Longitudinal Monitoring and Detection of Alzheimer’s Type Dementia from Spontaneous Speech Data

Saturnino Luz
Usher Institute of Population Health Sciences & Informatics
The University of Edinburgh
Edinburgh, UK
Email: s.luz@ed.ac.uk

Abstract—A method for detection of Alzheimer’s type dementia though analysis of vocalisation features that can be easily extracted from spontaneous speech is presented. Unlike existing approaches, this method does not rely on transcriptions of the patient’s speech. Tests of the proposed method on a data set of spontaneous speech recordings of Alzheimer’s patients (n=214) and elderly controls (n=184) show that accuracy of 68% can be achieved with a Bayesian classifier operating on features extracted through simple algorithms for voice activity detection and speech rate tracking.

Keywords—Dementia; Alzheimer’s Disease; Medical Informatics; Machine Learning; Speech Processing;

I. INTRODUCTION

Automatic monitoring of physical and cognitive well-being has become a focus of great research interest as well as practical relevance in the area of elderly care. Emerging applications aimed at promoting healthy ageing and improving care have built on advances in sensor technology and machine learning, which provide opportunities for collection and analysis of vast quantities of personal behavioural data. Due to its high prevalence, and the burden it places on carers and society, Alzheimer’s type dementia (ATD) has been the focus of numerous efforts in this area. A promising research direction is the automatic categorisation of patients with ATD based on cognitive markers. This type of categorisation has long been part of diagnostic procedures and of the overall study of the natural history of ATD [1]. Methods of diagnostic assessment of ATD based on cognitive markers include the mini mental state examination (MMSE) and a range of other neuropsychological tests. However, a need for improved methods for earlier detection of ATD symptoms has been identified by dementia researchers [2], and there is growing recognition that technologies that enable finer-grained personal monitoring in daily life might play an important role in the development of such methods.

Recently, there have been attempts to use machine learning for identification of ATD through analysis of the patient’s speech. Language dysfunction is characteristic of ATD, and it is believed that first semantics, then syntax and finally phonology are affected as the disease progresses [3]. Consequently, machine learning approaches have often employed lexical and sometimes syntactic structure as categorisation features [4].

While these new methods have achieved some success, with accuracy rates up to 81%, they often rely on the availability of speech transcripts. However, automatic speech recognition (ASR) for spontaneous speech produced in daily life remains a very challenging problem. We present an alternative approach based on patterns of vocalisations and other paralinguistic speech features [5] which does not depend on ASR. If successful, this approach may help pave the way for fine-grained, longitudinal monitoring of speech in daily life for the purposes of identifying cognitive changes that may be characteristic of ATD.

II. METHODS

We stipulate as a basic requirement that the input data for the proposed method must consist of features that can be either readily entered by the user or easily acquired in natural interactive settings with acceptable accuracy. Therefore we assume a scenario where the user enters personal information (e.g. age, gender) and the subsequent monitoring consists of tracking of low-level paralinguistic features and functionals of the user’s spontaneous speech, such as the timing and duration of vocalisations and pauses [5], [6] speaking rate, and voice quality measures. In this paper we focus on vocalisation events and speech rate. These features are easily extracted through basic signal processing and are reasonably robust to environmental noise and diarisation issues [6].

In order to test the potential usefulness of these features in predicting ATD from speech data, we define a categorisation task of differentiating between speech from participants with ATD (represented as \( C = a \), or \( a \)) and speech from control participants (\( \bar{a} \)). The feature set \( F = \{ F_1, \ldots, F_n \} \) consists of numerical features corresponding to summary statistics (mean, variance, minimum and maximum, entropy) for vocalisation events, speech rate and number of utterances over a discourse event (described below), and the abovementioned nominal features. As the purpose of this work was to assess baseline performance, we chose to use a simple probabilistic model rather than carry out extensive classifier comparisons. We have also not performed any feature selection procedures.
or parameter tuning. Formally, the probability of a patient being diagnosed with ATD is represented as:

$$P(a|F) \propto P(F_1 = v_1, \ldots, F_n = v_n|a)$$

$$= \prod_{i=1}^{n} P(F_i = v_i|a)$$ (1)

where (2) represents the conditional independence assumption, and $P(F_i = v_i|b)$ are modelled through Gaussian kernels, if $F_i$ is a continuous variable, or estimated by maximum likelihood and incorporated to a multinomial model, if $F_i$ is a nominal variable.

This model was tested on the Pitt data set [1], available through the DementiaBank1. Although the Pitt data set consists of speech recordings from different neuropsychological tests (see [1] for details), we use only the spontaneous speech data gathered for the Boston “cookie theft” picture description task. In this task, the subject is shown a picture and asked to describe the scene depicted. The 551 audio recordings of 247 different participants performing this task included recordings from elderly controls (n=242), people with probable Alzheimer’s Disease (n=235), and others (n=74). Data were gathered longitudinally, on a yearly basis. The data set also included timed, manually produced transcripts, which have been used in previous work on ATD prediction. [4]. We ignored these transcripts for the purposes of this experiment. After excluding recordings with missing data, and other conditions (MCI, vascular dementia etc), we used n=214 ATD and n=184 control recordings.

Vocalisations produced by the instructor were removed. The participant’s vocalisations were segmented using a simple silence filter based on the amplitude ratio ($\frac{R_a}{R_n}$) where $R_a$ and $R_n$ stand for the root mean square amplitude of signal and noise respectively) of the speech signal, setting the silence amplitude rate threshold empirically to -25dB and a duration threshold of 1 second, according to standard practice in vocalisation analysis [6]. Speaking rate was estimated based on a syllable nuclei detection algorithm [7].

Ten-fold cross validation was performed, and accuracy scores were computed averaging over the different folds.

III. RESULTS

Figure 1 shows the receiver operating characteristic (ROC) curve for the relation between sensitivity (the proportion of ATD patients correctly identified as such) and the probability of a false alarm (the complement of specificity) in the above described model. This model achieved an overall accuracy of 68%, with $F_1$ scores of 70% for the control class and 64% for the ATD class, as measured against a gold standard diagnostic established through a combination of neuropsychological and neurologic tests [1].

These results are 36% higher than the baseline, showing that simple paralinguistic features extracted from noisy audio files have predictive value for ATD diagnosis.

1http://www.talkbank.org/DementiaBank/

IV. CONCLUSION

While the results reported here are still some way from the accuracy levels required for reliable monitoring and prediction of ATD, they are nevertheless comparable to results reported for approaches that employ far richer lexical and syntactic features which cannot at present be realistically acquired in naturalistic settings. In fact, our results surpass the 58.5% accuracy reported in a recent study [4] when feature selection is not performed on their original set of 370 linguistic and acoustic features, even though by exploring the feature space accuracy up to 81% can be achieved.

Our results are promising in that they point towards possible practical implementations of systems for monitoring of cognitive health in daily life, similarly to the way aspects of physical health are already being monitored by a number of portable consumer devices.

REFERENCES


