International Medicine and Health Sciences Congress (IMHSC'24)
5-7 November 2024, Male-Maldives

Comparison of speech elicitation tasks for machine learning-based depression classification

Jonathan F. Bauer
Department for Clinical Psychology and Psychotherapy
Friedrich-Alexander-Universität
Erlangen, Germany
https://orcid.org/0000-0002-1549-6534

Maurice Gerzcuk
Chair of Embedded Intelligence for Health Care & Wellbeing
University of Augsburg
Augsburg, Germany
https://orcid.org/0000-0001-8293-6635

Björn Schuller
Munich Data Science Institute
Technical University Munich
Munich, Germany
https://orcid.org/0000-0002-6478-8699

Matthias Berking
Department for Clinical Psychology and Psychotherapy
Friedrich-Alexander-Universität
Erlangen, Germany
https://orcid.org/0000-0001-5903-4748

***Abstract*—Machine learning-based depression classification based on paralinguistic speech parameters yields a novel approach to detect depression. However, there is uncertainty about the effect of different types of speech recordings on classification accuracy. We suggest that recordings of free speech containing anti-depressive statements may be particularly suitable for depression classification.**

To test this hypothesis, we conducted Structured Clinical Interviews for DSM-5 to determine depression diagnoses on suitable candidates, resulting in a final sample of 48 clinically depressed individuals, 48 sub-clinically depressed individuals, and 48 non-depressed individuals. Participants from each group completed four different speech tasks: Participants read aloud neutral texts, they read aloud scripted depressive statements, they came up with and expressed anti-depressive statements, and 50% of participants read aloud scripted anti-depressive statements. Separate classification models aimed at classifying current depression were trained for each speech type and with two different state-of-the-art machine learning methods.

We found that training a depression classification model on recordings of anti-depressive statements was not superior to training models on other types of speech recordings. We only found a significantly better accuracy for the depression classification model trained on recordings of neutral read speech compared to the model trained on recordings of depressive read speech.

We could not confirm our hypothesis that recordings of anti-depressive statements would result in superior depression classification accuracy compared to recordings of neutral text reading. Eliciting depression-related speech may reduce affective variability in individuals' responses and therefore diminish depression-discriminative information. Our findings provide important directions for future research aimed to optimize speech elicitation tasks for depression classification.

Keywords—depression, speech, voice, machine learning, speech elicitation

I. Introduction

Depression is a debilitating disease that causes considerable suffering on the individual and the public health level [1]. The most common depression diagnosis is Major Depressive Disorder, characterized by depressed mood, loss of motivation or interest, and behavioral alterations such as reduced activity and disturbed sleep [2]. Particularly psychotherapeutic and pharmacotherapeutic interventions as well as their combinations constitute effective treatments of depression (e.g., [3]). However, a substantial amount of people, who actually meet criteria for depression, remain undiagnosed [4, 5], which can lead to continued suffering and chronification of the disorder [6]. Many patients with depression suffer from recurrent depressive episodes, making continuous monitoring an important measure to detect recurrence or remission of depressive symptoms. However, only 20% of practitioners use standardized screening methods for systematically assessing depression [7]. In addition, standardized screening methods rely on self-reported depressive symptoms, which can be subject to response or memory bias. Therefore, there is promise in developing alternative methods that allow accessible, time-efficient, and cost-effective assessments of depression. This may be achieved by identifying objective markers that are valid and reliable indicators of depression.

Previous studies have found altered speech patterns in individuals with current depression compared to non-depressed individuals, specifically for paralinguistic speech parameters, such as prosody, voice quality, and resonance [8]. As the speech apparatus is a highly complex muscular structure, it can be hypothesized that the physiological, neurofunctional, and cognitive changes associated with depression lead to specific changes in such speech parameters. Affective states are associated with specific patterns of physical states that influence the sound of speech, e.g., by alterations in muscular tension, in respiratory action, or articulatory motor control. Therefore, identifying these associations may allow the assessment of affective states, such as depression, by evaluating speech patterns [8].

Machine learning models allow the detection of speech patterns indicative of depression and can provide automated and objective depression assessments. Models in previous studies achieved accuracies in classifying depression ranging from 50% up to 96% [8]. This considerable variability in accuracy between studies may originate from a variety of factors such as differences in sample characteristics (e.g., sample size, diagnostic status of speakers), depression assessment methods determining input variables (e.g., self-reported symptoms, interview-based assessments), feature selection approaches, feature extraction methods, machine learning methods, recording setups and settings, and speech elicitation tasks. Most studies optimized their models by systematically varying machine learning methods or feature selection approaches. Only few studies systematically varied speech elicitation tasks to test whether structure and content of speech recordings affects depression classifications. Studies showed that free speech recordings were superior to read speech recordings for the development of accurate depression classification models [9–11]. There is further evidence that within free speech recordings, those with self-relevant content may be superior compared to those with generic content [11], although another study could only partially confirm this finding [10]. Surprisingly, emotional speech was not superior to neutral speech for depression classification [10, 11], showing that eliciting affective states with open questions about positive or negative affective topics does not increase depression-discriminant information. Arguably, tasks that are able to discriminate between depressed and non-depressed individuals may be better suited to increase depression-discriminative information in speech recordings. Therefore, regulating negative emotions in response to self-referent depressive thoughts may be an appropriate task to increase the saliency of depressive speech patterns. A task aimed at downregulating negative mood explicitly elicits depression-relevant speech and involves emotion regulation skills that have been shown to be deficient in individuals with depression [12, 13] and may therefore increase discriminatory power of the task. Thus, we expect that depression classification models are more accurate if trained on recordings from free speech elicitation tasks that incorporate emotion regulation compared to speech elicitation tasks with neutral content and/or with scripted responses.

II. Method

A. Participants

The present study included 144 participants, with 48 individuals with a current MDD diagnosis, 48 individuals with elevated yet non-clinical depressive symptoms, and 48 non-depressed individuals with no history of depression. We matched participants for age and gender across these three groups. Exclusion criteria were a current diagnosis of bipolar, psychotic, or substance-related disorders (except for nicotine) within the past six months, and psychotherapeutic treatment during the past six months. Participants had a mean age of 32.72 years (ranging from 20 to 63, SD = 11.02), 67% of participants were female, and 19% of participants had another psychiatric disorder than MDD.

B. Procedures

To determine a current depression diagnosis, we conducted the Structured Clinical Interview for DSM-5 (SCID; [14]) with participants in an initial diagnostic session. The subsequent session (taking place on average 13 days after the diagnostic session) included several speech assessments. First, participants had to read aloud a statement that would be typical for patients with depression (e.g., 'I am not as capable as others'). Then they were asked to give an anti-depressive response that would downregulate negative mood induced by the initial depressive statement. This was repeated five times and each depressive statement and each anti-depressive response was recorded separately. Subsequently, participants completed a reading tasks including two neutral texts ('The North Wind and the Sun' by Aesop and an excerpt from 'Homo Faber' by Max Frisch). The final session included a similar emotional speech elicitation task: In response to depressive statements, they were instructed to read aloud scripted anti-depressive statements. Each response was read three times and the task included ten depressive statements, followed by a response. The task in the final session was completed by 50% of participants that were randomly selected.

C. Speech analysis

We performed binary classification of depression (currently depressed/non-depressed) utilizing a machine-learning pipeline consisting of feature extraction and a linear Support Vector Machine (SVM) classifier. We selected distinct sets each covering different speech parameters for extracting features. The first type of speech features came in the form of the small handcrafted eGeMAPS [15] set of audio functionals, extracted with openSMILE [16]. It computes statistics over a number of low-level descriptors, including pitch, harmonic ratios, jitter, shimmer, loudness and spectral slope. For the second type of speech features, we utilized a pre-trained deep neural network, specifically the transformer-based wav2vec2 [17] as a feature extractor. The specific model we used has been fine-tuned for German automatic speech recognition [18].

We trained and evaluated linear SVMs for each of these feature sets in a 10-fold speaker-independent cross-validation (audio samples from one speaker never appear in the training and validation sets at the same time). We optimized the SVM's cost parameter on a logarithmic scale between 10-2 and 10-5 with an additional inner 5-fold cross-validation. We chose balanced accuracy as our main metric for evaluating the results and hyperparameter optimization.

D. Statistical analyses

In order to evaluate machine learning-based depression classifications, we calculated sensitivity ((number of true positives)/(number of true positive + number of false negatives)), specificity ((number of true negatives)/(number of true negatives + number of false positives)), and balanced accuracy ((sensitivity + specificity)/2). The diagnostic status of MDD according to the SCID served as a validation criterion.

To test for statistical differences, we calculated multilevel models to compare speech types. We calculated multilevel models, each predicting a performance metric (i.e. balanced accuracy, sensitivity, and specificity). We compared a model

including the interaction between method and speech type as a predictor, a model including both speech type and method as predictor, and a model including only speech type as a predictor to identify the model with the best fit. The results of each fold were nested within the respective experiment. For all three metrics, the models did not significantly differ; therefore, we calculated the least complex model including only speech type as predictor.

We used the R packages lme4 [19], lmerTest [20], and emmeans [21] to compare speech types and used t-tests with Satterthwaite's method for degree of freedom approximation. All tests were two-sided with critical alpha set at 5%.

III. Results

Table I. shows performance metrics for depression classification models across different speech elicitation tasks. Regarding balanced accuracy, we found that depression classification was more accurate for neutral read speech compared to depressive read speech ($t(76) = -2.28$, $p = 0.03$). No other comparisons between speech types were significant (all ps > 0.05). This means that the depression classification model that was trained on neutral speech recordings was more accurate compared to the model trained on depressive read speech recordings. Regarding sensitivity, we did not find any significant differences between speech elicitation tasks (all ps > 0.05). This means that no depression classification model was superior at correctly identifying individuals with depression. Regarding specificity, we found that depression classification was more sensitive for neutral read speech compared to depressive read speech ($t(76) = -2.23$, $p = 0.03$). No other comparisons between speech types were significant (all ps > 0.05). This means that the depression classification model that was trained on neutral speech recordings was superior at correctly identifying individuals without depression compared to the model trained on depressive read speech recordings.

TABLE I. Performance metrics of depression classification models across speech elicitation tasks

Machine learning method	**Speech types**			
	Anti-depressive free speech[a]	Anti-depressive read speech[b]	Depressive read speech[a]	Neutral read speech[a]
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
	Balanced accuracy			
eGeMAPS	0.61 (0.09)	0.58 (0.16)	0.55 (0.09)	0.63 (0.11)
wav2vec2	0.62 (0.07)	0.60 (0.12)	0.56 0.08)	0.65 (0.17)
	Sensitivity			
eGeMAPS	0.52 (0.23)	0.41 (0.39)	0.42 (0.22)	0.45 (0.18)
wav2vec2	0.50 (0.19)	0.46 (0.34)	0.42 (0.18)	0.51 (0.31)
	Specificity			
eGeMAPS	0.71 (0.10)	0.76 (0.21)	0.69 (0.13)	0.81 (0.11)
wav2vec2	0.74 (0.08)	0.74 (0.22)	0.70 (0.14)	0.78 (0.12)

[a]$n = 144$, [b]$n = 72$

IV. Discussion

We could not confirm our hypothesis that training depression classification models on recordings of anti-depressive statements would improve accuracy compared to a model trained on neutral speech recordings. Further, our results are in contrast to previous findings that showed superiority of depression classification from free speech recordings compared to read speech recordings [9–11]. Thus, incorporating depression-related content in speech tasks may in fact not be helpful for distinguishing between depressed and non-depressed speech. Arguably, open questions in free speech tasks may result in broader affective variability in participants' responses: Depression-associated biased cognitive processing with a tendency to focus their attention or interpretation on negative information [13] may result in negative affective responses in individuals with depression, whereas non-depressed individuals' affective responses may be more neutral or positive. However, in a free speech task eliciting depressive-relevant speech, the affective variability in responses may be reduced, diminishing depression-discriminative information compared to speech tasks with open questions allowing variable affective responses. This inference is supported by a study showing that speech recordings containing different types of affective valence (i.e. positive, negative, and neutral) enabled more accurate depression classifications than speech recordings containing only either positive, negative, or neutral valence [22].

Our results further suggest that the reading task containing depressive statements was inferior for depression classification compared to the neutral reading task. Although we would expect more pronounced depressed mood states in individuals with current depression, studies suggest that reading self-referent depressive statements leads to elevated depressed mood also in non-depressed individuals [23]. Thus, the reading of depressive self-statements may have induced a state of depressed mood in participants across diagnostic status, thereby diminishing depression-discriminative information in participants' voices. This suggests that recordings of read statements with depression-related content are not suited for depression classification.

A major limitation of the present study is that the different types of speech recordings differed in their duration and structure. The utterances elicited in the neutral readings task were substantially longer than the other speech types that contained relatively short statements. In addition, recordings from the neutral reading task contained uninterrupted text reading, whether the other speech types contained various statements that were recorded with interruptions. While there is evidence that also short, individually recorded utterances allow accurate depression classifications [24], it is unclear whether utterance length and structure may have influenced our results. Future studies should choose a methodological design that allows a systematical evaluation of potential effects of recoding duration and structure. It is of note that in the accuracies in this study are notably lower compared to maximum accuracies from previous studies (e.g., see [8]). However, we aimed to overcome some problems associated with previous research that likely led to higher accuracy but at the same time to lower ecological validity and comparability between studies. Most previous studies applied a variety of machine learning methods on the

same datasets and/or selected the features based on the same dataset that was used to train the model. Whereas these approaches improve the likelihood of reaching high accuracies, they increase the risk of overfitting models and overestimating accuracy. To minimize such risks and to develop models that allow accuracy comparisons between studies, we decided on using two state-of-the-art machine learning methods with high generalizability, allowing better comparisons between studies and datasets.

In conclusion, our results suggest that speech recordings of anti-depressive statements are not superior to speech recordings with neutral content. In fact, speech elicitation tasks that limit variability in affective responses might reduce accuracy of depression classification models compared to speech elicitation tasks with open questions that allow a broader range of affective responses. Thus, further studies should develop speech elicitation tasks with open questions about self-relevant topics incorporating emotion regulation that allow a broad range of affective responses and evaluate whether such tasks are better suited to optimize depression classification models.

REFERENCES

[1] R. C. Kessler and E. J. Bromet, "The epidemiology of depression across cultures," Annual review of public health, vol. 34, pp. 119–138, 2013.

[2] American Psychiatric Association, Diagnostic and statistical manual of mental disorders, 5th ed., 2013.

[3] P. Cuijpers, M. Berking, G. Andersson, L. Quigley, A. Kleiboer, and K. S. Dobson, "A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments," The Canadian Journal of Psychiatry, vol. 58(7), pp. 376–385, 2013.

[4] M. A. Craven, "Depression in primary care: current and future challenges," The Canadian Journal of Psychiatry, vol. 58(8), pp. 442–448, 2013.

[5] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis," The Lancet, vol. 374(9690), pp. 609–619, 2009.

[6] L. Ghio, S. Gotelli, A. Cervetti, M. Respino, W. Natta, M. Marcenaro, ... and M. B. Murri, "Duration of untreated depression influences clinical outcomes and disability," Journal of affective disorders, vol. 175, pp. 224–228, 2015.

[7] C.C. Lewis, M. Boyd, A. Puspitasari, E. Navarro, J. Howard, H. Kassab, ... and K. Kroenke, "Implementing measurement-based care in behavioral health: a review," JAMA psychiatry, vol. 76(3), pp. 324–225, 2019.

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

[9] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: a comparison between spontaneous and read speech," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7647–7551, May 2013.

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

[11] H. Long, Z. Guo, X. Wu, B. Hu, Z. Liu, and H. Cai, "Detecting depression in speech: Comparison and combination between different speech types," IEEE International Conference on Bioinformatics and Biomedicine, pp. 1052–1058, November 2017.

[12] J. Joormann, and C. H. Stanton, "Examining emotion regulation in depression: A review and future directions," Behaviour Research and Therapy, vol. 86, pp. 35–49, 2016.

[13] J. Joormann and I. H. Gotlib, "Emotion regulation in depression: Relation to cognitive inhibition," Cognition and Emotion, vol. 24(2), pp. 281–298, 2010.

[14] M. B. First, J. B. W. Williams, R. S. Karg, and R. L. Spitzer, "Structured clinical interview for DSM-5 disorders. SCID-5-CV," Arlington, VA: American Psychiatric Association Publishing, 2016.

[15] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," IEEE Transactions on Affective Computing, vol. 7(2), pp.190–202, 2016.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," Proceedings of the 18th ACM international conference on Multimedia, pp. 1459–1462, October 2010.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460, 2020.

[18] J. Grosman, "Fine-tuned XLSR-53 large model for speech recognition in German," https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german, 2021.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4" https://arxiv.org/abs/1406.5823, 2014.

[20] A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, "lmerTest package: tests in linear mixed effects models," Journal of statistical software, vol. 82(13), 2017

[21] R. V. Lenth, "emmeans: estimates marginal means, aka least-squares means," R package version 1.10.3, 2024.

[22] B. Stasak, J. Epps, and R. Goecke, "Automatic depression classification based on affective read sentences: opportunities for text-dependent analysis," Speech Communication, vol. 115, pp. 1–14, 2019.

[23] R. O. Frost and M. L. Green, "Velten mood induction procedure effects," Personality & Social Psychology Bulletin, vol. 8(2), pp. 341–347, 1982.

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