Emotion Detection for Dementia Patients Monitoring

Orith Toledo-Ronen and Alexander Sorin

IBM Research – Haifa, Haifa University Mount Carmel, Haifa 31905, Israel
{oritht, sorin}@il.ibm.com

Abstract

Emotion detection has many potential applications in the medical domain for diagnosis of medical conditions that affect human speech communication. In the context of a European project, called Dem@Care, we develop a monitoring system for patients with dementia. In this paper, we focus on the voice-based mood assessment part of the project and present results from our ongoing research on detecting depression using a data set of patients with dementia that was collected within the project. We review the techniques reported in the literature for detecting depression, as an example of emotional vocal biomarkers extraction for medical diagnosis. We describe the data set, which consists of more than 80 participants including both dementia patients and healthy control subjects, along with an indication of depression level of each participant based on self-assessment mood questionnaire. We present some promising results with verbal reaction time as an acoustic measure that correlates with depression.

1. Introduction

Patient monitoring systems are becoming important in patient care and can provide useful feedback on health-related issues for clinicians and caregivers. These systems can also be beneficial for the patients themselves and improve their quality of life and care. In an ongoing European project, called Dem@Care [1], we are dealing with assessment of dementia signs in monitored subjects using multiple sensing devices including wearable microphones, static and wearable cameras and some physiological sensors. In one use case, the patient is monitored at home in natural settings. In another use case, the patient is asked to perform certain cognitive and physical exercises during clinical assessment in the lab.

Mood analysis is an important part of monitoring the patient's condition. It is known that dementia progression is often accompanied by signs of apathy and irritation. Therefore, we are interested in recognizing the two basic emotions of sadness and anger, which are closely related to the two states of mood often observed in Dementia patients. Sadness and anger are also signs of depression, so recognizing depression in general is of great interest and relevance to this project and is the focus of this work.

Prior works on depression severity assessment through voice analysis exist and they all suggest that acoustic measures of patients’ speech may be useful for evaluation of depression. As indicated in [2], many medical conditions, including depression, autism, and schizophrenia, affect the communication patterns and for many years psychologists have been studying the speech and language of patients with these conditions. Recently, researchers have started to apply computational approaches for analyzing patients’ speech and automatically extract voice-based correlates of various medical conditions with the aim to provide useful information for the clinical stuff, the caregivers and also the patients themselves.

Alpert et al. [3] reviews the literature on acoustic measures of voice of patients with depression. In their study, they examined acoustic features measuring speech fluency and prosody in the speech of patients with depression for medical treatment assessment. Their findings show that the prosody of depressed patients is flatter comparing to the prosody of healthy subjects. A similar study by Kuny and Stassen on acoustic correlates to the recovery from depression is described in [4].

Mundt et al. [5] present a research work on voice acoustic measures that significantly correlate with depression severity based on recordings over the telephone of 35 patients under treatment during a 6-week study. Results show that patients responding to treatment had significantly higher pitch variability, had less pauses in their speech, and they spoke faster. This research was continued with a large scale study with 105 patients as described in a recent paper [6] indicating that vocal acoustic measures in speech may serve as biomarkers for depression severity and response to treatment.

The research [7] studies the vocal biomarkers related to prosody and speech rate for separating control and depressed patient groups. The study investigates the correlations between phonologically-based biomarkers and the clinical HAMD severity scores, for a 35-speaker free-response speech database. The study shows the correlation of the HAMD scores with acoustic features and discusses preliminary results of automatic classifications of the depression severity, based on phone-specific measures of speech rate. The automatic classifications is further studied and presented in [8]. In this study the baseline classifier is based on a Gaussian mixture model (GMM) trained discriminatively. The classifier is aimed to predict the HAMD total score.
Another discriminative model for classifying depression based on acoustic features is presented in [9].

Moore et al. [10] describe a study on using glottal features for classifying depression and show some promising results on a data set of 33 subjects (15 patients and 18 controls) using audio recordings of the subjects reading a short story. The glottal features extracted consist of timing, ratios, shimmer, and spectral features of the glottal waveform.

Research on using other modalities for detecting depression also exists. For example, in [11] a system for detecting depression and negative emotions is proposed using a multidimensional emotional model and explored with four physiological signals, and [12] compares facial expressions and vocal prosody features.

The remainder of this paper is organized as follows. In Section 2 we briefly describe the Dem@Care system and the data set collected for our experiments. In Section 3 we describe our proposed method for classifying depression, and Section 4 presents our experimental results. Finally, in Section 5 we provide some conclusions.

2. System Description

The Dem@Care system utilizes multiple sensors of different modalities to monitor the dementia patients. The sensors are attached to the patient’s clothing or mounted in the room to capture the visual, the acoustics, and some physiological signals from the patient and the ambient soundings. We are analyzing the audio input, which could be from one of three different scenarios. The three use cases of the system are: lab settings during clinical doctor visits, home environment, and nursing home. The experimental data that was collected in the lab settings is described next.

2.1. Data Collection

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 88 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 Mild Cognitive Impairment (MCI), and a control group of healthy people with similar age range and demographic attributes. Each participant being guided by a supervisor performed four different spoken tasks. These tasks include: looking at a picture and verbally describing it; describing a picture from memory after looking at it for a while; repeating short sentences after the supervisor; and a Diadochokinetic test [13] consisting of repeated pronunciation of the three syllables sequence “pa-ta-ka” as fast and as long as possible.

In this work we use the data from the third spoken task in which the participants had to repeat a sequence of sentences spoken by the supervisor one at a time. The entire task includes 15 different sentences and was recorded at once. Each sentence is spoken by the supervisor and repeated by the participant. The audio was recorded by a simple headset microphone worn by the participant and digitized at 22050 Hz sampling rate with 16 bits per sample. The task was recorded entirely including the supervisor’s speech and the participant's speech.

2.2. Mood Score

As ground truth for the mood assessment part, we used the Geriatric Depression Scale (GDS) score [14]. The GDS is a self-report assessment designed to identify depression in elderly people and is generally recommended as part of a comprehensive geriatric assessment. Two versions of the GDS mood questionnaire exist, the original one with 30 items and a shorter version with 15 items that was developed later [15]. Each item is a yes/no question, and one point is assigned to each answer in correspondence to a scoring grid. For example, the first question is “Are you basically satisfied with your life?”. If the patient answers “no”, then the score for that item would be 1. The total score is tested against a threshold to separate between depressed and non-depressed patients. In our study, we used the short version with 15 questions with the GDS threshold score of 4, meaning that a patient with a GDS score of 0-4 is considered normal.

3. Method

We are looking for task-specific manifestations of depression in the speech data. In particular, we investigate the conversational behavior of the patient by measuring the response time of the patient to the supervisor’s prompt in the sentence repetition task described in subsection 2.1. We derive the latency of the patient’s response by measuring the pause duration between the end of the supervisor’s utterance and the beginning of the patient’s response. We refer to this feature as the Verbal Reaction Time (VRT). Verbal response latency was used as indicator of the affective state for children with autism during interaction [16]. We measure all the VRTs between the supervisor and the patient turns of all the sentences repeated in the entire session, as illustrated in Figure 1 by showing the segmentation of a portion of an audio file. We can then produce some statistics of the VRTs, such as the mean, variance, minimum, maximum, and range of the VRT measurements.
4. Experimental Results

4.1. The Setup

As a preparation step for the feature extraction for the sentence repetition task, we determined the end-points of all the reference sentences uttered by the supervisor and the sentences repeated by the participant. End-pointing was performed using an energy-based voice activity detector to find the speech parts and exclude the silence segments, followed by speaker diarization techniques that identify segment the participant and the supervisor speech segments, as shown in Figure 1. The diarization algorithm we used is based on the technology described in [17]. Once the utterances of the recording session of each participant are segmented, we can extract the acoustic features described in Section 3, and in particular the mean VRT feature for which we present the classification results in the next subsection.

4.2. Results

In Figure 2, we can see the mean VRT feature in seconds for each participant in the data set. The green data points are for the 70 non-depressed participants (with GDS between 0-4) and the blue points are for the 18 depressed patients (with GDS above 4). The mean values of the two groups, shown in black, are well separated with a p-value smaller than 0.001.

In Figure 3, we see the classification results using the mean VRT as a single feature for classifying between depressed and non-depressed participants. The classification is based on applying a simple threshold on the mean VRT value. Using a global threshold on the entire test set, we measure the recall rate of the two classes, as shown by the green and blue curves for non-depressed and depressed groups respectively. The overall best performance achieved with this feature is: unweighted average recall rate of 77.4%.

Figure 1: Speaker segmentation (patient in green, supervisor in red, silence in yellow) for a portion of an audio signal from the sentence repetition task (shown in blue) with indication of the two corresponding VRT segments (shown in gray).

Figure 2: Mean VRT in seconds for depressed and non-depressed participants (GDS threshold is 4).

Figure 3: Recall rate in percent as a function of the mean VRT threshold for depressed and non-depressed participants (GDS threshold is 4).

5. Conclusions

In this paper, we show that the mean VRT as a single acoustic feature can be useful for classifying between depressed and non-depressed people. However, there are some limitations to our experimental setup that we would like to point out and address in future research. First, the use of automatic segmentation of the data using the speaker diarization technology may not be optimal and may introduce some errors to the VRT analysis. Therefore, in future experiments, we will perform manually segmentation of the data to exclude the influence of the automatic segmentation errors. Secondly, we would like to extend our feature set and explore additional acoustic measures that correlate with depression and apply machine learning techniques for classification.
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/