INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection

Fuxiang Tao[1], *Anna Esposito*[2], *Alessandro Vinciarelli*[1]

[1]University of Glasgow, UK
[2]Univeristà degli Studi della Campania Luigi Vanvitelli, Italy

Fuxiang.Tao@glasgow.ac.uk, Anna.Esposito@unicampania.it,
Alessandro.Vinciarelli@glasgow.ac.uk

Abstract

This paper presents the Androids Corpus, a new benchmark for speech-based automatic depression detection. The corpus is a collection of 228 recordings uttered by 118 native Italian speakers, including 64 who were diagnosed with depression by professional psychiatrists. Out of the 228 recordings, 112 contain read speech (all speakers read the same text) and 116 contain spontaneous speech (all speakers answer the same questions posed by an interviewer). For 110 speakers, including 58 diagnosed with depression, the corpus includes both read and spontaneous speech samples. Overall, the total duration of the material is 1 hour, 33 minutes and 49 seconds for read speech and 7 hours, 24 minutes and 22 seconds for spontaneous speech. Besides the data, the corpus includes experimental protocols that can be replicated, thus ensuring reproducibility of the experiments performed over it and comparison of the results.
Index Terms: depression detection, benchmark, computational paralinguistics

1. Introduction

Depression is one of the most common mental health issues. In 2015, the World Health Organization estimated that 4.4% of the world's population was affected by the pathology (roughly 300 million people at the time) [1]. Since then, an analysis of 19 countries showed that the consumption of antidepressants doubled, thus suggesting that the number of depression patients keeps increasing [2]. As a consequence, mental health services are under pressure to provide appropriate responses. However, because depression diagnosis is a long and time consuming process, many cases remain undetected or are addressed months, if not years, after the first onset of the pathology [3].

The computing community tried to address the problems above by making substantial efforts towards the development of depression detection approaches (see, e.g., [4, 5, 6, 7, 8, 9]). The results are encouraging and some works claim to achieve performances similar to those of General Practitioners [10], the first line of intervention against depression in many healthcare systems. However, the lack of data remains a bottleneck that prevents the field from progressing. One reason is that advanced approaches, especially since the advent of deep networks, require increasingly larger amounts of data to be trained. The second is that, without multiple corpora, it is not possible to test the robustness of an approach with respect to changes in setting, sensors, people, language and any other factors that can characterize a corpus.

For the reasons above, this article presents the Androids Corpus, a new, publicly available benchmark for speech-based depression detection. The corpus was designed to provide complementary and distinctive opportunities with respect to other available corpora (see, e.g., [11, 12, 13, 14, 15, 16]):

- The distinction between depressed and non-depressed speakers was made by professional psychiatrists and not through the administration of self-assessment questionnaires. This means that the data allows one to investigate depression detection and not prediction of scores obtained through questionnaires, known to be affected by multiple biases [17].
- The data were collected with the microphone of a laptop in the mental health centers where the depression patients are treated. Such an *in-the-wild* setting corresponds to the situation in which doctors and depressed participants meet for their therapeutic interactions. This allows one to develop an approach in conditions representative of a real-world application environment.
- For 110 out of its 118 participants, the corpus includes both *read* and *spontaneous* speech samples. This allows one to investigate the interplay between depression and type of speech.
- The populations of depressed and non-depressed participants (64 and 54 individuals, respectively) have the same distribution in terms of age, gender and education level. This limits speaking differences resulting from factors other than depression.
- In the case of spontaneous speech (collected during interviews), the data includes a manual segmentation into turns. This allows one to investigate conversational dynamics and turn-taking related features (e.g., turn length distribution, speaking time, etc.).
- The corpus includes reproducible experimental protocols (the split of the participants into subsets to be used for a k-fold setup). This allows rigorous comparisons across multiple results obtained over the data (including the baseline results provided in this paper).
- The corpus provides an OpenSmile [18] configuration file allowing one to extract a standard feature set from the speech recordings. This improves the reproducibility of the results obtained over the data.
- The data is in Italian, a language not widely represented in other publicly available corpora. This allows one to test cultural effects and to investigate whether approaches developed in one language can extend to other languages.

The use of the corpus is likely to lead to other, non anticipated research directions. Therefore, the list above should not be considered exhaustive.

The rest of this article is organized as follows: Section 2 describes the corpus in detail, Section 3 describes the baseline results and the final Section 4 draws some conclusions.

10.21437/Interspeech.2023-894

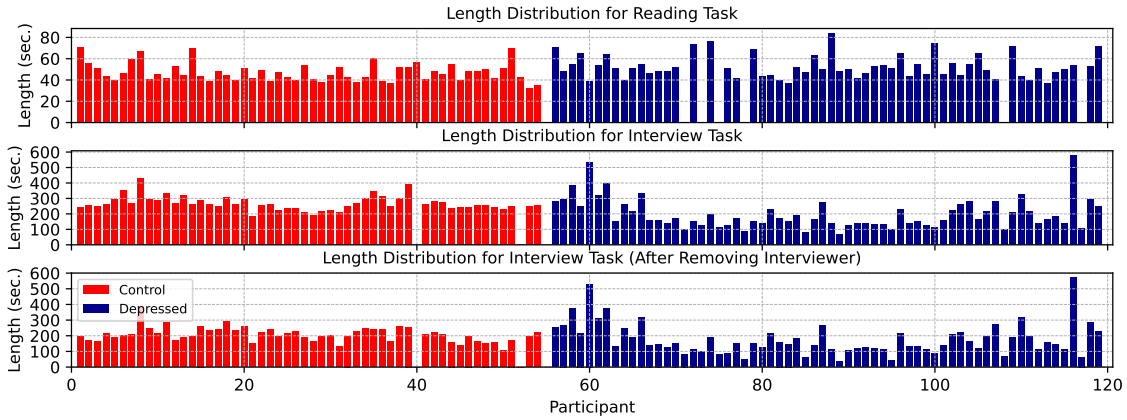

Figure 1: *Length distribution across participants. The top bar chart shows the length of the RT recordings for each participant, the middle one shows the same information for the IT recordings and the bottom one does the same for the IT data after removing the turns of the interviewer. Missing bars correspond to participants that did not perform one of the tasks.*

2. The Androids Corpus

The corpus includes samples from 118 native Italian speakers, including 64 who were diagnosed with depression (the *depression participants* hereafter) and 54 who never experienced mental health issues (the *control participants* hereafter). Every speaker was invited to perform two tasks referred to as *Reading Task* (RT) and *Interview Task* (IT). The first consists of reading a short fairy tale written by Aesop, namely "*The Wind of the North and the Sun*" (the text is provided in the corpus). The main motivation behind the choice of the text, extracted from a textbook for children, is that it is simple. Therefore, it can be read easily irrespective of one's education level. The IT corresponds to answering a few questions about everyday life (e.g., "*What did you do last weekend*") posed by an interviewer instructed to speak as little as possible. The goal of RT and IT is to produce read and spontaneous speech samples, respectively.

The psychiatrists involved in the data collection diagnosed the patients using the *Diagnostic and Statistical Manual of Mental Disorders 5* (DSM-5). The distribution across specific pathologies is as follows: 22 cases of Major Depressive Disorder, 15 cases of bipolar disorder in depressive phase or with last depressive episode, 8 cases of reactive depression, 7 cases of endo-reactive depression, 5 cases of anxiety-depressive disorder and 1 case of persistent depressive disorder. No specific pathology was indicated for the remaining 6 depression participants. The distribution shows that there is enough diversity to cover all aspects of depression and not just some of its forms.

All participants were involved on a voluntary basis and they all signed an informed consent letter. The data collection was performed according to the ethical regulations of countries and institutions involved in the work. In the case of the depression group, the participants were recorded while being treated, meaning that they were actually experiencing the pathology. This makes it likely that depression left traces in their speech samples. The control participants were recruited through a word of mouth process and they were selected to match the depression group in terms of age, gender and education level distribution (see below).

Table 2 provides demographic information about the participants. Most of them (110 out of 118) participated in both RT and IT, while some of them participated only in one of the two

Table 1: *Participant distribution across tasks. Acronyms RT and IT stand for* Reading Task *and* Interview Task*, respectively.*

Group	Only RT	Only IT	RT and IT	Total
Control	2	0	52	54
Depression	0	6	58	64
Total	2	6	110	118

Table 2: *Demographic information. Acronyms* F *and* M *stand for Female and Male, respectively. Acronyms* L *and* H *stand for Low (8 years of study at most) and High (at least 13 years of study) education level, respectively. The sum over the education level columns does not correspond to the total number of participants (118) because 2 of these did not provide details about their studies.*

Task		Age	M	F	L	H
RT	Control	47.1 ± 12.8	12	42	19	35
	Depression	47.4 ± 11.9	20	38	25	32
	Total	47.2 ± 12.3	32	80	44	67
IT	Control	47.3 ± 12.7	11	41	19	33
	Depression	47.5 ± 11.6	21	43	29	33
	Total	47.4 ± 12.1	32	84	48	66
Total	Control	47.3 ± 12.7	11	41	19	33
	Depression	47.4 ± 11.9	20	38	25	32
	Overall	47.3 ± 12.2	31	79	44	65

(see Table 1). For this reason, Table 2 provides the information not only for the whole set of participants, but also for the two tasks separately. According to a χ^2 there is no difference between depression and control participants in terms of gender and education level distribution. This applies to the corpus as a whole and to the subsets of participants who were involved in each task. The same can be said for age distribution, according to a two-tailed t-test. Overall, this means that speech differences between depression and control participants depend mostly on the pathology and not on other factors that might have an impact on speech. Female participants are roughly 2.5 times more than male ones because, according to epidemiological observations, women tend to experience depression more frequently than men roughly in the same proportion [19, 20].

Figure 1 shows the length of the recordings for each individual. The total duration of the RT recordings is 1 hour, 33 minutes and 49 seconds, while the average duration is 50.3 $\pm$ 10.3 seconds. When considering separately depression and control participants, the averages are 52.9 $\pm$ 10.9 and 47.4 $\pm$ 8.8 seconds, respectively. The difference is statistically significant ($p < 0.01$ according to a two-tailed t-test), in line with previous studies showing that depression patients tend to read slower [21]. In the case of the IT recordings, the total duration is 7 hours, 24 minutes and 22 seconds (the overall average and standard deviation are 229.8 $\pm$ 86.6 seconds). According to a two-tailed t-test, there is no statistically significant difference between depression and control participants (the averages are 198.8 $\pm$ 99.2 and 268.0 $\pm$ 45.3 seconds, respectively). In the IT recordings, the total duration of the turns for the participants is 6 hours, 8 minutes and 21.8 seconds, for an average of 190.5 $\pm$ 84.1 seconds. The average lengths for depression and control participants are 176.4 $\pm$ 103.1 and 207.9 $\pm$ 47.3 seconds, respectively.

2.1. Content of the Distribution

The corpus can be downloaded at the following link: `https://github.com/androidscorpus/data`. The distribution includes the following material:

- RT recordings (112 items);
- IT full recordings (116 items);
- IT participant turn recordings (874 items);
- Segmentation of IT recordings into turns (1 item);
- Reproducible experimental protocol (5 lists corresponding to the folds to be used in a 5-fold protocol);
- OpenSmile configuration file (1 item);
- Readme file (1 item).

The recordings are available in a standard *Waveform Audio File Format* (the extension is "*wav*"). The file naming convention was designed to provide all information available about a speaker. In particular, the name of the files is of the form *nn_XGmm_t.wav*, where *nn* is a unique integer identifier such that, *in a given group*, files with the same *nn* contain the voice of the same speaker (there is a trailing 0 for numbers lower than 10), *X* is an alphabetic character corresponding to the speaker's condition (*P* for depression patient and *C* for control), *G* is an alphabetic character that stands for the speaker's gender (*M* for male and *F* for female), *mm* is a two-digits integer number corresponding to the speaker's age, and *t* is an integer number between 1 and 4 accounting for the education level (1 corresponds to primary school and 4 corresponds to university). The letter *X* was used for the 2 participants who did not provide information about this aspect. There is no indication of the task because recordings corresponding to RT and IT are stored in different directories.

The segmentation files for the IT recordings are in *comma separated values* format (extension *csv*). The files provide start and end times of every turn of the participants, thus allowing one to eliminate the speech of the interviewer. The segmentation was performed manually so that the files can be used as a ground truth for automatic speaker diarization approaches.

The goal of the experimental protocols available in the distribution is to allow interested researchers to perform rigorous comparisons with the baseline approaches presented in Section 3. More generally, the protocols will allow rigorous comparisons between all results obtained using the corpus. The protocols include lists of recordings corresponding to the folds used in the experiments of Section 3. In a similar vein, The OpenSmile configuration file allows one to extract the same features used in the baseline approaches from the recordings. Using the same features will allow one to disentangle the effects of different features and different models.

The readme file is written in ASCII text (it can be read with any editor) and it provides information about the distribution (including the details above).

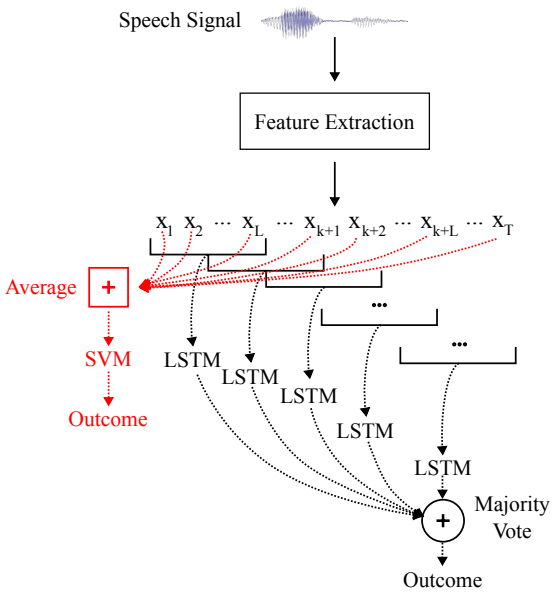

Figure 2: *Baseline approaches. The diagram shows the two baseline approaches used in the experiments. The symbol* $\boxplus$ *corresponds to the average, while the symbol* $\oplus$ *corresponds to the majority vote.*

3. Baseline Approaches and Results

The goal of the baseline approaches is to provide an initial set of results that can be used as a term of comparison in future works based on the corpus data. All experiments were performed using the experimental protocols available in the corpus distribution (see Section 2).

3.1. The Approaches

The two approaches (see Figure 2) include two main steps, namely *feature extraction* and *depression detection*. The first step is the same for both baseline approaches and it starts by segmenting the recordings into analysis windows of length 25 ms that start at regular time steps of 10 ms. Both values are standard in the literature and were set a-priori (no attempt was made to find values possibly leading to better performances). Every analysis window is converted into a feature vector $\boldsymbol{x}$ with OpenSmile [18]. The feature set is as follows:

- *Root Mean Square of the Energy* (1 feature): it provides a measure of how loud someone speaks and it was shown to account for depression in the literature [22, 23];
- *Mel-Frequency Cepstral Coefficients 1-12* (12 features): they provide a measure of the phonetic content of the speech signal and they were shown to be effective in a wide spectrum of applications aimed at inferring social and psychological

Table 3: *Depression detection results. The table shows the results obtained over RT and IT.*

	Acc.	Prec.	Rec.	F1
Reading Task				
Rand.	50.1	51.8	51.8	51.8
BS1	69.6±5.3	73.6±19.1	68.8±12.0	68.4±7.7
BS2	84.4±1.1	84.5±2.1	85.6±2.8	83.7±1.1
Interview Task				
Rand.	50.5	55.2	55.2	55.2
BS1	73.3±10.6	73.5±16.1	74.5±13.2	73.6±13.6
BS2	83.9±1.3	85.8±3.1	86.1±2.7	84.7±0.9

information from speech, including depression [24];

- *Fundamental Frequency* (1 feature): it is the frequency that carries the highest energy in the signal, known to convey depression-relevant information [25, 26, 27];
- *Zero-Crossing Rate* (1 feature): it is an alternative estimate of the fundamental frequency and it was shown to lead to 80% accuracy in predicting whether a naive listener considers a speaker depressed [28];
- *Voicing probability* (1 feature): it accounts for the probability of an analysis window corresponding to the emission of voice and it was shown to be effective for the inference of affective states [24, 29].

The feature set above was enriched with the regression coefficients, thus reaching a dimensionality $D = 32$. The feature set was initially designed for the Interspeech Emotion Recognition Challenge in 2009 [30] and it was used in a wide spectrum of applications, especially when it comes to the inference of social and psychological information from speech.

At the end of the feature extraction process, every recording is converted into a sequence of vectors $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$, where T is the total number of vectors (proportional to the length of the recording). The simplest baseline, referred to as *BS1* hereafter, corresponds to calculating the average of the $\boldsymbol{x}_k$ vectors and giving it as input to a Support Vector Machine with linear kernel (see red components in Figure 2).

A second baseline (see Figure 2), referred to as *BS2* hereafter, segments the X into subsequences of length $L = 128$, the *frames*, that start at regular steps of length $L/2$ (two consecutive frames overlap by half of their vectors). Each frame is given as input individually to a Long Short-Term Memory Network (LSTM) and assigned to one of the two possible classes (depression and control). The LSTM has one hidden layer with 32 hidden states (both parameters were set a-priori). Given that there are multiple frames per recording, the individual frame classification outcomes are aggregated through a majority vote (the recording is assigned to the class its frames are most frequently assigned to).

All baselines are compared to a random classifier that assigns each recording to class c with probability equal to the prior $p(c)$ of c. The accuracy of such a classifier is $\sum_{c \in \mathcal{C}} p(c)^2$, where $\mathcal{C}$ is the set of all classes. Precision, Recall and F1 Score of the random classifier are all equal to the prior of the positive class (*depression* in the experiments of this work).

3.2. Experiments and Results

All experiments were performed according to a k-fold protocol ($k = 5$). The data was split into k disjoint folds and the experiments were repeated k times. At every repetition, a different fold was used as a test set. The remaining folds were used for training (*no validation is necessary because all hyperparameters were set a-priori and no attempt was made to find values possibly leading to better alternatives*). The data of a given participant were all in one fold and, therefore, the experiments are *person independent*, meaning that the same speaker is never represented in both training and test set. In this way, it is possible to ensure that the approaches recognize depression and not the identity of the participants. The composition of the folds is available in the corpus distribution so that the experiments can be replicated (see Section 2).

The LSTMs were trained using the Adam [31] optimizer with cross entropy [32] as a loss function. The learning rate was set to λ=0.001 and the number of training epochs was T=100. At each of the k=5 iterations in the k-fold, the weights of the LSTMs were initialized differently through a random process. For this reason, all results are reported in terms of average and standard deviation over the k folds. The BS1 were implemented with scikit-learn (version in 0.23.2) [33], while the BS2 were implemented with PyTorch (version in 1.13.1+cu116) [34].

Table 3 shows the depression detection results in terms of Accuracy, Precision, Recall and F1 Score. All approaches improve over the random baseline to a statistically significant extent ($p < 0.01$ according to a two-tailed t-test). The results of the table can serve as a term of comparison for other approaches using the same data and, possibly, the same experimental protocol.

4. Conclusions

This article presented the Androids Corpus, a new publicly available benchmark for speech-based depression detection. Besides describing the data, the article provides reproducible results that can serve as a term of comparison (the corpus distribution includes information necessary to implement the same experimental protocols as those used in Section 3). In addition to depression detection, the corpus can be used to address different problems of interest, including the interplay between read and spontaneous speech, the relationship between depression and conversational dynamics, etc. In this respect, the Androids Corpus was designed to become an opportunity for expanding research on automatic approaches dealing with speech and depression.

The main limitation of the Corpus is that it is not multimodal. However, it is possible to transcribe it manually or automatically and this still gives the opportunity to develop multimodal approaches based on language and paralanguage. Another limitation is that the Corpus does not allow one to investigate how the pathology progresses over time because every participant is recorded only once. On the other hand, to the best of our knowledge, no publicly available corpora allow one to address such a problem. Despite the limitations above, the dataset has the advantage of being recorded in a real-world setting (the hospitals involved in the data collection) with standard laptop microphones. In this respect, the data can be considered representative of real-world conditions in which a depression detection would be used.

Acknowledgments The research leading to these results has received funding from the project ANDROIDS funded by the program V:ALERE 2019 Universit'a 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.

5. References

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

[2] L. Foulkes, *What mental illness really is*. Vintage, 2021.

[3] P. Wang, M. Angermeyer, G. Borges, R. Bruffaerts, W. Chiu, G. De Girolamo, J. Fayyad, O. Gureje, J. Haro, Y. Huang, R. Kessler, V. Kovess, D. Levinson, N. Yoshibumi, M. Oakley Brown, J. Ormel, J. Posada-Villa, S. Aguilar-Gaxiola, J. Alonso, S. Lee, S. Heeringa, B. Pennell, S. Chatterji, and T. Bedirhan Üstün, "Delay and failure in treatment seeking after first onset of mental disorders in the world health organization's world mental health survey initiative," *World Psychiatry*, vol. 6, no. 3, pp. 177–185, 2007.

[4] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 239–253, 2021.

[5] N. Cummins, V. Sethu, J. Epps, J. Williamson, T. Quatieri, and J. Krajewski, "Generalized two-stage rank regression framework for depression score prediction from speech," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 272–283, 2020.

[6] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, p. 103107, 2022.

[7] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," in *Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 283–288.

[8] Y. Yang, C. Fairbairn, and J. Cohn, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, 2012.

[9] 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.

[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] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the ACM International Workshop on Audio/visual Emotion Challenge*, 2013, pp. 3–10.

[12] M. Valstar, B. Schuller, J. Krajewski, J. Cohn, R. Cowie, and M. Pantic, "AVEC 2014 – The three dimensional affect and depression challenge," in *Proceedings of the ACM International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 1–9.

[13] 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.

[14] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 3–9.

[15] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, and S. Marsella, "The Distress Analysis Interview Corpus of human and computer interviews," in *Proceedings of the Language Resources and Evaluation Conference*, 2014.

[16] H. Cai, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, and Q. Zhao, "Modma dataset: a multi-modal open dataset for mental-disorder analysis," *arXiv preprint arXiv:2002.09283*, 2020.

[17] D. Paulhus and S. Vazire, "The self-report method," in *Handbook of Research Methods in Personality Psychology*, R. Robins, R. Fraley, and R. Krueger, Eds. Gilford, 2007, pp. 224–239.

[18] 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.

[19] L. Andrade, J. Caraveo-Anduaga, P. Berglund, R. Bijl, R. De Graaf, W. Vollebergh, E. Dragomirecka, R. Kohn, M. Keller, R. Kessler, N. Kawakami, C. Kiliç, D. Offord, T. Bedirhan Ustun, and H.-U. Wittchen, "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.

[20] R. Kessler, P. Berglund, O. Demler, R. Jin, D. Koretz, K. Merikangas, A. Rush, E. Walters, and P. Wang, "The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R)," *Journal of the American Medical Association*, vol. 289, no. 23, pp. 3095–3105, 2003.

[21] 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." in *Proceedings of Interspeech*, 2020, pp. 1828–1832.

[22] L. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2011.

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

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

[25] 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, pp. 829–837, 2000.

[26] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, and D. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

[27] T. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Proceedings of Interspeech*, 2012.

[28] A. Nilsonne and J. Sundberg, "Differences in ability of musicians and nonmusicians to judge emotional state from the fundamental frequency of voice samples," *Music Perception*, pp. 507–516, 1985.

[29] C. Gobl and A. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

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

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

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

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.