ELSEVIER

Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1 (2015) 112–124

Alzheimer's & Dementia

Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease

Alexandra König[a,b,*], Aharon Satt[c], Alexander Sorin[c], Ron Hoory[c], Orith Toledo-Ronen[c], Alexandre Derreumaux[a], Valeria Manera[a], Frans Verhey[b], Pauline Aalten[b], Phillipe H. Robert[a,d], Renaud David[a,d]

[a]*Research Unit CoBTeK - Cognition Behaviour Technology, Edmond & Lily Safra Research Center, University of Nice Sophia Antipolis, Nice, France*
[b]*Alzheimer Centre Limburg, Maastricht University Medical Center, School for Mental Health and Neuroscience, Maastricht, The Netherlands*
[c]*Speech Technologies, IBM Research, Haifa, Israel*
[d]*Centre Mémoire de Ressources et de Recherche, CHU de Nice, Nice, France*

Abstract

Background: To evaluate the interest of using automatic speech analyses for the assessment of mild cognitive impairment (MCI) and early-stage Alzheimer's disease (AD).
Methods: Healthy elderly control (HC) subjects and patients with MCI or AD were recorded while performing several short cognitive vocal tasks. The voice recordings were processed, and the first vocal markers were extracted using speech signal processing techniques. Second, the vocal markers were tested to assess their "power" to distinguish among HC, MCI, and AD. The second step included training automatic classifiers for detecting MCI and AD, using machine learning methods and testing the detection accuracy.
Results: The classification accuracy of automatic audio analyses were as follows: between HCs and those with MCI, 79% ± 5%; between HCs and those with AD, 87% ± 3%; and between those with MCI and those with AD, 80% ± 5%, demonstrating its assessment utility.
Conclusion: Automatic speech analyses could be an additional objective assessment tool for elderly with cognitive decline.

Keywords: Dementia; Alzheimer's; Mild cognitive impairment; Speech analyses; Assessment; Information and communication technology (ICT); Audio; Vocal task

1. Introduction

Various types of dementia affect human speech and language [1] and disorders or irregularities in the language domain could be a strong predictor of disease progression [2,3]. Considering this association, reason exists to explore speech analysis as a method for early dementia diagnosis. One avenue we investigated was the analysis of speech using software that takes as input the audio recording from a clinical consultation. Combined with other methods such as video monitoring [4] and actigraphy [5], the speech analysis tool has the potential to become a useful, noninvasive, and simple method for early dementia diagnosis [6]. These technologies enable rapid, accurate, and inexpensive monitoring over time. Noninvasive diagnosis methods will also reduce the burden on the healthcare system and improve the possibility of early dementia detection.

Alzheimer's disease (AD) is diagnosed when it has reached the stage at which cognitive (i.e., episodic memory impairment) and neuropsychiatric symptoms interfere with social functioning or activities of daily living. In addition to the clinical criteria, Dubois et al [7] recently suggested that pathophysiologic biomarker evidence is also needed. The dementia diagnosis is strongly based on clinical judgment, for which appropriate assessment instruments are of vital importance. Providing reliable additional methods to

*Corresponding author. Tel.: +33-65-202-1156; Fax: +33-49-352-9257.
E-mail address: a.konig@maastrichtuniversity.nl

http://dx.doi.org/10.1016/j.dadm.2014.11.012

assess dementia progression in patients is of high interest, because the cognitive domains other than memory have been increasingly recognized as important outcome measures in clinical practice. Information and communication technology (ICT), in particular, automatic speech analysis, is important, because it enables the capture of patient performance and actions to accurately evaluate patients in real time. Using real life situations and applying less intrusive methods that do not require specialized personnel would also be advantageous.

Speech analyses have already been used in patients with dementia [8–10] and those with Parkinson's disease [11] to find potential vocal markers. Studies have shown that one consistently found language abnormality in early AD is anomia, or impaired word finding [12,13], leading to circumlocution that is evidenced in poor word list generation, in particular, for words in a given semantic category [14]. Patients with AD have difficulty accessing semantic information intentionally, which manifests itself in a manner that appears to reflect a general semantic deterioration [1].

This difficulty can affect the temporal cycles during spontaneous speech production (speech fluency) and therefore can be detectable in the hesitation and uttering of a patient [15]. Additional affected speech characteristics in patients with AD seem to be those related to articulation (speed in language processing) [16], prosody in terms of temporal and acoustic measures, which includes alterations in rhythm (ability to vary pitch level, pitch modulation, reduced or fluctuating rate of language output, frequent word finding pauses, a lack of initiative, and slowness) [17,18], and eventually, in later stages, phonologic fluency [19]. Some of these characteristic alterations could be detected by automatic analysis by extracting speech parameters from patients, for example, by performing cognitive vocal tasks or even recording free speech.

Recently, Ahmed et al [2] tried to identify the features of speech that could be used to examine the longitudinal changes to profile impairment in patients with AD. Their study showed that progressive disruption in language integrity was detectable as early as the prodromal stage. Meilan et al [20] found that voiceless segments explained a significant portion of the variance in the overall scores obtained in the neuropsychological test. López-de-Ipiña et al [21] investigated the potential of applying artificial intelligence algorithms to patients' speech as a method to improve the diagnosis of AD and determine the degree of its severity, with promising results for early diagnosis and classification of patients with AD. Roark et al [22] studied different characteristics of spoken language that were automatically derived, such as pause frequency and duration. They demonstrated statistically significant differences between healthy subjects and subjects with mild cognitive impairment (MCI) for a number of measures. However, in these studies, the expected performance of an end-to-end voice-based dementia assessment system was not clearly demonstrated and was often limited by the sample size, method (i.e., manual transcriptions), and technologies.

The present study aimed to determine the value of automatic analyses of voice recordings during vocal tasks for the early diagnosis of AD. This was done through the Dem@Care project, a substudy of the European Community FP7 program. Within this project, ICTs are used for the assessment of patients with dementia in an ecological setting [23]. Specialized IBM speech researchers worked together in close collaboration with the clinical dementia researchers to develop a process to analyze automated speech recordings. To detect dementia-related characteristics in human voice and speech patterns, a classifier was developed using a support vector machine to analyze the statistical properties of vocal features [24]. The classifier determined which features characterize different states of the disease. The main research objective was to determine whether automated speech and voice pattern analyses increases the accuracy, reliability, and affordability of AD early detection.

The present study analyzed voice recordings collected during cognitive vocal tasks performed by patients with AD, MCI, and healthy elderly controls (HCs). The tasks were used to determine the effectiveness of speech processing technologies to support dementia assessment. The primary objective of the study was to investigate whether a method based on automatic speech analyses can detect differences among AD, MCI, and HC subjects. Furthermore, we sought to determine which speech features are the most sensitive to the deterioration due to the disease. This would allow us to obtain a profile for the different populations to improve the differential diagnosis. The secondary objective was to evaluate which tasks are the most appropriate ones to allow the detection of differences in these speech features.

2. Methods

2.1. Participants

Within the framework of the Dem@care project, speech recordings were conducted at the Memory Clinic in Nice, France. The Nice ethics committee approved the present study. Each participant gave informed consent before the first assessment. Participants aged 65 years or older were recruited through the Memory Clinic located at the geriatric department of the University Hospital.

2.2. Clinical assessment

The general cognitive status was assessed using neuropsychological tests, including the

- Mini-mental state examination (MMSE) [25]
- Five word test [26]
- Frontal assessment battery [27]
- Instrumental activities of daily living scale [28]

Additionally, neuropsychiatric symptoms were assessed using the neuropsychiatric inventory [29]. Apathy was

assessed using the apathy inventory [30] and the apathy diagnostic criteria [31]. These tests were performed for potential postanalysis of the audio recordings investigating the effect of neuropsychiatric symptoms on the voice parameters.

After the clinical assessment, the participants were categorized into 3 groups: HCs, who had complained about subjective memory problems but were diagnosed as cognitively healthy after the clinical consultation; patients with MCI; and patients who had been diagnosed with AD. For the AD group, the diagnosis was determined using the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association criteria [32]. For the MCI group, diagnosis was conducted using the Petersen criteria [33]. Participants were excluded if they had any major audition or language problems or a history of head trauma, loss of consciousness, or psychotic or aberrant motor behavior.

2.3. Recording protocol

Each participant performed four spoken tasks during a regular consultation with a general practitioner while being recorded as a part of an ongoing research protocol. The tasks consisted of a counting backward task, a sentence repeating task, an image description task, and a verbal fluency task (Table 1). These tasks were chosen from the findings of previous studies [24].

The vocal tasks were recorded using an Audio Technica AT2020 USB Condenser Microphone (16-mm diaphragm, cardioid polar pattern, frequency response 20 Hz to 16.000 Hz, Thomann 144 dB maximum sound pressure level) that was placed on a stand 10 cm from the participant. Each task was recorded entirely to extract specific vocal features, including pause length, verbal reaction time, and amount of silence. After recoding, the vocal features were extracted from each spoken task using both the open software tool Praat [34] and a set of purposefully developed signal processing tools.

2.4. Statistical analysis

Demographic variables are presented as the median and interquartile range. Before analysis, the data were verified for normality, potential outliers, and missing values. Intergroup comparisons for continuous variables were performed using a nonparametric Mann-Whitney U test, because the distribution of the data was not normal. α Error adjustments ($P < .05/3 = .016$) were performed using the Bonferroni correction method. Categorical testing for gender and education was calculated using the Fisher' exact test. All statistical analyses of the demographic and neuropsychological data were computed using SPSS, version 20.0, and are presented in detail in the subsequent sections.

2.5. Vocal features and analysis

After recording, numerous vocal features were extracted from each spoken task. These features presumably covered the task-specific manifestation of dementia in the speech data. Some were similar to features described in previous studies [2,20,35–37] and others were novel. Because of the developing language-independent technology, speech recognition was not included, and only nonverbal features were targeted.

2.5.1. Countdown and picture description

In cognitive task 1 of the protocol, the participants were asked to count backward from "305" to "285" as fluently as they could. In cognitive task 2, free speech was collected by asking the participants to describe a picture. Task 1 was therefore more cognitively demanding.

The tasks were analyzed for the continuity of speech (i.e., longer contiguous voice segments and shorter silent segments were sought as speech features for both the countdown and the picture description tasks). We expected greater continuity for HCs, lower continuity for those with MCI, and the lowest continuity for those with AD. Some speech features were based on techniques reported in previous studies [2,35–38], and others were new.

We derived continuity features from the length (duration) of the contiguous voice and silent segments and from the length of contiguous periodic and aperiodic segments. Voice or silence was detected using a voice activity detection algorithm, which is based on the energy envelop (intensity) of the recorded speech signal, as calculated using the Praat software [34]. Periodic versus aperiodic segments were detected using pitch contour (periodicity), calculated using the Praat software

Table 1
Vocal tasks of the protocol

No.	Task	Description
1	Countdown	Count backward 1 by 1 from 305 to 285 without making a mistake
3	Picture description	Look at a picture and describe it as detailed as you can in 1 minute
2	Sentence repeating	Repeat 10 short sentences after the clinician (1 at a time); the first 3 are "La montagne est enneigée en ce mois de mars" "Le chien a fait une longue promenade ce matin" "Le Schtroumpf grognon est très content aujourd'hui"
4	Semantic fluency (animals)	Name as many animals as you can think of as quickly as possible (1-minute time limit); this semantic fluency test is widely used in neuropsychological assessments to evaluate frontal lobe functions

[34]. Figure 1 shows the voice and silence segments and the periodic and aperiodic segments of a typical spoken task recording.

From that analysis we obtained four data types for each recording, reflecting each cognitive task (countdown or picture description):

- Voice segment length (in seconds)
- Silence segment length (in seconds)
- Periodic segment length (in seconds)
- Aperiodic segment length (in seconds)

For each set of the four data types (the set for the countdown and the set for the picture description), we calculated several vocal features:

- The mean of the durations (we expected longer durations of the voice and periodic segment lengths and shorter durations for the silence and aperiodic segment lengths): The comparison groups were the HC participants versus those with MCI and those with MCI versus those with AD
- The ratio mean of the durations (defined as the mean voice duration/mean silence duration, mean silence duration/mean voice duration, mean periodic duration/mean aperiodic duration, and mean aperiodic duration/mean periodic duration, for the four data types [voice, silence, periodic, and aperiodic])
- The median of the durations and the ratio median of the durations (the same as for the mean and ratio mean but computing for the median instead)
- The standard deviation of the durations and the ratio standard deviation of the durations (similar to the mean and ratio mean)
- The sum of the durations, ratio sum of the durations (similar to the mean and ratio mean)
- Segment count

These features were derived from techniques introduced in a previous study [24], other than the ratio features, which are novel to the field of speech analysis.

2.5.2. *Sentence repeating*

In cognitive task 3, the participants had to repeat a sequence of 10 sentences spoken by the clinician one at a time while being recorded. First, speaker separation was performed on the entire recording to detect the boundary points of the individual sentences. Next, a standardized signal processing technique (dynamic time warping) was used to evaluate the time alignment—calculating the alignment curve between pairs of corresponding waveforms (the sentence uttered by the clinician and the corresponding sentence repeated by the participant). Figure 2 shows the time alignment for two different cases. On the left side of Figure 2, a "successful" sentence repeating case is shown. The blue curve indicates a quite smooth and regular time alignment between the clinician's signal and the participant's repeated signal. This smooth and regular curve demonstrates that the repeated sentence followed the clinician's sentence closely. The linear and second-order approximations to the time-alignment curve (green line and red curve, respectively) closely match the time-alignment curve. The close matching to the line and a smooth second-order curve are measures of the time alignment "regularity." We defined the first- and second-order irregularity measures as the squared error between the time alignment (blue) curve and its linear and second-order approximations (green and red), respectively.

In contrast, the right side of Figure 2 shows the case of a poorly repeated sentence. A good match between the clinician's signal and the participant's repeated signal was only partially found along parts of the sentence. The nearly horizontal part of the blue curve represents an "insertion" by the patient: it is a segment of the participant's speech for which no alignment was found with any part of the clinician's uttered sentence. The nearly vertical part of the blue curve represents a "deletion" by the patient: it is a segment of the clinician's speech for which no alignment was found with any part in the patient's repeated sentence. Also, on the right side, the time alignment curve (blue) is highly irregular,

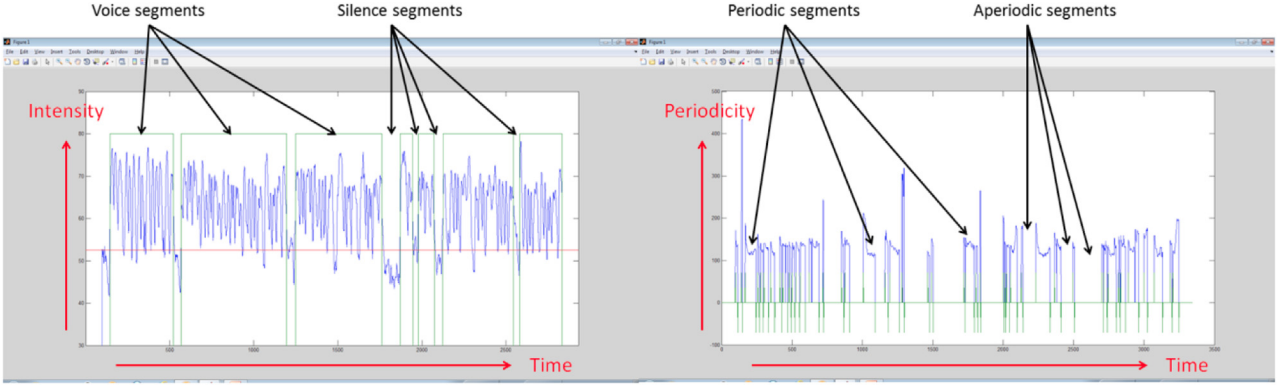

Fig. 1. Voice versus silence segments and periodic versus aperiodic segments of a typical spoken task recording. The horizontal axis designates time frames of 10 ms; the vertical axis on the left designates the signal intensity and that on the right designates the signal periodicity. Voice versus silence and periodic versus aperiodic were determined from the smoothed intensity and periodicity contours, respectively.

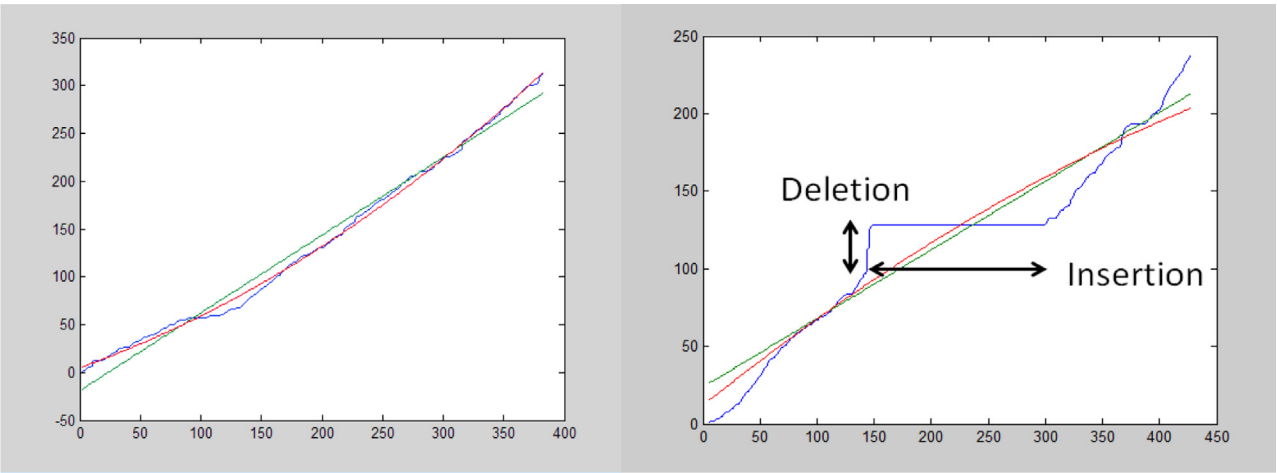

Fig. 2. Time alignment between the clinician's sentence and the participant's repeated sentence. The horizontal axis designates the time of the participant's signal (in 10-ms frames); the vertical axis, the time of the clinician's signal (in 10-ms frames). The blue curve shows the "best" match (alignment) between the two signals; the green line and red curve, the best linear and second-order approximations of the blue curve, respectively.

because it differs significantly from its linear and second-order approximations (green and red, respectively).

We expected fewer insertions, fewer deletions, and less irregularity for the HC participants compared with those with MCI and between those with MCI and those with AD.

We defined the following vocal measures for each pair of corresponding sentences:

- Vocal reaction time (in seconds)
- Relative length (patient sentence duration/clinician sentence duration)
- Amount of silence (0 to 1 continuous scale)
- Amount of insertions (0 to 1 scale)
- Amount of deletions (0 to 1 scale)
- Irregularity—first order (arbitrary units)
- Irregularity—second order (arbitrary units)

We calculated the vocal features of the entire sentence repeating task for each patient to compute the mean and standard deviations of the vocal measures across the different sentence pairs. These features and the use of dynamic time warping to derive the vocal features from spoken cognitive tasks are also novel to the field of speech analysis.

2.5.3. *Semantic fluency*

In cognitive task 4, the participants were asked to name as many animals as they could within 1 minute. This semantic verbal fluency test is widely used in neuropsychological assessments to evaluate frontal lobe functions [39].

Figure 3 shows the positions (in time) of the individual words detected from a sample of a participant's recording of task 4. The word positions were estimated from the signal's intensity and periodicity using a peak detector. The signal intensity information was used to locate the peaks, and the periodicity information was used to reject the irrelevant peaks. The intensity and periodicity were calculated using the Praat software [34]. The vocal features for the semantic fluency task were defined as follows: the distances in time of the second, third, fourth,… and until the ninth detected word positions from the first detected word position.

2.5.4. *Classification procedure*

Before running a classifier, the feature selection procedure was implemented. The purpose of the feature selection procedure was to select the most meaningful vocal features and discard the "noisy" features that contribute less to the classification accuracy and might, in fact, reduce the accuracy if used. This procedure was tested to verify the outcomes.

The feature selection techniques, known as wrapper or embedded methods [33], were found to perform poorly in terms of classification accuracy. This resulted from the limited size of the collected data; hence, the sparseness of the training feature vectors. The filter approach using the Mann-Whitney *U* test was found to perform well using our data.

Three different classification scenarios were evaluated, covering the three pairwise combinations of the three groups (HC, MCI and AD):

- HC versus AD: detecting AD from the mixed HC and AD population
- HC versus MCI: detecting MCI from the mixed HC and MCI population
- MCI versus AD: detecting AD from the mixed MCI and AD population

For each classification scenario, an optimal subset of vocal features was selected. The *P* value of the Mann-Whitney *U* test was calculated for each vocal feature to estimate its "value" for distinguishing between the two classes associated with the scenario. The vocal features with a *P* value less than the threshold were selected for classification, and the remaining features were ignored.

All the meaningful vocal features exhibited a property in which the feature values in one group tended to be greater than those in the other group. For example, the mean silence

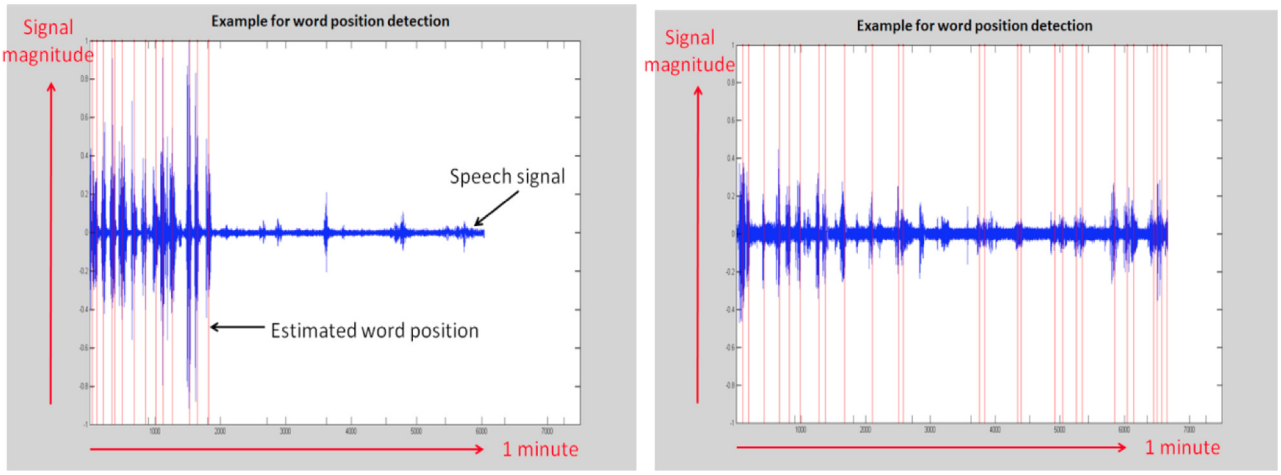

Fig. 3. The time positions of the individual words along a 1-minute recording. (Left) View of recording of healthy elderly control, demonstrating a faster rate of uttering words (assumed to be animal names) at least at the beginning of the task. (Right) View of a recording of a patient with Alzheimer's disease, demonstrating a slower rate of uttering words at the beginning of the task.

segment lengths tended to be smaller for the HC group than for the MCI group. We used the Mann-Whitney *U* statistical test, because it facilitated selecting the features when this property was present. The *P* value selection thresholds, for the three scenarios, were chosen to retain about 22 to 23 vocal features in each scenario and to ignore the rest; this quantity was found to yield good (low) classification error. Figure 4 demonstrates the different distributions of the mean silence segment lengths across the three groups: HC, MCI, and AD.

After feature selection, classification accuracy was evaluated using the support vector machine classifier and random subsampling based cross-validation. We report the classification accuracy in terms of the equal error rate (EER), which is the point at which the rate of type I error (α error rate, false alarm rate) equals the rate of type II error (β error rate, misdetection rate). For each of the three classification scenarios, the following procedure was implemented:

1. We randomly divided the entire data set (in the form of vocal feature vectors, containing the selected features) into test/train subsets
2. Applied regularization to the training set
3. Trained a support vector machine classifier using the regularized train set
4. Normalized the (original, not the regularized) test set according to the parameters derived from the training set
5. Ran the normalized test data through the classifier to evaluate the EER for the current random selection of test versus the training sets
6. Repeated steps 1 through 5 with different random selections of the test versus training sets
7. Calculated the mean and standard error (SE) of the EER and divided them by the EER values that corresponded to the different random selections

The training set regularization in step 2 helped to remove the outliers and increase the classification accuracy. Steps 1 to 5 were repeated 300 times to obtain stable results. We evaluated the results in terms of the EER, which is the point at which the false alarm rate equals the misdetection rate. The EER is equivalent to the point of equal specificity-sensitivity (specificity-sensitivity = 1 − EER/100).

3. Results

3.1. Participant characteristics

Because the distribution of the data was nonparametric, the results are reported as the median and interquartile range. The characteristics of the HC group (n = 15, age 72 years, interquartile range 60–79; MMSE 29, interquartile range 29–30), MCI group (n = 23, age 73 years, interquartile range 67–79; MMSE 26, interquartile range 25–27), and AD group (n = 26, age 80 years, interquartile range 71.75–86; MMSE 19, interquartile range 16.75–21.25) are presented in Table 2. Categorical testing using Fisher's exact test showed no significant differences in education level among the three groups ($P < .05$). However, if using $P < .1$, significant differences in the education levels were found between the HC and AD groups ($P = .062$) and MCI and AD groups ($P = .059$). Furthermore, after α error adjustments ($P < .05/3 = .016$) using the Bonferroni correction method, the HC, MCI, and AD groups did not significantly differ in age. The three groups had significantly different scores on the MMSE, instrumental activities of daily living scale, verbal fluency, neuropsychiatric inventory, and apathy inventory ($P < .05$). However, the frontal assessment battery score just differed between the MCI and AD groups and the HC and AD groups. The same results were obtained using the five word learning test. Using the apathy diagnostic criteria [31], a total of 20 participants had apathy (5 in the MCI and 15 in the AD group).

The results in Table 3 were obtained using the cross-validation procedure as described in the Methods section

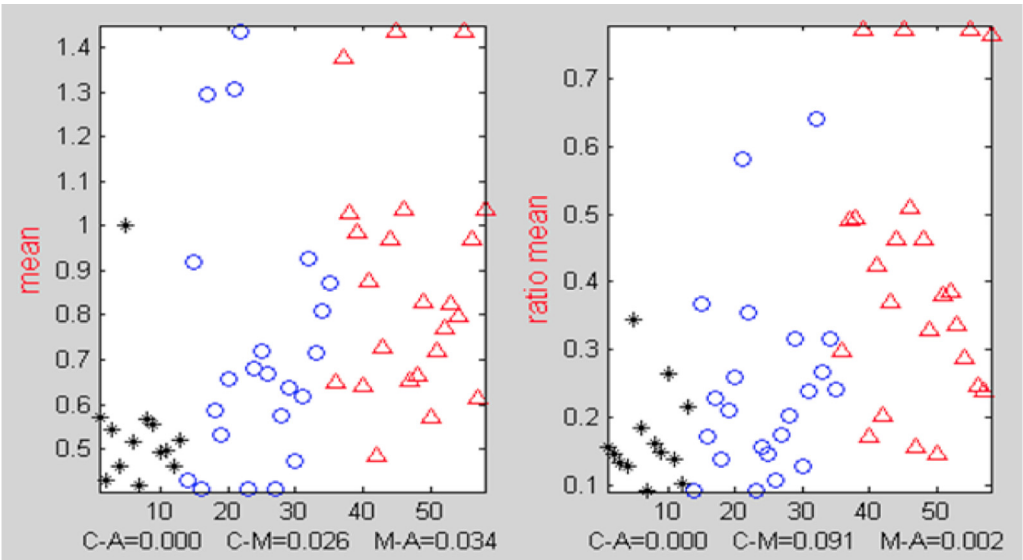

Fig. 4. Distributions and *P* values from Mann-Whitney *U* tests for silence durations. Horizontal axis designates the participant index. Black asterisks indicate healthy elderly controls; blue circles, those with mild cognitive impairment; and red triangles, those with Alzheimer's disease. The values for each class tended to be higher (or lower) than those in another class. Also shown are the *P* values for the three classification scenarios. The ratio mean (right) helped in distinguishing between those with mild cognitive impairment and Alzheimer's disease better than the plain arithmetic mean (left).

(section 2). These results reflect the average and SE of the EER values from 300 random selections of the test and train sets.

3.2. *Vocal feature selection and analysis*

Comparing the HC and MCI groups, 23 features were selected (14 from the countdown task and 9 from the picture description task). Comparing the MCI and AD groups, 23 features were also selected (12 from the countdown task, 5 from the picture description task, and 6 from the verbal fluency task). Comparing the HC and AD groups, 22 features were selected (all from the countdown task). The selection *P* value threshold, per classification scenario, was optimized to yield the highest average classification accuracy after

Table 2
Characteristics and comparisons for HC, MCI, and AD groups

Variable	All subjects (n = 64)	HC (n = 15)	MCI (n = 23)	AD (n = 26)
Gender				
Female	34	9	12	13
Male	30	6	11	13
Age (y)	76 (70–82)	72 (60–79)	73 (67–79)	80 (71.75–86)
Education category				
Primary	18/64*,†	2/15	6/23	10/26
Secondary	19/64*,†	4/15	4/23	11/26
College	14/64*,†	4/15	7/23	3/26
University	13/64*,†	5/15	6/23	2/26
MMSE	25 (19.25–28)‡,§,¶	29 (29–30)	26 (25–27)	19 (16.75–21.25)
FAB	15 (12–17)§,¶	17 (16–18)	15.5 (14.75–17)	11 (9–13.75)
IADL	4 (2–4)§,¶	4 (4–4)	4 (3–4)	2 (1–3)
5 Word test	9 (7–10)§,¶	10 (10–10)	9 (9–10)	7 (4.25–8)
Verbal fluency	13 (8.75–18)‡,§,¶	22.5 (17.75–25)	14 (11–14)	8.5 (6.75–11)
NPI total	3 (1–8)‡,§,¶	0 (0–1.25)	2 (1–6)	8 (4–16)
Apathy diagnostic	20/64	0/15	5/23	15/26
Apathy inventory	2 (2–4)‡,§¶	0 (0–0)	2 (0–3)	4 (2–6)

Abbreviations: AD, Alzheimer's disease; FAB, frontal assessment battery; HC, healthy elderly control; IADL, instrumental activities of daily living questionnaire; MCI, mild cognitive impairment; MMSE, mini-mental state examination; NPI, neuropsychiatric inventory.

NOTE. All values presented as median and interquartile range or number of subjects. Group comparisons were performed using the Mann-Whitney *U* test ($P < .05$). Categorical testing for education was analyzed using Fischer's exact test.

*$P < .1$ for MCI versus AD.
†$P < .1$ for HC versus AD.
‡$P < .05$ for HC versus MCI.
§$P < .05$ for MCI versus AD.
¶$P < .05$ for HC versus AD.

cross-validation. The sentence repeating task was inferior than the other tasks in contributing to the classification accuracy. Thus, no features from that task were selected.

3.2.1. Countdown and picture description

The optimal features for both the countdown and picture description tasks were those that reflected speech continuity, showing longer contiguous voice segments and shorter silence segments and longer contiguous periodic segments and shorter aperiodic segments. According to the cognitive states, greater continuity would be expected for HCs, lower continuity for those with MCI, and the lowest for those with AD. Only a small subset of the continuity-reflecting vocal features significantly contributed to the classification accuracy. The ratio features were found to help, emphasizing the separation among the different groups (HC, MCI, and AD), more than many other features.

3.2.2. Sentence repeating

The most relevant features we found were determined by comparing the pair of waveforms representing the sentence uttered by the clinician and the sentence repeated by the participant. The vocal reaction time was of little benefit for patient classification into the HC, MCI, and AD groups. Other features were more relevant. Although these vocal features were not selected using the Mann-Whitney *U* test, they can be useful when redesigning the recordings of the task. These features were powerful for some of the participants across all three groups.

3.2.3. Semantic fluency

From the many features we examined, the greatest contribution to the classification accuracy was obtained from the positions (in time) of the individual words at the first part of the task. The vocal features determined from the voice and silence segment durations, as described for tasks 1 and 2, were also useful for improving the classification accuracy. This task was particularly useful for distinguishing between those with MCI and those with AD and greatly improved the classification accuracy. Its contribution to separating the HC and MCI groups was not significant.

3.3. Classification procedure

The classification accuracy is presented in Table 3 in terms of the EER, corresponding to the point at which the false alarm rate equaled the misdetection rate. For the classification scenario of HC versus MCI, the EER was 20% ± 5% (SE). This corresponded to an equal specificity-sensitivity result of 0.80 ± 0.05 (SE). For the classification scenario of HC versus AD, the EER was only 13% ± 3% (SE), corresponding to a specificity-sensitivity of 0.87 ± 0.03 (SE). For the classification scenario of MCI versus AD, the EER was 19% ± 5% (SE), corresponding to a specificity-sensitivity of 0.81 ± 0.05 (SE).

Table 3
Classification accuracy results of voice-based analyses

Comparison	Equal error rate (%)	Equal specificity-sensitivity
HC versus MCI	21 ± 5	0.79 ± 0.05
HC versus AD	13 ± 3	0.87 ± 0.03
MCI versus AD	20 ± 5	0.80 ± 0.05

Abbreviations: AD, Alzheimer's disease; HC, healthy elderly control subjects; MCI, mild cognitive impairment.

Figure 5 depicts the receiver operating characteristics curves—the α-error rate (false alarm rate) against the β-error rate (misdetection rate)—for the three classification scenarios. The blue curves show the individual receiver operating characteristic curves for the different test/train set selections during the cross-validation procedure. The red curve shows the average receiver operating characteristic curve, and the point of the EER is highlighted with a red circle.

These results were further investigated statistically to evaluate whether they would generalize to new unseen data with the same statistical properties. The classifier's generalization error consists of three components [40]: (1) the classifier bias (reported in Table 3 as the mean EER); (2) the classifier variance (reported in Table 3 as the SE of the EER), and (3) the classifier generalization measure, which relates to the dependency between the test and train data across different random selections of test/train sets. The third component was very small compared with the mean EER and its SE. This implies that the reported classifier performance would generalize to a new unseen data set of the same statistical type.

4. Discussion

Speech, as the main channel of human communication, has great potential to monitor people with dementia, because speech and language characteristics could represent behavioral markers of dementia variants [41]. The present study has demonstrated that speech processing technology could be a valuable supportive method for clinicians responding to the need for additional objective and automated tools to enable assessment of very early-stage dementia. High classification accuracy, ≤81%, has been achieved, which can possibly be generalized to new unseen data with the same statistical properties. Assessing the voice and unvoiced segments tended to show the best discrimination results between the groups. However, certain features, namely the ratio features, were more informative than were the others for classification, depending on the vocal task.

The extracted vocal features that were significant to this classification were novel in the context of a dementia assessment from speech, such as the comparison of the sentence uttered by the clinician to that uttered by the patient. These were the ratio features in the countdown and picture description tasks, all the features were based on the time alignment in sentence repeating and the word location detection in semantic fluency.

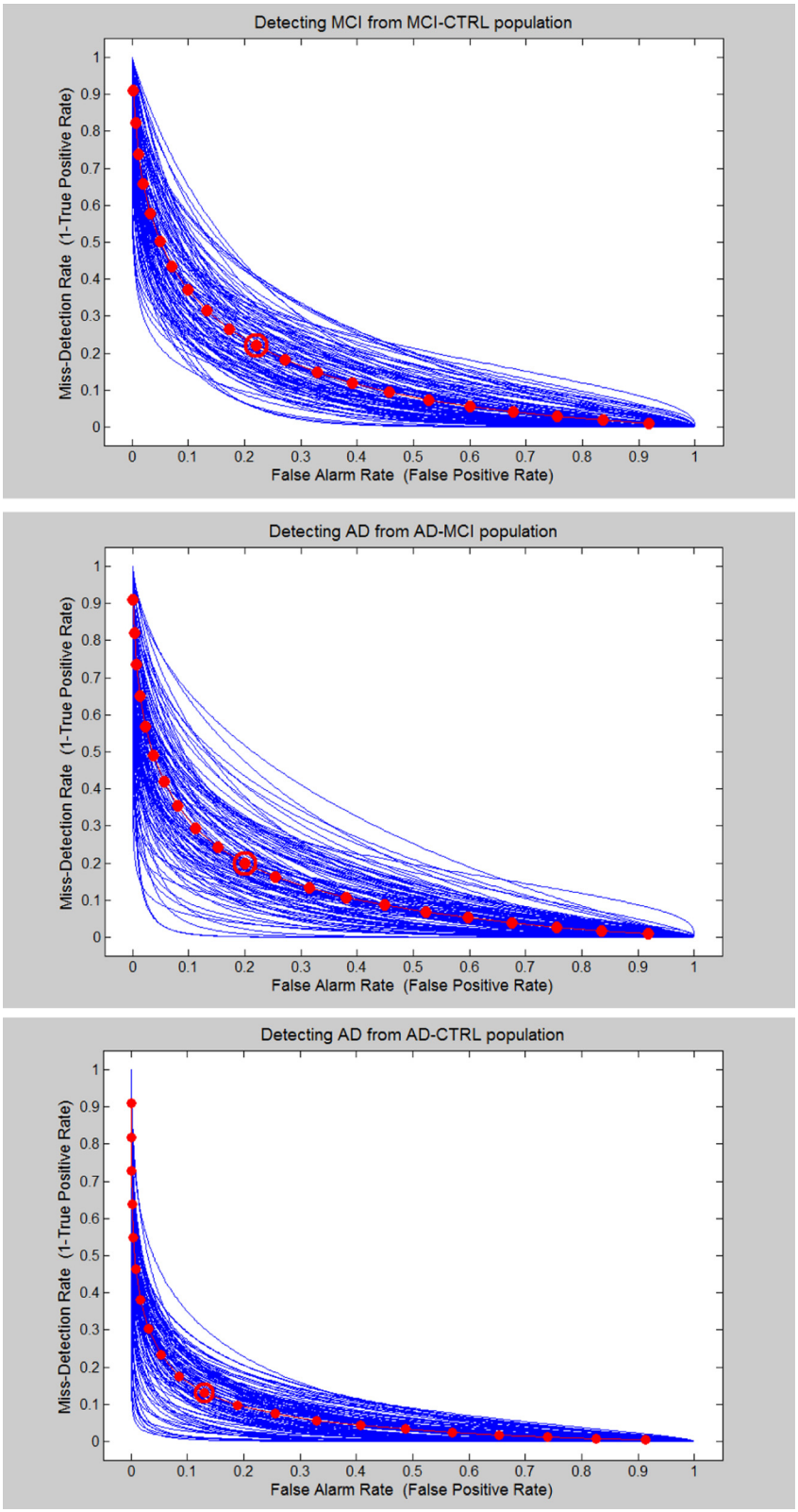

Fig. 5. Plots of the false alarm error probability (horizontal axis) versus the misdetection error probability (vertical axis), which was 1 minus the standard receiver operating characteristic curve.

Generally, the chosen vocal tasks, such as the countdown task, were appropriate and cognitively challenging. This provided the automatic speech processing technology with sufficient information to classify the participants into their diagnostic groups with precise accuracy. Good vocal tasks were optimized to find a balance between being challenging enough to detect the cognitive decline in those with early AD and MCI but not so challenging the HC subjects would fail, masking the differences among the populations. The countdown task is an example of such a balance. Verbal fluency represented a good balance for distinguishing between AD and MCI, but not between MCI and HC. Finally, the sentence repeating task requires reworking to provide that balance, perhaps by using a different spoken content to become a significant contributor to the classification process. These observations demonstrate that the applicability of the cognitive tasks is a main factor for the speech detection software to function effectively.

The present study analyzed the differences in vocal features shown by patients with AD compared with those with MCI and HCs to determine whether these differences represent the characteristics of disease progression. The diagnostic utility of speech measures has been previously demonstrated by Canning et al [9], who reported 88% sensitivity and 96% specificity for AD/HC discrimination for the category fluency. However, the results of the present study have illustrated the ability of automatic speech analyses of recorded vocal tasks to discriminate among HC, MCI, and AD groups.

The participants in each diagnostic group performed differently on each vocal task. These differences in vocal features were not perceptible to the ear of a clinician most of the time; however, they were detectable by the developed speech analysis algorithm. Our results are in line with the findings from Gayraud et al [42], who showed that patients with AD differ from controls in their process of discourse production, displaying more frequent silent pauses outside the syntactic boundaries, which might be a marker of planning difficulties. Meilan et al [20] found that the increase in voiceless segments in a patient's speech is a sign that explains more than 34% of the variance in scores obtained for a specific language and memory test. Thus, this increase could be related to AD and to some extent to the cognitive impairments associated with it. Similar to our results, they also found that the speech of patients with AD seemed characterized by a greater proportion and number of voice breaks [38].

Furthermore, Roark et al [37] demonstrated that using multiple complementary spoken language measures can help in the early detection of MCI, underlining that effective automatic vocal feature extraction of audio recordings is possible such that significant differences in the feature mean values between HC and MCI can be obtained. Similarly, López-de-Ipiña et al [21] and Singh and Cuerden [43] reported new approaches for dementia evaluation using automatic speech analysis. Singh and Cuerden [43] reported that the mean duration of pauses, standardized phonation time, and verbal rates are useful features for discriminating between HC and MCI.

The results of the present study could have several explanations. As reported in the review by Taler and Phillips [44], the commonly found language deficits in those with MCI and AD, shown clinically by alterations in speech production, could possibly be generated by deterioration of semantic knowledge. Certain studies have demonstrated that the pause rate in reading and spontaneous speech correlates with cognitive impairment in elderly individuals [35,37]. The hypothesis is that the number and duration of pauses indicate the cognitive load experienced by the person trying to continue the logical train of thought. D'Arcy [35] measured the pause rate on clinical and telephone speech recordings using very simplistic methods for pauses and breathing detection. These results imply that the more cognitively demanding the task, the more difficult the speech production for the patient reflected in longer pause and breathing rates, which seems in line with our results.

The symptoms that seem to define the language of patients with AD could stem from the presence of atrophy in the medial temporal lobe that might be present from the prodromal phase of the disease [15]. Hence, abnormalities in the expression of vocal features such as voiced/unvoiced segment frequency, periodicity, amplitude, and so forth might represent the direct characteristic consequences of such neurophysiologic changes. These prosodic variables can provide insight about the cognitive processes such as language planning, speech production and word naming, the access to lexicosematic memory, and the structural organization of semantic memory [15].

Early detection of these potential vocal markers could allow earlier treatment interventions that will possibly modify the progression of the disease before extensive and irreversible brain damage occurs. The existence of simple voice analysis programs could facilitate the oral language assessment in the specific parameters sought, even with specific technology for doing so. Thus, we would have to define the briefest language stimuli that best bring to light the variables that discriminate phonologic change, such as the counting down task.

One major advantage of the use of such techniques is the possibility to assess a patient's state in an unobtrusive, less stressful, and more natural method. For example, by recording the patient during a regular consultation or in an open discussion (e.g., during a telephone conversation), the use can be imagined in many different contexts and situations, possibly with a direct visualized output to provide immediate feedback to the patient.

To date, the most commonly used outcome measure in clinical AD trials has been the performance on cognitive tests. Nevertheless, it might be debatable whether commonly used screening tools are sensitive enough to detect small subtle changes, either deterioration or amelioration, in a patient's state [45] within a shorter period. As dementia research has progressed, the findings have demonstrated

that the disease does not just affect cognition, but also other functions, such as motor [46], vision [47], and speech [38], possibly even from the early stages. Thus, the results of the present study indicate we should consider other assessment methods in addition to the currently used methods. This would enable clinicians to cover the full spectrum of dementia when evaluating a patient, maximizing the detection capacities of sensitive changes in behavior. This could be of particular interest for pharmacologic dementia research and to monitor a patient's disease progression.

The present study was performed in French; however, nonverbal vocal features were assessed. Also, considering that similar results that were obtained from a previous proof-of-concept study [24], performed in Greek, we have concluded that multiple languages can be supported, optionally with the requirement for per-language classifier retraining. In the present study, the word content of speech was intentionally not used to aim for a language-independent solution that would be easily deployable. Tools to support classification and statistical analysis in the case of limited-in-size and sparse data were described. The small sample size represents the major drawback of the present study. Another limitation was the age differences among the groups and the choice of the vocal tasks. It can be argued that counting backward or repeating sentences are not very natural tasks and therefore would be strongly influenced by a patient's stress level. The different tasks also required different levels of cognitive effort. For instance, counting backward is cognitively more demanding than describing a picture; thus, these differences could partly explain the high sensitivity of the voice analyses. Furthermore, we did not recruit HC participants from the general elderly population. Instead, rather we limited the HC group to those who had presented for clinical consultation and had subjective complaints. Although this choice limited the HC population size, it reflects the expected scenario in which our technology is likely to be useful: those already experiencing some (subjective) level of cognitive or functional issues, although less than the level of clinical MCI. Finally, it would be very challenging to recruit younger patients with AD to age-match them with those with MCI, because they represent a very small population using our memory clinic. Finally, a relatively high number of patients with apathy were recruited, which could also have influenced the results. In the follow-on study, this aspect will be analyzed in more depth to investigate the effect of apathy and other relevant factors on the voice analysis results.

In future work, we aim to extend the research scope, collecting data on a wider scale using a newly developed application that ensures a standardized recording scenario and adding new spoken cognitive tasks that are more challenging, such as describing a positive memory. Special attention will be paid to the calibration of the audio measurements used to achieve a meaningful characterization of the status and progress of the person with dementia. We expect the new cognitive tasks to increase the classification accuracy even further.

Acknowledgments

This study was supported by grants from the FP7 Dem@care project (grant 288199), the Innovation Alzheimer Associations, the IBM Speech Laboratory in Haifa, Israel, the Cognition Behaviour Technology Research Unit from the Nice Sophia-Antipolis University (UNS), the Memory Resource and Research Centre Nice team, and by the platform patients of the Nice CHU member of the CIU-S.
Disclosure: The authors report no conflict of interests in this work for the past 5 years.

RESEARCH IN CONTEXT

1. Systematic review: From the findings of previous studies and data reviews, we noted that various types of dementia and MCI manifest as irregularities in human speech and language, even from the very early stages. Thus, they could represent strong predictors for disease presence and progression. However, until now, only a few studies have investigated the potential utility of using automatic speech analysis for the assessment and detection of early-stage AD and MCI.
2. Interpretation: The obtained group classification accuracy of the automatic audio analyses, which were based on vocal features extracted from recorded vocal cognitive tasks, was relatively high at up to 87% ± 3%. This demonstrates the value of such techniques for accurate automatic differentiation among HC, MCI, and AD.
3. Future directions: Additional studies are needed with larger population sizes to improve the classification accuracy and to investigate new and improved vocal tasks, signal processing tools, and pattern recognition tools.

References

[1] Braaten AJ, Parsons TD, McCue R, Sellers A, Burns WJ. Neurocognitive differential diagnosis of dementing diseases: Alzheimer's dementia, vascular dementia, frontotemporal dementia, and major depressive disorder. Int J Neurosci 2006;116:1271–93.
[2] Ahmed S, Haigh AM, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain 2013;136(Pt 12):3727–37.
[3] Forbes KE, Venneri A, Shanks MF. Distinct patterns of spontaneous speech deterioration: An early predictor of Alzheimer's disease. Brain Cogn 2002;48:356–61.
[4] Sacco G, Joumier V, Darmon N, Dechamps A, Derreumaux A, Lee JH, et al. Detection of activities of daily living impairment in Alzheimer's

disease and mild cognitive impairment using information and communication technology. Clin Interv Aging 2012;7:539–49.
[5] Yakhia M, Konig A, van der Flier WM, Friedman L, Robert PH, David R. Actigraphic motor activity in mild cognitive impairment patients carrying out short functional activity tasks: Comparison between mild cognitive impairment with and without depressive symptoms. J Alzheimers Dis 2014;40:869–75.
[6] López-de-Ipina K, Alonso JB, Travieso CM, Sole-Casals J, Egiraun H, Faundez-Zanuy M, et al. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. Sensors (Basel) 2013;13:6730–45.
[7] Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, et al. Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. Lancet Neurol 2014;13:614–29.
[8] Bucks RS, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. Aphasiology 2000;14:71–91.
[9] Canning SJ, Leach L, Stuss D, Ngo L, Black SE. Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. Neurology 2004;62:556–62.
[10] Forbes-McKay KE, Venneri A. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. Neurol Sci 2005;26:243–54.
[11] Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng 2012;59:1264–71.
[12] Barr A, Brandt J. Word-list generation deficits in dementia. J Clin Exp Neuropsychol 1996;18:810–22.
[13] Reilly J, Peelle JE, Antonucci SM, Grossman M. Anomia as a marker of distinct semantic memory impairments in Alzheimer's disease and semantic dementia. Neuropsychology 2011;25:413–26.
[14] Mendez MF, Cummings JL. Dementia: A Clinical Approach. 3rd ed. Boston: Butterworth-Heinemann; 2003.
[15] Hoffmann I, Nemeth D, Dye CD, Pakaski M, Irinyi T, Kalman J. Temporal parameters of spontaneous speech in Alzheimer's disease. Int J Speech Lang Pathol 2010;12:29–34.
[16] Cummings JL, Benson DF. Dementia: A Clinical Approach. 2nd ed. Boston: Butterworth-Heinmann; 1992.
[17] Horley K, Reid A, Burnham D. Emotional prosody perception and production in dementia of the Alzheimer's type. J Speech Lang Hear Res 2010;53:1132–46.
[18] Martinez-Sanchez F, Garcia Meilan JJ, Perez E, Carro J, Arana JM. Expressive prosodic patterns in individuals with Alzheimer's disease. Psicothema 2012;24:16–21.
[19] Henry JD, Crawford JR, Phillips LH. Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. Neuropsychologia 2004;42:1212–22.
[20] Meilan JJ, Martinez-Sanchez F, Carro J, Sanchez JA, Perez E. Acoustic markers associated with impairment in language processing in Alzheimer's disease. Span J Psychol 2012;15:487–94.
[21] López-de-Ipiña KA, Alonso JB, Solé-Casals J, Barroso N, Faundez M, Ecay-Torres M, et al. A new approach for Alzheimer's disease diagnosis based on automatic spontaneous speech analysis and emotional temperature. Ambient Assist Living Home Care 2012;7657:407–14.
[22] Roark B, Mitchell M, Hosom JP, Hollingshead K, Kaye J. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans Audio Speech Lang Process 2011;19:2081–90.
[23] Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support. Available at: http://www.demcare.eu/. Accessed March 20, 2015.
[24] Satt A, Sorin A, Toledo-Ronen O, Barkan O, Kompatsiaris I, Kokonozi A, et al. Evaluation of speech-based protocol for detection of early-stage dementia. In: Bimbot F, Cerisara C, Fougeron C, Gravier G, Lamel L, Pellegrino F, et al (eds): INTERSPEECH. Presented at the 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25–29, 2013; ISCA, 2013:1692–1696.
[25] Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–98.
[26] Robert PH, Schuck S, Dubois B, Lepine JP, Gallarda T, Olie JP, et al. Validation of the Short Cognitive Battery (B2C): Value in screening for Alzheimer's disease and depressive disorders in psychiatric practice. Encephale 2003;29(3 Pt 1):266–72.
[27] Dubois B, Slachevsky A, Litvan I, Pillon B. The FAB: A frontal assessment battery at bedside. Neurology 2000;55:1621–6.
[28] Mathuranath PS, George A, Cherian PJ, Mathew R, Sarma PS. Instrumental activities of daily living scale for dementia screening in elderly people. Int Psychogeriatr 2005;17:461–74.
[29] Cummings JL, Mega MS, Gray K, Rosemberg-Thompson S, Gornbein T. The neuropsychiatric inventory: Comprehensive assessment of psychopathology in dementia. Neurology 1994;44:2308–14.
[30] Robert PH, Clairet S, Benoit M, Koutaich J, Bertogliati C, Tible O, et al. The apathy inventory: Assessment of apathy and awareness in Alzheimer's disease, Parkinson's disease and mild cognitive impairment. Int J Geriatr Psychiatry 2002;17:1099–105.
[31] Mulin E, Leone E, Dujardin K, Delliaux M, Leentjens A, Nobili F, et al. Diagnostic criteria for apathy in clinical practice. Int J Geriatr Psychiatry 2011;26:158–65.
[32] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 1984;34:939–44.
[33] Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: Clinical characterization and outcome. Arch Neurol 1999;56:303–8.
[34] Boersma P and Weenink D. Praat: doing phonetics by computer. Available at: www.fon.hum.uva.nl/praat. Accessed May 15, 2013.
[35] D'Arcy S, Rapcan V, Pénard N, Morris ME, Reilly RB, Robertson IH. Speech as a means of monitoring cognitive function of elderly subjects. In: INTERSPEECH. Presented at the Ninth Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22–26, 2008.
[36] Rapcan V. The use of telephone speech recordings for assessment and monitoring of cognitive function in elderly people. In: INTERSPEECH. Presented at the 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6–10, 2009.
[37] Roark B. Automatically derived spoken language markers for detecting mild cognitive impairment. Presented at the Second International Conference on Technology and Aging (ICTA), Toronto, Canada, 2007.
[38] Meilan JJ, Martinez-Sanchez F, Carro J, Lopez DE, Millian-Morell L, Arana JM. Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? Dement Geriatr Cogn Disord 2014;37:327–34.
[39] Monsch AU, Bondi MW, Butters N, Salmon DP, Katzman R, Thal LJ. Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. Arch Neurol 1992;49:1253–8.
[40] Bengio Y. No unbiased estimator of the variance of K-fold cross-validation. J Machine Learn Res 2004;5:1089–105.
[41] Forbes-McKaya K, Shanksa MF, Venneria A. Profiling spontaneous speech decline in Alzheimer's disease: A longitudinal study. Acta Neuropsychiatr 2013;25:320–7.

[42] Gayraud F, Lee HR, Barkat-Defradas M. Syntactic and lexical context of pauses and hesitations in the discourse of Alzheimer patients and healthy elderly subjects. Clin Linguist Phon 2011; 25:198–209.

[43] Singh SB, Bucks RS, Cuerden J. Evaluation of an objective technique for analysing temporal variables in patients with Alzheimer's spontaneous speech. Aphasiology 2001;15:571–83.

[44] Taler V, Phillips NA. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. J Clin Exp Neuropsychol 2008;30:501–56.

[45] Velayudhan L, Ryu SH, Raczek M, Philpot M, Lindesay J, Critchfield M, et al. Review of brief cognitive tests for patients with suspected dementia. Int Psychogeriatr 2014;26:1–16.

[46] Kuhlmei A, Walther B, Becker T, Muller U, Nikolaus T. Actigraphic daytime activity is reduced in patients with cognitive impairment and apathy. Eur Psychiatry 2013;28:94–7.

[47] Verheij S, Muilwijk D, Pel JJ, van der Cammen TJ, Mattace-Raso FU, van der Steen J. Visuomotor impairment in early-stage Alzheimer's disease: Changes in relative timing of eye and hand movements. J Alzheimers Dis 2012;30:131–43.