Speech-Based Automatic and Robust Detection of Very Early Dementia

Aharon Satt [1], *Ron Hoory* [1], *Alexandra König* [2,4], *Pauline Aalten* [4], *Philippe H Robert* [3]

[1] IBM Research – Haifa, Israel
[2] University of Nice Sophia Antipolis, France
[3] Centre Mémoire de Ressources et de Recherche, CHU de Nice, Nice, France
[4] Maastricht University Medical Center, Maastricht, Netherlands

{aharonsa, hoory}@il.ibm.com, {a.konig, p.aalten}@maastrichtuniversity.nl,
robert.ph@chu-nice.fr

10.21437/Interspeech.2014-544

Abstract

We provide evidence to the potential use of simple spoken tasks for automatic assessment of very early dementia. Timely detection of dementia is required for effective psychological treatment and to enable patients to participate in new drug therapy research. The technology enables automatic, cheap, remote and wide-scale screening of dementia, typically a costly and complex procedure. It can aid clinicians in the diagnosis of very early dementia, as well as assessing the disease progression.

We describe the spoken tasks, and their respective language-independent vocal feature extraction, followed by classification accuracy evaluation. We use recordings from over 60 persons, diagnosed as healthy-control (CTRL) / mild-cognitive-impairment (MCI) / early-stage-Alzheimer-disease and early-mixed-dementia (AD).

We present a new data regularization technique to overcome data sparseness due to the limited data set size. Next, we present a comprehensive statistical analysis, showing that the suggested classifier generalizes, and revealing the role and the statistical importance of the different spoken tasks and their respective vocal features.

We demonstrate classification accuracy of about 80% for CTRL vs. MCI and MCI vs. AD, and 87% for CTRL vs. AD, all shown to generalize. This provides an evidence for potential use for automatic detection of very early dementia.

Index Terms: Dementia, Alzheimer, Vocal Features, Classification, Regularization, Data Sparseness, ROC Curve, Classifier Bias.

1. Introduction

Dementia, including Alzheimer Disease (AD), represents a major challenge for health care systems within the aging population. Clinicians and researchers seek robust, non-intrusive, simple and cheap tools for assessing the disease severity and progression, from its very early stages including the Mild Cognitive Impairment (MCI) condition which progresses to dementia at high probability. Early detection of dementia and MCI is necessary to optimize the patient's care and the support to caregivers, and to provide better tools for clinical research [1]. Research shows that psychological treatment to AD patients, which begins at the very early stage of the disease, alleviates psychological and behavioral symptoms [2].

Spoken language is the most spontaneous, intuitive and efficient method of communication revealing an individual's cognitive and emotional state. Various types of dementia affect human speech and language; deficits in those domains have proven to be strong predictors for the disease progression [3], [6].

The studies [3], [4], [5], [7], [8], [9] reported correlation between dementia and certain vocal features. The vocal features include non-verbal (acoustic) and verbal (linguistic) metrics. Non-verbal metrics that were shown to predict dementia include speech continuity, based on duration and proportion of silence segments within the speech, and pitch (periodicity) related parameters. Useful verbal metrics, which require automatic speech recognition for their evaluation, include speech richness based on the vocabulary size, and proportions of different parts of speech such as verbs and nouns [10]. However, in these studies, the expected performance of an end-to-end voice-based dementia assessment system was not clearly demonstrated.

In [11], a step towards demonstrating a system for voice-based dementia assessment is taken, by describing and analyzing a test protocol consisting of cognitive spoken tasks, followed by automatic analysis and classification. This proof-of-concept study does not include a comprehensive statistical analysis that would enable prediction of the system performance on large data sets.

This paper presents the speech analysis system that has being developed within the framework of the EU research project Dem@Care [12]. It significantly extends the work in [11], and provides clear evidence for the effectiveness of speech processing technologies in supporting dementia assessment.

The rest of the paper is organized as follows: in section 2 we describe the spoken cognitive tasks; in section 3 we present the vocal features; in section 4 we describe the feature selection method; in section 5 we present, the classification technique and the achieved classification accuracy.

2. The Spoken Task Protocol

Within the framework of the Dem@care project, speech recordings were conducted at the Memory Clinic in Nice, France, located at the Geriatric department of the University Hospital. Participants aged 65 or older were recruited at the Memory Clinic. Following clinical assessment, the participants (referred to as 'patients' for simplicity) were categorized in three groups:

- CTRL – participants that complained about subjective memory problems but were diagnosed as cognitively healthy
- MCI
- AD – patients that were diagnosed as suffering from early stage AD or early stage mixed dementia.

Table 1: patient distribution

Group	Number of patients
CTRL	15
MCI	23
AD	26

Table 2: the spoken cognitive tasks

Task	Description
1. Countdown	Count backwards from 305 down to 285
2. Picture description	Look at a picture and describe it as detailed as you can in one minute
3. Sentence repeating	Repeat ten short sentences after the clinician (one at a time)
4. Semantic fluency (animals)	Name as many animals as you can think of as quickly as possible, in one minute

3. The Vocal Features

We developed specific vocal features for each spoken task. Some are similar to features described in previous publications ([3], [4], [5], [7], [8], [9], [15]), and others are novel. In view of developing language-independent technology we avoided speech recognition, and considered non-linguistic features only.

3.1. Countdown and Picture Description

We found that speech-continuity related features contribute most significantly to reducing the classification error, for both the countdown task and the picture-description task.

We derived continuity features from the lengths (durations) of contiguous voice and silence segments, and from the lengths of periodic and aperiodic segments. Voice vs. silence are detected using simple voice activity detection algorithm, based on the pitch-synchronous energy envelop (intensity) of the recorded speech signal, as calculated using the PRAAT software [13].

Periodic vs. aperiodic segments are detected using the pitch contour (periodicity), calculated using the PRAAT software [13].

We smoothed out the pitch contour to eliminate rapid changes shorter than 30 milliseconds.

From the above analysis we get four **data types** for each recording of a spoken task: **voice segment lengths, silence segment lengths**, **periodic segment lengths** and **aperiodic segment lengths**.

For each data type we calculate several **vocal features**, as follows:

- **The mean of lengths**
- **The ratio mean of lengths**: defined as mean voice length / mean silence length, mean silence length / mean voice length, mean periodic length / mean aperiodic length, and mean aperiodic length / mean periodic length, for the four data types, respectively
- **The median of lengths, the ratio median of lengths**: similarly
- **The standard deviation of lengths, the ratio standard deviation of lengths**: similarly
- **The sum of lengths, the ration sum of lengths**: similarly
- **Segment count**

The 'ratio' features were found to emphasize the separation between the different groups: CTRL, MCI and AD, beyond the other features. These features are novel.

3.2. Sentence Repeating

We first performed speaker separation to detect the boundary points of the individual sentences. Then, we used standard Dynamic Time Warping (DTW) technique for time-alignment, based on Mel-Frequency Cepstral Coefficients (MFCCs), to calculate the alignment curve between pairs of corresponding waveforms. Figure 1 depicts the alignment curves between pairs of waveforms:

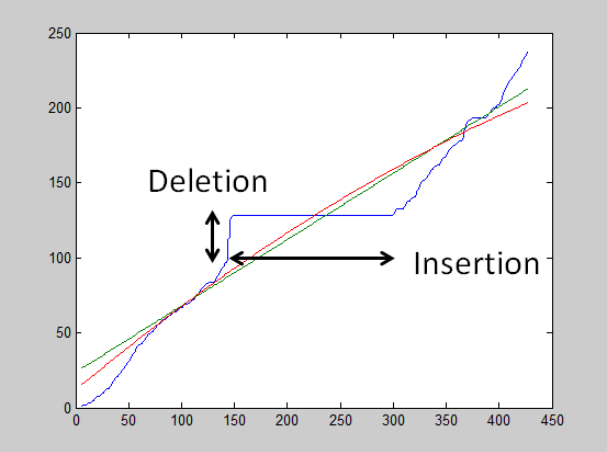

Figure 1: an example for time alignment curve

Figure 1 shows the time alignment between the clinician's uttered sentence (the vertical axis) and the patient's repeated sentence (the horizontal axis). Figure 1 demonstrates two specific points of interest:

- The nearly horizontal part of the blue curve, a segment of the patient's speech for which no alignment is found to the clinician's uttered sentence, represents an 'insertion' by the patient
- The nearly vertical part of the blue curve, a segment of the clinician's speech for which no alignment is found to the patient's repeated sentence, represents a 'deletion' by the patient.

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

- Vocal reaction time (in seconds)
- Relative length (patient's sentence duration / clinician's sentence duration)
- Amount of silence (0 to 1 scale)
- Amount of insertions (0 to 1 scale)

- Amount of deletions (0 to 1 scale)

We calculated the **vocal features** of the entire sentence repeating task for each patient, as the means, medians and standard deviations of the above vocal measures, across the different sentence pairs. The use of time-alignment based vocal features, as well as sentence repeating task, are novel.

3.3. Semantic Fluency

Figure 2 depicts an example for the individual word positions, using a recording of the semantic fluency task.

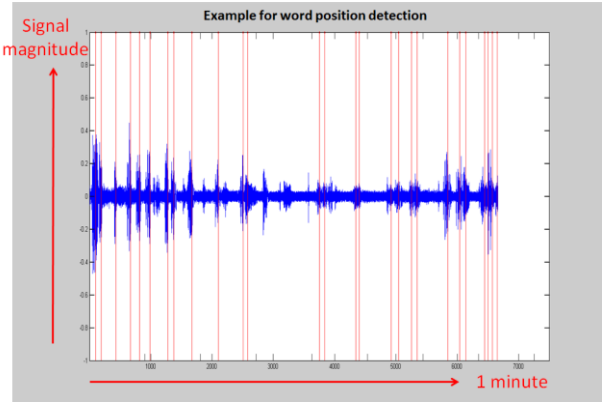

Figure 2: word location detection

Figure 2 shows the estimated individual word positions, based on a peak detector that uses the signal intensity information for locating peaks and the periodicity information to reject irrelevant peaks. The intensity and the periodicity are calculated using the PRAAT software [13].

We found the following vocal features to significantly contribute to the classification accuracy of the semantic fluency task:

- The distances in time of the second, third, fourth,… and until the ninth detected word positions, from the first detected word position

These features, as well as the use of the task, are novel.

4. The Feature Selection

Different feature selection techniques are described in the scientific literature. An overview, with comparative performance study for medical applications, is given in [14]. The feature selection techniques known as wrapper or embedded methods [14], were found to perform poorly in terms of classification accuracy, in our case. This is due to the limited size of the data we collected, hence the sparseness of the training feature vectors. We found the filter approach based on the Mann–Whitney test [14] to perform well based on our data.

We evaluated three classification scenarios:

1. Classification of CTRL vs. MCI
2. Classification of MCI vs. AD
3. Classification of CTRL vs. AD.

For each classification scenario we evaluated the p-values of the different vocal features, using the Mann-Whitney test. We selected the features with low p-values, below a threshold.

Table 3: number of selected vocal features

Task & Scenario	Count down	Picture descript.	Sentence repeating	Fluency
CTRL vs. MCI	14	9		
MCI vs. AD	12	5		6
CTRL vs. AD	22			

The selection p-value threshold, per classification scenario, was optimized to yield the highest average classification accuracy, following cross-validation.

The sentence repeating task was found inferior, comparing with other tasks, in terms of contributing to the classification accuracy, and in fact no features from this task were selected; we believe however, that this task can be revised, using for example different spoken content, to become a significant contributor to the classification process.

5. The Classification

5.1. Training Data Regularization

We tried several common classification techniques, including Naïve Bayes and Support Vector Machine (SVM) with common kernels, and obtained poor results from all of them. Figure 3 provides an explanation.

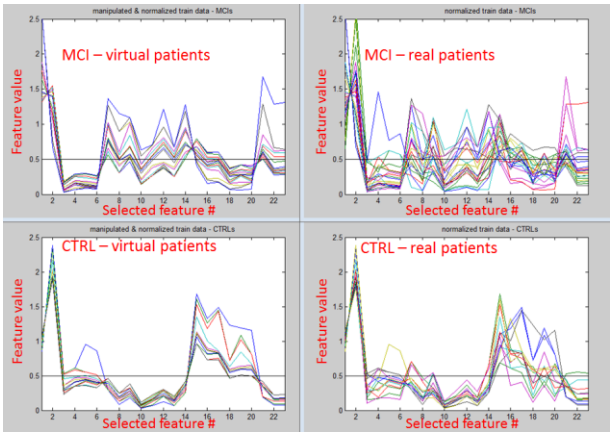

Figure 3: regularization of the training feature vectors

In figure 3, each curve in a different color represents the vocal feature vector of a different patient. The horizontal axis shows the index of the selected vocal feature, and the vertical axis shows the normalized feature value.

On the right hand side of figure 3, the original feature vectors of the patients (referred to as **real patients**) are shown: MCI at the top, CTRL at the bottom. The significant overlapping between the two classes, or the significant number of outliers in each class, is evident. The small number of participants (table 1) results in feature vector sparseness, and lack of sufficient data for statistically training a robust classifier.

To overcome this problem, we implemented a novel data regularization technique, which eliminates outliers, or overlapping, between the two classes. The left hand side of figure of figure 3, the regularized feature vectors (which

correspond to '**virtual patients**') are shown: MCI at the top, CTRL at the bottom. It is clearly seen that the regularized feature vectors, corresponding to the virtual patients, maintain complete separation across the two groups, while preserving the overall structure, per group, of the real patients' feature vectors. This artificial separation performs better than the soft-margin separation of common SVM in our case of "small data" (table 4).
The data regularization is based on two assumptions:

1. The features are statistically independent
2. For each feature, the values tend to be lower for one group and higher for the other.

While the first assumption is in general untrue, leading to potentially suboptimal classifier design, we gain more from overcoming the feature overlapping, than we lose from assuming independence. The second assumption is evident for the selected features, as these features were selected based on low p-values of the Mann-Whitney test.
Following the above assumptions, we reorder the individual vocal features independently, as follows:

- We assign all the worst feature values (in terms of poor cognitive performance) across all the real MCI patients to the first virtual MCI patient; this virtual patient then performs most poorly, across all the cognitive tests
- Next, we assign the next-worst feature values across all the real MCI patients to the second virtual MCI patients
- Finally, we assign the best feature values across all the real MCI patients to the last virtual MCI patient; this virtual patient performs most strongly across all the cognitive tests
- Effectively, we created a set of virtual MCI patients, sorted from worst performer to best performer, across all the selected features
- We assign feature values in an analogous manner to the virtual CTRL patients, creating sorted virtual CTRL patients, from the best performer to the worst performer
- To conclude the process, we trim the feature values of the feature vectors to avoid overlapping; this affects only a small number of virtual patients.

This process results in trimming out the overlapping parts of the vocal features, while preserving the bulk of the feature vector structure.

5.2. The Classifier Description

The classification process, for each classification scenario, consists of the following steps:

1. Selecting the optimal subset of vocal features
2. Randomly dividing the entire data set (in the form of vocal feature vectors, containing the selected features) into train set and test set
3. Applying regularization to the train set and training a "standard" SVM classifier on the regularized train set
4. Normalizing the (original, not regularized) test set according to parameters derived from the train set
5. Running the normalized test data through the classifier to evaluate the Equal Error Rate (EER) for the current random selection of test vs. train sets
6. Repeating steps 2-5 above with different random selections of test vs. train sets
7. Calculating the mean and the standard error of the EER

Table 4 shows the classification accuracy. It is given in terms of Equal Error Rate (EER), corresponding to the point where the false-alarm probability equals the misdetection probability.

Table 4: classification results

Scenario	EER (mean ± standard error) with new regularization	EER (mean ± standard error) using plain SVM without regularization
CTRL vs. MCI	20 ± 5 %	24 ± 6 %
MCI vs. AD	19 ± 5 %	27 ± 6 %
CTRL vs. AD	13 ± 3 %	14 ± 4 %

6. Conclusions and Future Work

This study clearly demonstrates that speech-processing technology, as part of a wider MM technology, can fill the gap in dementia treatment and research: answer the need for objective and automatic tool to support the clinicians' assessment of very early stage dementia. The speech-processing technology provides the answer to the cognitive part of the assessment.

We demonstrated good classification accuracy, which generalizes to new unseen data of the same statistical properties.

The research was carried out in French; however, we used non-verbal vocal features, and considering similar results that were obtained from a previous proof-of-concept study, done in Greek, we conclude that multiple languages can be supported, optionally requiring a per-language classifier retraining.

We described tools to support classification and statistical analysis in cases of limited-in-size and sparse data.

We aim to extend the research scope, collecting data at a wider scale, and adding new spoken cognitive tasks. We expect the new cognitive tasks to increase the classification accuracy further.

7. Acknowledgement

This work is supported by the Dem@Care FP7 project [12], partially funded by the EC under contract number 288199.

8. References

[1] P. Robert at al., 2013, Recommendations for ICT Use in Alzheimer's Disease Assessment: Monaco CTAD Expert Meeting, *Nutr Health Aging, 2013, 17(8): pp. 653-60*

[2] I. Hallikainen at al., 2013, Progression of Alzheimer's disease during a three-year follow-up using the CERAD-NB total score: Kuopio ALSOVA study, *International Psychogeriatrics / Volume 25 / Special Issue 08 / August 2013, pp 1335-1344*

[3] B. Roark at al., 2007, Automatically Derived Spoken Language Markers for Detecting Mild Cognitive Impairment, *Proceedings of 2nd International Conference on Technology* and Aging (ICTA), Toronto, Canada, June, 2007

[4] S. D'Arcy et al., 2008, Speech as a Means of Monitoring Cognitive Function of Elderly Subjects, *Proceedings of Interspeech 2008, Brisbane, Australia, 26 September 2008*

[5] V. Rapcan at al., 2009, The Use of Telephone Speech Recordings for Assessment and Monitoring of Cognitive Function in Elderly People, *Proceedings of Interspeech 2009, Brighton, UK, September 2009*

[6] J. Reilly at al., 2010, Cognition, language, and clinical pathological features of non-Alzheimer's dementias: an overview. *J Commun Disord, 2010. **43**(5): p. 438-52*

[7] S. Ahmed at al., 2013, Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease, *Brain, 2013. **136**(Pt 12): p. 3727-37*

[8] JJ. Meilan at al., 2012, Acoustic markers associated with impairment in language processing in Alzheimer's Disease. *Spanish J Psychology, 2012. **15**(2): p. 487-94*

[9] KA López-de-Ipiña at al., 2012, New approaches for Alzheimer's disease diagnosis based on automatic spontaneous speech analysis and emotional temperatur,. *Ambient Assist. Living Home Care, 2012. **7657**: p. 407–414*

[10] C. Thomas at al., 2005, Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech, *Mechatronics and Automation, 2005 IEEE International Conference (Volume:3)*

[11] A. Satt et al., 2013, Evaluation of Speech-Based Protocol for Detection of Early-Stage Dementia. in *INTERSPEECH. 2013*

[12] http://www.demcare.eu/

[13] http://www.fon.hum.uva.nl/praat/

[14] C. Christin at al., 2013, a Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics, *Molecular Cellular Proteomics. 2013 January; 12(1): 263–276*

[15] J. Grothendieck et al., 2009, Social correlates of turn-taking behavior, *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009*