Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 1266–1274
https://doi.org/10.54941/ahfe1005694

Optimizing Speech Elicitation Tasks for Machine Learning-Based Depression Assessment

Jonathan Bauer[1], Maurice Gerczuk[2], Björn Schuller[2], and Matthias Berking[1]

[1]Department for Clinical Psychology and Psychotherapy, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91052 Erlangen, Germany
[2]Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, 86159 Augsburg, Germany

ABSTRACT

Theoretical Background: The field of machine learning-based speech analysis may provide unobtrusive, time-efficient, and cost-effective ways of automated depression assessment. Systematically optimizing speech elicitation tasks may further improve the accuracy of this approach. We hypothesized that machine learning-based depression classification would perform better if trained on recordings of individuals reading anti-depressive statements with the instruction to intone them as convincingly as possible compared to readings of anti-depressive statements without instructions regarding intonation.
Methods: To test this hypothesis, we recruited a sample of 48 clinically depressed individuals, 48 sub-clinically depressed individuals, and 48 non-depressed individuals. Participants from each group were randomly allocated to either the experimental or the control condition. In both conditions, participants read aloud scripted anti-depressive self-statements. Participants in the experimental condition received instructions to heighten the prosodic expression of conviction in their voice, whereas participants in the control condition received no such instructions. Separate classification models aimed at detecting current depression were trained for each condition and with a selection of different machine learning methods.
Results: We found that models trained on data from the experimental condition were more accurate and reliable than those trained on data from the control condition. While the former models reached balanced accuracies between 65–76%, the latter only reached balanced accuracies between 36–61%.
Discussion: Our results suggest that features of speech elicitation tasks have substantial influence on model performance for automated depression classification. The present findings highlight that speech elicitation tasks including voice modulation instructions can improve the validity and reliability of machine learning-based depression classification.

Keywords: Depression assessment, Machine learning, Voice, Speech, Major depressive disorder

INTRODUCTION

Depression is a debilitating disease causing substantial suffering on the individual and the public health level (Kessler & Bromet, 2013). The most common depression diagnosis is Major Depressive Disorder (MDD), which is characterized by depressed mood, loss of motivation or interest, and behavioral alterations such as reduced activity and disturbed sleep (APA, 2013). There are effective treatments available, including psychotherapeutic and pharmacotherapeutic interventions as well as their combinations (e.g., Cuijpers et al., 2013). However, studies suggest that many people who actually meet criteria for MDD remain undiagnosed (Craven et al., 2013; Mitchell et al., 2009), which can lead to continued suffering and chronification of the disorder (Ghio et al., 2015). As depression can be a recurrent disease, continuous monitoring is an important measure to detect recurrence or remission of depressive episodes. Although many screening methods for the systematic assessments are available, only 20% of practitioners use those (Lewis et al., 2019). Therefore, there is a need to develop 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.

Theoretical Models of Depression

Depression is a heterogeneous disease, involving cognitive, affective, and somatic symptoms. Most commonly, depression is described with the cognitive model that states that dysfunctional beliefs about the self (e.g., 'I am worthless'), the world (e.g., 'No one likes me'), and the future (e.g., 'My future is hopeless') are at the root of depression (Beck and Bredemeier, 2016). The seminal theory of interacting cognitive subsystems (ICS) proposed by Teasdale and Barnard (1993) extends the cognitive model by suggesting different subsystems that process cognitive, sensory, proprioceptive, and somatosensory information. For the genesis of an affective state, coherent information must be processed on multiple subsystems (e.g., thinking about one's worthlessness, perceiving the low state of energy in the body, and hearing the sound of one's own feeble voice). There are feedback loops between subsystems that can lead to self-perpetuating interlocked configurations. Individuals affected with depression are stuck in an interlock between depressogenic cognitions (i.e., dysfunctional beliefs) and (somato-)sensory states (e.g., vocal expression, posture, and arousal) that re-activate each other reciprocally. This interlock makes it particularly difficult for depressed individuals to change their affective state. Studies have identified body states that are characteristic for depression (e.g., typical speech parameters; Cummins et al., 2015, or gait patterns; Adolph et al., 2021). For example, during a therapeutic intervention (e.g., uttering 'There are people who appreciate me'), a depressed individual may use a depressed, unconvincing voice (e.g., slow speech rate, high pause frequency; Jiang & Pell, 2017) that invalidates the content of the utterance and the individual may fail to elevate their mood.

Current Approaches of Machine Learning-Based Depression Assessment From Speech

While most common depression assessments are based on self-reported symptoms, research efforts have been made to assess depression based on speech parameters (Cummins et al., 2015). Technical innovations and the rising availability of smartphones have led to developments in machine learning-based assessments of depression from speech. Machine learning enables the development of algorithmic models that use a large number of input variables ('features', e.g., speech parameters) to classify output variables ('labels', e.g., diagnostic status of depression). Machine learning models for classifying depression have been developed in numerous studies, achieving accuracies ranging from 50% up to 96% (Cummins et al., 2015). The origin of these differences in accuracy between studies is yet unclear since the variability in methodology between studies is considerable. Variability may originate from differences in sample characteristics (e.g., sample size, diagnostic status of speakers), depression assessment methods determining input variables, feature selection approaches, feature extraction methods, machine learning methods, recording setups and settings, and speech sampling tasks. Most studies systematically varied methods for machine learning and feature selection and extraction. However, few studies aimed to optimize speech elicitation tasks, although data suggests that what a person says greatly affects how they say it (Filippi et al., 2017). Studies testing accuracy differences in depression classification from speech with or without emotional content only found small (Long et al., 2017) or no differences (Jiang et al., 2017). Arguably, emotional content may only affect classification accuracy if it contains information that is relevant to the self and/or to depression. Accordingly, studies showed superior accuracy in models trained with spontaneous speech compared to those trained with read speech (Alghowinem et al., 2013). However, while spontaneous speech tasks probably include more depression- and self-relevant information, the way someone responds to such a task may vary greatly between individuals. Response behavior can be influenced by variables unrelated to depression, such as personality traits (Freudenthaler & Neubauer, 2007) or motivational state (Klehe & Latham, 2008), and may therefore lead to biased depression classifications. Thus, combining self-relevant content and emotional content in a speech task and simultaneously minimizing bias from response behavior may optimize the performance of depression classification models.

The Present Study

Measuring what participants can do at their best instead of measuring participants' typical behavior is an approach to minimize variability coming from response behavior in psychological assessment. Inducing depressed mood and then asking participants to utter anti-depressive statements as convincingly as possible may uncover a person's capability to overcome a state of depressed mood. As described above, we expect that this is particularly difficult for depressed individuals, due to the interlocked configuration between their cognitions and their body state. Therefore,

this task may amplify differences in speech parameters between those who are able to overcome a state of depressed mood (i.e. non-depressed individuals) and those who have difficulties overcoming this state (i.e. depressed individuals). Thus, we hypothesized that machine learning-based depression classification would perform better if trained on speech samples that show the ability to use a convincing voice when uttering anti-depressive self-statements compared to speech samples from individuals reading aloud the same statements without the instruction to modulate the voice.

METHOD

Participants

We included 144 participants in the present study. To ensure variance across different levels of depressive symptom severity, we recruited 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. Participants were matched 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 any exposure to 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 in addition to MDD.

Procedures

In an initial diagnostic session, the Structured Clinical Interview for DSM-5 (SCID; First et al. 2016) was conducted with participants. For the subsequent experimental session (taking place on average 30 days after the diagnostic session), participants were randomly allocated to an experimental or control condition. The experimental session was designed to resemble a psychotherapeutic session, focusing on invalidations of depressogenic self-statements (Phase 1), followed by validations of anti-depressive self-statements (Phase 2). Initially and in-between the training parts, depressed mood was induced with a validated mood induction procedure (Diedrich et al. 2016; Velten 1968). After each mood induction, statements (depressogenic statements in Phase 1, anti-depressive statements in Phase 2) were presented consecutively and participants were asked to select one of three possible scripted coping responses (invalidations in Phase 1, validations in Phase 2). Participants were instructed to read this coping response aloud three times. In the experimental condition, participants received additional instructions and ongoing feedback from the experimenter to modulate their voice to sound as convincing as possible (focusing on loudness, emphasis, and intensity). In the control condition, participants received identical instructions about the intervention procedures and occasional encouraging feedback, but no instructions on voice modulation.

Speech Analysis

We performed binary classification of depression utilizing a traditional machine-learning pipeline, which consists of feature extraction and a linear Support Vector Machine (SVM) classifier. For feature extraction, we selected distinct sets each covering different speech parameters. First, we constructed a set of speech rhythm statistics based on previous works (Dellwo et al., 2006; Grabe & Low, 2008; Ramus et al., 1999). These features can be grouped into four categories: The total (speech) duration and the number of phonemes per total (speech) duration are considered global speech rate features. Local variability features are computed as the raw and normalized average durational difference between vocalic/consonantal segments (Grabe & Low, 2008). On the other hand, global variability features consider the total duration and standard deviation of vocalic/consonantal segments and pauses and are further normalized by total duration to remove the effect of speech rate (Dellwo et al., 2006; Ramus et al., 1999). We finally included the total number of phonemes and vowels as phonological features. The second type of speech features came in the form of the small handcrafted eGeMAPS (Eyben et al., 2016) set of audio functionals. It computes statistics over a number of low-level descriptors (LLDs), including pitch, harmonic ratios, jitter, shimmer, loudness and spectral slope. Finally, we utilized a pre-trained deep neural network, specifically the transformer-based wav2vec2.0 (Baevski et al., 2020) as a feature extractor. The specific model we used has been fine-tuned for German automatic speech recognition (Grosman, 2021).

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). Furthermore, 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.

Statistical Analyses

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

We employed multilevel models to test the effect of condition on performance metrics. To account for the nested structure of the folds, random intercepts were included for the models. The models included all three machine learning methods and were tested for balanced accuracy, sensitivity, and specificity, respectively. We determined significance based on the z distribution and set the significance level at $\alpha = 0.05$. We used the lme4 package in R to fit the model.

RESULTS AND DISCUSSION

Table 1 shows an overview of performance metrics of different classification methods, trained with the dataset from the experimental condition or the dataset from the control condition, respectively. Multilevel modeling revealed that compared to classification models trained on data from the control condition, classification models trained on data from the experimental condition were significantly more accurate (estimated effect = 0.19, $p = .034$) and significantly more sensitive (estimated effect = 0.35, $p = .024$), but no difference was found for specificity (estimated effect = 0.03, $p = .531$). Descriptively, the approach with the best balanced accuracy was feature selection with wav2vec2.0, followed by feature selection with eGeMAPS. Only using rhythm features resulted in superior performance when training the model with recordings from the experimental compared to the control condition. Variances between folds were comparable between models trained on recordings from the experimental condition to models trained on recordings from the control condition.

In sum, results show that depression classification models were more accurate and sensitive if trained on speech samples elicited in a conviction maximization speech task compared to speech samples elicited without voice modulation instructions.

Table 1. Balanced accuracies, sensitivities, and specificities across machine-learning methods and experimental conditions.

Method	Balanced Accuracy		Sensitivity		Specificity	
	EC *M (SD)*	CC *M (SD)*	EC *M (SD)*	CC *M (SD)*	EC *M (SD)*	CC *M (SD)*
Rhythm	0.65 (0.10)	0.36 (0.08)	0.58 (0.22)	0.00 (0.00)	0.72 (0.24)	0.71 (0.17)
eGeMAPS	0.72 (0.12)	0.60 (0.16)	0.67 (0.31)	0.44 (0.36)	0.77 (0.15)	0.76 (0.16)
wav2vec2.0	0.76 (0.11)	0.61 (0.11)	0.69 (0.29)	0.46 (0.31)	0.82 (0.14)	0.75 (0.16)

EC = Experimental condition; CC = Control condition.

Strengths of our study are the high reliability and validity of clinical data ensured by the use of the SCID as the state-of-the-art method for diagnosing MDD (Stuart et al., 2014). Further, we increased external validity by including non-depressed participants, participants with subclinical depressive symptoms and clinically depressed participants. Arguably, this is more representative for patients in clinical practice than sample selections that either used only participants from a general population, from clinical populations, or comparing a clinical population with a non-clinical population, excluding subclinically depressed participants.

There are several limitations of the current study: First, there was an average time lag of 30 days between the diagnostic session and the speech recordings. In this time, some depressed participants may have gone into remission or symptoms of subclinically-depressed participants may have worsened to the point where they would qualify for MDD. However,

participants did not receive any additional treatment in the meantime and 96% of depressed participants received psychotherapy after the experiment, suggesting continued depression in those initially meeting criteria for MDD. Second, the performances of depression classification models in this study may not yet suffice for implementing them in clinical practice. Pettersson and colleagues proposed that depression detection instruments should have a sensitivity of at least 80% and a specificity of at least 70% (Pettersson et al., 2015). While almost all models achieved a specificity of over 70%, none had a sensitivity of above 69%. This means all models would classify too many people as non-depressed although they are currently depressed. Third, this kind of speech task requires substantial effort from participants, diminishing the efficiency and unobtrusiveness of automated depression classification from speech. However, since this task also has therapeutic effects (Bauer, Schindler-Gmelch et al., 2024), it may allow simultaneous depression assessment and treatment. Finally, our task only included readings of scripted coping responses, whereas previous studies have suggested the superiority of free speech recordings for depression classification (Alghowinem et al., 2013). However, by designing this highly structured task, we could make sure that the prosody modulation instruction was responsible for the improvement in classification accuracy. In a next step, prosody modulation may be implemented in a free speech task, which most likely will allow the development of models with even better performance metrics.

CONCLUSION

Automated depression classification models based on speech analysis could bring significant advances for clinical practice if models can provide valid, reliable, and accurate classifications. This study suggests that speech elicitation tasks can be optimized by including conviction maximization instructions during the reading of anti-depressive self-statements to achieve better results in automated depression classification. We suggest that our approach amplifies the differences in speech patterns between depressed and non-depressed individuals and thereby increases sensitivity and accuracy for detecting a current depression diagnosis. Building on these findings and further systematically optimizing methods may allow the development of models with sufficient accuracy to be implemented in clinical practice for automated depression assessment.

ACKNOWLEDGMENT

This study was supported by grants from the German Research Foundation (Project IDs: BE4510/11-1, SCHU2508/8-1, KR3698/9-1). The authors would like to acknowledge our participants for their commitment and our student assistants Lisa Meisel, Lena Grüttner, Cem Akin, and Marleen Bogdanov for their support in data collection.

REFERENCES

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., & Parker, G. (2013). Detecting depression: a comparison between spontaneous and read speech. In IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7547–7551, Vancouver, BC

APA. (2013). Diagnostic and Statistical Manual of Mental Disorders (5th ed.). American Psychiatric Association.

Adolph, D., Tschacher, W., Niemeyer, H., & Michalak, J. (2021). Gait patterns and mood in everyday life: A comparison between depressed patients and non-depressed controls. Cognitive Therapy and Research, 1–13.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460.

Bauer, J. F., Gerczuk, M., Schindler-Gmelch, L., Amiriparian, S., Ebert, D. D., Krajewski, J., Schuller, B., & Berking, M. (2024). Validation of Machine Learning-Based Assessment of Major Depressive Disorder from Paralinguistic Speech Characteristics in Routine Care. Depression and Anxiety, 2024(1), 9667377.

Bauer, J. F., Schindler-Gmelch, L., Gerczuk, M., Schuller, B., Berking, M. (2024). Prosody-Focused Feedback Enhances the Efficacy of Anti-Depressive Self-Statements in Depressed Individuals - A Randomized Controlled Trial. [Manuscript submitted for publication].

Beck, A. T., & Bredemeier, K. (2016). A unified model of depression: Integrating clinical, cognitive, biological, and evolutionary perspectives. Clinical Psychological Science, 4(4), 596–619.

Craven, M. A., & Bland, R. (2013). Depression in primary care: current and future challenges. The Canadian Journal of Psychiatry, 58(8), 442–448.

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

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. Speech communication, 71, 10–49.

Dellwo, V., Karnowski, P., & Szigeti, I. (2006). Rhythm and speech rate: A variation coefficient for deltaC. In P. Karnowski & I. Szigeti (Eds.), Language and language-processing (pp. 231–241). Peter Lang.

Diedrich, A., Hofmann, S. G., Cuijpers, P., & Berking, M. (2016). Self-compassion enhances the efficacy of explicit cognitive reappraisal as an emotion regulation strategy in individuals with major depressive disorder. Behaviour research and therapy, 82, 1–10.

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

Filippi, P., Ocklenburg, S., Bowling, D. L., Heege, L., Güntürkün, O., Newen, A., & De Boer, B. (2017). More than words (and faces): evidence for a Stroop effect of prosody in emotion word processing. Cognition and Emotion, 31(5), 879–891.

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

Freudenthaler, H. H., & Neubauer, A. C. (2007). Measuring emotional management abilities: Further evidence of the importance to distinguish between typical and maximum performance. Personality and Individual Differences, 42(8), 1561–1572.

Ghio, L., Gotelli, S., Cervetti, A., Respino, M., Natta, W., Marcenaro, M.,... & Murri, M. B. (2015). Duration of untreated depression influences clinical outcomes and disability. Journal of affective disorders, 175, 224–228.

Grabe, E., & Low, E. L. (2008). Durational variability in speech and the Rhythm ClassHypothesis. In C. Gussenhoven & N. Warner (Eds.) Laboratory Phonology 7 (pp. 515–546). De Gruyter.

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

Jiang, H., Hu, B., Liu, Z., Yan, L., Wang, T., Liu, F.,... & Li, X. (2017). Investigation of different speech types and emotions for detecting depression using different classifiers. Speech Communication, 90, 39–46.

Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. Speech Communication, 88, 106–126.

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. Annual review of public health, 34, 119–138.

Klehe, U. C., & Latham, G. P. (2008). Predicting typical and maximum performance with measures of motivation and abilities. Psychologica Belgica, 48(2-3), 67–91.

Lewis, C. C., Boyd, M., Puspitasari, A., Navarro, E., Howard, J., Kassab, H.,... & Kroenke, K. (2019). Implementing measurement-based care in behavioral health: a review. JAMA psychiatry, 76(3), 324–335.

Long, H., Guo, Z., Wu, X., Hu, B., Liu, Z., & Cai, H. (2017). Detecting depression in speech: Comparison and combination between different speech types. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1052–1058, Kansas City, MO, USA.

Mitchell, A. J., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: a meta-analysis. The Lancet, 374(9690), 609–619.

Pettersson, A., Boström, K. B., Gustavsson, P., & Ekselius, L. (2015). Which instruments to support diagnosis of depression have sufficient accuracy? A systematic review. Nordic journal of psychiatry, 69(7), 497–508.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. Cognition, 73(3), 265–292.

Stuart, A. L., Pasco, J. A., Jacka, F. N., Brennan, S. L., Berk, M., & Williams, L. J. (2014). Comparison of self-report and structured clinical interview in the identification of depression. Comprehensive psychiatry, 55(4), 866–869.

Teasdale, J. D., & Barnard, P. J. (1993). Affect, Cognition, and Change: Re-Modeling Depressive Thought. Psychology Press.

Velten, Emmett (1968). A laboratory task for induction of mood states. Behaviour Research and Therapy 6 (4), 473–482.