Natural Language Features for Detection of Alzheimer's Disease in Conversational Speech

Ali Khodabakhsh, Serhan Kuşcuoğlu and Cenk Demiroğlu
Electrical and Computer Engineering Department, Ozyegin University, Istanbul, Turkey
{ali.khodabakhsh,serhan.kuscuoglu}@ozu.edu.tr,cenk.demiroglu@ozyegin.edu.tr

***Abstract*— Automatic monitoring of the patients with Alzheimer's disease and diagnosis of the disease in early stages can have a significant impact on the society. Here, we investigate an automatic diagnosis approach through the use of features derived from transcriptions of conversations with the subjects. As opposed to standard tests that are mostly focused on memory recall, spontaneous conversations are carried with the subjects in informal settings. Features extracted from the transcriptions of the conversations could discriminate between healthy people and patients with high reliability. Although the results are preliminary and patients were in later stages of Alzheimer's disease, results indicate the potential use of the proposed natural language based features in the early stages of the disease also. Moreover, the data collection process employed here can be done inexpensively by call center agents in a real-life application using automatic speech recognition systems (ASR) which are known to have very high accuracies in recent years. Thus, the investigated features hold the potential to make it low-cost and convenient to diagnose the disease and monitor the diagnosed patients over time.**

I. INTRODUCTION

With the aging population, Alzheimer's disease is becoming more widespread in the world. The World Health Organization estimated that in 2005, 0.379% of people worldwide had dementia, and that the prevalence would increase to 0.441% in 2015 and to 0.556% in 2030 [1]. Because there is no treatment to cure the disease, years of healthcare costs become a significant economic burden on the governments as well as the patients and their families. Thus, simplifying the healthcare processes and reducing the costs through the use of automated monitoring systems can make significant economic impact.

Diagnosis of the disease is not easy. Even in the later stages, recognition or evaluation of the disease by clinicians fail 50% of the time [2]. Moreover, even if the disease is diagnosed correctly, monitoring the progression of the disease by a clinician over time is costly. Thus, patients cannot visit the clinicians frequently and what happens between the visits is largely unknown to clinicians. Furthermore, early diagnosis enables getting early treatment which is known to help patients have better quality of life.

Telephone-based automated measures for detection and/or monitoring of the disease can be a low-cost solution to the diagnosis problem. Patients who do not feel comfortable visiting a doctor, or cannot afford a doctor visit, can do private self-test. Moreover, diagnosed patients can be monitored frequently by the system to detect changes with minimal cost and convenience for the patients.

Typically, clinicians use tests such as Mini-Mental State Examination (MMSE) and linguistic memory tests. Linguistic memory tests are based on the recall rates of word lists and narratives and they are typically more effective than the MMSE tests. None of those typical practices, however, consider the speech signal in diagnosing the disease. Moreover, they are hard to do over the telephone line since both patients and elderly people often fail to use such sophisticated technology. Analysis of speech signal has been considered for Alzheimer's detection in [3][4]. However, in those works, speech signal is recorded during the standard clinical tests. Moreover, most of the focus is on the high level structural processing of the spoken language for a specific language. A more limited study with one patient and a focus on the prosodic features of speech, which determines stress, intonation, and emotion, is reported in [5]. Problems of speech production that are related to central nervous system problems are also noted in [6]. Speech-based features are investigated in [7] to detect fronto-temporal lobar degeneration with promising results.

Here, we focus on extracting features from spoken language using natural language processing techniques to detect Alzheimer's disease. Because the patients are not necessarily able to take automated tests or carry a structured conversation over the phone, the application scenario here is based on free-flow conversational speech. That way, subject's speech can be recorded in the most natural and effortless way by a person with minimal technical or clinical skills. For example, the conversation can be carried by a person at a call center and recorded and transcribed automatically which is substantially lower-cost compared to a hospital visit. Such conversational data has been investigated in [8][9], however only high level linguistic complexity of a specific language are analyzed in [8][9]. Similarly, conversational data has been investigated in [7][10] for linguistic and speech dysfluency features by measuring the correlation of those features with the disease and without and attempt to do diagnosis using them. Correlation of linguistic capability with the Alzheimer's disease was also shown in [11].

We investigated a wide range of linguistic features derived from the transcriptions of patients and healthy subjects. We have investigated the prediction power of each feature as well as combination of features using SVM, LDA and decision tree methods. We have shown that, even though individual prediction accuracy of the features are not very high, some of the proposed features can be used as strong markers of

978-1-4799-2131-7/14/$31.00 ©2014 IEEE

the disease when used together.

II. Natural Language Features

Seventeen natural language features are extracted from the transcriptions of the recorded conversations of test subjects. The features are geared towards detecting problems with the flow of the conversation, such as hesitations and repetitions, syntactic features using part-of-speech (POS) tagging, intelligibility of speech, diversity and complexity of the words used, and how well the subject can understand the question or carry the conversation without getting puzzled. The features investigated here are described below.

A. Hesitation and Puzzlement Features

During recordings, we have found that patients tend to hesitate more, forget what they were talking about, and have a harder time finding the right words or remembering details about their pasts. They also sometimes get puzzled about why they cannot remember the details or forget the context of the conversation. Those observations led us to propose features that will be able to capture those patterns in transcriptions.

1) Question Rate: Patients are more likely to forget details in the middle of conversation, not understand the questions or forget the context of the question. In those cases, they tend to ask the interviewer to repeat the question or ask themselves what the detail was. The question words such as "'which"', "'what"' etc. are tagged and the total number of questions are calculated in each conversation. The question rate is computed by dividing the number of question words over the total number of words in the conversation.

2) Confusion Rate: Because patients sometimes forget the context of conversation and get unsure of details that they are describing, they tend to use confusion words as an alternative to question words. These are words such as "'maybe"', "'perhaps"', "'cannot remember exactly but"' etc. The rate of those words are measured by the ratio of the confusion words over the total number of words in each recording.

3) No Answer Count: In some cases, the patient does not respond to a question of the interviewer. In those cases, interviewer repeats the question or moves onto another question. The no answer count feature is calculated by the number of such no answer events in each conversation.

4) Rate of Pauses in Utterances: Patients tend to stop more in the middle of sentences to think about what to say next. These pauses can be used as a marker for cognitive problems. Here, total number of occurrence of pauses inside each utterance is extracted. Utterances are assumed to be segments where the subjects talk without having very long pauses or an interruption from the interviewer. Pause rate is computed by dividing the number of pauses to number of segments. Rate for the whole recording is computed by averaging the pause rate computed in each utterance.

5) Filler Sounds: Filler sounds such as "'ahm"', "'ehm"' etc. are used by people in spoken language when they think about what to say next. These are used by everybody but here we hypothesize that they are used more frequently by the patients.

B. POS based Feature

Part of speech (POS) tags can be used as a marker of syntactic problems in speech. Adjectives can indicate more colorful and descriptive use of language. Also proportion of frequency of each POS tag can be different between healthy subjects and patients.

POS Tags can be automatically added for each word separately. Each of the word classes can be used as a feature by computing the ratio of the occurrence in a segment over the total number of words in that segment. In this research the verb, noun, pronoun, adverb, adjective, particle, and conjunction are used.

C. Unintelligible Word Rate

The number of unintelligible words in each segment: Detection of the unintelligible can be easily done with check of existence of the word in a wide range dictionary of POS. any word that is not in a predefined dictionary can be assumed to be a malformed word or a meaningless word or a failed trial to pronounce the target word. The ratio of the events over total number of words is extracted for this feature.

D. Complexity Features

1) Phonemes per Word: The number of phonemes in each word can be a measure of complexity of the words. Thus, for this feature, the average number of phonemes in each word spoken by the subject is used.

2) Words per Recording: How long the subject can continue talking without a pause can be a measure of cognitive health. As a measure of duration, number of words that are used without any pause is used.

3) Standardized Word Entropy: One of the earliest damage occurs in the cortex of the brain that deals with language ability. We hypothesize that this will result in the variety of words or word combination that the subject can use. Moreover, in the recordings we have observed that subjects tend to repeat the some words throughout the conversation. Because the repetitions tend to be local in the conversation, entropy of each utterance measured independently and then average entropy is calculated for each recording.

4) Phone Entropy: Due to cognitive damage, patients tend to use some phonemes more and have trouble saying some of the other phonemes. Each word in the recording is broken into the phones and the entropy is then computed for the whole recording.

III. Experiments

Conversational speech recordings of 20 patients and 20 healthy subjects were manually transcribed. The transcriptions are done by two persons and then double checked. The transcribers and the subjects are native Turkish speakers. The data has been collected in healthcare facilities at Istanbul. For each subject, approximately 10 minutes of conversation have been recorded using a high-quality microphone. The text is then segmented based on speech turns between interviewer and subject.

After feature extraction, support vector machines (SVM), linear discriminant analysis (LDA), and decision trees are used for classification. Linear Kernel is used for SVM. Fisher LDA is used. Due to small sample size, leave-one-out method is used in training and testing. In each test, features are normalized to have zero mean and unit variance using the features extracted from the training data. Logarithm of the following features are used for classification: filler rate, unintelligible word rate, question rate, words per recording, confusion rate, rate of pauses in utterances, no answer rate.

IV. Results and Discussion

Classification results for each individual feature are separately shown in figures 1, 2 and 3 along with confidence intervals. For each classifier, only the best performing features are plotted. False alarm (FA), missed detections (MD) and prediction accuracy of the features are shown as performance metrics. Even though each of the classifier has a different objective function, on average, classification rates are similar.

Order of best performing features change drastically for each classifier. Filler rate and confusion rate features perform well both with decision tree and SVM. A scatter plot of those features is shown in Fig. 4. Both mean and variance of confusion and filler rates are significantly higher in the patients. Moreover, patients seem to prefer one of the strategies when there is a strong marker. The rates of confusion and fillers are similar to healthy people when they are used together. Thus, the marker seems to become more useful in classification when the confusion rate or filler rate is unusually high.

For feature selection, all possible combinations of features are investigated for each classifier and the best performing features are found. Results are shown in Table I. Best performance of 90% is achieved by SVM and decision tree classifiers. LDA did not work as well as others because Gaussianity assumption in Fisher LDA does not hold in most of the features. Performance of the systems begin to drop with addition of more features after a certain feature number. This is expected since the amount of data is not enough to develop classifiers with good generalization capability when the number of features is high.

Even though SVM and decision tree achieved the same performance, decision tree achieved it with only three features. The features used by the decision tree to get the 90% prediction accuracy are word entropy, phone entropy, and rate of pauses in utterances. Interestingly, when these features are plot in combinations of two, they are far from being linearly separable. For example, scatter plot of word entropy and rate of pauses is shown in Fig. 5. However, when combined, taking advantage of the nonlinear separation capability of the decision trees, high performance can be achieved as show in Table II.

An analysis of the way the three features are used by the decision tree are shown in Fig. 6. Phoneme entropy is the first, and therefore most important, feature used by the tree. Subjects with a phoneme entropy above a threshold are easily classified as healthy. This result is surprising because we were expecting word entropy to be lower in the patients

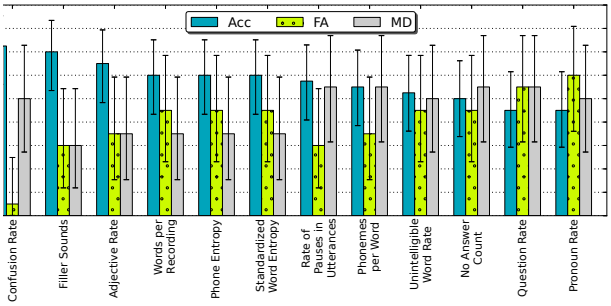

Fig. 1: Performance of the decision tree classifier for individual features.

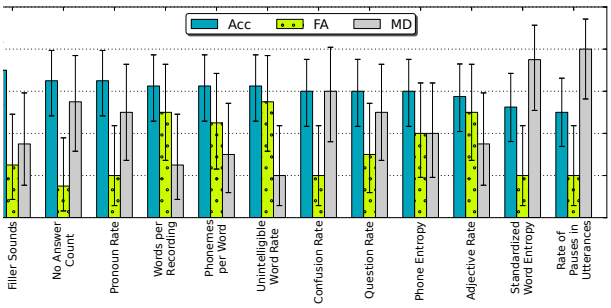

Fig. 2: Performance of the SVM classifier for individual features.

rather than the phoneme entropy. Our finding can indicate that deterioration in the cortex has a more effect in the variety of phonemes the subject uses than the variety in words. In the recordings, we could clearly hear that patients had difficulty with producing some of the sounds. Neurological basis for this calls for a multi-disciplinary investigation.

At the second layer of the tree, if the rate of silences is above a threshold, the subject is classified as patient. High silence rate with the patients was observed in the recordings, and the use of that by the decision tree is not surprising. The last two layers of the tree uses the word entropy feature. In fact, these two layers shed some light on why word entropy is the least important feature used in the tree. Entropy of patients are either above or below the entropy of healthy subjects. Even though some of the subjects could not respond to the questions very well, some subjects talked with a good flow especially when they were talking about their past experiences and life stories.

V. Conclusion and Future Work

In this work, we have proposed an extensive set of features derived from the transcriptions of Alzheimer's patients and

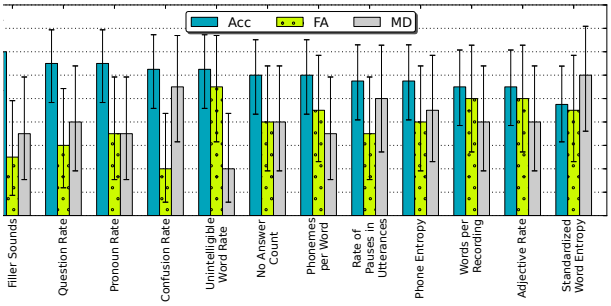

Fig. 3: Performance of the LDA classifier for individual features.

	1	2	3	4	5	6	7	8	9	10	11	12
LSVM	70.00%	80.00%	77.50%	80.00%	82.50%	87.50%	**90.00%**	85.00%	80.00%	77.50%	70.00%	65.00%
LDA	70.00%	77.50%	80.00%	80.00%	80.00%	80.00%	80.00%	80.00%	75.00%	75.00%	70.00%	67.50%
CTree	72.50%	80.00%	**90.00%**	**90.00%**	**90.00%**	**90.00%**	**90.00%**	87.50%	82.50%	75.00%	62.50%	52.50%

TABLE I: Classification accuracy obtained with best combination of features for each classifier. Results are shown when different number of features are used.

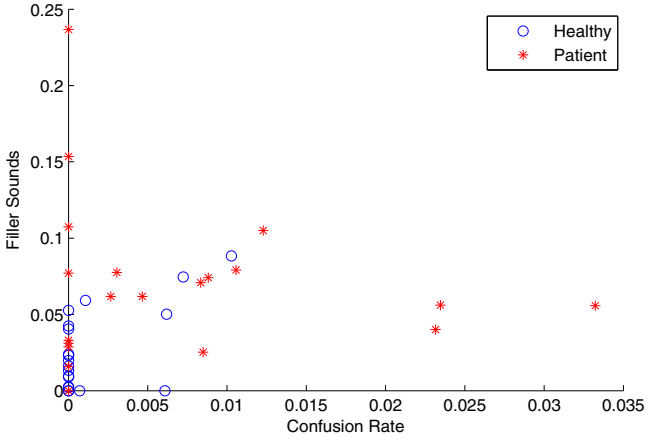

Fig. 4: Scatter diagram of the filler sound and confusion rates.

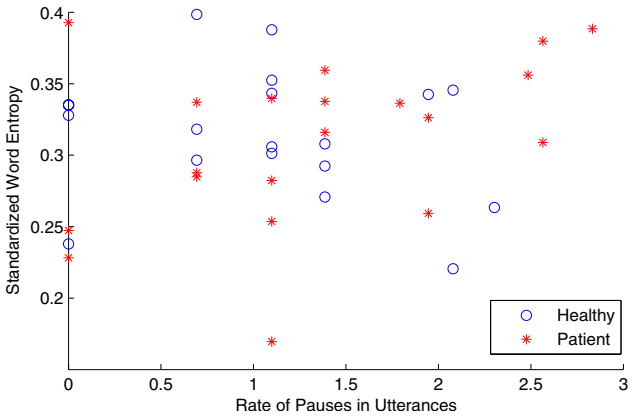

Fig. 5: Scatter diagram of the word entropy and rate of pauses inside utterances.

healthy elderly subjects. It is already known that part of the brain cortex that deals with linguistic abilities are among the first parts that deteriorate with the disease. Our work explored how that deterioration is reflected in the patient's spoken language and if there are markers that can be effectively detected using natural language processing and machine learning techniques. Our results indicate that 90% prediction accuracy can be obtained using only phone entropy, silence rate per utterance, and word entropy with a decision tree classifier. Especially the success of phone entropy may indicate a potential damage in the part of the brain that controls the articulatory organs for speech production. One of our goals in future work is to combine the linguistic features with the acoustic features where speech production problems can be measured directly.

Our experiments are preliminary and effectiveness of the markers that we have found should be measured with early stage patients where the signals are more subtle and more subjects are needed to reach statistically significant results. However, our experiment results and manual observations

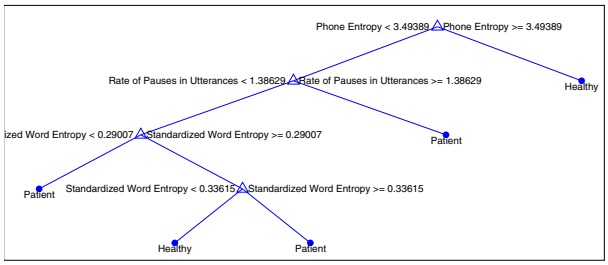

Fig. 6: Decision Tree for the best performed combination of 3 features.

	Low	Mid	High
Accuracy	76.34%	90.00%	97.21%
Missed Detection	0.13%	5.00%	24.87%
False Alarm	3.21%	15.00%	37.89%

TABLE II: Word entropy, phone entropy, and rate of pauses in utterances. Low and high indicate the end points of the confidence interval. Mid indicates the average performance.

from the data are encouraging and we will start collecting data from early stage patients in a near future.

REFERENCES

[1] World Health Organization (2006). Neurological Disorders: Public Health Challenges. Switzerland: World Health Organization. pp. 204207. ISBN 978-92-4-156336-9.

[2] L. Boise, M. B. Neal, and J. Kaye, Dementia assessment in primary care: results from a study in three managed care systems, J. Gerontol. A Biol. Sci. Med. Sci., vol. 59, no. 6, pp. M621626, Jun. 2004, PMID: 15215282.

[3] B. Roark, M. Mitchell, J. Hosom, K. Hollingshead, and J. Kaye, Spoken language derived measures for detecting mild cognitive impairment, vol. 19, no. 7, pp. 20812090, 2011.

[4] B. Roark, J.-p. Hosom, M. Mitchell, and J. A. Kaye, Automatically Derived Spoken Language Markers for Detecting Mild Cognitive Impairment, ser. Proc. 2nd Int. Conf. Technol. Aging (ICTA), 2007.

[5] G. Tosto, M. Gasparini, G. Lenzi, and G. Bruno, Prosodic impairment in alzheimers disease: Assessment and clinical relevance, J Neuropsychiatry Clin Neurosci, vol. 23, no. 2, pp. E21E23, Mar. 2011.

[6] Vassiliki Iliadou and Stergios Kaprinis, Clinical psychoacoustics in alzheimers disease central auditory processing disorders and speech deterioration, Annals of General Hospital Psychiatry, vol. 2, p. 12, Dec. 2003, PMID: 14690547 PMCID: PMC317473.

[7] I. Hoffmann, D. Nemeth, C. Dye, M. Pkski, T. Irinyi, and J. Klmn, Temporal parameters of spontaneous speech in alzheimers disease, pp. 528532, Feb. 2010, PMID: 20380247.

[8] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K.Wilcock, Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance, ser. Aphasiology, 2000, vol. 14.

[9] C. Thomas and N. Cercone, Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech, 2005.

[10] H. LEE, F. Gayraud, F. Hirsch, and M. Barkat-Defradas, Speech dysfluencies in normal and pathological aging: a comparison between alzheimer patients and healthy eldery subjects, in the 17th International Congress of Phonetic Sciences (ICPhS), 2011, p. 11741177.

[11] D. A. Snowdon, S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein, and W. R. Markesbery, Linguistic ability in early life and cognitive function and alzheimers disease in late life. findings from the nun study, JAMA, vol. 275, no. 7, pp. 528532, Feb. 1996, PMID: 8606473.