DEPRESSION SPEAKS: AUTOMATIC DISCRIMINATION BETWEEN DEPRESSED AND NON-DEPRESSED SPEAKERS BASED ON NONVERBAL SPEECH FEATURES

F.Scibelli[1], *G.Roffo*[2], *M.Tayarani*[2], *L.Bartoli*[3], *G.De Mattia*[4], *A.Esposito*[1] *and A.Vinciarelli*[2]

[1]Università degli Studi della Campania "L. Vanvitelli" and IIASS (Italy) - [2]University of Glasgow (UK)
[3]UOSM Angri-Scafati, Asl Salerno (Italy) - [4]UOSM Santa Maria Capua Vetere, Asl Caserta (Italy)

ABSTRACT

This article proposes an automatic approach - based on nonverbal speech features - aimed at the automatic discrimination between depressed and non-depressed speakers. The experiments have been performed over one of the largest corpora collected for such a task in the literature (62 patients diagnosed with depression and 54 healthy control subjects), especially when it comes to data where the depressed speakers have been diagnosed as such by professional psychiatrists. The results show that the discrimination can be performed with an accuracy of over 75% and the error analysis shows that the chances of correct classification do not change according to gender, depression-related pathology diagnosed by the psychiatrists or length of the pharmacological treatment (if any). Furthermore, for every depressed subject, the corpus includes a control subject that matches age, education level and gender. This ensures that the approach actually discriminates between depressed and non depressed speakers and does not simply capture differences resulting from other factors.

Index Terms— Depression, Social Signal Processing, Feature Selection, Computational Paralinguistics, Nonverbal Communication

1. INTRODUCTION

Depression is a mood disorder characterized by physical (e.g., appetite disturbance, insomnia or ipersomnia, loss of energy, psychomotor agitation or retardation), emotional (e.g., depressed mood, anhedonia), cognitive (e.g., low self-esteem, diminished ability to think, concentrate and make decisions) and behavioural symptoms (e.g., social isolation) "*causing significant distress or severely [impacting] social, occupational or other important life areas*" [1]. It is one of the most common mental disorders in the world (more than 300 million patients only in 2015 [2]), it is the second cause of disability after ischaemia and it is the most important suicide risk factor for elderly people [3].

As a result of the above, depression requires prolonged and expensive medical treatments that result into a significant economic burden for both patients and society [4]. The development of automatic approaches for the detection of depression can help to reduce such costs by supporting the activity of clinicians and, in particular, by reducing the time they need to diagnose a patient as depressed. For this reason, this paper proposes a speech-based automatic approach for the discrimination between depressed and non-depressed individuals. The experiments show that it is possible to automatically discriminate between depressed and not depressed speakers with an accuracy of over 75%.

Current methodologies for the assessment of depression are based on clinical interviews - e.g., the *Hamilton Rating Scale for Depression* (HRSD) [5] - or self-reporting questionnaires - e.g., the *Beck Depression Inventory II* (BDI-II) [6]. In the first case, there is a risk of biases due to training and theoretical orientation of the clinicians, while in the second case, there is a risk of biases due to the patients' motivation and ability to recognize or express their symptoms. For these reasons, automatic approaches based on measurable aspects of behaviour can be useful not only to reduce healthcare costs (see above), but also to provide quantitative information that can complement more established instruments like HRSD and BDI-II, possibly reducing the effect of their limitations.

Probably because of the above, the problem of depression detection has attracted significant attention in the computing community and it has been the subject of several international benchmarking campaigns (see, e.g., [7]). The proposed approaches are often multimodal (see, e.g., [8, 9]), but methodologies based on the sole speech are not uncommon (see [10] for an extensive review). One possible explanation is that the collection of speech data is, on average, less invasive and makes it easier to preserve the anonymity of the subjects. The most common approach is to adopt features inspired by affective computing - e.g., Mel Frequency Cepstral Coefficients, pitch, formants, energy, turn-taking characteristics (in case the data is conversational) - and to feed the resulting feature vectors to classifiers. The tasks most commonly addressed are the inference of how severe is the depression of an individual and the discrimination between depressed and non-depressed speakers (see, e.g., [11, 12, 13, 14, 15, 16]).

This work is supported by the Engineering and Physical Sciences Research Council (EPSRC) through the projects "Socially Competent Robot" (EP/N035305/1) and "School Attachment Monitor" (EP/M025055/1), and by the EU Horizon 2020 research and innovation programme through the project EMPATHIC (Grant Agreement No.: 769872).

The task addressed in this work is the discrimination between depressed and non-depressed speakers (these latter are called *control* subjects hereafter). For this reason, the corpus adopted for the experiments - never used in the literature so far - includes both people that have been diagnosed as depressed (62 subjects in total) and control speakers (54 subjects in total). In particular, for every depressed subject of a certain gender, education level and age, the corpus includes a control subject of the same gender, the same education level and comparable age. The goal is to ensure that the approach actually detects depression and not speaking differences that might result from other factors. Furthermore, the depressed subjects have been diagnosed as such by professional psychiatrists and, at least at the moment of the data collection, they were being treated for depression in medical structures. This represents a major advantage with respect to corpora where the subjects fill questionnaires like the BDI-II, but have not actually been diagnosed as depressed. The reason is that it makes it possible to avoid the risk to deal with data that does not reflect actual psychiatric problems [17, 18] and, hence, it leads to more realistic estimates of the approaches' performance.

The rest of the paper is organized as follows: Section 2 presents the corpus used in the experiments, Section 3 describes the approach, Section 4 reports on experiments and results, and Section 5 draws some conclusions.

2. THE DATA

The corpus used in this work includes 54 healthy control individuals (42 female and 12 male) and 62 depressed subjects (42 female and 20 male). The control subjects have been recruited via a *word-of-mouth* process, while the depressed subjects have been recruited among the patients of six medical centers - where they were being treated for their condition - located in Naples and its surrounding area (Italy)[1]. The psychiatrists of the centers have applied the criteria of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV) [1] and have diagnosed one of the following pathologies for each depressed subjects: major depressive disorder (26 cases), bipolar disorder in depressive phase or with last depressive episode (14 cases), reactive depression (8 cases), endo-reactive depression (8 cases) and anxiety-depressive disorder (6 cases). The diagnosis has been performed as part of the regular clinical activities of the doctors and not specifically for the collection of the data. All subjects are native Italian speakers and the recordings are in a such a language.

During the data collection process, the subjects have been asked to perform two tasks. The first is called *Diary* and aims at obtaining samples of spontaneous speech. During the task, an experimenter asks the subjects to talk about their last week end, their family, job, hobbies, etc. The second task is called *Tale* and it targets the collection of read speech samples. During the task, the subjects are asked to read aloud a tale written by Aesop (*The North Wind and the Sun*). For every subject, the experimenters have collected demographic information (age, gender, marital status, education level, type of employment, eyes problems and other physical diseases) and have administered the BDI-II. At the beginning of the tasks, the experimenters have asked the subjects to provide their informed consent by signing a document explaining the study and its goals.

The speech samples were recorded with clip-on microphones (Audio-Technica ATR3350), with external USB sound card, at a sampling rate of 16 kHz, with every sample represented with 16 bits. The data was collected in non-controlled settings where the level of environmental noise is significant and it was not possible to respect a rigorous experimental protocol. In this respect, the collection has been performed in the wild and the data can be considered challenging not only for the inherent difficulties of the task, but also because the data has been collected *in the wild*. A few samples had to be discarded because of technical difficulties and, as a result, the data available for the two tasks described above is as follows: 62 depressed subjects and 52 controls for Diary, and 57 depressed subjects and 54 controls for Tale.

3. THE APPROACH

The proposed approach includes three main steps. The first is the automatic segmentation of a sample (a recording corresponding to one subject performing either the Diary or the Tale task) into speech and non-speech intervals. The second is the application of a feature extraction approach aimed at converting every speech interval into a vector. The third step is the classification: all the vectors corresponding to the speech intervals of a sample are fed to a linear kernel SVM that assigns them to one of the two possible classes, namely *Depressed* (D) or *Control* (C). The classifications made at the level of every speech interval are combined through a majority vote to provide the final outcome of the process (the classification of a given subject). During the training process, the approach makes use of a feature selection technique to identify the subset of features expected to maximize the performance of the classification step.

The feature extraction is performed with OpenSmile [19], a publicly available tool commonly adopted for the inference of social and psychological phenomena from speech. The feature set - known as *IS09* [20] - is extracted following the methodologies of computational paralinguistics. First 16 short-term features are extracted from 25 ms long analysis windows that span the whole sample being processed at regular time steps of 10 ms. The 16 features are *Root Mean Square* (RMS) of the energy, the first 12 *Mel Frequency Cepstrum Coefficients* (MFCC), the *Zero Crossing Rate* (ZCR),

[1] The private clinic *Villa Camaldoli* in Naples and the Public Healthcare Centers of Santa Maria Capua Vetere, San Cipriano, Atripalda, Scafati, and Angri.

the *Voicing Probability* (VP) and the *pitch* ($F0$). After these features are available, it is possible to extract their Δ's, i.e., the differences between consecutive frames, thus leading the total number of features to 32. Finally, once these 32 features have been extracted from all short-term windows, their distribution across the speech sample is represented through the following 12 statistics: minimum, maximum, range, positions of the windows where maximum and minimum have been extracted, mean, slope and offset of the linear approximation of the contour, difference between linear approximation and actual contour, standard deviation, third and fourth order moment. As a result, the feature vector representing a sample includes $32 \times 12 = 384$ features.

The vectors are used to construct a linear predictor. This latter is learned with linear kernel Support Vector Machines (SVM) by fitting the model to the available training data, i.e., by minimizing an objective function that balances a quadratic regularizer and the hinge-loss. The hyper-parameter of the linear kernel SVM, aimed at trading-off regularizer and loss, is determined using a held-out validation subset of the training set. In order to reduce the dimensionality of the feature vectors, the approach uses the *Infinite Latent Feature Selection* (ILFS) approach [21] which allows one to distill the subset of the features expected to be most likely to discriminate between depressed and control individuals. The ILFS is a probabilistic latent graph-based feature selection algorithm that performs the ranking step while considering all the possible subsets of features, as paths on a graph. The gist of the approach is that it aims to discover an abstraction behind low-level sensory data, that is, relevancy. Relevancy is modeled as a latent variable in a PLSA-inspired generative process that allows the investigation of the importance of a feature when injected into an arbitrary set of cues.

4. EXPERIMENTS AND RESULTS

The experiments follow a leave-one-out approach: All subjects except one are used to perform the feature selection and to train the SVM aimed at doing the actual classification. The resulting system is then tested over the left-out subject. The process is iterated as many times as there are subjects and, at each iteration, a different subject is left out. Table 1 shows the performance for the Tale and Diary tasks, both with and without feature selection. The number of subjects is different for the two tasks because of technical problems encountered during the collection of the data (see Section 2).

The accuracy $\hat{\alpha}$ of a random classifier that assigns a sample to class C or D with probability p_C or p_D, respectively (where p_C and p_D are the fractions of samples that belong to class C and D) can be estimated as follows:

$$\hat{\alpha} = p_C^2 + p_D^2. \tag{1}$$

In the experiments of this work, the accuracy $\hat{\alpha}$ is 50.0% and 50.4% for Tale and Diary, respectively. Thus, according to a two-tailed t-test, the accuracies reported in Table 1 differ from chance to a statistically significant extent ($p << 0.01$ in all cases). There are no major performance differences across the tasks and the feature selection brings an improvement only in the case of Diary. The probable reason is that such a task involves spontaneous speech and, hence, there is more variability in the data, possibly not related to the depression condition of the subjects.

Task	D	C	π	ρ	α
Tale	57	54	75%	74%	76%
Diary	62	52	66%	60%	68%
Tale-FS	57	54	74%	80%	77%
Diary-FS	62	52	74%	65%	74%

Table 1. The table reports the classification results in terms of *Precision* π (percentage of subjects attributed to a class that actually belong to such a class), *Recall* ρ (percentage of samples belonging to a given class that have actually been attributed to that class) and *Accuracy* α (percentage of times the classification is correct. The acronym FS stands for *Feature Selection*, while D and C stand for number of Depressed and non-depressed (Control), respectively.

Task	Diary		Diary-FS	
	D	C	D	C
D	74.2%	25.8%	80.6%	19.4%
C	40.4%	59.6%	34.6%	65.4%
	Tale		Tale-FS	
	D	C	D	C
D	77.2%	22.8%	73.7%	26.3%
C	25.9%	74.1%	20.4%	79.6%

Table 2. The table shows the confusion matrices corresponding to the four tasks addressed in the experiments. The element p_{ij} of ever matrix is the probability that a sample belonging to class i has been assigned to class j. Element p_{ii} is the accuracy for the samples of class i.

The confusion matrices corresponding to the four classification tasks (Diary and Tale with and without feature selection) are shown in Table 2. Both classes are recognized beyond chance in all cases, but the accuracy difference between depressed and control subjects is larger in the case of the Diary tasks. One possible explanation is that such tasks require spontaneous speech and, hence, they are more difficult not only for the depressed subjects, but also for the control ones. Possible difficulties in planning what to say might make the speaking style of some control subjects similar to the one of depressed subjects and, hence, the classification task becomes more difficult. In the case of the Tale tasks, the control subjects do not face any more the difficulty of planning what to say - the task involves only read speech - while the depressed subjects still maintain the difficulties inherent to their condition. As a result, the difference between control and depressed subjects becomes more evident and the classification task easier to address, especially when it comes to the correct classification of the control subjects.

The information at disposition about the subjects includes gender and, for the depressed subjects, diagnosis and amount of time since the pharmacological treatment, if any, has started. This makes it possible to check whether there is a relationship between any of the factors above and the chances of misclassification. In the case of gender - an information available and relevant to all subjects - it is possible to perform a χ^2 test to check whether the subjects belonging to one of the two genders are misclassified more frequently. The results show that this is not the case and the chances of misclassification are the same for both female and male subjects. This suggests that the approach classifies the depression condition and not the gender, even if most of the depressed subjects (40 out of 22) are female.

The χ^2 test can be adopted to check whether the diagnosis of the psychiatrists for the depressed patients (major depressive disorder, bipolar disorder in depressive phase or bipolar disorder with last depressive episode) makes it more or less likely for a subject to be misclassified. The result is negative and this suggests that the chances of misclassification are the same, at least for the data at disposition, irrespectively of the diagnosis. In the case of the length of the pharmacological treatment, a t-test shows that there is no statistically significant difference between misclassified subjects and the population of the depressed subjects. Thus, at least for the data at disposition, there seems to be no relationship between such a factor and the chances of misclassification.

The feature selection approach adopted during the experiments (see Section 3) provides further insight about the results. Since the experiments follow a leave-one-out approach, the feature selection is applied over a different training set at every iteration. Since it is a statistical process, the subset of the selected features is likely to change at every iteration. Thus, every feature will be selected only in a fraction of the iterations and the higher such a fraction, the higher the chance that the feature acts as a depression marker, i.e., that it makes the difference between depressed and control subjects. The number of features selected at least 90% of the times for the Diary and Tale tasks is 210 and 371 (out of 384), respectively. This probably explains why the feature selection does not improve the performance in the case of the Tale task (see Table 1).

The figures above suggest that the feature selection approach - shown to be effective on the most important benchmarks of the literature [21] - can discard only a limited number of features in the case of the Tale task. In other words, the number of features that act as depression markers is significantly larger in read speech than how it is in spontaneous speech. This probably explains why the performance over the two tasks is similar even if the number of features discarded in the case of the Diary task is larger and, hence, the classifier should be trained more effectively. Overall, this finding seems to suggest that read speech captures more effectively the difference between depressed and control subjects (see beginning of this section). Furthermore, it probably explains why the control subjects are recognized better in the Tale task than how they are in the Diary one.

5. CONCLUSIONS

This article proposes an approach for the automatic discrimination between depressed and non-depressed speakers. The experiments have been performed over one of the largest corpora available for the task in the literature, including 62 and 54 depressed and control subjects, respectively. Unlike other benchmarks in the literature, the depressed patients of the corpus have been diagnosed as such by psychiatrists during their actual clinical activity. Furthermore, all subjects have been recorded when performing two tasks, one based on spontaneous speech (Diary) and the other on read speech (Tale). The results show that it is possible to achieve an accuracy of more than 75% for both spontaneous and read speech. However, in the case of this latter depressed and control subjects are classified correctly with the same performance while in the case of spontaneous speech, depressed subjects are detected with higher accuracy than the control ones.

The error analysis shows that the accuracy is the same for both male and female subjects, for the depressed subjects diagnosed with different types of pathology and for the depressed subjects with different length of pharmacological treatment. This suggests that the effect of these factors on the accuracy, if any, is too weak to be observed with the data at disposition. The adoption of a feature selection approach shows that the number of retained features tends to be high (around 66% in the case of the Diary task) and, in the case of the Tale task, only a minority of the features is discarded (less than 5%). Given that the feature selection approach has been shown to be effective on the main benchmarks adopted for this type of methodologies, this suggests that the task addressed in the experiments requires a large number of features to be performed effectively. In other words, speech carries a large number of depression markers, but no small subset of them can be identified that actually makes the difference between depressed and control subjects.

The future work will focus on conversational features (in the Diary task the subjects interact with an experimenter) like length and number of turns, speaking time, etc. that, in the current approach, have not been considered. Furthermore, the experiments of this work have made use of a linear SVM, but it is possible to explore the use of other classifiers. In this respect, the results obtained in this work can be considered as a lower bound of the performance that it is possible to achieve.

6. REFERENCES

[1] ”AA. VV.”, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*, American Psychiatric Association, 2013.

[2] ”AA. VV.”, “Depression and other common mental disorders: global health estimates,” Tech. Rep., World Health Organization, 2017.

[3] K. Wahlbeck and M. Mäkinen, “Prevention of depression and suicide: Consensus paper,” Tech. Rep., European Communities, 2008.

[4] J. Olesen, A. Gustavsson, M. Svensson, H-U Wittchen, and B. Jönsson, “The economic cost of brain disorders in europe,” *European Journal of Neurology*, vol. 19, no. 1, pp. 155–162, 2012.

[5] M. Hamilton, “A rating scale for depression,” *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 23, no. 1, pp. 56, 1960.

[6] A.T. Beck, R.A. Steer, and G.K. Brown, *Manual for the Beck Depression Inventory-II*, San Antonio TX: The Psychological Corporation, 1996.

[7] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M Pantic, “AVEC 2016: Depression, mood, and emotion recognition workshop and challenge,” in *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3–10.

[8] J.F. Cohn, T.S. Kruez, I. Matthews, Y. Yang, M.H. Nguyen, M.T. Padilla, F. Zhou, and F. De la Torre, “Detecting depression from facial actions and vocal prosody,” in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2009, pp. 1–7.

[9] S. Scherer, G. Stratou, and L.-P. Morency, “Audiovisual behavior descriptors for depression assessment,” in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2013, pp. 135–140.

[10] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T.F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Communication*, vol. 71, pp. 10–49, 2015.

[11] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, “Investigating voice quality as a speaker-independent indicator of depression and ptsd.,” in *Proceedings of Interspeech*, 2013, pp. 847–851.

[12] N. Cummins, J. Epps, and E. Ambikairajah, “Spectro-temporal analysis of speech affected by depression and psychomotor retardation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7542–7546.

[13] J.C. Mundt, P.J. Snyder, M.S. Cannizzaro, K. Chappie, and D.S. Geralts, “Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology,” *Journal of Neurolinguistics*, vol. 20, pp. 50–64, 2007.

[14] B.S. Helfer, T.F. Quatieri, J.R. Williamson, D.D. Mehta, R. Horwitz, and B. Yu, “Classification of depression state based on articulatory precision.,” in *Proceedings of Interspeech*, 2013, pp. 2172–2176.

[15] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, “Detecting depression: A comparison between spontaneous and read speech,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7547–7551.

[16] H. Jiang, B. Hu, Z. Liu, L. Yan, T. Wang, F. Liu, H. Kang, and X. Li, “Investigation of different speech types and emotions for detecting depression using different classifiers,” *Speech Communication*, vol. 90, pp. 39–46, 2017.

[17] P.C. Kendall, S.D. Hollon, A.T. Beck, C.L. Hammen, and R.E. Ingram, “Issues and recommendations regarding use of the beck depression inventory,” *Cognitive Therapy and Research*, vol. 11, no. 3, pp. 289–299, 1987.

[18] A.A. Stone, C.A. Bachrach, J.B. Jobe, H.S. Kurtzman, and V.S. Cain, *The science of self-report: Implications for research and practice*, Psychology Press, 1999.

[19] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in Opensmile, the Munich open-source multimedia feature extractor,” in *Proceedings of the ACM International Conference on Multimedia*, 2013, pp. 835–838.

[20] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, “Cooperative learning and its application to emotion recognition from speech,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 115–126, 2015.

[21] G. Roffo, S. Melzi, U. Castellani, and A. Vinciarelli, “Infinite latent feature selection: A probabilistic latent graph-based ranking approach,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.