View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Springer - Publisher Connector

Khodabakhsh *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:9
DOI 10.1186/s13636-015-0052-y

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech

Ali Khodabakhsh*, Fatih Yesil, Ekrem Guner and Cenk Demiroglu

Abstract

Automatic diagnosis and monitoring of Alzheimer's disease can have a significant impact on society as well as the well-being of patients. The part of the brain cortex that processes language abilities is one of the earliest parts to be affected by the disease. Therefore, detection of Alzheimer's disease using speech-based features is gaining increasing attention. Here, we investigated an extensive set of features based on speech prosody as well as linguistic features derived from transcriptions of Turkish conversations with subjects with and without Alzheimer's disease. Unlike most standardized tests that focus on memory recall or structured conversations, spontaneous unstructured conversations are conducted with the subjects in informal settings. Age-, education-, and gender-controlled experiments are performed to eliminate the effects of those three variables. Experimental results show that the proposed features extracted from the speech signal can be used to discriminate between the control group and the patients with Alzheimer's disease. Prosodic features performed significantly better than the linguistic features. Classification accuracy over 80% was obtained with three of the prosodic features, but experiments with feature fusion did not further improve the classification performance.

Keywords: Alzheimer's disease; Speech processing; Linguistic features; Prosodic features; Machine learning

1 Introduction

As the worldwide elderly population increases, the incidence of Alzheimer's disease is becoming more widespread. It is estimated that 7% of the world's population over 65 years old has Alzheimer's or a related dementia [1]. Moreover, only one in four patients has been diagnosed [1]. Because there is no treatment to cure the disease, years of healthcare costs are becoming a significant economic burden on governments as well as patients and their families. The global cost of Alzheimer's and dementia is estimated to be $605 billion, which is equivalent to 1% of the entire world's gross domestic product [2]. The problem intensifies each year with the aging world population. Thus, simplifying healthcare processes and reducing costs through the use of automated systems can make a significant socio-economic impact.

Diagnosis of the disease is costly and difficult. Moreover, even if the disease is diagnosed correctly, monitoring the progression of the disease by a clinician over time further increases the cost. Thus, patients cannot visit clinicians frequently and what happens between the visits is largely unknown to clinicians.

Typically, clinicians use tests such as Mini-Mental State Examination (MMSE) and linguistic memory tests [3]. Linguistic memory tests are based on the recall rates of word lists and narratives, and they are typically more effective than the MMSE. Moreover, individual's medical and family histories are used, along with MRI scans to test for other brain-related conditions, such as stroke. Biomarkers showing the level of beta-amyloid accumulation in the brain or the neurons that are injured or actually degenerating can also be used in combination with the other tests [4].

None of those typical practices consider the speech signal in diagnosing the disease even though the part of the brain cortex that processes language abilities is one of the earliest parts to be affected by the disease [5]. For example,

*Correspondence: ali.khodabakhsh@ozu.edu.tr
Electrical and Computer Engineering Department, Ozyegin University, Orman Street, 34794, Istanbul, Turkey

Springer

© 2015 Khodabakhsh et al.; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

narrative retelling ability is found to be strongly correlated with the disease [6]. Similarly, linguistic features derived from the transcription of a narrative retelling task were found to be significantly correlated with primary progressive aphasia, which is a type of dementia [7]. Similarly, analysis of the speech signal has been shown to be useful for Alzheimer's detection in [8,9]. However, in those works, the speech signal is recorded during the administration of standard clinical tests. Moreover, most of the focus is on the high-level structural processing of spoken language for a specific language. For example, in [10], features such as moderate word finding difficulty, reduced phrase length, and reduced comprehension are manually tagged by humans and shown to contain information complementary to standardized tests. Correlation of linguistic capability with Alzheimer's disease was also shown in [11]. Speech-based features are investigated in [5] to detect fronto-temporal lobar degeneration with promising results. Speech is recorded in a semi-structured interview setting in [5]. The frequency and ratio of syntactic categories such as pronouns and adverbs are found to be markers of the disease.

In addition to natural language processing (NLP) features, speech acoustics have also been studied and reported in the literature. A limited study with one patient with a focus on the prosodic features of speech such as stress, intonation, and emotion is reported in [12]. Problems of speech production that are related to central nervous system problems are also noted in [13]. In [14], dysfluency cycles in speech are measured using the length and frequency of hesitations in speech. Subjects with dementia were found to have patterns different from those of control subjects.

Here, we focus on extracting an extensive set of acoustic and linguistic features from spoken language to detect Alzheimer's disease. Because the patients with Alzheimer's are usually not able to take automated tests or to carry on a structured conversation, data collection is done during unstructured conversational speech. In this way, a subject's speech can be recorded in the most natural and effortless way by a person with minimal technical or clinical skills. Semi-structured conversational data has been investigated in [15,16], but only linguistic features are analyzed, and speech features are not considered. Similarly, conversational data has been investigated for limited sets of linguistic and speech dysfluency features by [5,17], who measured the correlation of those features with the disease and attempt to use the features for diagnosis.

In our work, we focused on evaluating the effectiveness of a large set of features for detecting Alzheimer's disease in unstructured conversations. The data was collected in Turkish, which has not been studied to the extent of languages such as English. We propose 20 prosodic features extracted automatically from the recordings and 18 linguistic features derived from the transcriptions of patients and control subjects. We have investigated the predictive power of each feature as well as combination of features using support vector machines (SVM), nearest neighbor (NN) classifiers, naive Bayesian classifiers, and classification trees (CTree). Our results indicate that some of the investigated features are strong predictors of the disease with high statistical significance independent of the age, education, and gender of the subjects. Prosodic features were more successful than the linguistic features. In fact, only two of the linguistic features were found to be significant. Accuracy of greater than 80% was obtained with three of the prosodic features. Silence ratio, which is defined as the rate of silences in speech regardless of their durations, was found to be the most useful feature. Feature fusion did not improve the performance, which indicates that the features are not complementary to one another.

2 Linguistic features

A list of the linguistic features that are extracted from the transcriptions of the recorded conversations with the test subjects is shown in Table 1. The features are geared towards detecting problems with the flow of the conversation and measuring how well the subject can understand the question or carry on the conversation without getting confused. Recordings are manually transcribed. Each recording is first split into conversation turns. Then, the turns where the subjects speak are further split into utterances that are segments where the subjects talk without interruption by the interviewer and without long silences. The splitting mechanism is shown in Figure 1 where a voice activity detector is used for detecting silences. Only the turns of subjects are used in feature extraction.

2.1 Hesitation and confusion features

During recording, we 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 confused 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.

2.1.1 Question ratio

Patients are more likely to forget details in the middle of conversation, to not understand the questions, or to forget the context of the question. In those cases, they tend to ask the interviewer to repeat the question or they get confused, talk to themselves, and ask further questions about the details. The question words such as 'which,' 'what,' etc. are tagged automatically in each conversation. The full list of question tags that were used here is shown in Table 2. The question ratio of a subject is computed by dividing

Table 1 Lists of linguistic and prosodic features and their IDs

	Category	ID	Features
Linguistic features	Hesitation and puzzlement features	1.1	Question ratio
		1.2	Filler ratio
		1.3	Incomplete sentence ratio
	POS-based features	2.1	Verb freq.
		2.2	Noun freq.
		2.3	Pronoun freq.
		2.4	Adverb freq.
		2.5	Adjective freq.
		2.6	Particle freq.
		2.7	Conjunction freq.
		2.8	Pronoun-to-noun ratio
	Intelligibility	3	Unintelligible word ratio
	Complexity features	4.1	Standardized word entropy
		4.2	Suffix ratio
		4.3	Number ratio
		4.4	Brunet's index
		4.5	Honore's statistic
		4.6	Type-token ratio
Prosodic features	Voice activity-related features	5.1	Response time
		5.2	Response length
		5.3	Silence ratio
		5.4	Silence to utt. ratio
		5.5	Long silence ratio
		5.6	Avg. silence count
		5.7	Silence rate
		5.8	Cont. speech rate
		5.9	Avg. cont. word count
	Articulation-related features	6.1	Avg. abs. delta energy
		6.2	Dev. of abs. delta energy
		6.3	Avg. abs. delta pitch
		6.4	Dev. of abs. delta pitch
		6.5.1	Avg. abs. delta formant 1
		6.5.2	Avg. abs. delta formant 2
		6.5.3	Avg. abs. delta formant 3
		6.5.4	Avg. abs. delta formant 4
		6.6	Voicing ratio
	Rate of speech-related features	7.1	Phoneme rate
		7.2	Word rate

the total number of question words by the number of utterances spoken by the subject.

2.1.2 Filler ratio

Filler sounds such as 'ahm' and 'ehm' are used by people in spoken language when they think about what to say next. We hypothesize that they may be used more frequently by the patients because of slow thinking and memory recall processes. Patients tend to forget what they are talking about and to use fillers more often than the control subjects. The filler ratio is computed by dividing the total number of filler words by the total number of utterances spoken by the subject.

2.1.3 Incomplete sentence ratio

One of our observations of the patients is their inability to complete sentences. They seem to either forget what they were going to say or to completely change the context and start talking about a different topic. Incomplete sentences are manually labeled for each conversation. To compute this feature, the ratio of incomplete sentences to the total number of the sentences is calculated.

2.2 POS-based features

Part of speech (POS) tags can be used to extract markers for detecting the disease. For example, frequent adjectives can indicate more colorful and descriptive use of language, while frequent adverbs can indicate the ability to relate different utterances to each other. The frequency of each POS tag can also be a useful identifier of patients with Alzheimer's disease.

POS tags are added automatically to each word using a Turkish stemmer [18]. In cases where a word can have multiple alternative POS tags, equal weights are given to all possibilities. For instance, if a word can be either a noun or an adverb, depending on the sentence, that word is counted as half adverb and half noun in computation. The following POS tag frequencies are used as features:

- Verb frequency
- Noun frequency
- Pronoun frequency
- Adverb frequency
- Adjective frequency
- Particle frequency
- Conjunction frequency
- Pronoun-to-noun ratio

Frequency of a POS tag is computed by dividing the total number of words with that tag by the total number of words spoken by the subject in the recording. Pronoun-to-noun ratio is the ratio of the total number of pronouns to the total number of nouns.

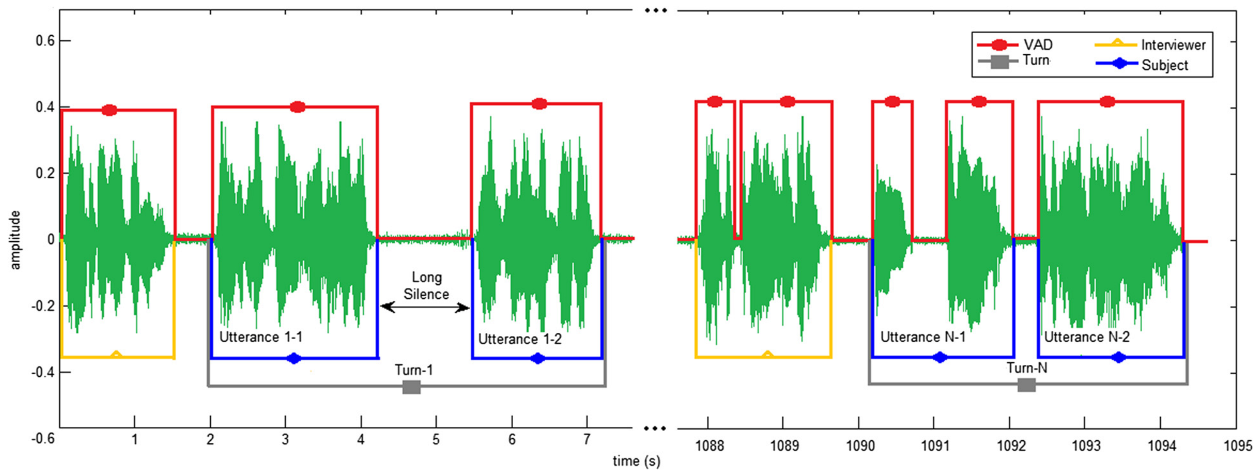

Figure 1 Conversation turns between a patient and the interviewer are shown. Each conversation turn contains one or more utterances. Voice activity detector (VAD) is used to detect the long and short silences.

2.3 Unintelligible word ratio

During the conversations, some of the words spoken by the patients were unintelligible. These are mostly because patients could not produce the words correctly, they mumbled, or they were thinking while talking, which reduced intelligibility. Unintelligible word ratio is the ratio of unintelligible words to all words spoken by the subject.

Annotation of unintelligible words was done manually by three listeners for each conversation. A word was marked as unintelligible only when at least two of the three listeners could not understand it.

2.4 Complexity features

2.4.1 Standardized word entropy

One of the earliest parts of the brain to be damaged by Alzheimer's disease is the part of the brain that deals with language ability [5]. We hypothesize that this may cause a degradation in the variety of words and word combinations that a patient uses. Standardized word entropy, i.e., word entropy divided by the log of the total word count, is used to model this phenomenon. Because the aim is to compute the variety of word choice, stemming is done, and only the stems of the words are considered.

Table 2 Question tags that were used in computing the question ratio

Word	Translation
Efendim	Excuse me
Hangi[si]	Which [one]
Hani	So where's/Why ... not .../You remember
Kaç[ı]	How many/much [of]
Kim[in]	Who[se]
M[i|ı|u|ü]	Question suffix
Nasıl	How
Ne[suffix+]	What
Nere[suffix+]	Where
Niye	Why

Some of the question tags occur with one or more additional suffixes. Those are indicated with [.] symbol.

2.4.2 Suffix ratio

The standardized word entropy feature focuses on the variety of the stem words while ignoring the suffixes. However, suffixes can also be strong indicators of the complexity of a sentence. Turkish, in particular, has a rich and complex morphological structure [19]. Hundreds of different words can be generated from the same stem word by appending suffixes to it. Thus, we investigated whether the patients tend to construct simpler words than the control subjects by analyzing the suffixes they used. The suffix ratio of a subject is calculated by dividing the total number of suffixes by the total number of words spoken by the subject.

2.4.3 Number ratio

During conversations, subjects give details about their birth dates, how many kids they have, and other numerical information. Such use of numbers in a sentence can be a measure of recall ability. The number ratio feature is calculated by dividing the total count of numbers by the total count of words the subject used in the conversation.

2.4.4 Brunet's index

Brunet's index (W) quantifies lexical richness [20]. It is calculated as $W = N^{V^{-0.165}}$, where N is the total text length and V is the total vocabulary. Lower values of W correspond to richer texts. As with standardized word entropy, stemming is done on words and only the stems are considered.

2.4.5 *Honore's statistic*

Honore's statistic [21] is based on the notion that the larger the number of words used by a speaker that occur only once, the richer his overall lexicon is. Words spoken only once (V_1) and the total vocabulary used (V) have been shown to be linearly associated. Honore's statistic generates a lexical richness measure according to $R = 100 \times \log(N/(1 - V_1/V))$, where N is the total text length. Higher values correspond to a richer vocabulary. As with standardized word entropy, stemming is done on words and only the stems are considered.

2.4.6 *Type-token ratio*

A pattern that we noticed in the recordings of the Alzheimer's patients is the frequency of repetitions in conversation. Patients tend to forget what they have said and to repeat it elsewhere in the conversation. The metric that we used to measure this phenomenon is type-token ratio [22]. Type-token ratio is defined as the ratio of the number of unique words to the total number of words. In order to better assess the repetitions, only the stems of the words are considered in calculations.

3 Prosodic features

A total of 20 prosodic features were extracted and evaluated for detecting Alzheimer's disease. A list of all prosodic features used here is shown in Table 1. Descriptions of the prosodic features are given below. All prosodic feature computations are performed over the locution of the subject. Locution is the total response period of the subject which is the sum of all of the subject's speech turns. Each speech turn includes utterances, long silences, and short silences, as shown in Figure 1.

3.1 Voice activity-related features

Silence and speech segments are automatically labeled in each conversation with a voice activity detector (VAD). 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, the energy distribution has two modes, both of which can be modeled with a Gaussian distribution. The bimodal distribution of silence and speech is trained using the expectation-maximization (EM) algorithm. The mode that has a lower mean is used to represent silence, and the mode that has a higher mean is used to represent speech.

Energy of each short-time speech frame in the recording is classified as either speech or silence using the likelihood ratio test (LRT). Because the test treats each frame independently, a second processing step is used where silence and speech segments that were shorter than four frames are removed.

The transcriptions of recordings were available and could be used for VAD through forced alignment using an automatic speech recognition system. However, the VAD described above worked well and more sophisticated VAD techniques were not required.

3.1.1 *Response time*

When the interviewer asks a question, it often takes some time before the subject gives a response. It is hypothesized that this time can be an indicator of the disease since it is expected to be related to cognitive processes such as attention and memory. The time it takes the subject to answer a question is calculated in each segment as the response time measure.

3.1.2 *Response length*

Response length is the average length of a subject's response in seconds to the interviewer's question. Beginning and trailing silences are removed.

3.1.3 *Silence ratio*

The plan-and-execute cycle in speech production was found to be distinctly different in patients compared to control subjects as noted in [14]. In our data, we also observed that patients tend to stop more in the middle of sentences to think about what to say next. The silence ratio is computed by dividing the total number of silences in the whole locution by the total number of words in the locution. Dividing by the number of words, we reduce the variability that arises from different speaking rates.

3.1.4 *Silence-to-utterance ratio*

Silence-to-utterance ratio is the ratio of the total number of silences to the total number of utterances. Similar to silence ratio, it is a measure of the hesitation rate of the subject.

3.1.5 *Long silence ratio*

Patients sometimes pause for a long time while answering a question. They do not use fillers during these long periods, and the interviewers did not interrupt these periods of silence. We hypothesized that these pauses may correspond to moments when the subject is retrieving information which is expected to be longer for the Alzheimer's patients. Similarly, confusion may also lead to long silences. The rate of such long hesitation events, defined as silences longer than approximately one second, is used to detect the disease. This feature is computed as the ratio of the total count of long silences to the total number of words.

3.1.6 *Average silence count*

This feature specifies the average number of silences produced by a speaker in one second of speech. It is calculated by dividing the total number of silences by the duration of the locution.

3.1.7 Silence rate

The silence rate measures the silence as a proportion of the whole locution. It is computed by dividing the total duration of all silence segments by the duration of the locution.

3.1.8 Continuous speech rate

This feature measures how long the subject speaks until the next long silence, which is considered to be a thinking or recalling state. It is defined as the average duration of continuous speech segments over the whole locution.

3.1.9 Average continuous word count

As mentioned above, the thinking process longer for patients than for the control subjects. The silence rate features discussed above try to exploit this long thinking process. Another way to measure it is to compute the average number of consecutive words that are spoken without intervening silences. First, the number of words for each continuous segment is computed. Then, the mean of these counts is used as the feature.

3.2 Articulation-related features

The voice activity-related features discussed above are related to cognitive thought processes. However, it is also important to measure how the subject uses his or her voice articulations during speech. For example, if the subject becomes 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 subject's health. If the formants do not change fast enough or are not distinct enough, sounds may become harder for listeners to identify, leading to the perception of mumbling. In order to see the impact of these effects on classification of the disease, pitch and formant trajectories are extracted, and the following features are derived over the whole locution.

3.2.1 Average absolute delta energy

Energy variations can convey information about the mood of the subject. Changing energy significantly during speech may indicate a conscious effort to stress words that are semantically important or a change in mood related to the content of the speech. The absolute value of each frame-to-frame energy change is measured, and the average of these changes over the whole locution is computed.

3.2.2 Deviation of absolute delta energy

In addition to changes in energy, changes in the delta energy, which is the acceleration of energy, can be used. The standard deviation of the average absolute delta energy is used to further investigate the possible impacts of the disease on the energy change rate.

3.2.3 Average absolute delta pitch

The average of the absolute delta pitch shows the rate of variations in pitch. This feature is highly correlated with the emotions carried through the speech signal.

3.2.4 Deviation of absolute delta pitch

The standard deviation of the absolute delta pitch is also used as a feature to further analyze the possible impacts of the disease on the pitch change rate. A monotonic increase or decrease in the pitch may simply be related to routine changes in sentence structure. However, acceleration of pitch, measured with the standard deviation of absolute delta pitch, can capture unusual pitch events in speech.

3.2.5 Average absolute delta formants

The average of the 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 the tongue and lips. Reduction of control over these organs related to damage in the brain caused by Alzheimer's disease can create speech impairments such as mumbling. In this case, formants do not change quickly and speech becomes less intelligible [23]. Changes in the first four formants are used as features in this research.

3.2.6 Voicing ratio

Another speech impairment is the loss of voicing in speech. In this case, the subject loses the ability to control the vibrations of the vocal cords, which results in breathy and noisy speech. The ratio of the total duration of voiced speech to the total duration of speech in the locution is used to detect any potential impairment in the vocal cords.

3.3 Rate of speech-related features

3.3.1 Phoneme rate

A basic identifier of rate of speech is the average number of phonemes spoken per second. The phoneme rate of a subject is computed by dividing the number of phonemes by the duration of the locution.

3.3.2 Word rate

Similar to phoneme rate, word rate is used to measure the rate of speech at the word level. Word rate is computed by dividing the number of words by the duration of the locution.

4 Experiments

Conversational speech recordings of 32 patients and 51 age and education-matched control subjects were collected and manually transcribed. Recordings from four patients were neglected because they were either unintelligible for the most part or they did not talk much. Thus, recordings from a total of 28 patients were used in experiments. The Alzheimer's patients and the control subjects were recruited from the same healthcare facility, but the control subjects were receiving treatment for injuries or illnesses other than Alzheimer's disease. Gender, age, and education details of the subjects are shown in Table 3. The age range is between 60 and 90 in both control subjects and patients.

Unstructured conversations were carried with the subjects where questions were asked depending on the flow of the conversation. Thus, different topics and different questions were used to make the subjects feel comfortable. The transcriptions were produced by one person and then reviewed by another person. The transcribers and the subjects were native Turkish speakers. In order to annotate unintelligible words, a third person also listened to the recordings.

The data was collected at elderly healthcare facilities in Istanbul. For each subject, approximately 10 min of conversation was recorded using a high-quality microphone. The recording was then manually segmented into speech turns between the interviewer and subject. In each speech turn, only the subject or the interviewer speaks. Segments of speech where both the subject and the interviewer talk were not used in the analysis.

After linguistic and prosodic features are extracted, SVM, NN classifiers, naive Bayesian classifiers, and CTree are used for classification. A linear Kernel is used for the SVM. For the NN classifier, Euclidean distance is used. For the CTree, nodes are split to minimize within-node impurity. Impure nodes that contain samples both from patients and control subjects are split only if they have more than nine samples.

There is more data available for the control subjects than for the Alzheimer's patients because the number of subjects that were in the healthcare facilities and willing to provide data was larger. Even though equal amounts of data from both groups could be used in the experiments to have a balance, all of the available data was used with special care while training the classifiers as discussed below.

Table 3 Gender, age, and years of education for the patient and control subjects

	AD ($n = 28$)	Control ($n = 51$)
Male/Female	18/10	31/20
Age	75 (6)	75.9 (6.4)
Education	11.6 (4.9)	11.4 (6)

Age and education data are presented in mean (standard deviation) format.

Data imbalance can become a problem for the SVM, NN, and decision tree algorithms, where the data points are used directly, as opposed to the naive Bayes approach, where the distribution of the data is used. For the NN, SVM, and decision tree classifiers, a random subsampling approach is used, in which a subset of the control subjects is randomly selected such that there is an equal number of data points for the control subjects and the patients. For each test case, the subsampling procedure is repeated ten times, and the average performance is reported.

In the first phase of testing, each feature is tested separately to assess the classification power of individual features. Then, in the second phase, combinations of features are used to increase the classification power of the algorithms. Features are normalized to have zero-mean and unit variance.

Because there is a limited number of subjects in the data set, a leave-one-out evaluation strategy is used, in which one of the subjects is left out and the classifier is trained with the rest of the subjects and tested on the left-out subject.

5 Results and discussion

The age, education level, and gender of the subjects can significantly affect performance in classification tests. Therefore, initial testing is done to control for the effect of age, education, and gender on the performance. Only features that have significant performance in all three control tests are reported as significant markers. Significance is measured using the paired t-test. A given feature may not have significant performance with all classifiers. In this case, the feature is reported as significant if it can pass the significance test with at least one of the classifiers.

Age-, education-, and gender-controlled experimental results are discussed below. Analysis and discussion of the features and combination of features with statistically significant discriminative power are reported in Section 5.4.

5.1 Age-controlled experiments

The age-controlled linguistic features are shown in Table 4. All features related to POS tags other than nouns and pronouns were found to be insignificant with this control variable. Incomplete sentences and unintelligible word ratios were found to be age related and not disease related. Similarly, all features that are related to the complexity of the language also became insignificant when age was used as a control variable.

The age-controlled prosodic features are shown in Table 5. The significance of these features was found to be less related to age compared to the linguistic features. In particular, formant and voicing features that are related to articulation were found to be age related and not disease related. Patients in the older age group had a harder time controlling their vocal cords and other articulatory

Table 4 Accuracy (%) of each classifier using the linguistic features in age-controlled experiments

		Age between 60 and 75					**Age between 76 and 90**				
		(*p* = 15, *c* = 25)					**(*p*= 13, *c* = 26)**				
ID	**Features**	**SVM**	**Bayes**	**CTree**	**NN**	***p* value**	**SVM**	**Bayes**	**CTree**	**NN**	***p* value**
1.1	Question ratio	***70.0***	***67.5***	***70.0***	60.0	**0.011**	***66.7***	35.0	38.5	43.6	**0.037**
1.2	Filler ratio	52.5	65.0	57.5	47.5	0.058	48.7	38.5	46.2	53.9	0.631
1.3	Incomplete sentence ratio	60.0	57.5	52.5	60.0	0.206	61.6	***66.7***	59.0	64.1	**0.037**
2.1	Verb freq.	37.5	62.5	47.5	50.0	0.114	***69.2***	61.5	53.9	59.0	**0.016**
2.2	Noun freq.	***67.5***	65.0	60.0	60.0	**0.027**	59.0	***71.8***	53.9	59.0	**0.006**
2.3	Pronoun freq.	***72.5***	***72.5***	65.0	47.5	**0.004**	***69.2***	***66.7***	56.4	59.0	**0.016**
2.4	Adverb freq.	52.5	62.5	52.5	52.5	0.114	59.0	41.0	38.5	43.6	0.262
2.5	Adjective freq.	57.5	47.5	45.0	42.5	0.343	51.3	43.6	51.3	51.3	0.873
2.6	Particle freq.	52.5	60.0	52.5	55.0	0.206	48.7	61.5	***77.0***	***66.7***	**0.001**
2.7	Conjunction freq.	62.5	62.5	65.0	52.5	0.058	53.9	56.4	56.4	46.2	0.423
2.8	Pronoun-to-noun ratio	***70.0***	***70.0***	***67.5***	62.5	**0.011**	***71.8***	***66.7***	61.5	46.2	**0.006**
3	Unintelligible word ratio	60.0	62.5	55.0	55.0	0.114	***82.1***	***84.6***	***84.6***	***82.1***	**<0.001**
4.1	Standardized word entropy	55.0	65.0	45.0	57.5	0.058	***74.4***	64.1	59.0	56.4	**0.002**
4.2	Suffix ratio	45.0	52.5	55.0	60.0	0.206	***69.2***	***74.4***	64.1	53.9	**0.002**
4.3	Number ratio	***70.0***	***67.5***	***70.0***	65.0	**0.011**	48.7	61.5	51.3	48.7	0.150
4.4	Brunet's index	62.5	57.5	***67.5***	65.0	**0.027**	48.7	59.0	61.5	56.4	0.150
4.5	Honore's statistic	40.0	45.0	65.0	47.5	0.058	43.6	61.5	61.5	***71.8***	**0.006**
4.6	Type-token ratio	47.5	45.0	42.5	45.0	1.000	***79.5***	***79.5***	64.1	***76.9***	**<0.001**

Number of patients (*p*) and control subjects (*c*) in each group is given in parenthesis. Classifiers with significant results are denoted in italics. Features with *p* values less than 0.05 are considered significant and are shown in bold. *p* value of the best classifier is reported for each feature.

Table 5 Accuracy (%) of each classifier using the prosodic features in age-controlled experiments

		Age between 60 and 75					**Age between 76 and 90**				
		(*p* = 15, *c* = 25)					**(*p* = 13, *c* = 26)**				
ID	**Features**	**SVM**	**Bayes**	**CTree**	**N**	**p-value**	**SVM**	**Bayes**	**CTree**	**NN**	**p-value**
5.1	Response time	60.0	60.0	52.5	45.0	0.206	61.5	59.0	48.7	48.7	0.150
5.2	Response length	62.5	57.5	52.5	65.0	0.058	***66.7***	41.0	51.3	61.5	**0.037**
5.3	Silence tatio	***70.0***	***70.0***	62.5	***72.5***	**0.004**	***94.9***	***94.9***	***94.9***	***84.6***	**<0.001**
5.4	Silence to Utt. ratio	***67.5***	65.0	60.0	62.5	**0.027**	***76.9***	***71.8***	***69.2***	59.0	**0.001**
5.5	Long silence ratio	57.5	***70.0***	57.5	55.0	**0.011**	***74.4***	***74.4***	***69.2***	***69.2***	**0.002**
5.6	Avg. silence count	***82.5***	***80.0***	***82.5***	***80.0***	**<0.001**	***74.4***	***74.4***	64.1	59.0	**0.002**
5.7	Silence rate	***67.5***	***75.0***	62.5	***72.5***	**0.002**	***74.4***	***79.5***	***74.4***	59.0	**<0.001**
5.8	Cont. speech rate	***80.0***	***80.0***	***75.0***	***75.0***	**<0.001**	***74.4***	***71.8***	***66.7***	64.1	**0.002**
5.9	Avg. cont. word count	***67.5***	62.5	60.0	55.0	**0.027**	***87.2***	***92.3***	***94.9***	***84.6***	**<0.001**
6.1	Avg. abs. delta energy	***67.5***	***75.0***	***67.5***	57.5	**0.002**	***71.8***	***71.8***	***79.5***	64.1	**<0.001**
6.2	Dev. of abs. delta energy	***75.0***	***72.5***	62.5	55.0	**0.002**	***69.2***	***71.8***	41.0	51.3	**0.006**
6.3	Avg. abs. delta pitch	***67.5***	62.5	55.0	47.5	**0.027**	61.5	***82.1***	59.0	64.1	**<0.001**
6.4	Dev. of abs. delta pitch	45.0	55.0	55.0	***72.5***	**0.004**	59.0	***69.2***	56.4	56.4	**0.016**
6.5.1	Avg. abs. delta formant 1	62.5	57.5	55.0	62.5	0.114	***74.4***	***71.8***	***69.2***	61.5	**0.002**
6.5.2	Avg. abs. delta formant 2	55.0	62.5	52.5	52.5	0.114	***69.2***	***66.7***	***66.7***	***71.8***	**0.006**
6.5.3	Avg. abs. delta formant 3	52.5	47.5	62.5	55.0	0.114	***66.7***	***66.7***	***76.9***	59.0	**0.001**
6.5.4	Avg. abs. delta formant 4	45.0	65.0	55.0	57.5	**0.002**	***74.4***	***76.9***	***66.7***	***69.2***	**0.001**
6.6	Voicing ratio	***75.0***	***75.0***	62.5	***75.0***	**0.002**	59.0	59.0	53.9	43.6	0.262
7.1	Phoneme rate	***67.5***	***75.0***	***77.5***	60.0	**0.001**	***82.1***	***82.1***	***76.9***	58.0	**<0.001**
7.2	Word rate	***72.5***	***72.5***	***67.5***	60.0	**0.004**	56.4	61.6	***66.7***	48.7	**0.037**

The number of patients (*p*) and control subjects (*c*) in each group is given in parenthesis. Classifiers with significant results are denoted in italics. Features with *p* values less than 0.05 are considered significant and are shown in bold. *p* value of the best classifier is reported for each feature.

organs, but this was not a significant discriminative factor for the younger group of subjects.

5.2 Education-controlled experiments

As in the age-controlled experiments, linguistic features performed poorly in the education-controlled experiments, shown in Table 6. Most notably, pronoun frequency and pronoun-to-noun ratio were significant indicators of the disease independent of the education level. Patients use pronouns more often than nouns. This is surprising since we hypothesized that patients would use pronouns less often since they are used to refer to nouns mentioned earlier in the conversation which we assumed would require more cognitive effort. Analyzing the transcripts in more detail, we have found that patients use pronouns without necessarily referring to a specific noun. Sometimes it is hard, even impossible, for the interviewer to understand what a pronoun is referring to. Patients seem to prefer using pronouns instead of actual nouns, which are not always specified in the conversation.

Prosodic features were less dependent on the education level compared to age, as shown in Table 7. Response length was found to be insignificant in education-controlled experiments. Some of the patients with higher education either do not talk much or talk significantly more than the control group. However, such speakers do not exist in the lower education group. Hence, on average, response length was not found to be significant.

The average absolute delta pitch and average absolute delta formant-2 features were also found to be dependent on the education level. These two features are particularly interesting since they also have high correlation with the display of mood and depression [23]. Patients in the high education group sometimes displayed exaggerated emotions which increased the pitch variability. Interestingly, some of the subjects in the same group tend to have lower second formant deviations, which can be a sign of depression. Those two patterns, however, were not observed in the younger patients. Thus, they were not found to be significant markers of the disease.

5.3 Gender-controlled Experiments

Gender-controlled experiments were performed to measure the performance of features for each gender separately. Results are shown in Tables 8 and 9. There are three features that performed well in the age-controlled and

Table 6 Accuracy (%) of each classifier using the linguistic features in education-controlled experiments

		High school and below (p = 18, c = 24)					College and above (p = 10, c = 27)				
ID	**Features**	**SVM**	**Bayes**	**CTree**	**NN**	***p* value**	**SVM**	**Bayes**	**CTree**	**NN**	***p* value**
1.1	Question ratio	54.8	38.1	59.5	45.2	0.217	***73.0***	***75.7***	***70.3***	62.2	**0.002**
1.2	Filler ratio	40.5	35.7	35.7	38.1	1.000	62.2	59.5	***70.3***	62.2	**0.014**
1.3	Incomplete sentence ratio	52.4	54.8	57.1	59.5	0.217	56.8	59.5	64.9	***67.6***	**0.033**
2.1	Verb freq.	61.9	61.9	61.9	47.6	0.123	64.9	54.1	64.9	***73.0***	**0.005**
2.2	Noun freq.	57.1	40.5	59.5	50.0	0.217	***78.4***	***78.4***	***86.5***	***73.0***	**<0.001**
2.3	Pronoun freq.	***76.2***	***71.4***	61.9	54.8	**0.001**	***75.7***	***73.0***	64.9	51.4	**0.002**
2.4	Adverb freq.	38.1	35.7	50.0	40.5	1.000	***70.3***	***81.1***	59.5	62.2	**<0.001**
2.5	Adjective freq.	52.4	45.2	54.8	57.1	0.355	54.1	51.4	56.8	56.8	0.411
2.6	Particle freq.	54.8	47.6	61.9	50.0	0.123	43.3	62.2	54.1	62.2	0.139
2.7	Conjunction freq.	42.9	35.7	35.7	42.9	1.000	***73.0***	***73.0***	***75.7***	64.9	**0.002**
2.8	Pronoun-to-noun ratio	***73.8***	***71.4***	57.1	59.5	**0.002**	***78.4***	***73.0***	51.4	62.2	**0.001**
3	Unintelligible word ratio	64.3	57.1	57.1	57.1	0.064	***78.4***	***78.4***	62.2	***67.6***	**0.001**
4.1	Standardized word entropy	64.3	***69.1***	57.1	57.1	**0.014**	***70.3***	56.8	59.5	56.8	**0.014**
4.2	Suffix ratio	57.1	57.1	59.5	57.1	0.217	62.2	37.8	54.1	62.2	0.139
4.3	Number ratio	57.1	50.0	50.0	35.7	0.355	59.5	56.8	56.8	43.3	0.250
4.4	Brunet's index	40.5	61.9	54.8	52.4	0.123	***67.6***	54.1	48.7	62.2	**0.033**
4.5	Honore's statistic	40.5	52.4	45.2	54.8	0.537	56.7	56.8	54.1	48.7	0.411
4.6	Type-token ratio	59.5	64.3	45.2	50.0	0.064	***67.6***	37.8	59.5	56.8	**0.033**

The number of patients (p) and control subjects (c) in each group is given in parenthesis. Classifiers with significant results are denoted in italics. Features with p values less than 0.05 are considered significant and are shown in bold. p value of the best classifier is reported for each feature.

Table 7 Accuracy (%) of each classifier using the prosodic features in education-controlled experiments

		High school and below					College and above				
		(p = 18, c = 24)					(p = 10, c = 27)				
ID	**Features**	**SVM**	**Bayes**	**CTree**	**NN**	***p* value**	**SVM**	**Bayes**	**CTree**	**NN**	***p* value**
5.1	Response time	52.4	61.9	54.8	***69.1***	**0.014**	***70.3***	59.5	62.2	54.1	**0.014**
5.2	Response length	47.6	42.9	38.1	42.9	1.000	***73.0***	51.4	46.0	***67.6***	**0.005**
5.3	Silence ratio	***73.8***	***71.4***	64.3	***73.8***	**0.002**	***91.9***	***97.3***	***94.6***	***91.9***	**<0.001**
5.4	Silence to utt. ratio	***69.1***	***69.0***	57.1	59.5	**0.014**	***83.8***	***81.1***	***73.0***	***75.7***	**<0.001**
5.5	Long silence ratio	64.3	***66.7***	59.5	***66.7***	**0.031**	***70.3***	***78.4***	54.1	51.4	**0.001**
5.6	Avg. silence count	***83.3***	***85.7***	***83.3***	***71.4***	**<0.001**	***75.7***	***70.3***	***73.0***	***67.6***	**0.002**
5.7	Silence rate	***66.7***	***73.8***	***83.3***	***76.2***	**<0.001**	***75.7***	62.2	***78.4***	59.5	**0.001**
5.8	Cont. speech rate	***83.3***	***83.3***	***78.6***	***73.8***	**<0.001**	***75.7***	***70.3***	59.5	***73.0***	**0.002**
5.9	Avg. cont. word count	59.5	***71.4***	57.1	59.5	**0.005**	***89.2***	***91.9***	***91.9***	***83.8***	**<0.001**
6.1	Avg. abs. delta energy	***69.1***	57.1	64.3	61.9	**0.014**	***78.4***	***78.4***	***75.7***	48.7	**0.001**
6.2	Dev. of abs. delta energy	***69.1***	***66.7***	50.0	42.9	**0.014**	***73.0***	***75.7***	59.5	56.8	**0.002**
6.3	Avg. abs. delta pitch	45.2	45.2	38.1	40.5	1.000	64.9	***83.8***	***73.0***	56.8	**<0.001**
6.4	Dev. of abs. delta pitch	38.1	50.0	47.6	42.9	1.000	56.8	62.2	54.1	***67.6***	**0.033**
6.5.1	Avg. abs. delta formant 1	61.9	64.3	61.9	61.9	0.064	***67.6***	64.9	48.7	46.0	**0.033**
6.5.2	Avg. abs. delta formant 2	61.9	64.3	54.8	47.6	0.064	***67.6***	***70.3***	***73.0***	***78.4***	**0.001**
6.5.3	Avg. abs. delta formant 3	57.1	47.6	35.7	38.1	0.355	***75.7***	62.2	64.9	62.3	**0.002**
6.5.4	Avg. abs. delta formant 4	***69.1***	***66.7***	57.1	***66.7***	**0.014**	54.1	62.2	***70.3***	62.3	**0.014**
6.6	Voicing ratio	64.3	64.3	61.9	***69.0***	**0.014**	***67.6***	64.9	64.9	56.8	**0.033**
7.1	Phoneme rate	***66.7***	***66.7***	47.6	42.9	**0.031**	***83.8***	***86.5***	***83.8***	***70.3***	**<0.001**
7.2	Word rate	***66.7***	64.3	61.9	***66.7***	**0.031**	64.9	***73.0***	***81.1***	64.9	**<0.001**

The number of patients (*p*) and control subjects (*c*) in each group is given in parenthesis. Classifiers with significant results are denoted in italics. Features with *p* values less than 0.05 are considered significant and are shown in bold. *p* value of the best classifier is reported for each feature.

Table 8 Accuracy (%) of each classifier using the linguistic features in gender-controlled experiments

		Male					Female				
		(p = 18, c = 31)					(p = 10, c = 20)				
ID	**Features**	**SVM**	**Bayes**	**CTree**	**NN**	**p-value**	**SVM**	**Bayes**	**CTree**	**NN**	**p-value**
1.1	Question ratio	47.1	50.0	***69.0***	60.8	**0.007**	***70.7***	***68.0***	58.7	61.0	**0.022**
1.2	Filler ratio	43.7	40.2	37.1	39.4	1.000	***66.7***	51.0	63.7	***69.0***	**0.031**
1.3	Incomplete sentence ratio	46.1	43.9	54.5	52.9	0.295	53.7	47.7	47.7	52.0	0.379
2.1	Verb freq.	51.4	55.5	45.5	51.0	0.242	46.7	54.0	52.0	48.0	0.335
2.2	Noun freq.	48.6	46.7	55.3	***64.7***	**0.031**	***80.7***	***70.7***	***76.0***	***69.7***	**0.001**
2.3	Pronoun freq.	59.2	59.7	***63.6***	52.8	**0.016**	59.0	***71.5***	65.5	50.0	**0.012**
2.4	Adverb freq.	47.1	46.7	50.8	48.0	0.458	57.0	45.0	44.7	41.7	0.183
2.5	Adjective freq.	53.7	52.2	54.7	***65.3***	**0.030**	36.0	42.0	44.0	50.7	0.473
2.6	Particle freq.	***65.3***	***66.9***	***66.7***	59.2	**0.010**	39.0	38.0	41.7	41.0	1.000
2.7	Conjunction freq.	45.3	***65.3***	50.8	49.2	**0.021**	53.0	***71.0***	64.7	***69.7***	**0.010**
2.8	Pronoun-to-noun ratio	***65.0***	61.1	39.4	39.4	**0.032**	***67.0***	***72.0***	65.0	***69.5***	**0.014**
3	Unintelligible word ratio	***66.7***	***66.1***	58.0	62.2	**0.022**	59.7	58.0	57.0	***72.0***	**0.009**
4.1	Standardized word entropy	57.1	60.8	42.2	46.7	0.092	56.0	53.7	52.7	58.7	0.148
4.2	Suffix ratio	56.9	48.6	46.1	42.4	0.196	***69.0***	***73.0***	***66.7***	58.7	**0.008**
4.3	Number ratio	***65.1***	***65.3***	59.4	55.3	**0.031**	51.7	58.7	60.0	49.0	0.120
4.4	Brunet's index	41.7	52.9	51.6	62.4	0.061	48.0	51.0	42.0	41.7	0.459
4.5	Honore's statistic	47.1	53.3	55.1	52.9	0.249	39.7	40.7	39.0	41.0	1.000
4.6	Type-token ratio	***64.1***	62.0	59.0	51.6	**0.041**	51.7	46.7	54.0	51.7	0.308

The number of patients (*p*) and control subjects (*c*) in each group is given in parenthesis. Classifiers with significant results are denoted in italics. Features with *p* values less than 0.05 are considered significant and are shown in bold. *p* value of the best classifier is reported for each feature.

Table 9 Accuracy (%) of each classifier using the prosodic features in gender-controlled experiments

		Male ($p = 18, c = 31$)					Female ($p = 10, c = 20$)				
ID	Features	SVM	Bayes	CTree	NN	p value	SVM	Bayes	CTree	NN	p value
5.1	Response time	***68.6***	58.4	***69.0***	***64.5***	**0.005**	50.7	50.0	44.7	42.7	0.468
5.2	Response length	50.0	46.1	***63.9***	***64.1***	**0.033**	51.0	50.7	59.0	50.0	0.114
5.3	Silence ratio	***86.1***	***83.3***	***83.9***	***80.8***	**<0.001**	66.0	***68.0***	***73.7***	***72.7***	**0.001**
5.4	Silence to utt. ratio	60.2	62.9	54.1	***66.9***	**0.017**	***83.0***	***82.0***	***71.7***	53.7	**<0.001**
5.5	Long silence ratio	***63.9***	***63.3***	55.9	48.0	**0.048**	51.7	63.0	47.0	53.7	0.077
5.6	Avg. silence count	***79.0***	***79.4***	***73.9***	***72.9***	**<0.001**	***76.0***	***72.0***	***80.0***	62.7	**<0.001**
5.7	Silence rate	***63.7***	***66.3***	***70.0***	***63.1***	**0.003**	59.0	60.7	58.0	***78.7***	**<0.001**
5.8	Cont. speech rate	***82.9***	***79.0***	***83.9***	***69.2***	**<0.001**	***73.0***	***72.0***	***71.7***	47.0	**0.007**
5.9	Avg. cont. word count	***86.1***	59.0	***84.1***	***77.1***	**<0.001**	***68.7***	62.7	***68.0***	50.7	**0.014**
6.1	Avg. abs. delta energy	***69.4***	***70.0***	***81.0***	***84.1***	**<0.001**	53.0	46.7	36.0	41.0	0.380
6.2	Dev. of abs. delta energy	***69.4***	***69.8***	49.2	51.0	**0.008**	55.7	57.0	42.7	38.7	0.176
6.3	Avg. abs. delta pitch	***67.8***	***74.7***	***63.1***	58.4	**0.001**	60.7	45.0	63.7	45.0	0.060
6.4	Dev. of abs. delta pitch	***76.1***	***72.2***	***63.9***	55.5	**<0.001**	48.0	38.7	***70.7***	***67.7***	**0.007**
6.5.1	Avg. abs. delta formant 1	59.4	***66.3***	***64.5***	53.9	**0.016**	53.0	52.0	46.7	48.0	0.374
6.5.2	Avg. abs. delta formant 2	55.9	60.2	57.8	52.0	0.108	39.0	38.7	42.7	55.7	0.259
6.5.3	Avg. abs. delta formant 3	***66.1***	***65.9***	60.2	50.8	**0.022**	52.7	42.7	***69.7***	48.0	**0.007**
6.5.4	Avg. abs. delta formant 4	***69.2***	***67.1***	***64.1***	58.6	**0.009**	42.7	53.7	56.7	52.7	0.209
6.6	Voicing ratio	***75.3***	***75.1***	***70.2***	***69.8***	**0.001**	57.7	41.0	58.7	50.7	0.111
7.1	Phoneme rate	***63.9***	***71.0***	***65.1***	***67.8***	**0.004**	***71.7***	***68.0***	48.0	46.0	**0.017**
7.2	Word rate	***70.0***	***70.2***	53.3	45.5	**0.005**	35.0	***66.7***	48.0	46.7	**0.014**

The number of patients (p) and control subjects (c) in each group is given in parenthesis. Classifiers with significant results are denoted in italics. Features with p values less than 0.05 are considered significant and are shown in bold. p value of the best classifier is reported for each feature.

Table 10 Overall accuracy (%), missed detection (%), and false alarm (%) rates of features with statistically significant performance

		Accuracy				Missed detection				False alarm			
ID	Features	SVM	Bayes	CTree	NN	SVM	Bayes	CTree	NN	SVM	Bayes	CTree	NN
2.3	Pronoun freq.	***73.4***	72.2	59.5	54.4	***39.3***	39.3	46.4	57.1	***19.6***	21.6	37.3	39.2
2.8	Pronoun-to-noun ratio	***72.2***	70.9	65.8	58.2	***39.3***	39.3	50.0	60.7	***21.6***	23.5	25.5	31.4
5.3	Silence ratio	***83.5***	81.0	79.8	78.5	***35.7***	35.7	28.6	25.0	***5.9***	9.8	15.7	19.6
5.4	Silence to utt. ratio	***73.4***	69.6	62.0	69.6	***35.7***	42.9	42.9	39.3	***21.6***	23.5	35.3	25.5
5.6	Avg. silence count	***81.0***	78.5	69.6	68.4	***21.4***	17.9	32.1	39.3	***17.7***	23.5	29.4	27.5
5.7	Silence rate	***73.4***	72.2	72.1	68.4	***64.3***	67.9	35.7	39.3	***5.9***	5.9	23.5	27.5
5.8	Cont. speech rate	***78.5***	76.0	68.4	62.0	***28.6***	25.0	25.0	39.3	***17.7***	23.5	35.3	37.4
5.9	Avg. cont. word count	76.0	***82.3***	76.0	70.9	28.6	***32.1***	28.6	35.7	21.6	***9.8***	21.6	25.5
7.1	Phoneme rate	69.6	***79.8***	60.8	57.0	32.1	***46.4***	46.4	50.0	29.4	***5.9***	35.3	39.2
7.2	Word rate	62.0	***65.8***	55.7	50.6	25.0	***32.1***	50.0	53.6	45.1	***35.3***	41.2	47.1

Features with statistically significant performance in age-, education-, and gender-controlled experiments. Performance of the classifier with highest accuracy is shown in italics for each feature. Note that because the number of patients and control subjects are not equal, sum of average error, which is (missed detection+false alarm)/2, and accuracy is not 100%.

Table 11 Correlation of features with statistically significant accuracies in age-, education-, and gender-controlled experiments

ID	Features	2.3	2.8	5.3	5.4	5.6	5.7	5.8	5.9	7.1	7.2
2.3	Pronoun freq.	—									
2.8	Pronoun-to-noun ratio	0.94	—								
5.3	Silence ratio	0.12	0.21	—							
5.4	Silence to utt. ratio	0.24	0.29	0.60	—						
5.6	Avg. silence count	−0.30	−0.36	−0.55	−0.38	—					
5.7	Silence rate	0.17	0.22	0.54	0.39	−0.68	—				
5.8	Cont. speech rate	0.33	0.34	0.38	0.31	−0.87	0.45	—			
5.9	Avg. cont. word count	0.04	−0.01	−0.67	−0.45	0.41	−0.40	−0.27	—		
7.1	Phoneme rate	−0.01	−0.14	−0.70	−0.44	0.52	−0.61	−0.23	0.49	—	
7.2	Word rate	−0.00	−0.06	−0.59	−0.38	0.47	−0.56	−0.19	0.42	0.90	—

education-controlled experiments but not in the gender-controlled experiments. Those features are: deviation of absolute delta energy, average absolute delta energy, and long silence ratio.

All three features performed well for males but not for females. In the recordings, we have found that males displayed less emotion which resulted in less expressive speech compared to male subjects in the control group. However, that pattern was not as strong with the female speakers. Moreover, the number of males is significantly larger than the number of females which makes it easier to get statistically significant results in classification experiments for the male subjects.

5.4 Analysis of significant features

Features that have significant performance in education-, age-, and gender-controlled tests are shown in Table 10, along with missed detection and false alarm rates. SVM and naive Bayes classifiers always outperform the CTree and NN classifiers. SVM has the best accuracy among all classifiers. In particular, SVM classifier with the silence ratio feature has the highest accuracy among all features and classifiers.

Missed detection rates are significantly higher than the false alarm rates in the best performing classifiers, as shown in Table 10. Even though more data from the control group is available, the subsampling method is used to

Table 12 Accuracy, missed detection, and false alarm rates of the best performing features

ID		Accuracy	Missed detection	False alarm
1 feature				
5.3	SVM	83.5%	35.7%	5.9%
		(73.5 to 90.9)	(18.6 to 55.9)	(1.2 to 16.2)
7.1	Bayes	79.8%	46.4%	5.9%
		(69.2 to 88.0)	(27.5 to 66.1)	(1.2 to 16.2)
2.3	SVM	73.4%	39.3%	19.6%
		(62.3 to 82.7)	(21.5 to 59.4)	(9.8 to 33.1)
2 features				
5.3 and 7.1	SVM	83.5%	35.7%	5.9%
		(73.5 to 91.0)	(18.6 to 55.9)	(1.2 to 16.2)
2.3 and 5.3	Bayes	82.3%	32.1%	9.8 %
		(72.1 to 90.0)	(15.9 to 52.4)	(3.3 to 21.4)
2.3 and 7.1	Bayes	78.5%	39.3%	11.8%
		(67.8 to 87.0)	(21.5 to 59.4)	(4.4 to 23.9)

Performance of best performing single feature in each feature category, as well as performance of combinations of these features. Results are reported only for the best classifier. Note that because the number of patients and control subjects are not equal, sum of average error, which is (missed detection + false alarm)/2, and accuracy is not 100%. Confidence intervals are shown in parenthesis.

ensure an equal number of patient and control subjects in the training datasets, as discussed in Section 4. Thus, the results show that significantly more patients were classified as healthy compared to control subjects classified as patients. It also indicates that features extracted from some patients are significantly different from some of the other patients and most of the control subjects, which helps in classification.

Voice activity-related features are particularly good at identifying the disease as shown in Table 10. Interestingly, features 5.3 and 5.6, which are related to the rate of silences, independent of the silence duration, were found to be more powerful discriminators than 5.4 and 5.7, which are related to long silences and duration of silences. Thus, frequency of silences during speech was found to be more important than the duration of silences. Features 5.8 and 5.9 indicate how long the subject can talk without a long silence. These two features are also highly correlated with the silence rate features, as shown in Table 11, and they had good performance in the classification experiments. Similarly, phoneme rate was strongly correlated with the rate of silences, and it performed well in experiments. Even though word rate is strongly correlated with the phoneme rate, it is not as strongly correlated with the silence rate as the phoneme rate, and it was not as successful in prediction of the disease.

Linguistic features did not perform as well as the prosodic features, as discussed in the previous section. Only pronoun rate and pronoun-to-noun ratio were strong indicators of the disease, but their prediction powers are not as strong as the prosodic features. However, their missed detection and false alarm rates are more balanced compared to prosodic features.

Features within each feature category are strongly correlated with each other, as shown in Table 11. Performances of the features with highest accuracy from each category are compared with confidence intervals in Table 12.

Feature fusion is used in an attempt to further boost the performance. In that approach, classifiers were trained with two features instead of a single feature. Because of high within-category correlations, feature fusion experiments were done by using the best performing feature in each category. Results are shown in Table 12. Not only statistically significant improvement over the best single feature could not be achieved but also performance slightly degraded with feature fusion.

Scatter diagrams of the features and decision boundaries for classification are shown in Figure 2. Silence ratio had better discrimination power than the other features. Unfortunately, the information in the other features failed to correct the errors made with silence ratio. Similarly, phoneme rate was found to be a powerful feature, but the pronoun frequency could not correct the errors it generates, as shown in the Figure 2C.

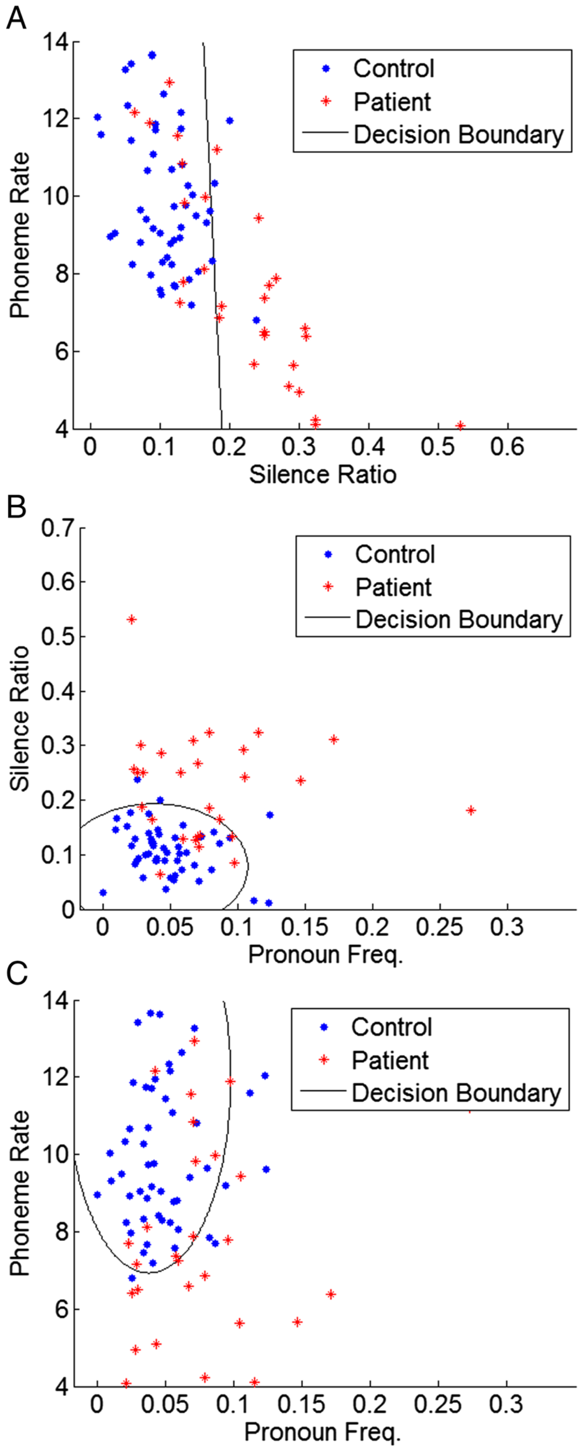

Figure 2 Scatter diagram of features in Table 12. Decision boundary of the best performing classifier is shown in each figure. Note that even though the features were normalized before classification, they are not normalized in these figures to make interpretation easier. **(A)** Scatter diagram for silence ratio vs. phoneme rate features. **(B)** Scatter diagram for pronoun frequency vs. silence ratio features. **(C)** Scatter diagram for pronoun frequency vs. phoneme rate features.

Note that increasing the size of feature vectors can in fact degrade the performance of classifiers due to the curse of dimensionality that occurs when there is not enough training data and the classifier cannot generalize and perform well on test data. That effect may be partly responsible for not observing an improvement with the feature fusion approach. For the same reason, feature fusion with larger number of features was not investigated.

6 Conclusions

We have investigated an extensive set of features derived from the speech signal and transcriptions of Alzheimer's patients and control subjects. It is already known that the part of the brain cortex that deals with linguistic abilities is one of the first to deteriorate with the onset of the disease. Our work explored how that deterioration is reflected in the patient's speech prosody and spoken language, and whether there are markers that can be effectively detected using machine learning techniques. Our results indicate that a prediction accuracy higher than 80% can be obtained with high confidence using the proposed features, independent of the age, education level, and gender of the subjects. Prosodic features were substantially better than the linguistic features. In fact, only two of the linguistic features were found to be strong markers of the disease.

Classification experiments were also done with combinations of features. However, using more than one feature did not outperform the best single feature. This may be a result of limited amounts of data used in training the classifiers which causes generalization problems when the number of features increases.

Our experiments are with late-stage patients, and the effectiveness of the markers that we have found should be measured with early-stage patients, where the signals are more subtle and more subjects may be needed to reach statistically significant results. However, our experimental results and manual observations from the data are encouraging, and we will start collecting data from early-stage patients in the near future.

Another topic that we will investigate in the future work is a cross-lingual study of the proposed features. Features that are independent of language can provide important clues about the neural degeneration process during the disease or perhaps can enable deeper understanding of neural networks in the brain that are responsible from cognition of language and speech production.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

The authors would like to thank the anonymous reviewers for their invaluable comments and suggestions to improve the quality of the paper. They are also grateful to Prof. Emily T. Prud'hommeaux of Rochester Institute of Technology for her invaluable comments during the revision of this paper.

Received: 1 February 2014 Accepted: 24 February 2015
Published online: 25 March 2015

References

1. M Prince, M Guerchet, M Prina, World Alzheimer report 2013: Journey of caring: an analysis of long-term care for dementia. http://www.alz.co.uk/research/world-report-2013 Accessed 2015-03-13
2. R Schmelzer, Roche Joins The Global CEO Initiative on Alzheimer's Disease. http://www.ceoalzheimersinitiative.org/node/71 Accessed 2014-08-20
3. MF Folstein, SE Folstein, PR McHugh, "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J. Psychiat. Res. **12**(3), 189–198 (1975)
4. LM Bloudek, DE Spackman, M Blankenburg, SD Sullivan, Review and meta-analysis of biomarkers and diagnostic imaging in Alzheimer's disease. J. Alzheimer's Dis. **26**(4), 627–645 (2011)
5. RS Bucks, S Singh, JM Cuerden, GK Wilcock, Analysis of spontaneous, conversational speech in dementia of Alzheimer type evaluation of an objective technique for analysing lexical performance. Aphasiology. **14**(1), 71–91 (2000)
6. ET Prud'hommeaux, B Roark. Extraction of narrative recall patterns for neuropsychological assessment. Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech) (Florence, Italy, 2011), pp. 3021–3024
7. KC Fraser, JA Meltzer, NL Graham, C Leonard, G Hirst, SE Black, E Rochon, Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. Cortex. **55**, 43–60 (2014)
8. B Roark, M Mitchell, JP Hosom, K Hollingshead, J Kaye, Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2081–2090 (2011)
9. G Tosto, M Gasparini, GL Lenzi, G Bruno, Prosodic impairment in Alzheimer's disease: assessment and clinical relevance. J. Neuropsychiat. Clin. Neurosci. **23**(2), 21–23 (2011)
10. DS Knopman, S Weintraub, VS Pankratz, Language and behavior domains enhance the value of the clinical dementia rating scale. Alzheimers Dement. **7**(3), 293–299 (2011)
11. SV Pakhomov, GE Smith, S Marino, A Birnbaum, N Graff-Radford, R Caselli, B Boeve, DS Knopman, A computerized technique to assess language use patterns in patients with frontotemporal dementia. J. Neurolinguistics. **23**(2), 127–144 (2010)
12. V Iliadou, S Kaprinis, Clinical psychoacoustics in Alzheimer's disease central auditory processing disorders and speech deterioration. Ann. Gen. Hosp. Psychiat. **2**(1), 12 (2003)
13. I Hoffmann, D Nemeth, CD Dye, M Pakaski, T Irinyi, J Kalman, Temporal parameters of spontaneous speech in Alzheimer's disease. Int. J. Speech Lang. Pathol. **12**(1), 29–34 (2010)
14. SV Pakhomov, EA Kaiser, DL Boley, SE Marino, DS Knopman, AK Birnbaum, Effects of age and dementia on temporal cycles in spontaneous speech fluency. J. Neurolinguistics. **24**(6), 619–635 (2011)
15. C Thomas, V Keselj, N Cercone, K Rockwood, E Asp, in *Mechatronics and Automation 2005 IEEE International Conference*. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech, vol. 3 (Niagara Falls, Canada, 2005), pp. 1569–15743
16. LEE H, F Gayraud, F Hirsch, M Barkat-Defradas. Speech dysfluencies in normal and pathological aging: a comparison between Alzheimer patients and healthy elderly subjects. the 17th International Congress of Phonetic Sciences (ICPhS) (Hong Kong, China, 2011), pp. 1174–1177
17. DA Snowdon, SJ Kemper, JA Mortimer, LH Greiner, DR Wekstein, WR Markesbery, Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study. JAMA. **275**(7), 528–532 (1996)
18. K Oflazer, S Inkelas, in *Proceedings of the EACL Workshop on Finite State Methods in NLP*. A finite state pronunciation lexicon for Turkish, vol. 82 (Budapest, Hungary, 2003), pp. 900–918
19. K Oflazer, Two-level description of Turkish morphology. Literary Linguist. Comput. **9**(2), 137–148 (1994)
20. v Brunet. Le Vocabulaire De Jean Giraudoux : Structure Et évolution : Statistique Et Informatique Appliquées à L'étude Des Textes à Partir Des

Données Du Trésor De La Langue Française. Le Vocabulaire des grands écrivains français (Genève, Slatkine, 1978). ASIN: B0000E99PZ

21. A Honore, Some simple measures of richness of vocabulary. Assoc. Literary Linguistic Comput. Bull. **7**, 1979
22. D Biber, S Conrad, G Leech, *The Longman student grammar of spoken and written English*, (Harlow: Longman, 2002). ISBN: 0 582 237262
23. E Moore, MA Clements, JW Peifer, L Weisser, Critical analysis of the impact of glottal features in the classification of clinical depression in speech. Biomed. Eng. IEEE Trans. **55**(1), 96–107 (2008)

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com