Detecting depression using voice signal extracted by chatbots: A feasibility study

Alexandros Roniotis, and Manolis Tsiknakis

Department of Informatics Engineering
Technological Educational Institute of Crete
Heraklion, Greece
alexandros.roniotis@gmail.com, tsiknaki@staff.teicrete.gr

Abstract. This work aims at proposing a novel framework for detecting depression, like commonly met in cancer patients, using prosodic and statistical features extracted by voice signal. This work presents the first results of extracting these features on test and training sets extracted from the AVEC2016 dataset using MATLAB. The results indicate that voice can be used for extracting depression indicators and developing a mobile application for integrating this new knowledge could be the next step.

Keywords: virtual coach; cancer; detecting depression; machine learning; MFCCs

1 Introduction

Cancer is a major health problem in developed countries and accounts for almost 15% of all deaths [1]. Apart from physical exhaustion, knowledge about the forthcoming death has a serious impact on the psychological condition of the patient, resulting in increased identification of depression [2-4]. Clinical depression can only be diagnosed by a professional psychologist or clinician and treated through antidepressants or psychotherapy [5-7]. Depression in cancer patients seems to accelerate disease progression [8]. However, proper assessment and recognition of mental disorders requires intensive training and experience [9]. Developing an automatic machine learning mechanism for automatically detecting signs of depression could prove handful to clinicians during the early detection and start of psychotherapy [10-11].

Developing e-health applications for smartphones or tablets is a rapidly growing sector [12]. Automated health monitoring using a software that interacts with the patient, called virtual coach, can help in self-treatment of the patient, reducing monitoring costs in a clinical environment and the timely notification of the supervising clinician [13-14]. Not only the cost of using a system of this type is low, but also the familiarity of patients with modern devices increases over the years.

By installing a virtual coach in the smart device of a patient, the patient could monitor his or her mental state at anytime, anywhere and without the use of additional equipment or the need to get monitored within any hospital. Moreover, the technical requirements of modern phones allow the processing of complex data, such as voice signals, as they currently have powerful processors, large storage space and memory [15].

This work presents some first results on extracting signal features from voice for the purpose of detecting depression in cancer patients (or in general cases). The program has been applied on real voice segments, provided by the AVEC2016 dataset (www.avec16.com). The first results indicate that the features extracted are correlated to depression and we could move to the next future step; to apply the detection algorithms to cancer patients through mobile application and augmented reality chatbots.

2 Background

Clinical depression affects mood, thinking, behavior and physical condition [16]. Especially, voice and articulation of a person are directly affected by mental state [17], therefore voice features can be used as biomarkers of depression [18]. Voice features correlate with the presence and grade of depression and are often used to develop automatic classifiers. These features are classified as normal, mild, moderate, severe or very severe mental disorder according to the Hamilton Rating Scale for Depression (HAM-D), the 9-item Patient Health Questionnaire (PHQ-9) or Beck Depression Index (BDI) [19-21].

Depression affects speech production by differentiating stimulation of muscles and vocal cords [22] and altering respiratory rate [23]. Therefore, the quality of the produced sound is affected and is objectively measurable [24]. Some features that are used for the classification of emotional state are categorized as prosodic or spectral. Prosodic features include the rate of speech, the fundamental voice frequency (*f0*), the intensity and the energy of the voice and glottal features [25-26]. Some frequent spectral features include the formants (the eigenfrequencies of the vocal organ), the power spectral density (PSD) and Mel Frequency Cepstral coefficients (MFCCs) [24, 27-28].

3 Objectives

High costs and modest effectiveness of health system is often attributed to lack of patient's engagement at home [29]. The effective engagement is considered the "trillion opportunity" [30] and many companies have invested in developing m-Health applications for smart devices to involve in self-monitoring their health state, following their therapeutic scheme, reporting symptoms, etc. However, the devotion of patients proved moderate to low, mainly because they were not prompted by a third party to keep using the application. Instead, when using the application under the constant presence of a clinician, the results turned very encouraging [31]. Thus, it appears that the existence of a coach during the usage of the application could improve its effectiveness [32].

After the usage increase of the first years, applications for smart devices have lost their initial momentum [33]. Now the new generation of applications is considered that of chatbots, i.e. automated communication programs where the user chats with the device [29]. Chatbots are considered the next challenge for healthcare applications where a virtual coach will discuss with the patients, encourage them to pursue an action, raise questions and guide the discussion according to the answers received and processed [29, 34-35]. Such an application could extract depressive biomarkers.

4 The framework

The proposed scientific work is divided into two main sections. At first the extraction of audio data is performed for each subject, parallel to filling a depression questionnaire. Then, data is processed and feature vectors are extracted to compose the training set.

In the second stage, the speech of a subject is recorded and is then processed to generate a feature vector with the same features of the first stage. The vector is then classified into a depression class, according to the training set of the previous stage. Then, depending on the results of the classification, the virtual coach urges the patient to perform more psychotherapeutic activities, adhere to the therapeutic scheme, and notify the supervising clinician when depression scale is classified as severe.

4.1 Training Set

Initially, an application for mobile devices will be developed in order to generate some data. The application will be installed on a mobile device such as a mobile phone or tablet. The first time it runs, the program will show the patient a BDI questionnaire [20]. The answers will be stored for processing at a later stage.

Then, the virtual coach appears, developed with the open source environment BotLibre (www.botlibre.com), which, based on Positive psychology theory [36-37], will start one interactive chat with the individual. The speech signal is then stored on the device for post-processing.

The answers on the BDI questionnaire are used as the ground truth towards defining the depression scale of the training set. Speech signal from the user is used to extract feature vectors. The features include the Mel-frequency cepstral coefficients (MFCCs) [36], speech duration, the duration of pauses, speech rate and the response-to-question delay and several more described in the next section. These features and the gender of the user form the training set.

4.2 Classification of Depression Scale

During the second stage a user is classified in one of the fore-mentioned depression scales. The user’s responses to the chatbot’s queries are recorded. The resulting signal will be recorded for producing the feature vector to be classified using the training set.

Another stage is to assess the effectiveness of classification. Towards evaluation the leave-one-out method is used, where each vector of the training set is sorted after removing the same vector. Finally, the resulting Confusion Matrix is used for estimating accuracy [39].

5 Methods

5.1 Extracting the Audio Features

The first step of our study before developing the framework is to evaluate the performance of voice features for detecting signs of depression. Therefore, the audio data provided by AVEC has been used, accompanied by their respective transcripts. The dataset consists of a series of pre-extracted features using the COVAREP toolbox at 10-ms intervals over the entire recording (*f0*, NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, Rd, Rd conf, MCEP 0-24, HMPDM 1-24, HMPDD 1-12, and Formants F1-F3).

For the purposes of the present work, the resulting time series data were submitted to additional preprocessing steps as follows: First, the participant's voice was isolated using the time stamps in the transcripts. Segments with values of "<synch>", "<laughter>", "[laughter]", "<sigh>" and "scrubbed entry" were ignored as non informative segments. Next, segments containing unvoiced segments (VUV=0) were removed from the final concatenated time series. Furthermore, to correct for instances of apparently inaccurate annotation analyses were restricted to continuous voiced segments lasting > 5 ms. To control for speaker dependency, the *f0* was normalized to a scale of 0 to 1, and the deltas and delta-deltas were extracted for *f*0 and MFCCs.

The main analyses consisted in computing three sets of features to be used in subsequent classification approaches using Matlab Treebagger (n=100 trees) classifier. The first set of audio features consisted of a series of statistical descriptors for each pre-extracted descriptor, while the second set consisted of Discrete Cosine Transform (DCT) coefficients for each descriptor. The first 10 values of the DCT were retained, reducing the number of parameters and therefore complexity. The third set of audio features consisted of 8 high level features which were computed for the entire duration of the concatenated time-series. The Pause Ratio was extracted measuring the frequency of pauses during the participant's speech. Pauses were detected automatically using a pause detector, which relied on a low loudness detection function based on the Perceptual Quality measure.

Some other features include the Voiced Segment Ratio (computed as the number of voiced segments divided by the length of the entire speech segment) and the Speaking Ratio (computed as the number of speaking instances that there is participant's speech), divided by the total number of selected recorded segments, as

$$\text{SpeakingRatio} = (\#\text{allinstants} - \#\text{pauses}) / \#\text{allinstants}$$

Some more are the Mean Laughter Duration, defined as the duration of laughter segments divided by the total number of laughter instances; the Mean delay in response to chatbot's questions; the Mean duration of pauses; the Maximum duration of pauses; The Fraction of pauses in overall time.

Finally, the former two sets of features were individually combined in feature level with the high level features into single feature vectors. The final set of statistical descriptors with high-level features was of size 494, and the set of DCTs with high-level features was of size 1278.

5.2 Classification

Gender-based classification for depression seems to substantially improve depression detection. In the present work, gender-based classification was implemented by building two different classifiers, one for men and another for women. The classifier for men was trained on feature-sets extracted from data of male participants and the women classifier from data of female participants. The classifier used was Matlab's treebagger using a forest of 100 trees.

6 Experimental Results

Performance of classification was evaluated through training with the training set and subsequent testing with the development and test sets provided by AVEC. In addition, the algorithms were assessed using the leave-one-out procedure on the joined training and development sets. The gender-based approach outperformed gender-independent models with the audio statistical descriptors.

More specifically, the F1-score for gender-independent classification was 0.24 for depression and 0.75 for non-depression. On the other hand, F1-score for depressed is 0.59 and for non-depressed is 0.87.

7 Conclusion

The F1-score of the gender-specific audio feature classification depicts that there is correlation of voice to depression, as expected by literature. However, it is interesting to study if the features could be fused with more features extracted by more modalities, such as video or text transcripts.

Future works include integrating the feature selection and classifications algorithms in a mobile application, where a chatbot will chat with patients. Patient replies will be recorded and post-processed for depression detection.

Acknowledgments. This research is supported by IKY scholarships programme and co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the action ”Reinforcement of Postdoctoral Researchers” in the framework of the Operational Programme ”Human Resources Development Program, Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) 2014 – 2020.

References

1. World Health Organization, "World Cancer Report 2014," 2014.
2. S. Dalton, T. Laursen, P. Mortensen, and C. Johansen, "Risk for hospitalization with depression after a cancer diagnosis: a nationwide, population-based study of cancer patients in Denmark from 1973 to 2003.," Journal of Clinical Oncology, vol. 27, no. 9, pp. 1440-1445, March 2009.
3. Greek National Research Institute, "Mental Health - Contemporary Approaches and Reflections", Athens, 2011.
4. G. Moussas, A. Papadopoulos, A. Christodoulaki, and A. Karkanias, "Psychological and psychiatric problems in patients with cancer: relationship with the localization of the disease," Psychiatry, vol. 23, no. 1, pp. 46-60, Athens, 2012.
5. American Psychiatric Association, "Diagnostic and statistical manual of mental disorders, Fourth Edition, Text Revision: DSM-IV-TR," American Psychiatric Publishing Inc., Washington, DC, 2000.
6. D. Karapoulios, I. Getsios, V. Rizou, A. Tsiklitara, S. Kostopoulou, Ch. Balodimou, and N. Margari, "Anxiety and depression in patients with lung cancer under chemotherapy. Evaluation with the Hospital Anxiety and Depression Scale HADS," Asclepios Step, pp. 428-440, Athens, 2013.
7. C. Mathers, T. Boerma, and D. Ma Fat, "The global burden of disease: 2004 update," WHO, Geneva, Switzerland, 2008.
8. American Cancer Society, "Depression Increases Cancer Patients' Risk Of Dying," 2009.
9. Fotiadou, F. Priftis, and S. Kiprianos, "The Role of Primary Health Care in the treatment of people with mental disorder,» Brain, vol. 41, no. 1, Athens, 2004.
10. J. Cesar, and F. Chavoushi, "Depression, WHO - Priority Medicines for Europe and the World (2013 Update), 2013.
11. S. Kampakis, and Th. Tsironis, "The role of engineering Learning in Clinical Psychiatry - Application on depressed patients data," Thessaloniki, 2011.
12. V. Gay, and P. Leijdekkers , "A Health Monitoring System Using Smart Phones andWearable Sensors," International Journal of ARM, vol. 8, June 2007.
13. Wissen, C. Vinkers, and A. Halteren, "Developing a Virtual Coach for Chronic Patients: A User Study on the Impact of Similarity, Familiarity and Realism," Proceedings of the 11th International Conference on Persuasive Technology, 2016.
14. T. Ellis, N. Latham, T. DeAngelis, C. Thomas, M. Saint-Hilaire, and T. Bickmore, "Feasibility of a Virtual Exercise Coach to Promote Walking in Community-Dwelling Persons with Parkinson Disease," American Journal of Physical Medicine & Rehabilitation, vol. 92, no. 6, pp. 472-485, June 2013.
15. C. Free, G. Phillips, L. Watson, L. Galli, L. Felix, P. Edwards, V. Patel, and A. Haines, "The Effectiveness of Mobile-Health Technologies to Improve Health Care Service Delivery Processes: A Systematic Review and Meta-Analysis," PLoS Medicine, vol. 10, 15 January 2013.
16. T. Albrecht, and C. Herrick, "100 Questions and Answers About Depression," Jones & Bartlett Publishers, 2010, p. 212.
17. S. Sahu, and C. Espy-Wilson, "Effect of depression on syllabic rate of speech," Journal of the Acoustical Society of America, vol. 138, p. 1781, 2015.

18. Ozdas, R. Shiavi, S. Silverman, and D. Wilkes, “Analysis of fundamental frequency for near term suicidal risk assessment,” in Systems, Man, and Cybernetics, 2000 IEEE International Conference, 2000.
19. H. Hamilton, “A rating scale for depression,” Journal of Neurology, Neurosurgery and Psychiatry, vol. 23, pp. 56-62, 1960.
20. Beck, R. Steer, and G. Brown, “Beck Depression Inventory-II,” The Psychological Corporation, 1961-1996.
21. K. Kroenke, R. Spitzer, and J. Williams, “The PHQ-9, Validity of a Brief Depression Severity Measure,” Journal of General Internal Medicine, vol. 16, pp. 606-613, 2001.
22. N. Roy, S. Nissen, and S. Sapir, “Articulatory changes in muscle tension dysphonia: evidence of vowel space expansion following manual circumlaryngeal therapy,” Journal of Communication Disorders, vol. 42, no. 2, pp. 124-135, 2009.
23. S. Kreibig, “Autonomic nervous system activity in emotion: a review,” Biological Psychology, vol. 84, no. 3, pp. 394-421, July 2010.
24. N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” Speech Communications, vol. 71, pp. 10-49, 2015.
25. A. Pampouchidou, O. Simantiraki, A. Fazlollahi, D. Manousos, Pediaditis M, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias, F. Yang, and M. Tsiknakis, “Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text,» 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, Netherlands, pp. 27-34, 2016.
26. D. France, R. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, “Acoustical properties of speech as indicators of depression and suicidal risk,” IEEE Transactions on Biomedical Engineering, vol. 47, no. 7, July 2000.
27. T. Laosaphan, and T. Yingthawornsuk, “Classification of Depressed Speakers based on MFCC in Speech Samples,” ICAEEE 2012, Pattaya, Thailand.
28. D. Sturim, P. Torres-Carrasquillo, T. Quatieri, N. Malyska, and A. McCree, “Automatic Detection of Depression in Speech using Gaussian Mixture Modeling with Factor Analysis,” Proceedings of Interspeech, 2011.
29. M. Yuan, “Chatbots: Building Intelligent Bots,” Addison-Wesley, 2016.
30. J. Greene, J. Hibbard, R. Sacks, V. Overton, and C. Parrotta, “When Patient Activation Levels Change, Health Outcomes And Costs Change, Too,” Health Affairs, vol. 34, no. 3, pp. 431-437, March 2015.
31. C. Bloss, N. Wineinnger, M. Peters, D. Boeldt, L. Ariniello, J. Y. Kim, J. Sheard, R. Komatireddy, P. Barrett, and E. Topol, “A prospective randomized trial examining health care utilization in individuals using multiple smartphone-enabled biosensors,” PeerJ, 2016.
32. D. Freeney, “Usability versus Persuasion in an Application Interface Design,” Institute for Innovation Design & Engineering, Mälardalen University, Eskilstuna, Sweden, 2014.
33. The Nielsen Company, “So Many Apps, So Much More Time for Entertainment,” 2016.
34. M. P. Zillman, “Healthcare Bots and Subject Directories,” 2016.
35. “Bots, the next frontier,” The Economist, 2016.
36. Versluis, B. Verkuil, P. Spinhoven, M. van der Ploeg, and J. Brosschot, “Changing Mental Health and Positive Psychological Well-Being Using Ecological Momentary Interventions: A Systematic Review and Meta-analysis,” Journal of Medical Internet Research, vol. 18, no. 6, June 2016.
37. R. Proyer, F. Gander, S. Wellenzohn, and R. Willibald, “Positive psychology interventions in people aged 50–79 years: long-term effects of placebocontrolled

online interventions on well-being and depression," Aging & Mental Health, vol. 18, no. 8, pp. 997-1005, 2014.

38. M. Sahidullah, and S. Goutam, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," Speech Communication, vol. 54, no. 4, pp. 543-565, 2012.
39. B. Dasarathy, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, 1991.