INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Speech Processing Approach for Diagnosing Dementia in an Early Stage

Roozbeh Sadeghian[1], *J. David Schaffer*[2], *Stephen A. Zahorian*[3]

[1]Department of Analytics, Harrisburg University of Science and Technology, Pennsylvania, USA
[2]Institute for Multigenerational Studies, Binghamton University, New York, USA
[3] Department of Electrical and Computer Engineering, Watson School, Binghamton University, New York, USA

rsadeghian@harrisburgu.edu, {dschaffe, zahorian}@binghamton.edu

Abstract

The clinical diagnosis of Alzheimer's disease and other dementias is very challenging, especially in the early stages. Our hypothesis is that any disease that affects particular brain regions involved in speech production and processing will also leave detectable finger prints in the speech. Computerized analysis of speech signals and computational linguistics have progressed to the point where an automatic speech analysis system is a promising approach for a low-cost non-invasive diagnostic tool for early detection of Alzheimer's disease.

We present empirical evidence that strong discrimination between subjects with a diagnosis of probable Alzheimer's versus matched normal controls can be achieved with a combination of acoustic features from speech, linguistic features extracted from an automatically determined transcription of the speech including punctuation, and results of a mini mental state exam (MMSE). We also show that discrimination is nearly as strong even if the MMSE is not used, which implies that a fully automated system is feasible. Since commercial automatic speech recognition (ASR) tools were unable to provide transcripts for about half of our speech samples, a customized ASR system was developed.

Index Terms: speech recognition, dementia, machine learning, MMSE

1. Introduction

Alzheimer's disease (AD) is the 6th leading cause of death in the United States [1] and a significant burden on the nation's and the world's health care systems, those who suffer from it, and their families. It is very difficult to diagnose, particularly in the early stages [2]. A common screening test often administered by physicians, is the mini mental state exam (MMSE) [3]. By itself, it is not diagnostic, but is often used to identify patients for referral to specialists for careful diagnosis. The MMSE is a simple pencil and paper test taking about 10 minutes and requiring only modest training. If an equally simple and short speech-based test could improve the accuracy of the MMSE this would seem to provide clinical value.

The idea that speech patterns might reveal early stage dementia has been investigated in [4], [5], and [6]. There are many relevant studies including those that attempt to establish specific voice-based features whose distributions are statistically different between those with dementia and normal controls [7]. Computerized analysis of speech signals and computational linguistics have progressed to the point where an automatic speech analysis system is a promising approach for a low-cost non-invasive diagnostic tool for early detection of Alzheimer's disease. In two recent studies [8] and [9], by analyzing spontaneous speech, some biomarkers were extracted as features. Machine learning algorithms have been developed to build diagnostic models using syntactic and lexical features resulting from verbal utterances of patients [10]. Some efforts have also tackled discriminations among dementia types and degrees of severity [11], [12], [13], [14].

In this paper, we provide results from our ongoing work into the feasibility of developing such a test. Particularly, we describe our automatic speech recognition (ASR) technology that is needed to make the test fully automatic. Our database of speech samples, the acoustic and linguistic features we extract (fully automated), and our results showing an improvement in diagnostic precision over the MMSE alone, are presented. We also compare the results when using manual transcripts and our newly automated transcripts.

2. Speech processing

2.1. Database

A standard protocol for collecting speech samples for aphasia work is to ask volunteers to describe what they see in a picture. They are able to view the picture while they speak (i.e. it is not a memory test). This paradigm was used for all speech samples used in this work. For this task, we collected 72 recordings using modern digital recording equipment and a new picture[1] [15]. A brief demographic summary of the participants is shown in Table 1. Clinical diagnoses (ground truth) were provided by treating physicians.

Table 1: *Demographic Summary*
AD = Alzheimer's disease, NL = normal control

Grp	n	Age (sd)	race % (white)	years_edu (sd)	MMSE (sd)
NL	46	71.43 (12.6)	98	13.28 (2.4)	28.70 (1.5)
AD	26	78.48 (10.9)	100	13.81 (2.3)	20.92 (6.6)
Total	72	74.04 (12.4)	99	13.48 (2.4)	25.89 (5.6)

[1] The interested reader may see our picture and listen to a voice sample at this web site: http://acoustics.org/2asp5-using-automatic-speech-recognition-to-identify-dementia-in-early-stages-roozbeh-sadeghian-j-david-schaffer-and-stephen-a-zahorian/

http://dx.doi.org/10.21437/Interspeech.2017-1712

The average recording sample length was 75.1 seconds (sd 61.0). Some modest preprocessing was performed on audio files, such as removing the beginning and ending pauses, click removal and signal strength normalization. These processes are straightforward to automate. The resulting acoustic speech files were processed directly for acoustic features such as pauses and pitch contours. A manual transcript was generated for each of the 72 recordings. Linguistic features (e.g. word counts, syntactic complexity, idea density) were extracted from manual and automatic transcripts, and used for experiments as described in later sections of this paper.

2.2. Acoustic feature extraction

Each wave file was processed by three methods for separating speech from pauses, one using pitch, one using energy, and one using a Voice Activity Detector (VAD) [16]. With the speech sample broken into pause and speech events, 22 metrics were computed including the total speech length, the number of pauses, the fraction of the speaking time that was pause, the fraction of pauses in certain time windows (e.g. less than 0.5 second, 0.5-1 second, …), and the fraction of the pauses in each quartile of the sample. In addition, the distribution of the voiced pitches in 10ms windows provided a mean, median, variance, minimum, and maximum that we hoped might provide an indication of emotive effect in the voice. Space limits preclude full description of all features.

2.3. Linguistic feature extraction

Each transcript was passed to the Charniak Parser [17] trained with the Penn Treebank Switchboard corpus. The raw text of the transcript, and the part-of-speech (POS) tagged parser outputs were used to compute a number of linguistic metrics. The syntactic complexity measures computed by Roark et al. [18] were computed, including a re-implementation of idea density [19]. A number of metrics that capture various aspects of vocabulary richness were also computed as well as counts of words related to the picture content. The Linguistic Inquiry Word Counts (LIWC) were also computed [20]. These and all the other features, such as speech pause and pitch features, were combined into a single feature vector for each subject. These 232 features from the speech samples were combined with demographic features and MMSE to give 237 total potential features.

3. Speech and punctuation recognizer

3.1. ASR system

In a fully automatic system, all the steps must be done automatically, including the crucial step of speech-to-text. Several attempts to apply commercial ASR tools revealed their limitations: these tools typically need training for each speaker and have restrictions on sample length. Since commercial ASR tools failed on about half our samples, we had to develop our own automatic speech recognition (ASR). There are some aspects which made the task more doable: limited domain vocabulary and no requirement for real-time ASR. In addition, ASR is eventually combined with easy to detect acoustic metrics, such as pauses, thus presumably reducing the burden of the ASR. However, there are also some challenges: limited training data, poorly articulated words, presence of non-speech sounds, and instances of word patterns difficult to predict by a language model, and difficulties associated with ASR for the elderly [21]. We initially did use the commercial ASR system Nuance Dragon Naturally speaking (version 13) but the system was not promising since we had to tune the software for each speaker which is not feasible in clinical practice. Also, even using this strategy, the overall Word Error Rate (WER) was quite high (about 35%), on the samples it could handle.

The first steps for designing a custom ASR system for this project were to prepare the dictionary (lexicon), make transcriptions to use for creating a language model, and to eliminate some problems in the speech data. Many of the data problems were due to errors in the manually transcribed words. Another problem was the Out of Vocabulary (OOV) words in transcripts. For the ASR acoustic models, we first created simple monophone models, then used those models to design triphone models, and finally incorporated a deep neural net (DNN) to compute posterior probabilities of the tied states in the triphone models. All models were built with 39 MFCC features, computed with 25ms frames spaced apart by 10ms. A bigram language model was developed based on the manually provided transcriptions. For monophone models, 3 state HMMs with 64 mixtures were used whereas for triphone models, 500 tied states were modeled with 8000 Gaussian mixtures.

The ASR was performed using the powerful and flexible Kaldi [22] toolbox. The DNN was implemented with two hidden layers, each having 1024 neurons. The initial learning rate α=0.015 and it was decreased to α=0.002 in the final step. The activation function was hyperbolic tangent. To speed up the training, a minibatch size of 128 was used. Since there are many non-speech events and silent sections in the speech files, a VAD (Voice Activity Detector) was used to remove them. We used a context window of 9 frames (4 behind, middle frame, 4 ahead). Since the amount of data was low, we trained on all speakers except one left out for testing, and repeated for all speakers (referred to as leave one out (LOO) method). The mean of the WER (averaged over all speakers) using this method was about 31%. Fig. 1 shows WER for all subjects.

The average test accuracy was 68.7%$\pm$16% with a maximum accuracy of 93.2% and minimum of 22.1%. The transcriptions for the tests speakers were used for the punctuation algorithm, described in a later section.

One way of checking both the overall correctness of the ASR system, and also to determine the potential accuracy if more training data were available, is to use components of the test data for training.

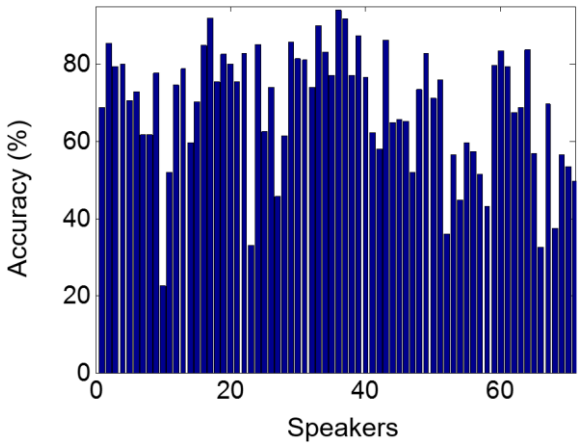

Figure 1: *The accuracy of ASR using LOO method*

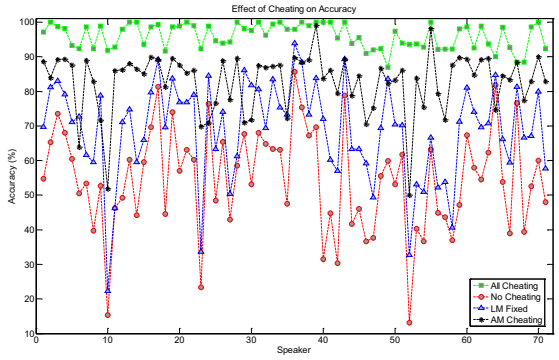

Figure 2: *ASR accuracy for each speaker with selective use of test data for training*

Ultimately, of course, the training and test data must be completely separate in order to be able to predict how well a system will perform on new unseen data. However, controlled insertion of test data into the training set can be used diagnostically to determine which aspects of training are most deficient. In this work, 4 different cases were evaluated: 1) No test data used for training; 2) for both the LM (Language Model) and AM (Acoustic Model), the training and test sets are identical; 3) for the AM, training and test sets are identical, LM has different training and test sets; 4) for the LM, training and test sets are identical, AM has distinct training and test sets. In *Figure 2,* we show the accuracy for each speaker for each condition. Case 2 (average accuracy of 96.0%) is an upper limit of performance if a really large database were available. Case 3 has an average accuracy of 82.9%, whereas case 4 has an average accuracy of 76.7%, thus indicating more training data for the AM would likely benefit accuracy more than increased data for the LM. Altogether, however, a larger database should be used. In future work, priority should be given to enlarging the database. For the present paper, except for Fig. 2, all results pertain to "honest" case 1.

3.2. Automatic Punctuation

One of the challenges in extracting the linguistic features is how to determine the punctuation of the automatically transcribed speech because ASR systems typically only recognize words, and ignore punctuation. However, punctuation is required for determining some linguistic features. To resolve this issue, the approach of Tilk and Alum [23] was used to punctuate the output of the ASR system. In this method, a bidirectional Recurrent Neural Network (RNN) with an attention mechanism is used to punctuate the text. In each of the recurrent layers, Gated Recurrent Units (GRU) are used to eliminate the effect of different time scales which appear as dependencies. To capture the relevance of the parts in a context, an attention mechanism was used whereby it chose which punctuation (period, comma or question mark) to use.

For training this RNN model, similar to the approach of Tilk and Alum [23], the English part of "Europalv7" [24] was used which contains more than 2 million sentences with around 53 million words from more than 800 speakers. There is a lexicon of all possible words in which assigned relevant number to each word and RNN will use these assigned numbers for further processing. The RNN was configured with two hidden layers, each with 256 neurons. The inputs of the network are (at maximum) 200 words of a sentence starting from the first word of the sentence. The outputs correspond to locations and type of punctuation.

It is very difficult to meaningfully quantify the accuracy of the automatic punctuation, partly because even the manually transcribed punctuation is highly subjective. We observed the automatic punctuation matched the manual punctuation (commas, periods, question marks) for approximately 50% of the cases. Extraneous punctuation occurred in about 10% of locations where there should have been none. The parser requires sentence boundaries, and some of the linguistic features use the punctuation, so the accuracy of the punctuation would be expected to affect the ultimate diagnostic classification performance, the most important figure of merit.

4. Experiments

The 72 subjects were divided 90/10 into training and validation sets and full 10-fold cross validation was performed.

We tested several approaches to feature selection and several classifiers, but here we report only the method using a best-first greedy algorithm with a multi-layer perceptron classifier. For this model, a NN with one hidden layer (containing 25 neurons) was used as a two-way classifier. The activation nodes were sigmoid. The inputs were features[2] to be evaluated (from training data) and the outputs were labels for each subject.

4.1. MMSE feature evaluation

The most informative single feature was generally the MMSE score alone. MMSE scores greater than or equal to 24 points (out of 30) indicates a normal cognition. Below this, scores can indicate severe (≤9 points), moderate (10–18 points) or mild (19–23 points) cognitive impairment [25]. To test the goodness of this feature, a two way NN (with similar specifications to above mentioned model) was trained as the AD/NL classifier using only this feature for the classifier. An accuracy of 70.8% was obtained; the confusion matrix given in Table 2.

Table 2: *Confusion matrix using MMSE score only as a feature*

		Estimated	
		AD	NL
Actual	AD	23	3
	NL	18	28

4.2. Using complete set of features

For the next set of experiments, a greedy approach was used whereby initially each of the 237 potential features was evaluated individually and the best performing feature was found. Best performance was determined by highest accuracy of the MLP on a group of test speakers. The decay parameter for this experiments was set to be 0.1 while the rate of dropout was set to 0.02 experimentally. The accuracy achieved was 94.4% using only five features, one of which was the MMSE score. The five features selected (in order of importance) were MMSE score, race, fraction of pauses greater than 10sec, fraction of speech length that was pause and LIWC

[2] All linguistic features were extracted from the ASR transcripts unless otherwise noted.

quantitative feature (words indicating quantities). The resulting confusion matrix is given in Table 3.

Table 3: *Confusion matrix using 5 best features selected from the complete features set*

		Estimated	
		AD	NL
Actual	AD	24	2
	NL	2	44

4.3. Demographic, Linguistic and acoustic features only

As described above, the accuracy of an AD/NL classifier is much higher if features are based on more than just the MMSE score (94.4% accuracy versus 70.8%). However, given that the MMSE score appears to be the most informative feature, but would be difficult to automate, a logical next step is to evaluate a system which does not include the MMSE score as a possible feature. If all possible features, except MMSE scores, are considered, detection accuracy of approximately 93.1%, for linguistic features derived from manual transcripts, and 91.7% for linguistic features derived from the ASR/automatic punctuation transcripts. The confusion matrix, based on the automatically generated transcripts, is given in Table 4. In order to achieve the 91.7% accuracy, 12 features were needed, as listed in Table.

Table 4*: Confusion matrix obtained using best 12 demographic, linguistic, and acoustic features*

		Estimated	
		AD	NL
Actual	AD	23	3
	NL	3	43

Table 5: *Features selected using all the features except MMSE*

Feature No.	Feature Name
1	Race
2	Speech rate (pitch based)
3	Content density
4	Fraction of pause greater than 1 sec
5	Speech rate
6	Total no. of pauses
7	LIWC_compare
8	Idea Density Ratio
9	Fraction of pause less than 0.5 sec
10	LIWC_we
11	LIWC_quant
12	LIWC_leisure

4.4. Demographic and acoustic features

For the last set of experiments, we considered the case where only demographic and acoustic features (no linguistic or MMSE) are in the initial candidate feature pool (81 features). Using the identical procedure as used for the previous two cases, an accuracy of 83.3% was obtained. This reveals that if only demographic and acoustic features (the "easy" ones) are considered, reasonably high accuracy is obtained, considerably higher than the 70.8% from the MMSE score alone, but much lower than the 91.7% possible if linguistic features are also included. The confusion matrix is given in Table 4. Table 5 lists the 7 features selected.

Table 4: Confusion matrix of using 7 best demographic and acoustic features

		Estimated	
		AD	NL
Actual	AD	15	11
	NL	1	45

Table 5: *Features selected using only demographic and acoustic features*

Feature No.	Feature Name
1	Race
2	Speech rate (using energy)
3	Speech rate (using VAD)
4	Speech less than 0.5 sec (pitch)
5	Total number of pauses (energy)
6	Total utterance length (pitch)
7	Total no. of pauses (pitch)

5. Conclusion and Discussion

There do appear to be strong patterns among the speech features that are able to discriminate the subjects with probable Alzheimer's disease from the normal controls.

The greedy algorithm combined with the neural network two-way classifier was very promising for both feature selection and final recognizer. In future work, the NN method could be improved in terms of more thorough searching by saving the top N (where N is some small number such as 5 to 10) choices at the end of each iteration, at the expense of some slowdown in speed.

We believe this study provides encouragement to seek speech patterns that could be diagnostic for dementia. The weaknesses of this study include the cross-sectional design that strives for a single pattern that works over the whole variety of subjects in each class. A longitudinal study would permit each subject to serve as his own control, helping to mitigate the large within-group variance in speaking patterns. The features used are by no means all the speech features that have been associated with dementia. The computational linguistics domain contains several additional interesting speech features that, with some effort, could be included in our basket.

The accuracy of 94% for diagnosing Alzheimer seems promising considering this small number of samples. Additionally, the results of manually and automatically transcribed systems are close to each other which shows that the ASR system worked in an acceptable range and the punctuator system was likely accurate enough.

6. References

[1] A. Association, "2016 Alzheimer's disease facts and figures," *Alzheimer's & Dementia,* vol. 12, no. 4, pp. 8-13, 2016.

[2] B. Dubois and e. al., "Timely diagnosis for Alzheimer's disease: A literature review on benefits and challenges," *Journal of Alzheimer's disease,* vol. 49, no. 3, pp. 617-631, 2015.

[3] M. Folstein, S. Folstein and P. R. McHugh, "Mini-mental state: A practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatr Res,* vol. 12, pp. 189-198, 1975.

[4] F. Cuetos, J. C. Arango-Lasprilla, C. Uribe, C. Vialencia and F. Lopera, "Linguistic changes in verbal expression: A preclinical marker of Alzheimer's disease," *Journal of the International Neuropsychological Society,* vol. 13, pp. 433-439, 2007.

[5] D. M. Jacobs, M. Sano, G. Dooneief, K. Marder, K. L. Bell and Y. Stern, "Neuropsychological detection and characterization of preclinical Alzheimer's disease," *Neurology,* vol. 45, pp. 957-962, 1995.

[6] B. B. Lowit, C. Dobinson and P. Howell, "An investigation into the influences of age, pathology and cognition on speech production," *Journal of Medical Speech Language Patholology,* vol. 14, pp. 253-262, 2006.

[7] A. Venneri, K. E. Forbes-Mckay and M. F. Shanks, "Impoverishment of spontaneous language and the prediction of Alzheimer's disease," *Brain,* p. 128, 2005.

[8] K. López-de-Ipiña, M. Ecay, J. Solé-Casals, A. Ezeiza, N. Barroso, P. Martinez-Lage and B. Beitia, "Feature selection for Spontaneous Speech Analysis to aid in Alzheimer's Disease diagnosis: A fractal dimension approach," *Computer Speech & Language,* vol. 30, no. 1, pp. 43-60, 2015.

[9] C. Thomas, V. Keselj, N. Cercone, K. Rockwood and E. Asp., "Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech," *In Proc. of the IEEE International Conference on Mechatronics and Automation,* pp. 1569-1574, 2005.

[10] S. O. Orimaye, J. S. Wong and K. J. Golden, "Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementias using Verbal Utterances," in *Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Baltimore, 2014.

[11] K. Fraser, G. Hirst, S. E. Black, N. L. Graham, E. Rochon and J. A. Meltzer, "Comparison of different feature sets for identification of variants in progressive aphasia," *Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality,* pp. 17-26, 2014.

[12] K. Fraser, F. Rudzicz, N. Graham and E. Rochon, "Automatic speech recognition in the diagnosis of primary progressive aphasia," *4th Workshop on Speech and Language Processing for Assistive Technologies,* pp. 47-54, 2013.

[13] A. Konig, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert and R. David, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia,* vol. 1, pp. 112-124, 2015.

[14] S. M. Wilson, M. L. Henry, M. Besbris, J. M. Ogar, N. F. Dronkers, W. Jarrold, B. L. Miller and M. L. Gorno-Tempini, "Connected speech production in three variants of primary progressive aphasia," *Brain,* vol. 133, pp. 2069-2088, 2010.

[15] R. Sadeghian, J. D. Schaffer and S. A. Zhorian, "Using Automatic Speech Recognition to Identify Dementia in Early Stages," *172th Acoustical Society of America meeting,* 2014.

[16] J. Sohn, N. S. Kim and W. Sung, "A Statistical model-based voice Activity Detection," *IEEE Signal Processing Letters,* vol. 6, no. 1, pp. 1-3, 1999.

[17] E. Charniak, "A maximum-entropy-inspired parser," *Proc. of the 1st North American chapter of the Association for Computational Linguistics conference,* pp. 132-139, 2000.

[18] B. Roark, M. Mitchell, J. Hosom, K. Hollingshead and J. Kaye, "Spoken Language Derived Measures for Detecting Mild Cognitive Impairment," *IEEE Tran. On Ausio, Speech and language Processing,* vol. 19, no. 7, 2011.

[19] S. Anderson and e. al., "Recognition of elderly speech and voice-driven document retrieval," *Proc. of IEEE International Conf. on. Acoustics, Speech, and Signal Processing,* vol. 1, 1999.

[20] J. W. Pennebaker, R. L. Boyd, K. Jordan and K. Blackburn, "The development and psychometric properties of LIWC2015," *University of Texas at Austin,* 2015.

[21] D. A. Snowdon, S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein and W. R. Marksbery, "Linguistic Ability in Early Life and Cognitive Function and Alzheimer's Disease in Late Life Findings from the Nun Study," *JAMA,* vol. 275, no. 7, pp. 528-532, 1996.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer and K. Vesely, "The Kaldi Speech Recognition Toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding,* 2011.

[23] O. Tilk and T. Alum, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," *INTERSPEECH '16,* 2016.

[24] P. Kohen, "Europarl: A parallel corpus for statistical machine translation," *MT summit,* vol. 5, pp. 79-86, 2005.

[25] D. Mungas, "In-office mental status testing: a practical guide," *Geriatrics,* vol. 46, no. 7, pp. 54–58, 63, 66, 1991.