Evaluation of Speech-Based Protocol for Detection of Early-Stage Dementia

Aharon Satt [1], *Alexander Sorin* [1], *Orith Toledo-Ronen* [1], *Oren Barkan* [1,2], *Ioannis Kompatsiaris* [3], *Athina Kokonozi* [3], *Magda Tsolaki* [4]

[1] IBM Haifa Research Lab, Israel
[2] School of Computer Science, Tel Aviv University, Israel
[3] Centre of Research & Technology – Hellas, Greece
[4] Aristotle University of Thessaloniki, Greek Association of Alzheimer's Disease and Related Disorders, Greece

{aharonsa, sorin, oritht, orenba}@il.ibm.com, {ikom, athko}@iti.gr, tsolakim1@gmail.com

Abstract

This paper describes a study of a protocol and a system for automatic detection and status tracking of early-stage dementia and Mild Cognitive Impairment (MCI), from speech and voice recordings. The research has been performed in the scope of the EU FP7 Dem@Care project. We describe the speech and voice recording protocol, different families of vocal features as derived from the recorded data, the statistical properties of the vocal features, a classifier based on support vector machine, and the classification results. The vocal features we used detect the manifestation of dementia in the human voice and speech, in three axes: the impact of cognitive deficit and slower brain processing, the impact of certain mood states often observed in dementia, and the impact of impairments of the neuromuscular mechanism of the speech production. Our analysis is based on recordings of over 80 diagnosed subjects; it yields dementia and MCI detection equal-error-rate below 20%, and demonstrates the high value of using speech and voice analysis for automatic screening and status tracking of dementia from the very early stage of MCI.

Index Terms: Dementia, Alzheimer Disease, Mild Cognitive Impairment, MCI, vocal biomarkers, vocal features, SVM, voice based diagnosis

1. Introduction

The increase in average lifespan across the world has been accompanied by an unprecedented upsurge in the occurrence of dementia with high socio-economic costs. Development of automatic solutions for dementia diagnosis and state assessment can provide effective means for dealing with the associated challenges. These solutions can enable rapid, more accurate and cheaper diagnosis and monitoring, reducing burden from the healthcare system and the caregivers and allowing them to better respond to the needs of individuals with dementia.

EU FP7 Dem@Care project [1] aspires to contribute to the timely diagnosis, assessment, maintenance and promotion of self-independence of people with dementia. As a part of the project scope we develop a system for clinical diagnosis/assessment of dementia. Based on the automatic analysis of audio, video and physiological sensors data the system provides to a clinician objective performance measures and a "second opinion" regarding the overall state of the observed subject.

A reduced system based purely on patient's speech audio data could offer a possibility of a remote automatic diagnosis/assessment over the phone using Interactive Voice Response systems or Smartphone applications with subsequent selective referral to clinical examination.

Various types of dementia significantly affect human speech and language [2], [3]. Therefore speech can be considered as a source of vocal biomarkers for dementia assessment. Several studies (see for example [4], [5], [6]) reported correlation between dementia and certain vocal features. However, the expected performance of an end-to-end voice-based dementia assessment system was not clearly demonstrated. Our aim is to help filling this gap based on a set of appropriate vocal features and state-of-the-art machine learning. We investigate automated diagnosis of early Alzheimer disease (AD) and a challenging task of detection of mild cognitive impairment (MCI), which is a precursor of dementia. In this study we intentionally do not use the word content of speech aiming at a language independent easily deployable solution.

Within the framework of Dem@Care project, we conducted speech recordings in the city of Thessaloniki, Greece, in co-operation with the Greek Association of Alzheimer's Disease and Related Disorders. The 89 participants in these recordings were elderly people that can be divided into three groups according to their medical profile: people with the early stage of Alzheimer disease (AD), people with MCI, and a control group of healthy people with similar age range and demographic attributes.

We recorded the participants as they performed several vocal tasks. We used the recorded audio data to extract vocal features and used the vocal features to train and test a classifier. The classifier aims at distinguishing between three states: early AD, MCI and healthy.

The rest of the paper is organized as follows. In Section 2 we describe the recording protocol. In Section 3 we describe the vocal features. In Section 4 we describe the experimental setup, present a statistical analysis of the features merits and the classification results. Finally, in Section 5 we present our conclusions and future work.

2. Recording Protocol

Our analysis is based on the recordings of 89 subjects from the 3 diagnosis groups, with gender and age attributes as specified in Table 1. Each participant being guided by a supervisor performed four spoken tasks presented in Table 2. These tasks were used in some of the previous studies. The audio was recorded by a simple headset microphone worn by the participant and digitized at 22050 Hz sampling rate, 16 bits per sample. Each task was recorded entirely including the supervisor's speech and the participant's speech. Extraction of some of the vocal features described in the next section

depends on the supervisor's speech. Although the supervisor's speech was captured by the microphone at a low SNR level we found it suitable for our purpose. This setup can be enhanced by using an additional microphone worn by the supervisor. In an envisaged phone system for the dementia assessment where the supervisor is replaced by an automatic agent and the user's speech is recorded at a separate channel this matter becomes irrelevant.

Table 1: The participant groups attributes

Group	Persons	Males	Females	Age mean	Age range
Control	19	4	15	67	56-84
MCI	43	12	31	73	52-88
AD	27	3	24	72	54-84

Table 2: The spoken tasks of the protocol

#	Task description
1	Verbally describe a picture while looking at it
2	Look at a picture, then describe it from memory
3	Repeat a short sentence after the supervisor. Done 15 times, each time with a different sentence
4	Pronounce repeatedly, fast and as long as possible the sequence of three syllables "pa-ta-ka", widely used in Diadochokinetic tests.

3. Vocal features

We derived a set of vocal features from the recordings of each spoken task in our protocol. The features presumably cover the task-specific manifestations of dementia in the speech data. Extraction of some of the features involved manual annotation steps. However all the features used can be extracted automatically using state of the art speech analysis techniques. In the next subsections we describe in detail the vocal features for each task.

3.1. Features for Tasks 1 and 2 – describing pictures

In the first two tasks of the protocol, the participants were asked to describe a picture presented while looking at it (Task 1) and from memory (Task 2). We define the same set of vocal features for these two tasks. These tasks solicit free speech while imposing certain cognitive load on a participant. We expect that speech of demented participant is less fluent and regular containing prolonged pauses comparing to the healthy subjects.

We hypothesize that for our needs the degree of speech fluency can be captured based on information containing in the occurrence and duration of speech pauses and voiced and unvoiced segments. Using a voice activity detector (VAD) we segment a speech signal to silence and speech segments. Silence segments longer than 60 milliseconds are considered pauses. Then the speech segments are further segmented to voiced and unvoiced segments using a pitch detector. Based on the above segmentation we computed the features listed in Table 3. While the first twelve features (1.1-1.12, 2.1-2.12) represent global statistics of the segments, the last two features capture the temporal structure of the speech and voiced segments respectively. To calculate the temporal regularity of (speech/voiced) segment durations, we used the sequence of the duration values, and calculated the real cepstrum of the sequence. We used DFT of length 32, hence real cepstrum with 16 coefficients. The sum of the absolute values of the higher order cepstrum coefficients (C8-C16) yields the measure of the temporal regularity of the sequence. In normal voices, the durations of the contiguous speech segments tend to be longer, in the order of a few seconds on average, and more "regular", which typically result in lower values of C8-C16. Conversely, for AD cases, the durations of the contiguous speech segments tend to be shorter, often tenths of a second, and less "regular", which typically result in higher average values of C8-C16.

Table 3: The list of vocal features for Tasks 1 and 2

Feature ID Task 1	Feature ID Task 2	Vocal feature description
1.1	2.1	Total speech duration incl. pauses
1.2	2.2	Median of speech segment durations
1.3	2.3	Stdev of speech segment durations
1.4	2.4	Median of pause durations
1.5	2.5	Stdev of pause durations
1.6	2.6	Median of voiced segment durations
1.7	2.7	Stdev of voiced segment durations
1.8	2.8	Median of unvoiced segment durations
1.9	2.9	Stdev of unvoiced segment durations
1.10	2.10	Percentage of pauses duration in the overall speech
1.11	2.11	Average number of speech segments per second
1.12	2.12	Average number of voiced segments per second
1.13	2.13	Temporal regularity of speech segment durations
1.14	2.14	Temporal regularity of voiced segment durations

3.2. Features for Task 3 – repeating sentences

In Task 3, the participants had to repeat a sequence of sentences spoken by the supervisor one at a time. The entire task including 15 sentences, each spoken by the supervisor and repeated by the participant was recorded at once. For this task we are looking for features that correlate with the brain reaction time, such as the response time, the speech rate and the correctness of the spoken sentences repetition. We expect to observe a later response, a lower speech rate, and more errors in repeating the sentences for dementia patients comparing to the healthy subjects, as a result of cognitive deficit and slower brain processing.

As a preparation step for the feature extraction we determined the end-points of all the reference sentences uttered by the supervisor and the sentences repeated by the participant. Although the end-pointing can be accomplished using VAD and speaker diarization (segmentation) techniques, we performed it manually to exclude the influence of the automatic segmentation errors.

The features extracted based on the above segmentation are listed in Table 4. The relative sentence duration is the duration of the repeated sentence divided by the duration of the corresponding reference sentence (as uttered by the supervisor). The verbal reaction time is the latency between the end of the supervisor's utterance and the start of participant's response.

Direct accuracy assessment of the sentence repetition would require a speech-to-text transcription or at least a forced alignment of the audio with the expected text and confidence scoring. Within the ASR-free approach we try to assess the accuracy indirectly, by measuring a distance between the reference and repeated utterances. To this end we convert the reference and repeated segments to cepstral representation. Then the two segments are aligned in time using the Dynamic Time Warping (DTW) transform, and Euclidean distance between the aligned segments is used.

Table 4: The list of vocal features for Task 3

Feature ID Task 3	Vocal feature description
3.1	Average relative sentence duration
3.2	Stdev of the relative sentence durations
3.3	Average verbal reaction time
3.4	Stdev of the verbal reaction times
3.5	Average DTW-aligned sentence distance

3.3. Features for Task 4 – diadochokinetic test

Repeated uttering of a *token* comprised of syllables "pa-ta-ka" is a common test for deficit in the neuromuscular mechanism of speech production used by speech-language pathologists [7]. To evaluate its value for detecting MCI and early-state AD we mainly look for irregularity in the pronunciation rate and for pronunciation errors. The features extracted from the task data are listed in Table 5.

The first five features (4.1-4.5) were extracted based on a manual annotation of the audio in terms of "expected syllable", "error" and "pause". The same could be done automatically using a phonetic alignment technique although the result is expected to be less accurate.

The last feature (the energy envelope periodicity, 4.5) was extracted automatically. In the case of a good performance of the test, the audio signal is expected to exhibit temporal periodicity where the period cycle corresponds to a single "pa-ta-ka" token. To measure the degree of the signal periodicity we use a temporal envelope of the audio signal in the form of instantaneous energy contour. We extracted the fine energy contour at a high frame update rate of 200 Hz and then down-sampled it to 60 Hz. Then the long-term DFT (we used a 5 seconds long interval usually containing from 10 to 20 tokens) was applied after the DC removal and windowing. This transform yields the *energy modulation spectrum* (EMS) which exhibits more or less harmonic structure depending on the test performance.

On the figures 1a and 1b two EMS examples are presented corresponding respectively to a good performance by a healthy subject and a poor performance by a demented subject. The EMS on the figure 1a exhibits a highly harmonic structure. The first high peak corresponds to the fundamental frequency of the EMS which approximately equals to 1.8 Hz in this example. It means that the subject uttered 1.8 "pa-ta-ka" tokens per second on average. The EMS example of the figure 1b has a chaotic structure indicating that the utterance in this case was irregular.

To estimate the signal regularity within a (5s long) analysis interval we used a harmonic-to-noise ratio measure calculated on the EMS extracted from the interval (HNR-EMS). The entire test performance normally lasts more than 8 seconds. To calculate the energy periodicity feature 4.5 for a given test performance we search for the analysis interval position providing the maximal HNR-EMS value which is considered the final value of the feature.

Table 5: The list of vocal features for Task 4

Feature ID Task 4	Vocal feature description
4.1	Average token duration
4.2	Stdev of the token durations
4.3	Total number of the tokens uttered
4.4	Average number of errors per token
4.5	Average number of pauses per token
4.6	Energy envelope periodicity

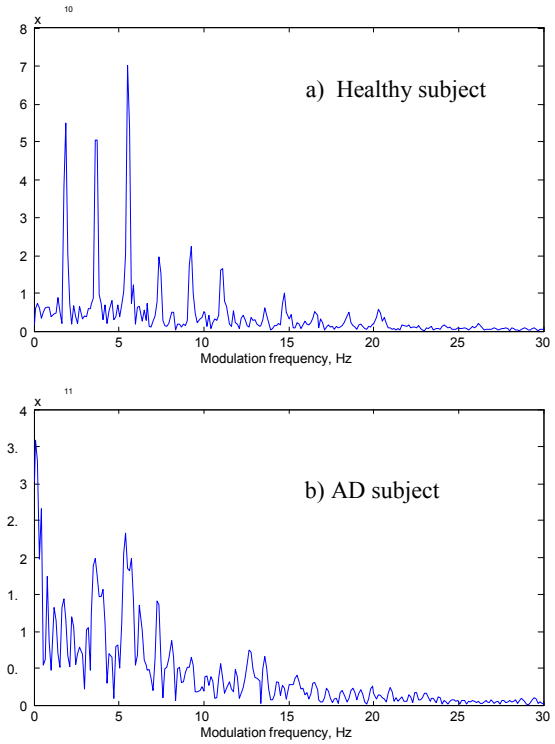

Figure 1: Energy modulation spectrum extracted in the diadochokinetic task: a) healthy subject; b) AD subject.

4. Classification

We studied the capabilities of the voice-based dementia diagnosis on the three binary classification experiments specified in Table 6.

Prior to implementing the classifier, we investigated the statistical significance of the difference between the classes observed in the individual vocal feature samples. For this purpose, we calculated for each feature three single-tailed p-values (P1,P2,P3) corresponding respectively to the three classes specified by the three tests from Table 6. We used single-tailed p-values, as all the presented vocal features are positive. Table 7 lists the p-value results.

Table 6: Three tests

Test	Class 1	Class 2
1	Control	MCI and AD
2	Control	MCI
3	Control	AD

Table 7: p-values of the vocal features for three tests

ID	P1	P2	P3	ID	P1	P2	P3
1.1	0.16	0.17	0.15	2.1	**0.14**	**0.14**	**0.12**
1.2	0.42	0.40	0.43	2.2	0.39	0.41	0.37
1.3	**0.11**	0.15	**0.12**	2.3	0.23	0.22	0.26
1.4	0.38	0.41	0.35	2.4	0.44	0.40	0.44
1.5	0.32	0.34	0.30	2.5	0.40	0.38	0.41
1.6	0.33	0.35	0.29	2.6	0.28	0.31	0.27
1.7	**0.042**	**0.045**	**0.041**	2.7	**0.038**	**0.036**	**0.041**
1.8	0.23	0.23	0.29	2.8	0.22	0.25	0.17
1.9	**0.061**	**0.06**	**0.063**	2.9	**0.059**	**0.061**	**0.041**
1.10	**0.12**	0.15	**0.10**	2.10	**0.095**	**0.11**	**0.07**
1.11	0.19	0.18	0.20	2.11	0.16	0.21	**0.14**
1.12	0.20	0.17	0.25	2.12	0.17	0.15	0.19
1.13	0.30	0.31	0.26	2.13	0.33	0.34	0.30
1.14	**0.047**	**0.044**	**0.061**	2.14	**0.045**	**0.057**	**0.033**
3.1	0.31	0.40	0.33	4.1	**0.12**	0.16	0.15
3.2	0.33	0.37	0.35	4.2	**0.015**	**0.013**	**0.01**
3.3	0.16	**0.15**	0.17	4.3	**0.071**	**0.11**	**0.075**
3.4	0.29	0.35	0.24	4.4	**0.021**	**0.028**	**0.013**
				4.5	**0.009**	**0.004**	**0.011**
				4.6	**0.01**	**0.016**	**0.008**

As it is revealed by the results in Table 7, the class separation capability of the individual features varies significantly. The lower the p-value, the better is the separation capability. **Bold font** designates the p-values below 0.15. Of course, no one single feature is expected to provide a perfect separation but the p-values can lead to meaningful weighting of the individual features within a classification framework. This matter is reserved for our future work. In the present work we used simple heuristic feature selection.

Next, we performed three classification experiments corresponding to the three tests of Table 6 using a subset of the features. The task-specific features that correspond to the audio data produced by each participant, and aggregated into a feature vector, represent the participant. We used a Support Vector Machine (SVM) binary classifier with Gaussian radial basis function kernel. We used the open-source libsvm package version 2.91 [8]. We used feature selection scheme based on the N-lowest p-values, with selection on N optimized separately for each of the three tests to minimize the equal error rate (EER). On average, the lowest EER was obtained using N~20. For each experiment we used a 4-fold cross-validation repeated 10 times with random sampling. Overall 40 SVMs were trained and evaluated per test. As the classification accuracy measure, we computed the EER, which is the point where the miss classification rates are equal for both of the classes. Table 8 summarizes the results of the classification experiments showing the average EER results over the 10 repeated cross validations of each test and its corresponding confidence interval. The results demonstrate the capability of the MCI and early AD detection at the accuracy above 80%.

Table 8: Classification results

Test	Equal error rate	Confidence interval
1	18.0%	6%
2	17.0%	6%
3	15.5%	5%

5. Conclusions and future work

The study reveals that vocal features extracted from the audio recorded in a controlled environment during performance of simple vocal tasks enable quite accurate classification of healthy and demented subjects. The classification accuracy is relatively high even for the challenging task of MCI detection. The results encourage moving towards an automatic system for dementia assessment in a clinical lab or remotely over the phone.

The next work is planned along following directions: matching the reported classification performance with purely automatic feature extraction process; inclusion of additional vocal features, e.g. Voice Onset Time in the diadochokinetic test; feature weighting based on the p-value measures; inclusion of additional spoken tasks, e.g. counting, question answering; dementia status assessment on a continuous scale using a soft decision measure output from SVM or another classifier.

6. Acknowledgements

This work is supported by the Dem@Care FP7 project, partially funded by the EC under contract number 288199.

7. References

[1] EU FP7 Dem@Care website: http://www.demcare.eu/

[2] Appell, J., Kertesz, A., Fisman, M., "A study of language functioning in Alzheimer patients", Brain Language, 17: 73–91, 1982.

[3] Reilly, J., Rodriguez, A., Lamy, M., Neils-Strunjas, J., "Cognition, language, and clinical pathological features of non-Alzheimer‘s dementias: an overview", Journal of Communication Disorders, 43(5): 438-452, 2010.

[4] Roark, B., Hosom, J. P., Mitchell, M., Kaye, J. A., "Automatically Derived Spoken Language Markers for Detecting Mild Cognitive Impairment," Proceedings of 2nd International Conference on Technology and Aging (ICTA), Toronto, Canada, June, 2007

[5] D'Arcy S., Rapcan V., Pénard N., Morris M.E., Reilly R.B., Robertson I.H., "Speech as a Means of Monitoring Cognitive Function of Elderly Subjects", Proceedings of Interspeech 2008, Brisbane, Australia, 26 September 2008.

[6] Rapcan, V., D'Arcy, S., Penard, N., Robertson, I. H., Reilly, R. B., "The Use of Telephone Speech Recordings for Assessment and Monitoring of Cognitive Function in Elderly People", Proceedings of Interspeech 2009, Brighton, UK, September 2009.

[7] Gadesmann, M., Miller, N., "Reliability of speech diadochokintetic test measurement", Int. Journal of Language and Communication Disorders, 43(1): 41-54, 2008.

[8] http://www.csie.ntu.edu.tw/~cjlin/libsvm/