Detection of Alzheimer's disease using prosodic cues in conversational speech

Ali Khodabakhsh, Cenk Demiroğlu

Electrical and Computer Engineering Department, Ozyegin University, Istanbul, Turkey

{ali.khodabakhsh,cenk.demiroglu}@ozyegin.edu.tr

Abstract

Automatic diagnosis of the Alzheimer's disease as well as monitoring of the diagnosed patients can make significant economic impact on societies. We investigated an automatic diagnosis approach through the use of speech based features. As opposed to standard tests that are mostly focused on memory recall, spontaneous conversations are carried with the subjects in informal settings. Prosodic speech features extracted from speech could discriminate between healthy people and the patients with high reliability. Although the patients were in later stages of Alzheimer's disease, results indicate the potential of speech-based automated solutions for Alzheimer's disease diagnosis. Moreover, the data collection process employed here can be done inexpensively by call center agents in a real-life application. Thus, the investigated techniques hold the potential to significantly reduce the financial burden on governments and Alzheimer' patients.

Index Terms: speech analysis, Alzheimer's detection, support vector machines

1. Introduction

Alzheimer's disease (AD) is becoming more widespread with the aging population in the developed countries. Thus, it is a significant economic burden on the governments as well as the patients and their families. Simplifying the healthcare processes and reducing the costs through the use of technology for this disease can make a 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 [1]. 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.

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-tests. Moreover, diagnosed patients can be monitored frequently by the system 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 because of problems in user-interface design, the need for high accuracy speech recognition systems which still are not good enough to meet the demands of such an application. Furthermore, both patients and elderly people often fail to use such sophisticated technology.

Analysis of speech signal has been considered for Alzheimer's detection in [2] [3]. However, in those works, speech signal is recorded during the standard clinical tests. Moreover, most of the focus is on the spoken language itself, which requires manual transcription, rather than the speech signal. 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 [4]. Problems of speech production that are related to central nervous system problems are also noted in [5]. Speech-based features are investigated in [6] to detect fronto-temporal lobar degeneration with promising results.

Here, we propose a system where a spontaneous conversation is carried with the patient. Thus, contents of the conversations are not predetermined. The goal of this approach is to keep the subject comfortable during speech without constraining the conversation. Moreover, lack of structure is appealing to the subjects since this requires minimal effort in terms of cognition. 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 automatically which is substantially lower-cost compared to a hospital visit. Such conversational data has been investigated in [7] [8], however only linguistic features are analyzed and speech features are not considered in [7] [8]. Similarly, conversational data has been investigated in [6] [9] for linguistic and speech dysfluency features by measuring the correlation of those features with the disease and without an attempt to do diagnosis using them. Correlation of linguistic capability with the Alzheimer's disease was also shown in [10].

Prosodic features can be used to identify the affect in speech as well as problems in articulation. Moreover, hesitations or long response times in speech can also indicate problems in the cognition processes. Here, 13 prosodic speech features are extracted automatically from the recordings and disease detection is done with linear discriminant analysis (LDA), support vector machines (SVMs), and decision tree classifiers. Proposed speech features, especially using the SVM algorithm, seems to be particularly good at distinguishing between healthy and sick people.

2. Prosodic Features

Because of the amount and nature of background noise in the recorded files, finding a robust VAD (Voice Activity Detector) was an important task. The VAD used here is based on the distribution of the short-time frame energy of the speech signal. Because there is both silence and speech in the recordings, energy distribution has two modes, both of which can be modeled with Gaussian distributions. Thus, based on the energy of the speech frames, they are classified as either speech or silence.

The features that are used to detect Alzheimers disease are

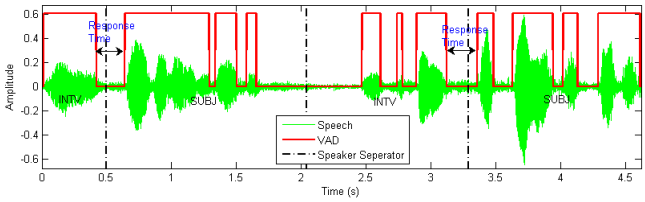

Figure 1: *Sample speech waveform from the interviews and the voice activity detector (VAD) output is shown. Response time indicates the amount of time it takes the subject to answer a question. Silence segments indicate the level of noise in the signal.*

extracted from the conversational speech recordings. A total of 13 features have been extracted and evaluated for detecting Alzheimers disease. A description of the speech features are given below.

2.1. Voice Activity Related Features

Silence and speech segments are labeled in each recording for feature extraction as discussed in Section 3.1. Using the voice activity information, following features are extracted from each recording:

(a) *Logarithm of the response time:* When the interviewer asks a question, it takes some time before the subject gives an answer. It is hypothesized that this time can be an indicator of the disease since it is expected to be related to the cognitive processes such as attention and memory. Logarithm of the average time it takes the subject to answer a question is calculated in each recording to extract this feature.

(b) *Silence-to-speech ratio:* ratio of the total silence time over the total amount of speech is a measure of the hesitations during speech.

(c) *Logarithm of continuous speech:* duration of each continuous speech segment indicates how long the subject can talk without pausing.

(d) *Logarithm of continuous silence:* duration of each continuous silence segment indicates how long the subject pauses in the middle of the speech.

(e) *Logarithm of pauses per second:* Besides the length of hesitations during speech, frequency of them can also be an indicator of the disease. Logarithm of the average number of pauses per seconds is used in this feature.

(f) *Average absolute delta energy:* similar to pitch, energy variations also convey information about the mood of the subject. Changing speech energy significantly during speech may indicate a conscious effort to stress words that are semantically important or changes of mood related to the content of the speech. Average of the energy changes are used in this feature.

2.2. Articulation Related Features

The voice activity related features discussed above are related cognitive thought processes. However, it is also important to measure how the subject uses his/her voice articulators during speech. For example, if the subject gets too emotional, significant changes in the fundamental frequency (pitch) can be expected. Similarly, changes in the resonant frequencies (formants) of speech can be a strong indicator of the subjects health. If the formants do not change fast enough or are not distinct enough, sounds may become harder to identify which can indicate mumbling in speech. To see the effects of these in classification of the disease, pitch and formant trajectories are extracted and following features are derived or each recording:

(a) *Average absolute delta pitch:* average of absolute delta pitch indicates the rate of variations in pitch. This feature has high correlation with the communication of emotions through the speech signal.

(b) *Average absolute delta formants:* average of absolute delta formant frequencies indicates the rate of change in the formant features. Formants are related to the positions of the vocal organs such as tongue, lips etc. Reduction of control over these organs because of a damage in the brain, such as AD, can create speech impairments such as mumbling. In this case, formants do not change quickly and speech becomes less intelligible.

(c) *Logarithm of voicing ratio:* another speech impairment is the loss of voicing in speech which results in smokers voice. In this case, the subject loses the ability to vibrate the vocal cords which results in breathy and noisy speech. Average duration of voiced speech is compared with the unvoiced speech to detect any potential impairment in the vocal cords due to AD.

(d) *Logarithm of voicing per second:* this feature measures the ratio of voiced speech to unvoiced speech per second.

(e) *Logarithm of the mean-duration of continuous voiced speech:* besides the ratio of voiced and unvoiced speech, for how long the subject maintain voicing without pausing is measured.

2.3. Rate of Speech Related Features

Using an automatic Turkish phoneme recognizer trained with a Turkish broadcast speech database, recordings are transcribed into phonemes. Following features are extracted using the phonemic transcriptions:

(a) *Phonemes per second:* number of phonemes generated per second is used to represent the rate of speech of the subject.

(b) *Logarithm of variance of Phoneme Frequency:* 42 phonemes are recognized by the automatic system in this study. However, distribution of the recognized phonemes change from subject to subject. Higher variance in phoneme frequency distribution may indicate clarity in speech and may be correlated with using a bigger dictionary during speech. Because the dictionary size and speech clarity is related to cognition and articulatory organs, this feature may be useful in detecting AD.

3. Experiments

In this research, conversational speech recordings of 27 patients (17 male, 10 female) with late stage Alzheimers disease, and 27 healthy elderly people (12 male, 15 female) have been used. All subjects were native speakers of Turkish. The age range is between 60 and 80 in both healthy and elderly subjects. For each subject, approximately 10 minutes of conversation have been recorded using a high-quality microphone with 16 KHz sampling rate. Subjects were directed casual/conversational questions which are not fixed between different patients. The data has been collected in healthcare facilities at Istanbul and then hand-labeled to split question and response parts. The labels are later refined using an automated voice activity detector (VAD)

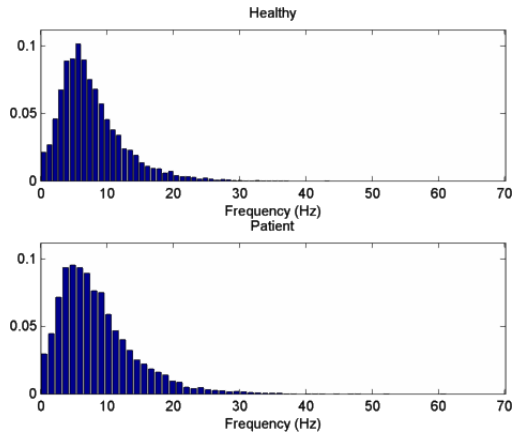

Figure 2: *Average Absolute Delta Pitch distribution.*

to accurately mark the beginning and end times of phonations in the conversations. Patients refused to use noise-cancelling microphones which require installation on their clothes. Therefore, there is some background noise in the recordings but the signal-to-noise ratio is high and speech features could be extracted reliably.

For classification of patients and healthy subjects, three classifiers are used: linear discriminant analysis (LDA), support vector machines (SVM), and decision trees. Quadratic kernel is used in the SVM classifier.

In the first phase of testing, each feature is tested separately to assess the classification power of individual features. Then, combinations of features are used to increase the classification power of the algorithms. Increasing the number of features used in classification, brute force search is done using all possible combinations of features. Feature sets that give the best results for each feature number are found using such brute force approach.

Because there is limited number of subjects in the test, leave-one-out strategy is used where one of the subjects is left out and the classifier is trained with the rest of the subjects. Then, testing is done on the left-out subject. All subjects are tested using this strategy. %95 confidence intervals of the classification scores are also computed.

3.1. Results and Discussion

Classification performance of the individual features is shown in Fig. 3. Performance results are reported in terms of detection rate (probability of detecting the disease in a patient), false alarm (probability of diagnosing the disease in a healthy subject), and the total accuracy (deciding between healthy and ill subjects correctly). %95 confidence intervals are also shown in Fig. 3. Individual features are not particularly strong at diagnosing the disease. Highest performance is obtained with the logarithm of voicing ratio, average absolute delta feature of the first formant, and average absolute delta pitch feature. Logarithm of voicing ratio is lower in the patients compared to healthy subjects. Similarly, average absolute delta formant feature is lower in the patients. Lack of voicing and smaller variations in the formants may indicate a correlation between ability/desire to control articulatory organs and the disease. In contrast, average absolute delta pitch feature has a higher variance in the patients as shown in Fig 2. That indicates more variations of emotion and emphasis during speech in the patients speech.

After testing the predictive power of each individual feature, combinations of features are used to improve the classification performance. All possible pairs of the 13 features are tested with the three classifiers. Highest scoring features are shown in Fig. 4. Using feature pairs improved the performance of LDA and SVM algorithms. However, performance of the decision tree algorithm is degraded with more features. Decision tree could not generate enough leaf nodes given the fixed set of features and therefore could not perform well in this task.

		4	5	6	7	8
LDA	Total	83 73:93	81.1 71:92	81.1 71:92	83 73:93	84.9 75:95
	%D	77.8 62:93	74.1 58:91	77.8 62:93	85.2 72:99	81.5 67:96
	%FA	11.5 0:24	11.5 0:24	15.4 2:29	19.2 4:34	11.5 0:24
SVM	Total	**86.8** 78:96	**88.7** 80:97	**92.5** 85:100	**94.3** 88:100	**90.6** 83:98
	%D	92.6 83:100	88.9 77:100	96.3 89:100	92.6 83:100	92.6 83:100
	%FA	19.2 4:34	11.5 0:24	11.5 0:24	3.8 0:11	11.5 0:24
DT	Total	84.9 75:95	84.9 75:95	83 73:93	81.1 71:92	81.1 71:92
	%D	85.2 72:99	85.2 72:99	85.2 72:99	88.9 77:100	85.2 72:99
	%FA	15.4 2:29	15.4 2:29	19.2 4:34	26.9 10:44	23.1 7:39

Table 1: Scores with combinations of different number of features using LDA, SVM and Decision Tree classifiers. For each number of features, best performing features are found using a brute force search method. In each cell, mean value is shown on the top and the lower and upper limits of confidence is shown on the bottom.

Similar to single feature case, SVM algorithm outperformed the LDA and decision tree algorithms for the feature pairs. Moreover, combination of the average absolute delta pitch and logarithm of voicing ratio feature has the best performance. Higher pitch variation and lower voicing ratio can clearly separate the data into two classes which can be classified with a quadratic decision surface in most cases. Moreover, the data points that are misclassified are close to the decision surface which indicates that those can also be easily recovered if more features are available.

All possible combinations of three features are also evaluated and the best performing 3-tuples are shown in Fig. 4. Performance of the best performing systems with increasing number of features are shown in Table. 1. Performance of all systems increase with the number of features until a point where overfitting occurs and classification performance gets lower with more features. Average accuracy of the SVM system goes up to %94 for the 7-feature case with confidence interval having a lower-bound of %88. Thus, in the worst case, the SVM system can achieve a total accuracy of %88.

4. Conclusion

To conclude, proposed acoustic features extracted from conversational speech are effective at detecting late-stage Alzheimer's disease especially when used with an SVM classifier. Thus, same features extracted from a patient at different time intervals may be useful for following the progression of the disease which will be the focus of future investigation.

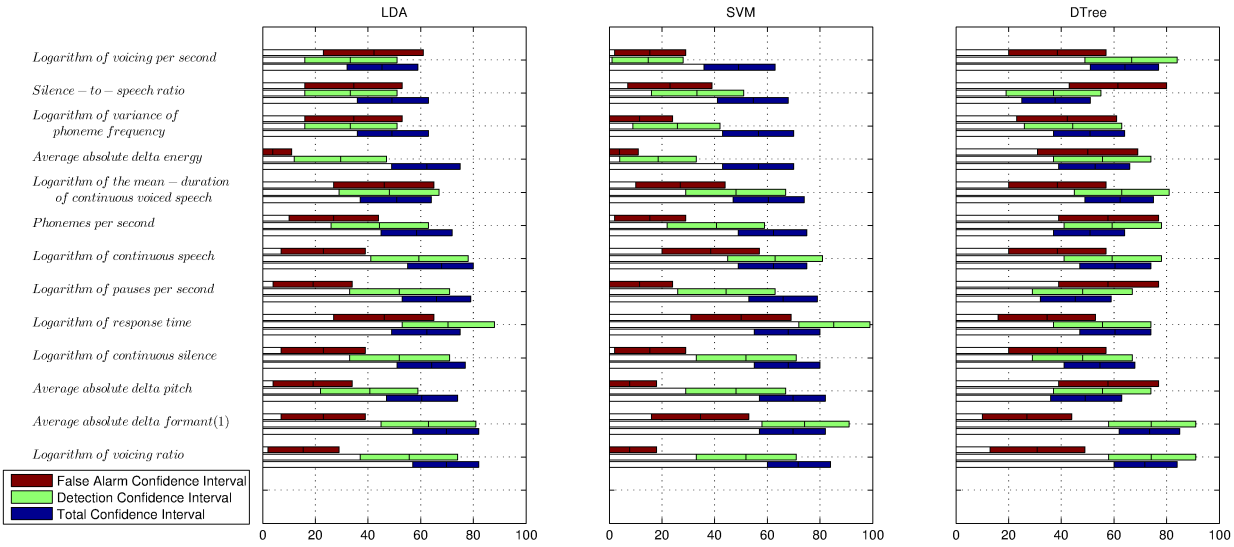

Figure 3: *Performance of each feature individually using different classification algorithms sorted by SVM performance.*

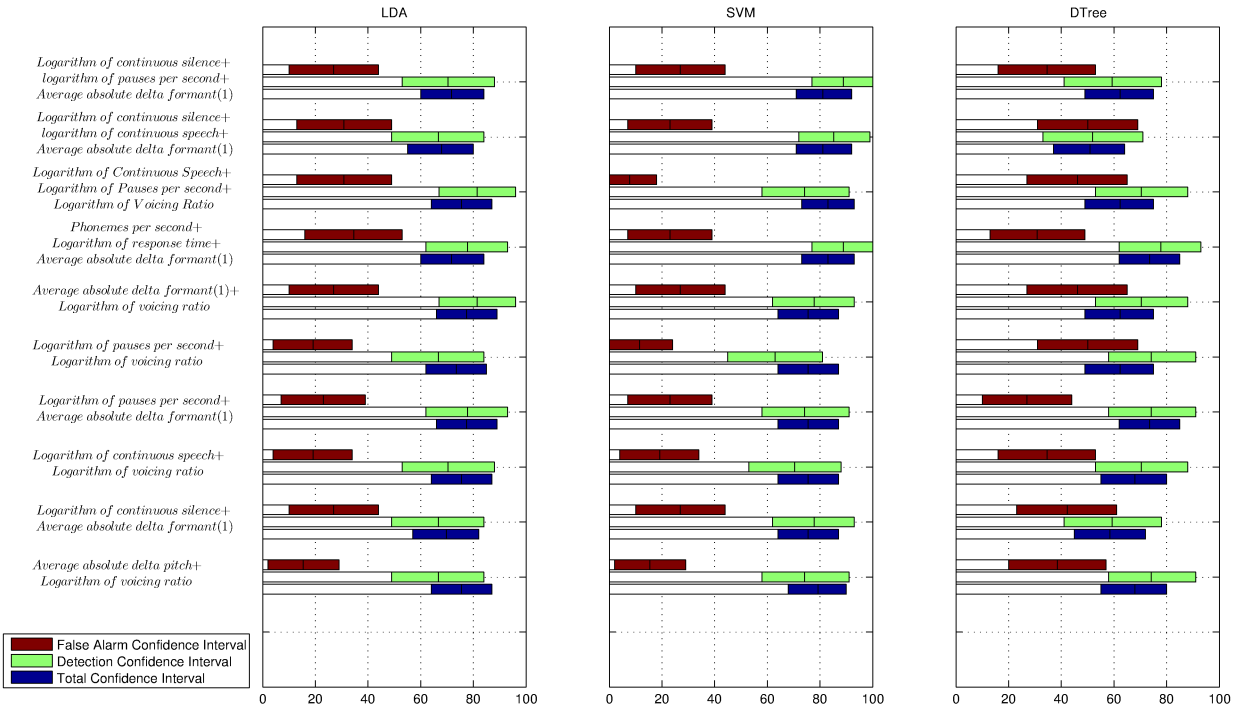

Figure 4: *Scores of combinations of two and three features using different classification algorithms. Systems are sorted by SVM performance. Combinations with scores less than 75% for two features and 80% for tree features are not shown.*

5. References

[1] 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. M621–626, Jun. 2004, PMID: 15215282.

[2] 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. 2081–2090, 2011.

[3] 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.

[4] G. Tosto, M. Gasparini, G. Lenzi, and G. Bruno, "Prosodic impairment in alzheimer's disease: Assessment and clinical relevance," *J Neuropsychiatry Clin Neurosci*, vol. 23, no. 2, pp. E21–E23, Mar. 2011.

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

[6] I. Hoffmann, D. Nemeth, C. Dye, M. Pkski, T. Irinyi, and J. Klmn, "Temporal parameters of spontaneous speech in alzheimer's disease," pp. 528–532, Feb. 2010, PMID: 20380247.

[7] 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.

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

[9] 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.

[10] 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 alzheimer's disease in late life. findings from the nun study," *JAMA*, vol. 275, no. 7, pp. 528–532, Feb. 1996, PMID: 8606473.