An Alternative Approach to Depression Diagnosis: Predicting Individual Symptoms Through Speech and Text Analysis

Karim M. Ibrahim, Antony Perzo, Larsen D'hiet

Emobot, France

research@emobot.fr

Abstract

We propose a symptom-based approach to depression diagnosis by predicting individual symptoms through speech and text analysis. By combining traditional acoustic features with emotion recognition metrics—specifically, valence and arousal from speech and text—our model enhances transparency and accuracy in assessing depression severity [1]. Using the DAIC-WOZ dataset, we demonstrate improved prediction of individual symptoms compared to traditional methods. The clear distinction in valence across depression severity levels underscores the utility of the proposed features. Our method can be seamlessly integrated into existing clinical workflows, offering clinicians a non-invasive and interpretable tool for diagnosis and monitoring. Future work will focus on incorporating additional non-speech-related symptoms and validating the approach in clinical settings to further enhance its applicability and effectiveness.

Index Terms: depression, diagnosis criteria, clinical usefulness, emotion recognition

1. Introduction

Major Depressive Disorder (MDD) is a widespread condition with significant societal and economic impacts, driving the need for innovative approaches to its diagnosis and treatment [2, 3]. Among various digital health initiatives, speech has emerged as a promising indicator due to its hierarchical structure and potential to reflect mental health conditions [4]. Research has shown that speech from individuals with depression exhibits acoustic changes, such as reduced speech rate, decreased pitch variability, and changes in energy, which correspond to key depressive symptoms like diminished emotional expression and psychomotor retardation [5, 6].

Despite these findings, the clinical adoption of speech-based systems for depression diagnosis has been limited. Key challenges include the suboptimal performance of current models, small dataset sizes that limit generalizability, and a lack of transparency, which hinders clinical trust in such tools [1, 7]. Addressing these issues requires an approach that not only improves accuracy but also offers clear insights into the reasoning behind predictions.

In this study, we propose a method that predicts individual symptoms of depression through the analysis of speech and text features. Our approach leverages traditional acoustic features along with symptom-based features, such as valence and arousal, derived from emotion recognition and sentiment analysis [8]. By focusing on individual symptoms rather than a binary diagnosis, we aim to enhance transparency and provide clinicians with detailed, interpretable insights that can improve treatment strategies.

The paper is structured as follows: in section 2, the DAIC-WOZ dataset design is presented, explaining the collection procedure and the depressive symptoms included through the PHQ-8 Questionnaire [9], along with the preprocessing steps for feature extraction. In section 3, we explain our methodology for combining established acoustic measures with proposed symptom-based features, particularly valence and arousal from speech and text, aiming at predicting the depressive symptoms. Finally, in section 4, we present the experiments and evaluation results obtained using our proposed approach.

2. Dataset

Our analysis is built on the DAIC-WOZ dataset, obtained from the USC Institute for Creative Technologies[1]. This dataset, comprising 189 participants, is derived from the larger Distress Analysis Interview Corpus (DAIC), extensively discussed in existing literature [10, 11]. Its primary purpose is to diagnose psychological distress conditions, such as depression, using computer-agent-conducted interviews. The duration of interaction sessions varies from 7 to 33 minutes, averaging at approximately 16 minutes per session. The interviews have been transcribed and annotated within the dataset. Alongside interview transcripts, the dataset provides PHQ-8 scores, a widely employed diagnostic measure for depression assessment.

The dataset includes the answers for each of the individual questions in the PHQ-8 questionnaire. Each of these questions relate to one of the MDD symptoms, namely: 1) lack of Interest, 2) sadness, 3) sleep disruption, 4) tiredness, 5) appetite disruption, 6) feelings of failure, 7) troubles concentrating, and 8) psychomotor retardation. Each question has one of four answers about the frequency of the symptom: 1) Not at all, 2) several days, 3) more than half the days, and 4) nearly everyday. Our objective is to predict these symptoms independently using tailored features.

The PHQ-8 score distribution within the dataset, illustrated in Figure 1, reveals a predominance of sub-clinical and mild depressive symptoms, with a ratio of approximately 3:1 favoring non-depressed over depressed individuals. This distribution pattern suggests an inherent challenge in achieving high sensitivity in model predictions, given the lower prevalence of depressive symptoms among the participants. Additionally, the PHQ-8 scores are split into four categories to represent the severity: 1) below 5 is classified as `low depression severity`, 2) 5 to 10 as `mild depression`, 3) 10 to 15 as `moderate depression`, and 4) above 15 as `severe depression`. These are the four categories we use for classi-

[1]USC Institute for Creative Technologies, http://dcapswoz.ict.usc.edu/

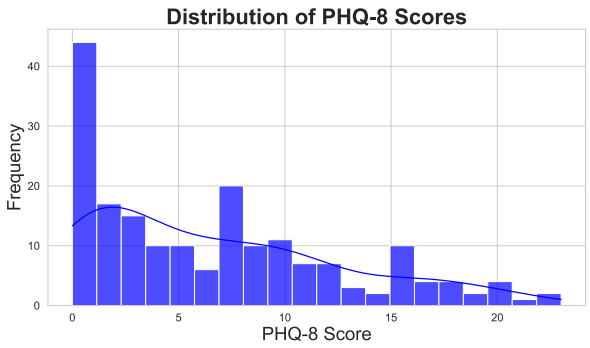

Figure 1: *The PHQ-8 score distribution within the DAIC-WOZ dataset. The distribution is highly skewed towards subclinical cases.*

fication throughout this study.

One of the main challenges in this dataset is the presence of the interviewer's speech in some of the recordings. We address this by applying a pre-processing step of speaker diarization to extract the patient's excerpts. We rely on using the open-source toolkit Pyannote [12, 13] to achieve this. We first apply speech segmentation on the entire interview to extract the speech excerpts. Then, we apply speaker identification to separate between the interviewer and interviewee. We select the most common speaker in the interview to be assigned as the patient, and extract the acoustic and textual features for the corresponding excerpts.

The dataset is divided into training (107 individuals) and development (35 individuals) sets to aid in the evaluation of machine learning models. However, to test the generalizability of our findings across the dataset, we implemented 5-fold cross-validation on the entire dataset such that each validation fold contained at least one sample for every PHQ-8 score in order to counter inherent sampling bias present in the dataset. We used the created splits to train and assess various classifiers, namely Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF). Out of which, SVM provided the best results, which are presented thereafter.

3. Methodology

We utilize both traditional acoustic features and novel symptom-based features to predict depressive symptoms from speech and text. This section outlines the key components of our feature extraction process and the machine learning models used for classification.

3.1. Acoustic features

We rely on the eGeMAPS feature set [14], a standardized set of acoustic features developed to capture emotional expression and affective states in speech. It includes descriptors related to frequency, energy, and spectral balance. To complement this, we include the first 13 Mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives, which are widely used in speech emotion and depression detection [15, 16]. All features are aggregated using the mean and standard deviation across the patient's speech excerpts from the interviews.

3.2. Symptoms-based features

In addition to traditional acoustic features, we propose several symptom-based features designed to capture long-term markers of depression, such as diminished emotional expression and psychomotor retardation [8]. These features focus on sentiment analysis in speech and text, as well as speech rate.

Speech-based sentiment: One core symptom of depression is a persistent low mood and loss of interest or pleasure (anhedonia) [8]. Recent advancements in speech analysis have provided opportunities to detect depression through short-term emotional states, such as valence (degree of pleasantness) and arousal (level of excitement) [5, 17]. Individuals with depression exhibit distinct emotional reactivity patterns, which can be effectively captured through speech, making emotional state tracking a valuable tool for early detection and personalized treatment [18, 19]. Using Speech Emotion Recognition (SER) systems, specifically pretrained models like Wav2vec 2.0, has shown significant promise for estimating valence and arousal from speech [20, 21]. Wav2vec 2.0 employs a convolutional neural network (CNN) to encode raw audio into low-level representations, followed by transformer layers to capture contextual information, and an optional quantization module to discretize the representations into units as needed for certain tasks [20]. In this study, we fine-tune Wav2vec 2.0 for the SER tasks, as suggested by approaches such as [22, 23], and use the MSP-Podcast dataset for fine-tuning [24]. The model achieves strong performance with concordance correlation coefficients (CCC) of 0.635 for valence and 0.745 for arousal on the MSP-Podcast test set [25]. After processing patient speech excerpts, we compute the mean and standard deviation for valence and arousal across all excerpts to correlate these features with depressive symptoms.

Text-based sentiment: Another important indicator of depression is the content and choice of words in speech [8]. To compute the valence and arousal values from patient transcripts in the DAIC-WOZ dataset, we use the NRC Valence, Arousal, and Dominance (VAD) Lexicon [26], a comprehensive resource containing emotional ratings for around 20,000 English words. The NRC VAD Lexicon was developed to facilitate emotion analysis, sentiment analysis, and related fields, helping to discern emotional content from textual data. Its ratings are based on human judgments collected via crowdsourcing, ensuring that the emotional associations are grounded in widespread human perception. This makes it a valuable tool for research across various domains, including customer feedback analysis, social media monitoring, and psychological studies. For each transcript excerpt, we extract the valence and arousal values for individual words and calculate the average values across the excerpt. We then compute the mean and standard deviation of these values across the entire interview to capture the patient's overall emotional expression.

Rate of speech: One of the main symptoms of depression is moving and speaking more slowly [8]. We aim at measuring this symptom using 2 features: *The total number of words in an utterance* and *the rate of speech*, which is calculated as the mean of the number of words divided by the length (in seconds) of each utterance. The unit is thus words/s. Utterances shorter than 7 words, chosen empirically, are not taken into account for a more reliable estimate. Similar to the previous features, we compute the mean and standard deviation across the whole interview for the patients' excerpts.

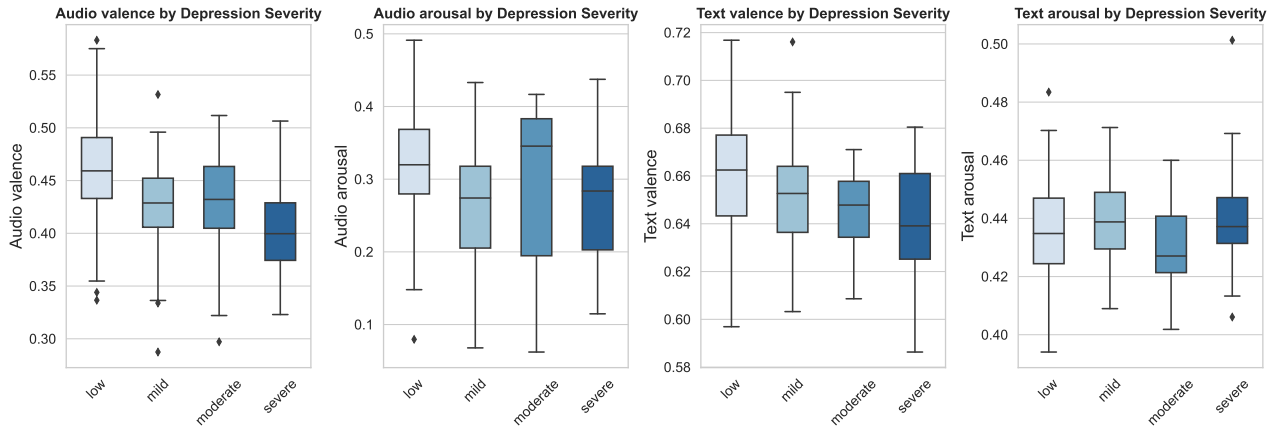

Figure 2: *Distribution of the speech and text valence values across different depression categories. We observe a clear distinction between low depression severity and severe depression in the values of the valence.*

3.3. Model setup

To evaluate the effectiveness of these features, we use a Support Vector Machine (SVM) with a radial basis function (RBF) kernel as our classification model. The eGeMAPS features, combined with the proposed symptoms-based features, are used as input to the model. Features are standardized prior to training, and feature selection is applied to reduce dimensionality and enhance performance. The number of selected features and the SVM's regularization factor are tuned as hyperparameters to optimize the model's accuracy and generalization. This approach enables us to assess the contribution of both traditional acoustic and symptoms-based features in predicting depressive symptoms.

4. Results and Discussion

4.1. Distribution analysis

First, we examine the distribution of valence and arousal in both speech and text across varying levels of depression severity, as illustrated in Figure 2. The valence values (on a scale from -1 to 1, where lower values indicate more negative emotions) show a consistent decrease as depression severity increases. This trend is particularly noticeable in the distinction between low depression severity and severe depression, where individuals with higher depression scores exhibit notably lower valence values in their speech. This finding suggests that as depression becomes more severe, the emotional content of speech shifts toward more negative expressions. Similarly, the text valence values also drop with increasing depression severity, indicating that participants use more negatively connoted language as their depression worsens. This supports the idea that speech and text analysis can effectively capture emotional states linked to depression. In contrast, the arousal values (measured on a scale from 0 to 1, where lower values indicate calmer or less excited speech) present a more overlapping distribution across different depression severity levels. While there is some variation in arousal, the overlap suggests that arousal may be less strongly correlated with depression severity compared to valence. This could be because arousal reflects a broader range of emotional and physiological states that may not be as tightly linked to depression as valence.

4.2. Performance evaluation & Feature importance analysis

We evaluated the performance of the models by comparing two sets of features: acoustic features alone and a combination of acoustic and symptoms-based features. Table 1 presents the results for predicting the severity of each depression symptom, with metrics reported as mean ± standard deviation across a 5-fold cross-validation process. The metrics used are accuracy, recall, precision, and F1-score, with all values scaled between 0 and 1, where higher scores indicate better model performance.

The F1-score is a performance metric that combines precision (the proportion of true positive predictions among all positive predictions) and recall (the proportion of true positive predictions among all actual positive cases) into a single score. It is particularly useful when the data is imbalanced, as it provides a balance between precision and recall, helping to assess the model's effectiveness in cases where false positives and false negatives are both important.

The inclusion of symptoms-based features led to improvements in nearly all metrics across various symptoms. Overall, accuracy in detecting specific symptoms using only acoustic features ranged from 0.34 in sleep disruption to 0.71 in psychomotor retardation, whereas the combined feature set allowed the accuracy for psychomotor retardation to rise to 0.75. Similarly, the F1-score for depression severity classification improved from 0.32 to 0.36. These results demonstrate that integrating emotional and symptom-specific features with traditional acoustic features enhances the model's ability to capture the nuanced patterns associated with different depression symptoms.

In contrast, for symptoms such as sleep disruption and feelings of failure, the improvements were more modest, and in some cases, performance slightly decreased. For example, accuracy for sleep disruption remained at 0.34, and the F1-score showed no significant improvement in these cases. This may indicate that speech-based features alone may not fully capture certain symptoms, particularly those that are not directly observable through vocal characteristics, such as physical symptoms like sleep and appetite disruptions.

The results also highlight the importance of valence and arousal in symptom prediction. The clear distinction in valence

Table 1. *Evaluation results on the test set for predicting the severity of each symptoms of depression using acoustic features and proposed features. Evaluation using accuracy , recall, precision, and F1-score. results are presented as mean ± std. across 5-fold cross validation.*

	Acoustic Features				Acoustic + Symptoms-based Features			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
Lack of Interest	0.44 ± 0.11	0.28 ± 0.08	0.27 ± 0.07	0.27 ± 0.07	0.42 ± 0.13	0.30 ± 0.09	0.31 ± 0.10	**0.30 ± 0.10**
Sadness	0.41 ± 0.06	0.34 ± 0.12	0.37 ± 0.15	0.34 ± 0.13	0.42 ± 0.05	0.36 ± 0.08	0.40 ± 0.05	**0.35 ± 0.04**
Sleep Disruption	0.34 ± 0.06	0.33 ± 0.08	0.35 ± 0.05	**0.33 ± 0.06**	0.35 ± 0.09	0.32 ± 0.05	0.33 ± 0.04	0.31 ± 0.05
Tiredness	0.35 ± 0.06	0.27 ± 0.06	0.29 ± 0.04	0.26 ± 0.04	0.37 ± 0.04	0.30 ± 0.05	0.38 ± 0.11	**0.30 ± 0.06**
Appetite Disruption	0.38 ± 0.08	0.28 ± 0.03	0.27 ± 0.04	0.26 ± 0.03	0.35 ± 0.06	0.30 ± 0.04	0.29 ± 0.03	**0.29 ± 0.03**
Feelings of Failure	0.40 ± 0.08	0.31 ± 0.10	0.38 ± 0.14	**0.31 ± 0.11**	0.39 ± 0.08	0.30 ± 0.02	0.32 ± 0.04	0.30 ± 0.03
Concentrating Troubles	0.39 ± 0.02	0.33 ± 0.11	0.30 ± 0.04	0.30 ± 0.05	0.51 ± 0.05	0.32 ± 0.07	0.32 ± 0.10	**0.31 ± 0.08**
Psychomotor Retardation	0.71 ± 0.06	0.31 ± 0.07	0.30 ± 0.10	0.30 ± 0.07	0.75 ± 0.04	0.34 ± 0.04	0.39 ± 0.08	**0.34 ± 0.04**
Depression Severity	0.41 ± 0.09	0.35 ± 0.10	0.36 ± 0.06	0.32 ± 0.09	0.49 ± 0.07	0.38 ± 0.07	0.39 ± 0.12	**0.36 ± 0.07**

across depression severity levels (especially between low and severe categories) suggests that valence is a strong indicator of depression severity. On the other hand, the limited differentiation in arousal implies that it may be less effective as a standalone predictor, though it likely contributes to the overall emotional profile when combined with other features.

4.3. Limitations and future work

While the results are promising, there are important limitations to consider. One of the key challenges lies in the reliance on self-reported PHQ-8 scores, which, while widely used in clinical settings for depression screening, may introduce biases. Individuals may underreport or overreport their symptoms due to recall bias or difficulties in recognizing and articulating their emotional states. This self-assessment process may not fully capture the complexity of depression, particularly in cases where symptoms fluctuate over time or where patients experience difficulty in emotional awareness. The PHQ-8's focus on the frequency of symptoms over the past two weeks may also fail to account for individuals with high variability in their symptoms, potentially overlooking critical nuances in their mental health. Moreover, the PHQ-8 questionnaire predominantly addresses cognitive and affective symptoms, such as mood, interest, and concentration, while placing less emphasis on physical symptoms like sleep disturbances and appetite changes, which are crucial aspects of depression for some individuals. As a result, it may not provide a complete picture of the disorder for those whose depression manifests primarily through physical symptoms. [8]

Additionally, the dataset used in this study primarily emphasizes speech-related symptoms, which may leave out other crucial non-speech-related symptoms, such as sleep disturbances and appetite changes, that play an integral role in understanding depression comprehensively. These symptoms are not directly observable through speech analysis, yet they are essential indicators of mental health and significantly influence diagnosis and treatment planning.

Future research should aim to address these limitations by adopting multimodal approaches that incorporate not only speech features but also physiological markers (e.g., heart rate variability, sleep patterns) and behavioral data (e.g., movement tracking, social interaction metrics). Such multimodal data would improve the robustness and accuracy of depression diagnosis models by capturing a more holistic view of the patient's mental state. Expanding the scope of analysis to include these non-speech symptoms would enrich our understanding of depression and enhance the precision of diagnostic tools, ultimately leading to more tailored and effective treatment interventions.

5. Conclusion

The research presented in this paper demonstrates the potential of a symptom-based approach to depression diagnosis, which focuses on predicting individual depressive symptoms rather than providing an overall binary classification. By leveraging a combination of traditional acoustic features and emotionally-driven features such as valence and arousal, we achieved more accurate and interpretable predictions of specific symptoms of depression. In particular, the clear differentiation in speech valence between low and severe depression severity highlights the utility of these features in identifying emotional states associated with depressive disorders.

This approach addresses some of the challenges associated with current depression diagnosis methods, including the need for greater transparency and clinical interpretability. By offering insight into specific symptoms, clinicians can better tailor their interventions, leading to more personalized and effective treatment strategies.

However, this study also underscores the need for further exploration. The integration of non-speech-related symptoms, such as sleep patterns and appetite changes, will be essential for developing a more comprehensive diagnostic tool. Future research should focus on multimodal approaches that incorporate not only speech and text analysis but also physiological and behavioral data to capture a broader range of depression symptoms. Such developments could significantly enhance the precision and clinical usefulness of depression diagnosis systems.

In conclusion, this study offers a promising direction for improving the assessment of depression through the prediction of individual symptoms, contributing to the broader goal of developing transparent, interpretable, and clinically applicable diagnostic tools in mental health.

6. References

[1] V. P. Martin and J.-L. Rouas, "Estimating symptoms and clinical signs instead of disorders: the path toward the clinical use of voice and speech biomarkers in psychiatry," in *Proceedings of the International Conference on Acous-*

tics, Speech, and Signal Processing (ICASSP), 2024.

[2] F. Matcham, C. Barattieri di San Pietro, V. Bulgari, G. De Girolamo, R. Dobson, H. Eriksson, A. Folarin, J. Haro, M. Kerz, F. Lamers *et al.*, "Remote assessment of disease and relapse in major depressive disorder (radar-mdd): a multi-centre prospective cohort study protocol," *BMC psychiatry*, vol. 19, pp. 1–11, 2019.

[3] C. G. Fairburn and V. Patel, "The impact of digital technology on psychological treatments and their dissemination," *Behaviour research and therapy*, vol. 88, pp. 19–25, 2017.

[4] N. Topooco, H. Riper, R. Araya, M. Berking, M. Brunn, K. Chevreul, R. Cieslak, D. D. Ebert, E. Etchmendy, R. Herrero *et al.*, "Attitudes towards digital treatment for depression: a european stakeholder survey," *Internet interventions*, vol. 8, pp. 1–9, 2017.

[5] 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.

[6] D. J. France, R. G. 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.

[7] M. Milling, F. B. Pokorny, K. D. Bartl-Pokorny, and B. W. Schuller, "Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell," *Frontiers in digital health*, vol. 4, 2022.

[8] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Publishing, 2013. [Online]. Available: https://books.google.fr/books?id=-JivBAAAQBAJ

[9] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[10] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014.

[11] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014.

[12] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, 2023.

[13] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, 2023.

[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, and K. P. Truong, "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, 2016.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010.

[16] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *Proceedings of the international conference on acoustics, speech and signal processing (ICASSP)*, 2014.

[17] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[18] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[19] S. Koops, S. G. Brederoo, J. N. de Boer, F. G. Nadema, A. E. Voppel, and I. E. Sommer, "Speech as a biomarker for depression," *CNS & Neurological Disorders Drug Targets*, vol. 22, no. 2, pp. 152–160, March 2023. [Online]. Available: https://doi.org/10.2174/1871527320666211213125847

[20] 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*, 2020.

[21] P. Dagum, "Digital biomarkers of cognitive function," *npj Digital Medicine*, vol. 1, no. 1, p. 10, March 2018. [Online]. Available: https://doi.org/10.1038/s41746-018-0018-4

[22] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, 2021.

[23] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[24] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[25] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[26] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the 56th annual meeting of the association for computational linguistics*, 2018.