MULTI-LOCAL ATTENTION FOR SPEECH-BASED DEPRESSION DETECTION

Fuxiang Tao[1], *Xuri Ge*[1*], *Wei Ma*[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

This article shows that an attention mechanism, the Multi-Local Attention, can improve a depression detection approach based on Long Short-Term Memory Networks. Besides leading to higher performance metrics (e.g., Accuracy and F1 Score), Multi-Local Attention improves two other aspects of the approach, both important from an application point of view. The first is the effectiveness of a confidence score associated to the detection outcome at identifying speakers more likely to be classified correctly. The second is the amount of speaking time needed to classify a speaker as depressed or non-depressed. The experiments were performed over read speech and involved 109 participants (including 55 diagnosed with depression by professional psychiatrists). The results show accuracies up to 88.0% (F1 Score 88.0%).

Index Terms— Depression detection, read speech, attention mechanisms, multi-local attention

1. INTRODUCTION

According to the World Health Organization, depression was affecting 4.4% of the world's population before 2017 [1]. However, COVID-19 further aggravated such a situation and the number of patients is now estimated to be up to seven times greater [2]. For these reasons, the computing community is making major efforts towards the development of automatic depression detection technologies. This article contributes to these efforts by showing that an attention mechanism [3], the *Multi-Local Attention* (MLA), can improve the performance of Long Short-Term Memory Networks (LSTM) [4] in detecting depressed speakers. In addition, the approach improves the effectiveness of a confidence measure at identifying the speakers more likely to be classified correctly and, furthermore, it reduces the amount of data necessary to classify a speaker.

* Corresponding Author

The research leading to these results has received funding from the project ANDROIDS funded by the program V:ALERE 2019 Università della Campania "Luigi Vanvitelli", D.R. 906 del 4/10/2019, prot. n. 157264,17/10/2019. The work of Alessandro Vinciarelli was supported by UKRI and EPSRC through grants EP/S02266X/1 and EP/N035305/1, respectively.

Previous work shows that Support Vector Machines, fed with the average of all feature vectors extracted from a speech recording, detect depression with a performance that is limited, but above chance (e.g., F1 Score 68.2% over the same data as this work [5]). Such an observation suggests that vectors close to the average are likely to convey depression-relevant information and, therefore, should be "trusted" more. This is in line with recent work showing that attention mechanisms can be implemented by emphasizing feature vectors expected to carry task-relevant information [6, 7]. Correspondingly, this article proposes to emphasize feature vectors more similar to the local average, i.e., to the average in the *frame* they belong to (hence the name Multi-Local Attention). The frames are short segments extracted from a speech recording and the reason for considering the local averages is that they can better reflect possible changes over time.

To the best of our knowledge, this is the first attempt to use such an approach for depression detection. In fact, previous works using attention mechanisms for the task relied on including attention layers in Deep Networks, a less recent and more established approach (see, e.g., [8, 9, 10]). The detector in [8] is designed according to an encoder/decoder architecture and the decoder is an LSTM with an attention layer. The key-aspect of the experiments is that the parameters of the decoder, including the attention layer, can be obtained through transfer learning methodologies. A similar problem is addressed in [9], where the main issue is the increase in the number of parameters due to the attention layers. The proposed solution is, like in [8], the use of transfer learning. Finally, the experiments in [10] show that analyzing the weights of an attention gate allows one to test, in quantitative terms, whether it is language or paralanguage that influences most the outcome of a depression detection approach.

Overall, the experiments of this work involved 109 participants, including 55 diagnosed with depression by professional psychiatrists. All participants were asked to read the same fairy tale ("*The North Wind and the Sun*") and the results show that the proposed approach reaches an accuracy of 88.0% (F1 Score 88.0%). Most importantly, the experiments showed that MLA increases the accuracy of an LSTM from 84.3% to 88.0% (from 84.7% to 88.0% in terms of F1 Score), thus reducing the error rate by 23.5%.

The rest of this article is organized as follows: Section 2

Table 1. The table provides demographic information about the experiment participants. Acronyms M and F stand for male and female, respectively. Acronym L refers to Low education level (up to 8 years of study), while H corresponds to High education level (at least 8 years of study). The total for education level is 106 because 3 participants did not disclose information about their studies.

	Age	M	F	L	H
Control	47.6 ± 12.6	12	42	19	33
Depressed	47.6 ± 12.0	18	37	23	31
Total	47.6 ± 12.2	30	79	42	64

describes the data, Section 3 describes the approach used in the experiments, Section 4 reports on experiments and results, and the final Section 5 draws some conclusions.

2. THE DATA

The data used in this work were collected in five Mental Health Centers in Italy, where 109 persons were recorded while reading aloud a tale by Aesop ("*The North Wind and the Sun*"). The participants were involved on a voluntary basis and signed an informed consent formulated in accord with Italian and European privacy and data protection laws[1]. The main reason behind the choice of the fairy tale is that its text is simple and easy to understand (the Italian translation used in the experiments comes from a book for children). This ensures that the reading task is not an obstacle for people with lower education levels. The recordings were collected with a standard laptop microphone in the clinical consultation rooms of the Mental Health Centres involved in the study. In such a way, the recording conditions reproduce the normal setting in which depressed patients interact with doctors. Furthermore, the use of standard equipments limits costs and avoids changes in clinical practice.

The 109 participants include 55 persons diagnosed with depression by professional psychiatrists using the *Diagnostic and Statistical Manual of Mental Disorders 5* (DSM-5). The other 54 persons, referred to as *control* participants, were recruited among people that never experienced mental health issues. All participants are native Italian speakers. Table 1 shows the distribution of age, gender and education level. According to a two-tailed t-test, when comparing depressed and control participants, there is no statistically significant difference in age. Similarly, according to a χ^2 test, there are no statistically significant differences in terms of gender and education level distribution. This suggests that read speech differences between the two groups, if any, actually result from depression and not from other factors that can interplay with the way one reads.

[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.

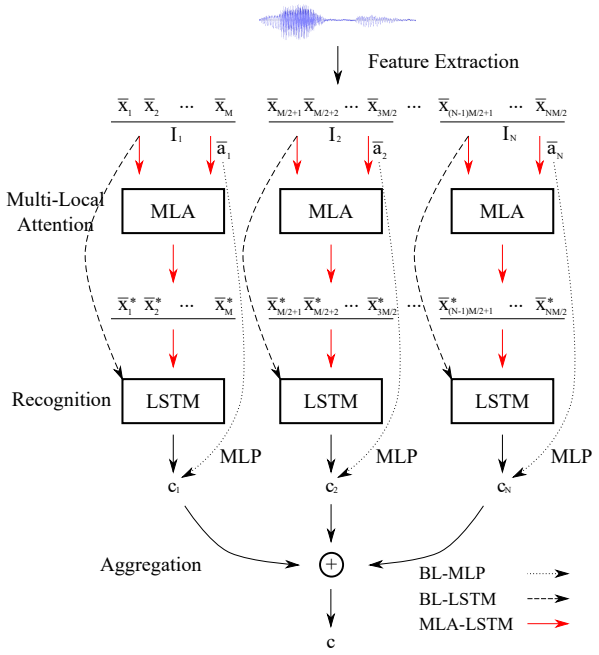

Fig. 1. The figure shows the main steps of the approach. Vectors $\vec{a}_k$ are the averages extracted from every frame, MLA stands for Multi-Local Attention, c_k is the classification outcome for frame I_k, the symbol $\oplus$ corresponds to the majority vote and c is the final classification outcome.

The number of female participants is 2.7 times greater than the number of male ones, in line with epidemiological observations showing that women tend to develop depression more frequently than men [11]. In a similar vein, there is a matching between the age range of the participants and the age range of people that tend to develop depression more frequently [10]. In this respect, the sample is expected to represent the general population of both depressed and non-depressed individuals.

The total duration of the recordings is 1 hour, 30 minutes and 58 seconds (the overall average is 50.1 seconds). The averages for depressed and control participants are 52.7 and 47.4 seconds, respectively. Such a difference is statistically significant ($p < 0.01$ according to a two-tailed t test) and this suggests that depressed people tend to read, on average, slower than control ones.

3. THE APPROACH

The proposed approach, referred to as MLA-LSTM, includes four main steps, namely *feature extraction*, *Multi-Local Attention*, *recognition* and *aggregation*. The approach is compared with two baselines that do not include the MLA step, referred to as BL-LSTM and BL-MLP, respectively (see Figure 1).

The goal of the feature extraction step is to convert the speech recordings into sequences of feature vectors $\vec{x}_k$. These are extracted at regular time steps of 10 ms from 25 ms long

analysis windows. Both values are standard in the literature and there was no attempt to find alternatives possibly leading to better results. The extraction was performed with OpenSMILE [12], a publicly available package widely applied in the literature. The features were designed for the *Interspeech 2009 Emotion Challenge* [13] and include *Root Mean Square Energy* (depressed speakers tend to show lower energy in speech [14]), *Mel-Frequency Cepstral Coefficients 1-12* (account for phonetic content and lead to good results in depression detection [15]), *Fundamental Frequency* (it has lower variability in depressed speakers [16]), *Zero-Crossing Rate* (it is another measure of the fundamental frequency), and *Voicing Probability* (it was shown to account for pauses that tend to be longer in depressed speakers [5]). The feature set was further expanded using the delta coefficients (difference between features in current and previous analysis window), thus reaching a dimension $D = 32$.

At the end of the feature extraction step, every recording is represented as a sequence $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$ of T feature vectors. Given that T is typically in the order of the thousands, the sequences are segmented into frames $I = \{I_k\}$ ($k \in \{1, \ldots, N\}$), where I_k includes $M = 128$ vectors. The frames start at regular steps of length $M/2$ and, therefore, two consecutive frames overlap by half of their vectors.

After the segmentation into frames, BL-LSTM and BL-MLP move to the recognition step (see dashed and dotted arrows in Figure 1), while MLA-LSTM performs the *Multi-Local Attention* step (see red arrows in Figure 1). This first calculates the average feature vector $\vec{a}_k$ of k-th frame and then transforms the other feature vectors of the same frame as follows: $\vec{x}_i^* = \vec{x}_i + \cos(\theta_i) \cdot \vec{x}_i$ ($i \in \{1, \ldots, M\}$), where $\cos\theta_i = \vec{a}_k\vec{x}_i/||\vec{a}_k||||\vec{x}_i||$ is the cosine of the angle between $\vec{a}_k$ and $\vec{x}_i$ ($\cos\theta_i$ is typically referred to as *cosine similarity*). The transform emphasizes the vectors that are more closely aligned with $\vec{a}_k$ by increasing their norm. A feature vector orthogonal or opposite to the average will be mapped into a null vector, while the average itself will see its norm multiplied by $\sqrt{2}$. Any other vector will be between such extremes.

After the MLA step, the proposed MLA-LSTM feeds the transformed vectors to an LSTM for the recognition step. The BL-MLP performs the recognition by feeding the averages extracted from the frames to a Multi-Layer Perceptron (see dotted arrows in Figure 1), while the BL-LSTM performs it by feeding the vectors of a frame to an LSTM. In all cases, the frame is assigned either to class *depressed* or to class *control*. Both baselines skip the MLA step before performing the recognition.

Given that there are multiple frames per recording, there are multiple classification outcomes too for all approaches. This makes it necessary an *aggregation* step that takes as input the N classification outcomes (N is the total number of frames in a recording) and performs a majority vote, i.e., it assigns a recording to the class its frames are most frequently assigned to (see Figure 1): $\hat{c} = \arg\max_{c \in \mathcal{C}} n(c)$,where $\mathcal{C}$ is the set of all possible classes (depression and control in the experiments of this work), $n(c)$ is the number of frames assigned to class c and $\hat{c}$ is the class assigned to the recording.

One of the main advantages of such an aggregation approach is that it is possible to define a confidence score s as follows: $s = n(\hat{c})/N$, where $n(\hat{c})$ is the number of frames assigned to the majority class, i.e., the class assigned to most frames in a recording. The main assumption behind such a definition is that the tendency to assign a greater fraction of frames to a given class should be associated to correct classification results.

Table 2. Recognition results in terms of Accuracy, Precision, Recall and F1 Score. The table includes not only the results obtained in this article, but also those obtained in previous studies involving the same speakers that participated in this work.

	Acc.	Prec.	Rec.	F1
[5]	84.5	84.5	84.6	84.5
[19]	77.0	74.0	80.0	77.0
[20]	67.6	71.7	72.2	72.3
BL-MLP	74.2±3.4	74.2±5.1	78.6±4.4	75.5±2.5
BL-LSTM	84.3±3.7	83.1±4.9	87.3±2.6	84.7±3.4
MLA-LSTM	**88.0±2.1**	**87.7±2.6**	**89.0±3.4**	**88.0±2.3**

4. EXPERIMENTS AND RESULTS

The experiments were performed according to a k-fold protocol ($k = 3$). The participants were first randomly split into k disjoint subsets and then $k - 1$ were used for training, while the remaining one was used for test. The process was iterated k times and, at each iteration, a different subset was used for test. Such a protocol is *person independent*, i.e., the same person never appears in both training and test set. This ensures that the proposed approach recognizes depression and not just the voice of the participants.

The number of hidden states in the LSTMs was set to 32, the learning rate to 10^{-3} and the number of training epochs to 300. The training was performed using the RMSProp optimizer [17] with categorical cross-entropy as a loss function [18]. Given that LSTMs require a random initialization, the experiments were repeated $R = 20$ times and, at each repetition, the weights were initialized differently. For such a reason, all performance metrics are reported in terms of average and standard deviation across the R repetitions. Parameters and training procedure were the same for both LSTM-based versions of the approach (BL-LSTM and MLA-LSTM). The number of hidden neurons in the MLP was 32 and the training process was the same as the LSTMs.

Table 2 shows the recognition results in terms of Accuracy, Precision, Recall and F1 Score. All approaches were compared to a random classifier assigning an unseen sample to class c with probability corresponding to the prior of c. The accuracy of such a classifier is $\hat{\alpha} = p(n)^2 + p(d)^2$, where $p(n)$

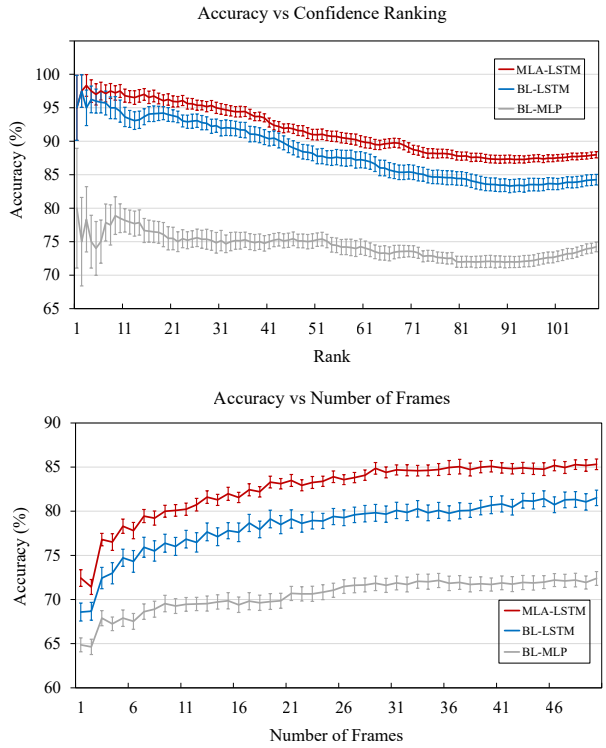

Fig. 2. The upper plot shows the accuracy obtained when considering only the speakers showing the r highest confidence scores. The lower plot shows the relationship between accuracy and the number of frames (50 is the maximum of frames that every participant has) used for depression detection. The vertical bars correspond to the standard error of the mean observed across R repetitions.

is the prior of class *control* and $p(d)$ is the prior of class *depressed* ($\hat{\alpha} = 50.0\%$ in this work). Precision, Recall and F1 Score correspond to the prior of the positive class (*depression* in these experiments) which is 50.4%.

According to a two-tailed t-test, all approaches outperform the random classifier to a statistically significant extent ($p < 0.001$ in all cases after Bonferroni correction). In addition, MLA-LSTM outperforms both BL-LSTM and BL-MLP and the difference is statistically significant ($p < 0.001$ according to a two-tailed t-test). Table 2 further shows that MLA-LSTM achieves results comparable to those obtained in previous studies involving the same speakers considered in this work. Overall, the results appear to confirm that vectors closer to the local average should be "trusted" more, the key-assumption behind the proposed Multi-Local Attention.

Section 3 shows that the approach associates a confidence score s to its classification outcomes. Therefore, it is possible to measure the performance of the approach when taking into account only the participants corresponding to the r greatest values of s, i.e., when taking into account only the top r ranking speakers in terms of the confidence score. The upper plot of Figure 2 shows how the accuracy changes as a function of r (the vertical bars correspond to the standard error of the mean over the R repetitions of the experiment). The curves show that, after the application of the Multi-Local Attention, the accuracy improves at any point of the ranking ($p < 0.001$ according to a two-tailed t-test). This means that the Multi-Local Attention improves the effectiveness of the confidence score at identifying speakers more likely to be classified correctly. In this way, it is possible to accept the response of the system when s is high enough, while requesting the attention of a doctor when s is too small. As a consequence, the workload for the doctors can be reduced while keeping the accuracy of diagnosis approach high enough.

The lower plot of Figure 2 shows how the accuracy changes as a function of the number of frames used to perform the classification. The comparison between the two curves shows that MLA-LSTM outperforms the other approaches to a statistically significant extent ($p < 0.001$ according to a two-tailed t-test) for every number of frames. Given that increasing the number of frames means to increase the amount of speech time used to classify a speaker, this means that MLA-LSTM can reach a predefined accuracy earlier than the baselines. This is important because depression patients tend to speak less and find it difficult to keep speaking for long time. Therefore, it is desirable to perform the classification as early as possible.

5. CONCLUSIONS

This work shows that the performance of a depression detector can improve after the application of the Multi-Local Attention, an attention mechanism emphasizing the input data expected to carry task-relevant information. Furthermore, the results show that the detector improves under two other important respects. The first is that the confidence measure accompanying detection outcomes becomes more effective at identifying correctly classified speakers. This is important because it makes it possible to identify participants for which the actual condition (depression or the lack of it) is evident enough to be recognized automatically. In this way, doctors can concentrate on difficult and ambiguous cases. The second is that the detector reaches its best accuracy by using less speech and this is important because depression speakers find it challenging to speak long time.

Another positive effect of the Multi-Local Attention is that the fraction of correctly classified frames tends to increase (hence the higher confidence scores for the correctly classified speakers). Such an observation is consistent with the thin slices theory (the tendency of people to manifest their inner state through short behavioral displays) [21] and it is the probable explanation behind the effectiveness of the majority vote. Given that such an aggregation approach is basic, future work will focus on the attempt to develop better methodologies to combine the classification outcomes obtained at the level of individual frames.

6. REFERENCES

[1] AA.VV., "Depression and other common mental disorders," Tech. Rep., World Health Organization, 2017.

[2] J. Bueno-Notivol, P. Gracia-García, B. Olaya, I. Lasheras, R. López-Antón, and J. Santabárbara, "Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies," *International Journal of Clinical and Health Psychology*, vol. 21, no. 1, pp. 100196, 2021.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] F. Tao, A. Esposito, and A. Vinciarelli, "Spotting the traces of depression in read speech: An approach based on computational paralinguistics and social signal processing," *Proceedings of Interspeech*, pp. 1828–1832, 2020.

[6] X. Ge, F. Chen, C. Shen, and R. Ji, "Colloquial image captioning," in *2019 IEEE International Conference on Multimedia and Expo*. IEEE, 2019, pp. 356–361.

[7] X. Ge, P. Wang, H. Han, J. M. Jose, Z. Ji, Z. Wu, and X. Liu, "Local global relational network for facial action units recognition," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2021, pp. 01–08.

[8] A. Harati, E. Shriberg, T. Rutowski, P. Chlebek, Y. Lu, and R. Oliveira, "Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7273–7277.

[9] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Hierarchical attention transfer networks for depression assessment from speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7159–7163.

[10] 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*, vol. 14, pp. 1585–1598, 2022.

[11] L. Andrade, J.J. Caraveo-Anduaga, P. Berglund, R.V. Bijl, R. De Graaf, W. Vollebergh, E. Dragomirecka, R. Kohn, M. Keller, and R.C. Kessler, "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] 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.

[13] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proceedings of Interspeech*, 2009.

[14] B. Schuller and A. Batliner, *Computational paralinguistics*, John Wiley & Sons, 2013.

[15] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Communication*, vol. 75, pp. 27–49, 2015.

[16] 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, no. 1, pp. 50–64, 2007.

[17] T. Tieleman, G. Hinton, et al., "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[18] R.Y. Rubinstein and D.P. Kroese, *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, Springer, 2004.

[19] F. Scibelli, G. Roffo, M. Tayarani, L. Bartoli, G. De Mattia, A. Esposito, and A. Vinciarelli, "Depression speaks: Automatic discrimination between depressed and non-depressed speakers based on nonverbal speech features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 6842–6846.

[20] R. Alsarrani, A. Esposito, and A. Vinciarelli, "Thin slices of depression: Improving depression detection performance through data segmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6257–6261.

[21] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis.," *Psychological Bulletin*, vol. 111, no. 2, pp. 256–274, 1992.