	Male	Female
[50, 55)	1	0	1	0
[55, 60)	5	4	5	4
[60, 65)	3	6	3	6
[65, 70)	6	10	6	10
[70, 75)	6	8	6	8
[75, 80)	3	2	3	2
Total	24	30	24	30

Figure 1: *ADReSS Training Set statistics. [8]*

	AD		non-AD	
Age Interval	Male	Female	Male	Female
[50, 55)	1	0	1	0
[55, 60)	2	2	2	2
[60, 65)	1	3	1	3
[65, 70)	3	4	3	4
[70, 75)	3	3	3	3
[75, 80)	1	1	1	1
Total	11	13	11	13

Figure 2: *ADReSS Test Set statistics. [8]*

3. ADRESS Challenge Dataset

The main objective of the ADReSS challenge [6] is to make available a benchmark dataset of spontaneous speech, which is acoustically pre-processed and balanced in terms of age and gender, defining a shared task through which different approaches to AD recognition in spontaneous speech can be compared. The data consists of speech recordings and transcripts of spoken picture descriptions elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [7] and is originally taken from the Pitt corpus [2] of DementiaBank dataset. The speech sample is basically related to conversation between an investigator and an older adult where the investigator asks the person to describe what is depicted in the picture. There is no specific time limit for the conversation. People can talk as long as they want. The participants were labelled as control (no cognitive decline) or AD (declined cognitively) based on their prior diagnostic test result.

The recorded speech has been segmented for voice activity using a simple voice activity detection algorithm based on signal energy thresholding. The average number of conversation length is 19.64 across training set and 12.29 in test set. Audio volume was normalised across all speech segments to control for variation caused by recording conditions such as microphone placement.

4. MMSE Prediction Task

present work uses a set of automatically-extracted lexicosyntactic, acoustic, and semantic features for estimating continuous MMSE scores on a scale of 0 to 30, using a several machine learning techniques for representing relationships between observed linguistic measures and underlying clinical scores. T Unlike classification, MMSE prediction is relatively uncommon in the literature, despite MMSE scores often being available. While the baseline published by ADRESSS Challenge [8], in our work, we experimented with both textual and audio data and compared our model performance with the baseline.

4.0.1. Data Preprocessing

ADRESS challenge released both transcripts and audio files for the prediction task. All speakers in all transcripts are anonymous to protect user privacy; each transcript and each speaker within a given transcript were instead linked to a unique ID (transcripts are represented using an interview number, and speakers are identified using participant numbers). Participant demographic data (e.g., age, gender, and interview date) can be extracted from accompanying metadata files using the participant and transcript IDs. 108 unique participants are represented among the 108 interview transcripts included in the corpus. We preprocessed the transcripts as we tokk only participant's utterance and cleaned the data removing numbers, punctuation and unwanted symbols.

4.1. Feature Extraction

Textual Features: As there is transcription available for each interview, we extracted following features:

- We extracted *n-gram* features for $n \in \{1, 2, 3\}$ from the entire training corpus, comprising all training utterances. We retained only those n-grams that appeared at least five times and at most 50 times across the training data, constructed a sparse feature vector for each utterance containing one dimension for each n-gram. Feature values were filled using TF-IDF counts for a given utterance, and each vector was L2-normalized with unit modulus. Vovabulary size is 621.
- We used *Demographic* feature like age and gender.
- We extracted some *nonverbal* features from the transcripts which was coded with special symbol. These feature are short_pause_count, long_pause_count, word_repetition_count, retracing_count (restarting the same phrase or segment again), filled_pause_count (e.g. *uh, umm*, incomplete_utterance_count and normalise by the number of words uttered in the conversation. We also added word count, utterance count. These hand crafted features added up to 8.
- *Psycholinguistic* features are linguistic properties of words that effect word processing and learnability [9]. Age of Acquisition, Familiarity, Imageability, Concreteness, word sentiment these 5 features are used.

Acoustic Features: Mel-frequency Cepstral Coefficients (MFCCS) are frequently used is speech processing and represent spectral information from the speech signal, using a scale known as the "mel-frequency scale," which is chosen to mimic the way humans perceive audio. The MFCCS are then calculated from the "mel log powers" of the first 14 coefficients calculated by the fourier transform of each segment5. Each segment from the original signal then produces 14 MFCCS, resulting in 14 MFCC distributions. We then calculate the mean, variance, skewness, and kurtosis of the first 14 MFCCS, representing spectral information from the speech signal. We did the same for the velocity and acceleration, where velocity is calculated as the delta between consecutive time steps and acceleration as the double-deltas. There are total 173 audio features for each segment.

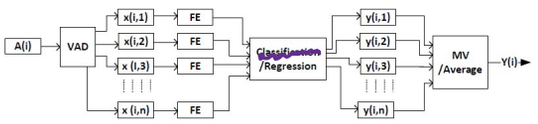

Figure 3: *System Architecture: A(i), the audio recording of ith subjects is segmented using voice activity detection (VAD) into n segments x(i, n). Feature extraction (FE) is performed at segment level. The output of classification or regression for the nth segment of the ith audio recording is denoted y(i, n). MV outputs the majority voting for classification, and Average the mean regression score. [8]*

The figure is really useful, but we should also explain it in this section.

4.2. Model

We have build models separately based on the textual features and acoustic features. As the audio files are segmented in 10-20s normalised chunks and divided into multiple chunks across a conversation, we used a segment based model with basic classifiers (Supprt Vector Regression (SVR) with polynomial kernel, gaussian process regression (GP, with a squared exponential kernel), [8]. For textual features, we used model that predict score for the entire conversation not a particular segment. We used the same basic classifiers (SVR, GP).

New section : Evaluation

4.3. Result

Briefly describe conditions

The regression results are reported as root mean squared error (RMSE) and R-squared value of the models in 1 for test data. These results show that the SVR (4.67) provides the best RMSE with R-squared value (0.449) using selected textual features showing promising performance. We also note that SVR using only *n-gram* features also shows reasonably good result with RMSE (4.99) with R-squared value (0.372). Using on the acoustic features based on the segment model, we could get RMSE (6.59) with R-squared value (0.093). The baseline mode using audio based segment model [8] could achieve RMSE (0.093) using Decision Tree (DT) model. From 1, we cansee that most of out text based models (SVR) outpeerforms that baseline with a descent margin. And we also added the R-squared value for each model which explains how much our features contribute to the regression task.

Update with challenge scores

Table 1: *MMSE prediction test results. The evaluation metric is **RMSE** and the value inside the brace is **R-Squared value** for the model*

Features	SVR	GP
ALL	5.43 (0.258)	6.31 (-0.001)
ALL-PSYCHOLINGUISTIC	5.01 (0.368)	6.35 (-0.015)
NGRAM	4.99 (0.372)	5.69 (0.185)
SELECTED-FEATURE	**4.67 (0.449)**	6.34 (-0.015)
ACOUSTIC	6.59 (0.093)	6.71 (0.135)

Table 2: *Top 10 features based on Randomforest regression classifier with 100 trees*

Style Name	Entities in a Paper
this	0.284
here	0.050
word_count	0.044
fall	0.037
well	0.034
laughs (non-verbal)	0.034
short_pause_count	0.021
in the	0.015
cookie jar and	0.014
it uh	0.013

4.4. Feature Analysis

To have some insights on which twxtual features are contributing most, we used feature selection technique based on Randomforest regression with 100 trees. Random forests consist of 4 –12 hundred decision trees, each of them built over a random extraction of the observations from the dataset and a random extraction of the features. Not every tree sees all the features or all the observations, and this guarantees that the trees are decorrelated and therefore less prone to over-fitting. Each tree is also a sequence of yes-no questions based on a single or combination of features. At each node (this is at each question), the three divides the dataset into 2 buckets, each of them hosting observations that are more similar among themselves and different from the ones in the other bucket. Therefore, the importance of each feature is derived from how "pure" each of the buckets is. For regression the measure of impurity is variance. Therefore, when training a tree, it is possible to compute how much each feature decreases the impurity. The more a feature decreases the impurity, the more important the feature is. In random forests, the impurity decrease from each feature can be averaged across trees to determine the final importance of the variable. We have added the top 10 features where *n-gram,nonverbal* features are included with the importance value based on which features contribute the most in splitting the trees in Randomforest classifier. There are total 90 text based features in our selected features and it surely increased performance of regression 1.

5. Discussion and Conclusion

As we could see from the moder evaluation and feature analysis, text based features revealed to be more informative in contributing he regression task. It is not easy to get reasonable result out of acoustic feature although the audio files are preprocessed and normalised to reduce biases. It seems that though acoustic features proved to be important in Dementia classification task along with textual features [10], it alone can not contribute much in MMSE regression task. Further investigation of fusing two types of features (textual and acoustic) may give us better insights of how to leverage from acoustic signals for this task to attain better accuracy. As from the feature selection, we see that some *nonverbal* features are proved to be important for regression, they might be extracted from acoustic signal rather than from transcripts. Also the segmented model can be applied on texts (per utterance) to see how it performs for regression task. While MMSE is one of the most widely used clinical tests for

cognitive ability, it is somewhat coarse, lacking sensitivity to subtle changes in cognition in the early stages of dementia, as well as having a high false-negative rate in addition to interannotator disagreement and test-retest variability [11, 12]. While automated prediction of the MMSE score may aid the screening process for AD by reducing the cost and time involved, and improving reliability, future work will explore more precise measures of cognitive decline. The Montreal Cognitive Assessent (MoCA) and the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) [13] are screening tests which have been shown to have higher sensitivity than MMSE to subtle changes in cognitive decline in populations with MCI and mild dementia [14]. Future studies are needed to assess the validity of automatic scoring of such tests as a more fine-grained measure of the progression of cognitive decline.

6. Bibliographical References

~~7. References~~

[1] E. S. Tom, A. R. Hubbard, K. P. Crane, J. S. Haneuse, J. Bowen, C. W. McCormick, S. McCurry, and B. E. Larson, "Characterization of dementia and alzheimer's disease in an older population: updated incidence and life expectancy with and without dementia," *American journal of public health*, pp. 408–413, 2015.

[2] B. F. L. O. L. S. J. Becker, J. T. and K. L. McGonigle, "The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51(6), pp. 585–594, 1994.

[3] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden, "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*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 78–87. [Online]. Available: https://www.aclweb.org/anthology/W14-3210

[4] F. Di Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 302–308. [Online]. Available: https://www.aclweb.org/anthology/P19-2042

[5] M. Yancheva, K. Fraser, and F. Rudzicz, "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*. Dresden, Germany: Association for Computational Linguistics, Sep. 2015, pp. 134–139. [Online]. Available: https://www.aclweb.org/anthology/W15-5123

[6] "ADRESS kernel description," http://www.homepages.ed.ac.uk/sluzfil/ADReSS/, accessed: 2010-09-30.

[7] C. Roth, *Boston Diagnostic Aphasia Examination*. New York, NY: Springer New York, 2011, pp. 428–430. [Online]. Available: https://doi.org/10.1007/978-0-387-79948-3$_8$68

[8] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[9] T. Salsbury, S. A. Crossley, and D. S. McNamara, "Psycholinguistic word information in second language oral discourse," *Second Language Research*, vol. 27, no. 3, pp. 343–360, 2011. [Online]. Available: https://doi.org/10.1177/0267658310395851

[10] V. Masrani, "Detecting dementia from written and spoken language," 2018.

[11] D. Molloy and T. Standish, "A guide to the standardized mini-mental state examination," *International psychogeriatrics / IPA*, vol. 9 Suppl 1, pp. 87–94; discussion 143, 02 1997.

[12] ?

[13] A. R. Loughan, S. E. Braun, and A. Lanoye, "Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary utility in adult neuro-oncology," *Neuro-Oncology Practice*, vol. 6, no. 4, pp. 289–296, 12 2018. [Online]. Available: https://doi.org/10.1093/nop/npy050

[14] C. Zadikoff, S. H. Fox, D. F. Tang-Wai, T. Thomsen, R. M. de Bie, P. Wadia, J. Miyasaki, S. Duff-Canning, A. E. Lang, and C. Marras, "A comparison of the mini mental state exam to the montreal cognitive assessment in identifying cognitive deficits in parkinson's disease," *Movement Disorders*, vol. 23, no. 2, pp. 297–299, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.21837