Noise Robust Recognition of Depression Status and Treatment Response from Speech via Unsupervised Feature Aggregation

Maurice Gerczuk[1], Shahin Amiriparian[1], Alexander Kathan[1], Jonathan Bauer[2], Matthias Berking[2], Björn W. Schuller[1,3]

***Abstract*— In the presented work, we utilise a noisy dataset of clinical interviews with depression patients conducted over the telephone for the purpose of depression classification and automated detection of treatment response. Compared to most previous studies dealing with depression recognition from speech, our data set does not include a healthy group of subjects that have never been diagnosed with depression. Furthermore, it contains measurements at different time points for individual subjects, making it suitable for machine learning-based detection of treatment response. In our experiments, we make use of an unsupervised feature quantisation and aggregation method achieving 69.2% Unweighted Average Recall (UAR) when classifying whether patients are currently in remission or experiencing a major depressive episode (MDE). The performance of our model matches cutoff-based classification via Hamilton Rating Scale for Depression (HRSD) scores. Finally, we show that using speech samples, we can detect response to treatment with a UAR of 68.1%.**

I. INTRODUCTION

With a twelve-month prevalence of $12.9\,\%$, Major Depressive Disorder (MDD) is one of the major mental health diseases in the general population [1], associated with a range of negative effects [2]. These can range from a permanently low mood and decreased motivation to thoughts of suicide. In addition to a deterioration in the quality of life for those affected, MDD is accompanied by far-reaching consequences for their social circle and a broader economic burden for society [3]. In high-income countries, depression is even estimated to be the leading cause of disability-adjusted life years by 2030 [4], resulting in a great need for early diagnosis and treatment.

Possible therapies for people suffering from MDD include acute treatment such as psychotherapy [5], psychopharmacology [6] or bright light therapy [7]. Even though these methods are effective, there are several challenges encountered in practice. On the one hand, many patients experience a relapse, resulting in rates up to $54\,\%$ after psychotherapy [8] and $68\,\%$ after pharmacotherapy [9]. On the other hand, the existing demand for treatment already cannot be met. Even in high-income countries, less than $30\,\%$ of people suffering from depression receive appropriate treatment [10]. Symptom monitoring and automatic depression detection can make a valuable contribution at this point to adjust the type and intensity of treatments according to the current symptom severity and thereby optimise the effectiveness of existing interventions [11], [12].

For a long time, self-assessment or clinician-rated questionnaires, such as the Patient-Health-Questionnaire (PHQ)-9 or the HRSD, were primarily employed to assess depression and its severity. These questionnaires are well-correlated with symptoms and thus, often form prediction targets for machine learning algorithms of digital health applications. However, a major disadvantage of these questionnaires is that they can be time consuming (HRSD) and further have to be regularly repeated to serve as effective monitoring tools. Instead, methods are needed that allow for automated continuous monitoring of symptoms.

Audio represents another promising data modality as it can be collected easily and inexpensively at any time, enabling an early detection [13]. In recent years, handcrafted features such as the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [14] were used, which was also successfully applied in other tasks including speech emotion recognition [15]. Other approaches rely on deep feature representations created using deep neural networks such as DeepSpectrum [16]. Furthermore, [17] extended the audio modality using both, audio and video data streams for predicting depression. However, transferring these approaches into real-world applications is associated with several challenges. One of them is that their usage in real-world scenarios is often accompanied by noisy conditions which results in a decreased performance. To solve this challenge, there is a need for approaches that can deal with noisy environments.

In this work, we explore a noise-robust depression detection and forecasting framework using solely speech data. In doing so, we rely on ComParE2016 audio functionals in combination with DeepSpectrum. In addition, we perform a feature aggregation approach using Bag-of-Audio-Words (BoAW) [18] and Bag-of-Deep-Features (BoDF) [19] quantisation, resulting in a depression detection framework which is more robust towards noisy environments.

II. DATA

We utilise a dataset from a previous psychological study which investigated the efficacy of web-based online trainings compared to traditional treatment – the protocol for its development can be found in [20]. It contains longitudinal data with follow-up assessments 3, 6, and 12 months after

[1]Chair of Embedded Intelligence for Healthcare and Wellbeing, University of Augsburg, Augsburg, Germany first.last@informatik.uni-augsburg.de

[2]Department of Clinical Psychology and Psychotherapy, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany first.last@fau.de

[3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK first.last@ic.ac.uk

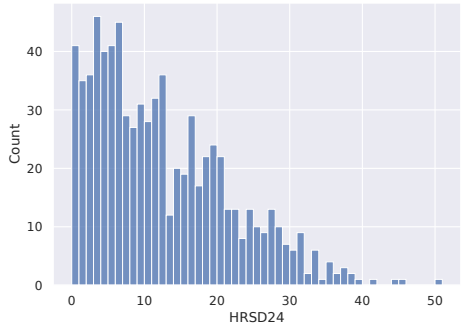

Fig. 1: Distribution of HRSD values over all time-points.

treatment. At these time points, clinical interviews were conducted and recorded via telephone. Of all the data collected in this study, a subsample of 325 (185 female) participants with a total of 799 telephone recordings forms the basis of the presented approaches. All included subjects have previously (in the last 6 months) been diagnosed with MDD. Current substance-, psychotic-, or bipolar disorder, as well as severe cognitive impairment, were chosen as exclusion criteria. However, comorbid diagnoses, such as panic or social anxiety disorders, appear for 42.2% of participants. In the subsample, participants were aged 18 to 69 with a mean age of 44.6 years (standard deviation 11.0).

HRSD [21] interviews were conducted for the assessment of depression severity while PHQ-9 [22] served as validation. HRSD includes a questionnaire of 17 clinician-rated items, each with scores ranging between $0-2$ or $0-4$, whereas PHQ-9 contains 9 items, self-rated on a Likert-scale of $0-3$. Both of these tests have been established as state-of-the-art depression assessments. For the downstream task of binary depression classification, we use both a discretisation of the HRSD scores (cutoff $\geq$ 10) as well the information if patients are currently experiencing an MDE or are in remission, as determined by the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) [23]. Furthermore, due to the longitudinal nature of the dataset described above, measurements of treatment responses between time points are available. Specifically, we consider a reduction of 50% in the HRSD-24 scores as response variable Response-24. As can be seen in Figure 1, HRSD scores are not normally distributed, but rather skewed towards lower values.

III. Methodology

A high-level overview of our approach for speech-based depression recognition is given in Figure 2. In the first step (preprocessing), we conduct speaker diarisation to extract speech segments of the patients and not the interviewers (cf. Section III-A). Subsequently, we extract a set of acoustic features and deep representations from each preprocessed audio recording on a segment level (cf. Section III-B). Using the obtained representations, we then create Bag-of-Features by generating a fixed-length histogram representation of each audio recording. This (quantisation) step is shown to be effective in filtering small amounts of noise available in the original feature space [19]. In the last step, we train machine learning models to obtain final predictions (cf. Section III-C).

A. Preprocessing

The audio recordings are sampled at 8 kHz and include both the speech of the patient as well as the interviewer. For the purpose of automatic paralinguistic analysis, we are only interested in the voice segments of the interviewees. Due to the dataset's size, we opted for an automatic diarisation of the speech recordings instead of a manual segmentation. We used the method described in [24] as implemented in the SpeechBrain library [25]. As the recordings are plagued with artefacts and noise mostly originating from the telephone transmission, the segmentation is expected to be imperfect. Afterwards, we manually sorted the obtained speaker clusters as belonging to either the interviewer or the patient and normalised the volume of the audio content.

B. Feature Extraction and Quantisation

We choose to extract both BoAWs [18] as well as BoDFs [19] representations from the obtained audio segments. In both approaches, input feature vectors are quantised according to a codebook that is obtained in an unsupervised fashion from the training data – in our case simple random sampling of a fixed number of input vectors. Afterwards, quantised segment-level vectors belonging to the same speech recording are aggregated into a fixed-length vector representation. This procedure helps us to deal with two challenging aspects of the dataset, namely the noisy nature of the telephone recordings, and the variation in the duration and the number of speech segments after speaker diarisation.

Our BoAWs are formed from 6 373 dimensional ComParE2016 audio functionals extracted with openSMILE [26] for each audio segment. For the BoDFs, we utilise DeepSpectrum [16] features generated by a DenseNet121 [27] image convolutional neural network (CNN) from Mel-spectrograms (128 Mel-bands, window size 5 seconds without overlap) of the audio content. For both, we randomly sample a codebook of 2 000 feature vectors from the input and quantise each vector based on its 50 nearest codebook vectors.

C. Experimental Setup

In our experiments, we perform automatic speech-based classification of depression-related diagnoses. Furthermore, the longitudinal nature of the dataset allows us to train machine learning algorithms to predict future depressive severity as well as detect changes occurring between time points, such as improvements due to a positive treatment response. In all of the considered experimental configurations, we apply a Leave-One-Speaker-Out cross-validation scheme (325 folds) in order to accurately assess the generalisation capabilities of the models. We employ Support Vector Machines with linear kernels for all experiments and optimise the complexity hyperparameter by a random inner cross-validation. Whenever we are dealing with the detection of a

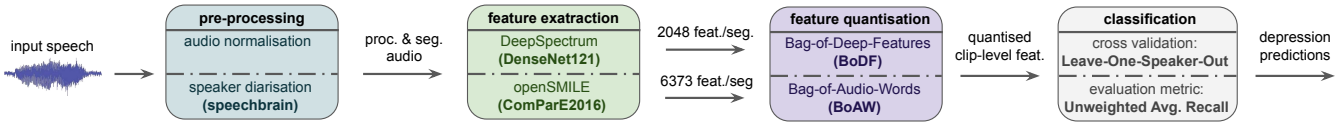

Fig. 2: High-level overview of our proposed depression detection approach comprising pre-processing, feature extraction/quantisation, and classification components. A detailed account of the approach is given in Section III.

change between two time points, we simply concatenate the respective feature vectors before feeding them to the machine learning models. We evaluate the accuracy of our machine learning models by the Unweighted Average Recall (UAR) as in many cases, the ground truth labels are not equally distributed across classes.

We consider a total of 5 classification experiments. First, we perform two binary classifications of current depression – detecting current MDE or remission status, and classifying HRSD scores into low (no depression) and high values (cutoff ≥ 10). Furthermore, we train models to predict the binary discretisation of HRSD scores at later time points from speech recorded at an earlier stage of the study. Finally, we investigate whether treatment responses between two time points can be detected by an automatic paralinguistic analysis. Here, we feed the models speech representations from two time points and train them to predict whether the HRSD 24 (*Response-24*) score has reduced by at least 50%.

IV. Results

Classification results obtained with the two machine learning approaches for all evaluated tasks can be found in Table I. In addition to the UAR values computed over all predictions, we report 95% confidence intervals obtained by bootstrapping (1000 repetitions of random sampling with replacement). Our systems achieve slightly below 70% UAR when tasked with detecting current depression diagnoses. Here, models trained with BoDFs fare better than those based on BoAWs, reaching a best UAR of 69.2% when classifying whether patients are currently experiencing an MDE or are in remission. The model achieves a sensitivity of 67.4% with a slightly higher specificity of 71.0%. Performance for both types of systems decreases when we instead consider the binarised HRSD scores (cutoff ≥ 10) as targets. Considering HRSD as a questionnaire targets the previous two weeks, this value will not indicate current depression quite as accurately as the SCID-I which specifically detects MDE.

Forecasting performance is negatively impacted by the fact that in our dataset, depression diagnoses are relatively stable in a majority of cases. Of the 330 sessions which have a follow up, only 72 see subjects changing from a depressed diagnosis to remission and vice-versa. If we compute UARs for these cases individually, our best model reaches slightly above chance-level at 53.9% while it matches performance for the remaining cases with its depression detection counterpart at 67.8%. We attribute this to the imbalance described above, not allowing our models to learn predictive features from the data.

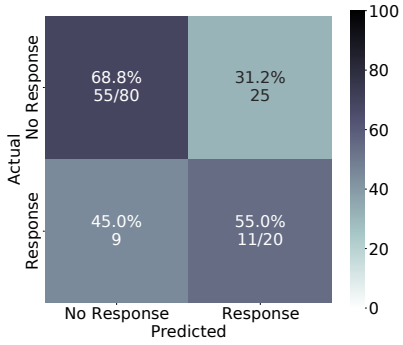

Fig. 3: Confusion matrix of our best system for automated classification of treatment response.

TABLE I: Speech-based classification results measured in UAR (chance-level 50%). We additionally report confidence intervals computed by bootstrapping on model predictions.

Measurement	BoAW	BoDF
MDD	67.4 (63.9 - 70.6)	69.2 (65.8 - 72.6)
HRSD	64.3 (60.8 - 67.7)	67.6 (64.2 - 70.8)
HRSD Forecasting	65.2 (60.5 - 70.7)	63.3 (58.1 - 68.1)
Response-24	68.1 (57.0 - 79.4)	61.9 (50.1 - 74.6)

On the other hand, directly targeting the detection of response leads to a better performance, achieving a UAR of 68.1% when classifying a 50% reduction in symptoms measured on the 24 item HRSD scale (cf. Figure 3). Contrasting BoDF's superiority in depression detection, here, BoAW seem to be more effective in capturing intra-individual variations in speech characteristics, such as reduced rate of speech or lower F1, which are related to depression.

In addition to reporting the classification results of our machine learning approaches, we compare the achieved performance to the accuracy of HRSD, PHQ-9, and QIDS-C for detecting remission and MDE as measured by the state-of-the-art SCID-I. Utilising cutoff values and comparing accuracies with the McNemar test, we found no difference in the number of correctly classified samples between our models' predictions and HRSD ($p = 0.356$). However, PHQ-9 was significantly ($p = 0.020$) more accurate than machine learning-based methods in classifying remission status.

V. Conclusion

In this paper, we utilised an unsupervised feature learning strategy to perform remission status classification of depressed and formerly depressed patients from noisy tele-

phone recordings, achieving 69.2% UAR, matching the performance of a cutoff-based classification with HRSD. As our dataset contains measurements at different time points, we further evaluated a classification of treatment response, reaching 68.1% UAR. However, there are some limitations to the work presented herein. First, while the noisy audio quality of the recordings is more representative to real-life scenarios than speech collected in laboratory settings, clinical interviews are not a natural choice for in-the-wild data. Another limitation lies with the statistics of our data in which depressive symptom severity was skewed towards the lower side and did not vary much between time points. Future work should therefore focus on utilising data sources, such as mood diaries, which can be integrated into traditional psychotherapy and further investigate the viability of automated treatment response detection over longer spans of time.

Ethical Approval

The ethics commitee of the Friedrich-Alexander-University approved all experimental procedures involving human subjects.

Acknowledgment

This research was partially supported by Deutsche Forschungsgemeinschaft (DFG) under grant agreement No. 421613952 (ParaStiChaD).

References

[1] G. Y. Lim, W. W. Tam, Y. Lu, C. S. Ho, M. W. Zhang, and R. C. Ho, "Prevalence of depression in the community from 30 countries between 1994 and 2014," *Scientific reports*, vol. 8, no. 1, p. 2861, 2018.

[2] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, 2018.

[3] R. P. Auerbach, J. Alonso, W. G. Axinn, P. Cuijpers, D. D. Ebert, J. G. Green, I. Hwang, R. C. Kessler, H. Liu, P. Mortier, *et al.*, "Mental disorders among college students in the world health organization world mental health surveys," *Psychological medicine*, vol. 46, no. 14, pp. 2955–2970, 2016.

[4] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS medicine*, vol. 3, no. 11, p. e442, 2006.

[5] P. Cuijpers, A. Van Straten, G. Andersson, and P. Van Oppen, "Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies." *Journal of consulting and clinical psychology*, vol. 76, no. 6, p. 909, 2008.

[6] P. Cuijpers, M. Sijbrandij, S. L. Koole, G. Andersson, A. T. Beekman, and C. F. Reynolds III, "The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: A meta-analysis of direct comparisons," *World psychiatry*, vol. 12, no. 2, pp. 137–148, 2013.

[7] D. Al-Karawi and L. Jubair, "Bright light therapy for nonseasonal depression: meta-analysis of clinical trials," *Journal of affective disorders*, vol. 198, pp. 64–71, 2016.

[8] J. R. Vittengl, L. A. Clark, T. W. Dunn, and R. B. Jarrett, "Reducing relapse and recurrence in unipolar depression: a comparative meta-analysis of cognitive-behavioral therapy's effects." *Journal of consulting and clinical psychology*, vol. 75, no. 3, p. 475, 2007.

[9] L. Pintor, C. Gastó, V. Navarro, X. Torres, and L. Fañanas, "Relapse of major depression after complete and partial remission during a 2-year follow-up," *Journal of Affective Disorders*, vol. 73, no. 3, pp. 237–244, 2003.

[10] D. Chisholm, K. Sweeny, P. Sheehan, B. Rasmussen, F. Smit, P. Cuijpers, and S. Saxena, "Scaling-up treatment of depression and anxiety: a global return on investment analysis," *The Lancet Psychiatry*, vol. 3, no. 5, pp. 415–424, 2016.

[11] J. C. Fortney, J. Unützer, G. Wrenn, J. M. Pyne, G. R. Smith, M. Schoenbaum, and H. T. Harbin, "A tipping point for measurement-based care," *Psychiatric services*, vol. 68, no. 2, pp. 179–188, 2017.

[12] S. Amiriparian, A. Awad, M. Gerczuk, L. Stappen, A. Baird, S. Ottl, and B. Schuller, "Audio-based recognition of bipolar disorder utilising capsule networks," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–7.

[13] M. Gerczuk, A. Triantafyllopoulos, S. Amiriparian, A. Kathan, J. Bauer, M. Berking, and B. Schuller, "Personalised deep learning for monitoring depressed mood from speech," in *Proceedings of the E-Health and Bioengineering Conference (EHB)*. Iași, Romania: IEEE, 2022, pp. 1–5.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[15] M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller, "Emonet: A transfer learning framework for multi-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, 2021.

[16] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 3512–3516.

[17] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[18] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Interspeech 2016*. ISCA, Sept. 2016, pp. 495–499.

[19] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, S. Pugachevskiy, and B. Schuller, "Bag-of-deep-features: Noise-robust deep feature representations for audio analysis," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–7.

[20] C. Vis, A. Kleiboer, R. Prior, E. Bønes, M. Cavallo, S. A. Clark, E. Dozeman, D. Ebert, A. Etzelmueller, G. Favaretto, A. F. Zabala, N. Kolstrup, S. Mancin, K. Mathiassen, V. N. Myrbakk, M. Mol, J. P. Jimenez, K. Power, A. van Schaik, C. Wright, E. Zanalda, C. D. Pederson, J. Smit, and H. Riper, "Implementing and up-scaling evidence-based eMental health in Europe: The study protocol for the MasterMind project," *Internet Interventions*, vol. 2, no. 4, pp. 399–409, Nov. 2015.

[21] M. Hamilton, "A RATING SCALE FOR DEPRESSION," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 23, no. 1, pp. 56–62, Feb. 1960.

[22] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[23] M. B. First, "Structured clinical interview for DSM-IV axis I disorders," *Biometrics Research Department*, 1997.

[24] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 3830–3834.

[25] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori, and Y. Bengio, "SpeechBrain: A General-Purpose Speech Toolkit," June 2021.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia - MM '10*. Firenze, Italy: ACM Press, 2010, p. 1459.

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, July 2017, pp. 2261–2269.