THIN SLICES OF DEPRESSION: IMPROVING DEPRESSION DETECTION PERFORMANCE THROUGH DATA SEGMENTATION

Rawan Alsarrani[1], *Anna Esposito*[2] *and Alessandro Vinciarelli*[1]

[1]University of Glasgow, Glasgow (UK)
[2]Università degli Studi della Campania "Luigi Vanvitelli", Caserta (Italy)

ABSTRACT

The computing community is making major efforts towards automatic detection of depression, a serious pathology that affects roughly 4.4% of the world's population. One of the main difficulties is the collection of data aimed at training models capable to learn differences between depressed and non-depressed people. In fact, data collection in the depression domain requires the respect of rigorous ethical constraints that, inevitably, limit the size of the corpora that can be collected. This article proposes to address the problem by using the thin slices theory, i.e., the possibility to detect the inner state of an individual (depression in this case) through very short samples of behavior. In particular, the article shows that the performance of data-driven models can be improved by segmenting the data at disposition into thin slices and then training data-driven models over them. This increases the amount of samples at disposition and allows a relative F1 Score improvement by up to 16.2%.

Index Terms: Computational paralinguistics, depression detection, social signal processing, read speech.

1. INTRODUCTION

The computing community makes major efforts towards automatic detection of mental health issues and depression is, by far, the pathology most commonly addressed [1]. The main reason is that, according to the World's Health Organization, depression affects almost one person out of twenty in the world [2], thus representing a major societal issue. Overall, the main effect of clinical depression is 'an *increase* in negative emotions and feelings and a *reduction* in positive emotions and feelings" [3]. As a consequence, depression patients experience "depressed mood, loss of interest or pleasure, decreased energy, feelings of guilt or low self-worth, disturbed sleep or appetite, and poor concentration" [4], enough to make of depression the most common cause of disability (7.5% of all years lived with disability in 2015 [2]) and suicide.

One of the main consequences of the problems above is that depressed people are a vulnerable population and, therefore, it is difficult to involve them in experiments. In particular, the number of participants, a key-factor in any experiment dealing with human behavior, tends to remain limited and it affects negatively the possibility to develop effective depression-detection approaches. For this reason, this article shows that it is possible to improve the performance of neural models (in particular Multi-Layer Perceptrons and Recurrent Neural Networks) by splitting available recordings into non-overlapping segments. These can then be used as individual training samples. The results show that such a strategy, while not increasing the number of participants, still allows one increase the size of the training material and to achieve a relative F1 Score improvement by up to 19.5%.

The experiments were performed over 125 interview recordings, 72 for depressed people and 53 for non-depressed control participants, for a total of 6 hours, 25 minutes and 12 seconds (see Section 2). Different models were trained using each recording as an individual input, then by splitting every recording into two halves used as individual inputs, and so on until every recording was split into 128 individual samples. The goal of such a process was to progressively multiply the number of training samples at disposition and test whether it had any positive impact on the performance of the models. The results show that there are improvements until the size of the training set was multiplied by 128.

One possible explanation behind the results above is the *Thin Slices Theory* [5], i.e., the indication that short behavioral observations can be sufficient to infer the inner state of an individual (emotions, personality traits, etc.). In the particular case of this work, this means that short speech samples (the shortest segments that were used are less than one second long) can be sufficient to reliably detect the condition of a speaker (depressed or non-depressed). For this reason, the approach used for the experiments follows the methodologies of Social Signal Processing [6] and Computational Paralinguistics [7], two computing areas shown to be effective in dealing with thin slices of behaviour. In addition, the majority of the thin slices are classified correctly and this allows one to make a correct decision about an individual through the application of a majority vote. In this respect, increasing the number of samples at disposition for an indidivual increases the chances that, for a given number of segments, the majority is correct.

To the best of our knowledge, most of the work presented

so far in the literature focused on depression detection effectiveness rather than on approaches for dealing with limited data (see, e.g., [8] for an extensive survey). However, previous work has shown that the use of short segments of speech collected on the phone can lead to accuracies of around 70% in predicting whether someone is depressed according to a self-assessment questionnaire [9]. Furthermore, previous experiments suggested that clauses (atomic linguistic units composed of a noun, a verb and a complement) carry enough depression-relevant information to allow depression detection (accuracies up to 80%) [1]. Finally, there was an attempt to increase the amount of training material by combining multiple available corpora (possibly including different languages and cultures), but the results were not encouraging [10]. Such a brief state-of-the-art suggests, one the one hand, that the use of short utterances it is not necessarily an obstacle and, on the other hand, that the use of multiple sources of data is not a viable solution. In this respect, the approach proposed in this work appears to be in line with the indications of the literature.

The rest of this article is organized as follows: Section 2 describes the dataset used for the experiments, Section 3 describes the approaches used in the experiments, Section 4 reports on experiments and results, and the final Section 5 draws some conclusions.

2. THE DATA

The experiments of this work were performed over a corpus of recordings in which 125 participants, all Italian native speakers, undergo an interview about aspects of everyday life (e.g., activities of the last week end or favorite movie). The questions were asked always in the same order and the interviewers were instructed to talk as little as possible. The turns of the interviewers were extracted manually so that the recordings used for the experiments include only the voice of the experiment participants. Out of the 125 participants, 72 were diagnosed with depression by a professional psychiatrist and, at the moment of the recordings, they were under treatment in one of the three Italian Mental Health Centers involved in the study. The diagnosis was made on the basis of the DSM-5, the version of the *Diagnostic and Statistical Manual of Mental Disorders* in use at the moment of the data collection. The remaining 53 people, referred to as *control* participants, were randomly recruited among people that never experienced mental health issues. All participants joined on a voluntary basis and signed an informed consent formulated in accord with Italian and European privacy and data protection laws[1].

Overall, it was ensured that there was no statistically significant difference between the two groups (depression and control) in terms of age and gender (full demographic information is available in Table 1). In such a way, it was possible to increase the chances that any observable difference in the way of speaking is actually due to depression and not to other confounding factors. Table 1 shows that female participants are 2.9 times more than male ones. The reason is that, in line with large-scale epidemiological observations [11], the number of women treated for depression in the Mental Health Centers above is significantly higher than the number of men. In this respect, the sample of depressed participants is representative of the depression patients at the moment the data were collected. Similar considerations apply to the age range that, according to the indications of the literature, matches the typical ages at which people develop depression [12, 13, 14].

Group	Female	Male	Avg. Age	Age Range
Depression	50	22	47.9±12.4	23-71
Control	43	10	48.0±12.4	19-68
Total	93	32	47.9±12.3	19-71

Table 1. The table provides demographic information for the two participant groups (rows "Depression" and "Control") as well as for the corpus as a whole (row "Total").

The overall average is 184.9 $\pm$ 79.4, while the averages for depressed and non-depressed participants are 173.4 $\pm$ 94.4 and 200.5 $\pm$ 49.1, respectively. According to a two-tailed t-test, such a difference is not statistically significant and it suggests that depressed participants tend to talk as much as the others. During the experiments, every recording was first used as a whole to perform classification experiments. Then, every recording was split into two halves so that the number of data samples at disposition was doubled (while keeping the number of participants constant). The halves were then used to train and test new models, thus leading to new recognition results. The whole process was then applied to the halves so that the initial number of samples was multiplied by four and so on until every interview recording was split into 128 segments of equal length. In the rest of the paper, the segments will be referred to as *slices*.

3. THE APPROACH

Section 2 shows that each interview recording was segmented into increasingly shorter slices. The goal was to increase the number of data samples at disposition and, during such a process, the initial number of recordings was multiplied by 2, 4, 8 and so on until 128 by splitting every slice into two equal length segments. For such a reason, the approach includes two main steps, namely the classification of individual slices and the aggregation of the classifications made at the level of the individual slices. The rest of this section presents the two steps in detail.

[1]The ethical committee of the Department of Psychology at Universitá degli Studi della Campania, "Luigi Vanvitelli", authorized the experiment with protocol number 09/2016.

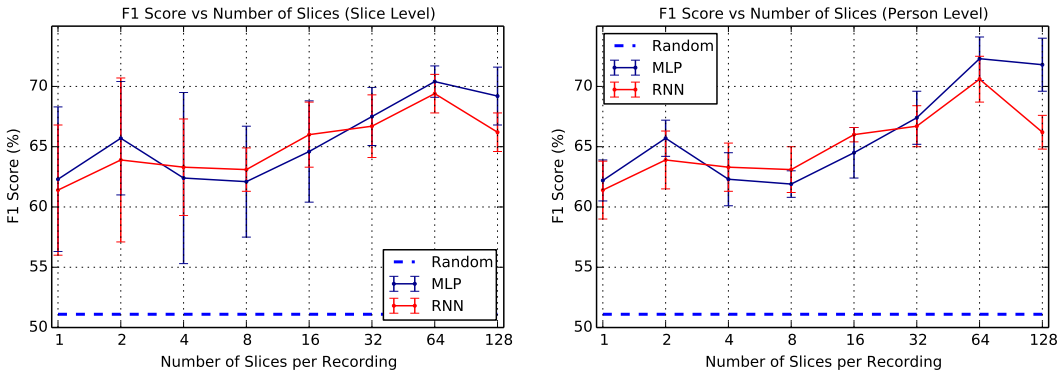

Fig. 1. The plot shows how the slice (left) and person (right) level accuracies change with the number of slices. The horizontal axis is logarithmic (every point corresponds to one of the powers of 2 between 0 and 8). The horizontal dashed line shows the F1 Score of a random system that assign samples to class c with probability corresponding to the prior of c itself. The vertical bars are the standard deviations over the R repetitions of the experiment.

3.1. Slice Recognition

At the beginning of the slice recognition process, the speech signal is converted into a sequence of feature vectors $\vec{x}_k$ extracted at regular time steps of 10 ms from 25 ms long analysis windows (the values of step and window length were selected because they are standard in the literature). The feature extraction was performed with OpenSMILE [15], a publicly available software package widely applied in the literature. The features that were extracted are as follows: Root Mean Square of the Energy (Energy, the feature accounts for how loud someone speaks), Mel-Frequency cepstral coefficients 1-12 (MFCC, these features account for the phonetic content of the data), Fundamental Frequency (F0, the feature corresponds to the frequency carrying the highest energy in the signal), Zero-Crossing Rate (ZCR, related to F0 and it provides indications about the frequency carrying most of the energy in the signal), and Voicing probability (VP, the probability of an analysis window to correspond to the emission of voice). The set above includes 16 features and, for each of them, the extraction step includes the difference between the value in the current analysis window and the value in the previous analysis window, thus leading to a final feature vector of dimension $D = 32$. The motivation behind the choice of this set is that it was designed to recognize emotions in speech [16] and one of the major symptoms of depression is the increase in negative emotions accompanied by a decrease in positive emotions (see beginning of Section 1) [3].

The classification is performed with two models, namely a *Multi-Layer Perceptron* (MLP) [17] and a *Recurrent Neural Network* (RNN) [18]. In the first case, all vectors extracted from a slice are averaged into an individual vector $\vec{x}$ that is fed to an MLP trained to discriminate between depressed and non-depressed speakers. In the second case, the vector sequences are split into frames of $L = 128$ vectors that start at regular steps of 64 vectors (two consecutive frames overlap by half their length) and fed to an RNN trained, like in the case of the MLP, to discriminate between depressed and non-depressed speakers. Given that a slice can include more than one frame, the final decision is made through a majority vote (the slice is assigned to the class its frames are most frequently assigned to). For the details about topology and training of the models, see Section 4.

3.2. Aggregation of Slice-Level Classifications

The previous section, shows how every slice is classified individually. Given that the number of slices per recording is k ($k = 1, 2, 4, 8, 16, 32, 64, 128$), the aggregation can be performed through a majority vote: $\hat{c} = \arg\max_{c \in \mathcal{C}} n(c)$, where $\hat{c}$ is the outcome of the aggergation (the class assigned to the person speaking in the recording from which the slices are extracted), $\mathcal{C}$ is the set of the classes (depression and control) and $n(c)$ is the number of slices assigned to class c. Given that the number of slices is even, it can happen that $n(c)$ is the same for both classes. In such a case, a recording is considered to be wrongly classified.

4. EXPERIMENTS AND RESULTS

The experiments were performed according to a k-fold experimental protocol ($k = 5$), meaning that the participants were randomly split into k disjoint subsets and, then, the data of the participants in $k - 1$ subsets were used for training, while the data of the participants in the remaining subset were used as a test set. The process was iterated k times and, at each iteration, a different subset was left out for test. In such a way, it is possible to test the approach over the whole corpus at disposition while still keeping separated training and test set. Since all the data of every participant is in one individual fold, the approach is *person independent*, i.e., the same person is never represented in both training and test set. This ensures that the

approach detects depression and does not simply recognize the voice of the participants.

The hyperparameters of the networks were set to standard predefined values and no attempts were made to find other, possibly better performing settings. For both MLP and RNN, the number of hidden neurons was set to $H = 32$, the learning rate was set to $\lambda = 10^{-2}$, and the number of training epochs was set to $T = 80$. The training was performed using the Adam optimizer [19] and the categorical crossentropy as a loss function [20]. Given that the training requires a random initialization, the experiments have were repeated $R = 10$ times and all performance metrics are reported in terms of average and standard deviation over the R repetitions.

Figure 1 shows the performance of the models as a function of the number r of slices, both at the level of individual slices and at the level of the participants (the one that actually matters from an application point of view). In both cases, the performance is compared with a random classifier that assigns a sample to class c with probability $p(c)$ corresponding to the *prior* of class c. The accuracy of such a classifier can be estimated as follows: $\alpha = \sum_{c \in \mathcal{C}} p(c)^2$. According to a two-tailed t-test, the performance of the approach is always better than such a value α (51.1% in the experiments of this work) to a statistically significant extent ($p < 0.01$ in all cases).

Overall, the plots in Figure 1 suggest that increasing the number of slices leads consistently to performance improvements. In particular, the difference between F1 Score when using only one slice and accuracy when using 128 slices is always statistically significant ($p < 0.01$ in all cases according to a two-tailed t-test). In this respect, the strategy of segmenting the data to increase the number of samples and, correspondingly, to tackle the lack of data, appears to be effective. There are several possible explanations behind such an observation. One is that, in line with the thin slices theory of behavior, people manifest their inner condition even in short time intervals (when $r = 128$, the average length of a slice is 1.44 seconds). Another one is that increasing the number of samples from 125 for $r = 1$ to 16,000 for $r = 128$ allows one to better train the models because the number of samples per parameter keeps increasing, even if samples extracted from the same recording are not necessarily independent (the topology of the models is always the same and, therefore, the number of parameters does not change) [21].

In the case of the participant level F1 Score (the one that actually matters from an application point of view), one further possible explanation is the application of the majority vote over the slices. In fact, the probability p_c of classifying a person correctly through the majority vote can be estimated as follows: $p_c = \sum_{n=r/2+1}^{r} \binom{r}{n} p^n (1-p)^{r-n}$, where n is the number of correctly classified slices, r is the total number of slices and p is the slice recognition accuracy (interpreted as the probability of correctly classifying a slice). Figure 2 shows how p_c changes with p and r and, in particular, it shows that increasing the number of slices tends to lead to higher probabilities to classify correctly an individual and, correspondingly, to higher F1 Scores.

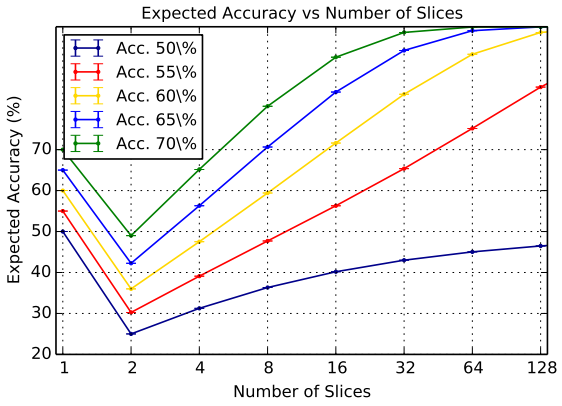

Fig. 2. The plot shows the probability of correctly classifying an individual through a majority vote over the slices when the number r of these later changes. Every plot corresponds to a different accuracy over individual slices.

5. CONCLUSIONS

Overall, the main indication of the experiments is that the performance can keep improving by increasing the number r of slices (see Figure 1). The limit to such a trend is the length of the slices that is inversely proportional to r. When such a length is not sufficient to extract enough feature vectors or, more generally, enough depression-relevant information, the advantages shown in the experiments of this work are no longer possible. In the case of this work, such a limit was reached at $r = 128$, given that the length of many recordings was not sufficient to move to $r = 256$. The only way to overcome such a limitation is to make the recordings longer by posing, e.g., more questions or by involving the participants in activities that require more time. However, such an approach might conflict with the difficulty of severely depressed individuals in interacting with people for extended periods of time.

The experiments performed until now were based on a "blind" segmentation, meaning that every slice was progressively split into two parts of the same length. Future work will focus on approaches aimed at performing a more intelligent segmentation of the data (e.g., by extracting turns or sentences). In alternative, it is possible to extract the emotional content of the slices and retain only those that correspond to negative emotions, well known to convey depression-relevant information [3].

6. REFERENCES

[1] N. Aloshban, A. Esposito, and A. Vinciarelli, "What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech," *Cognitive Computation (to appear)*, 2021.

[2] WHO Document Production Services, "Depression and other common mental disorders," World Health Organization, Tech. Rep., 2017.

[3] C. Irons, *Depression*. Palgrave, 2014.

[4] M. Marcus, T. Yasamy, M. Ommeren, D. Chisholm, and S. Saxena, "Depression: A global health concern," World Federation for Mental Health, Tech. Rep., 2012.

[5] N. Ambady, F. Bernieri, and J. Richeson, "Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream," in *Advances in Experimental Social Psychology*. Elsevier, 2000, vol. 32, pp. 201–271.

[6] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[7] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[8] 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 Communication*, vol. 71, pp. 10–49, 2015.

[9] Z. Huang, J. Epps, D. Joachim, and M. Chen, "Depression detection from short utterances via diverse smartphones in natural environmental conditions." in *Proceedings of Interspeech*, 2018, pp. 3393–3397.

[10] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. Cohn, "Cross-cultural depression recognition from vocal biomarkers." in *Proceedings of Interspeech*, 2016, pp. 1943–1947.

[11] L. Andrade, J. J. Caraveo-Anduaga, P. Berglund, R. V. Bijl, R. D. Graaf, W. Vollebergh, E. Dragomirecka, R. Kohn, M. Keller, R. C. Kessler *et al.*, "The epidemiology of major depressive episodes: results from the international consortium of psychiatric epidemiology (icpe) surveys," *International journal of methods in psychiatric research*, vol. 12, no. 1, pp. 3–21, 2003.

[12] J. Garber, C. Gallerani, and S. A. Frankel, "Depression in children," in *Depression in children*, I. Gotlib and C. Hammen, Eds. The Guilford Press, 2009, pp. 405–443.

[13] R. Kessler and E. Walters, "Epidemiology of DSM-III-R major depression and minor depression among adolescents and young adults in the national comorbidity survey," *Depression and Anxiety*, vol. 7, no. 1, pp. 3–14, 1998.

[14] F. McDougall, F. Matthews, K. Kvaal, M. Dewey, and C. Brayne, "Prevalence and symptomatology of depression in older people living in institutions in england and wales," *Age and Ageing*, vol. 36, no. 5, pp. 562–568, 2007.

[15] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[16] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[17] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer Verlag, 2012.

[18] E. Charniak, *Inreoduction to Deep Learning*. MIT Press, 2018.

[19] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *CoRR*, vol. abs/1103.0398, 2011. [Online]. Available: http://arxiv.org/abs/1103.0398

[21] F. Camastra and A. Vinciarelli, *Machine Learning for audio, image and video analysis: theory and applications*. Springer, 2015.