Emotional Dynamics in Semi-Clinical Settings: Speech Emotion Recognition in Depression-Related Interviews

Bakir Hadžić[1][a], Julia Ohse[2][b], Mohamad Eyad Alkostantini[1][c], Nicolina Peperkorn[2][d], Akihiro Yorita[3][e], Thomas Weber[1], Naoyuki Kubota[4][f], Youssef Shiban[2][g] and Matthias Rätsch[1][h]

[1]*Reutlingen University, Reutlingen, Germany*

[2]*Private University of Applied Sciences Göttingen, Göttingen, Germany*

[3]*Daiichi Institute of Technology, Kagoshima, Japan*

[4]*Tokyo Metropolitan University, Tokyo, Japan*

Keywords: Speech Emotion Recognition, Artificial Intelligence, Mental Health, Depression, Emotional Dynamics.

Abstract: The goal of this study was to utilize a state-of-the-art Speech Emotion Recognition (SER) model to explore the dynamics of basic emotions in semi-structured clinical interviews about depression. Segments of N = 217 interviews from the general population were evaluated using the emotion2vec+ large model and compared with the results of a depressive symptom questionnaire. A direct comparison of depressed and non-depressed subgroups revealed significant differences in the frequency of happy and sad emotions, with participants with higher depression scores exhibiting more sad and less happy emotions. A multiple linear regression model including the seven most predicted emotions plus the duration of the interview as predictors explained 23.7 % of variance in depression scores, with happiness, neutrality, and interview duration emerging as significant predictors. Higher depression scores were associated with lesser happiness and neutrality, as well as a longer interview duration. The study demonstrates the potential of SER models in advancing research methodology by providing a novel, objective tool for exploring emotional dynamics in mental health assessment processes. The model's capacity for depression screening was tested in a realistic sample from the general population, revealing the potential to supplement future screening systems with an objective emotion measurement.

1 INTRODUCTION

Emotions are an essential, and natural way of human communication and expression. They play a crucial role in defining the nature and dynamics of conversation among speakers. In clinical and semi-clinical psychological settings, understanding these emotional dynamics is particularly important for accurate clinical assessments as it can greatly enhance diagnostic accuracy. Recent advancements in Speech Emotion Recognition (SER) provide a very useful tool for automated emotion detection from voice, offering a more objective and standardized method to capture and interpret human emotions, especially for research purposes.

Based on this, the main aim of our approach was to utilise the state-of-the-art SER model on audio dataset collected in semi-clinical settings on depression-related interviews. By analysing emotional occurrences and their dynamics during the interviews, our study aimed to identify key emotional cues that are associated with depression. Such a novel application of machine learning models in the field of clinical psychology holds great potential for providing an objective, non-invasive and scalable approach to understanding the role of emotions in mental health assessments.

2 RELATED WORK

Speech emotion recognition (SER) has been a topic of growing interest in the field of affective computing and social signal processing (Wang et al., 2021). This is due to the potential applications of such technology in areas like human-computer interaction, customer service and psychological medical diagnosis. The term

[a] https://orcid.org/0009-0003-1197-7255
[b] https://orcid.org/0009-0005-3344-4753
[c] https://orcid.org/0009-0003-8421-4689
[d] https://orcid.org/0009-0008-9481-9354
[e] https://orcid.org/0000-0003-4733-4553
[f] https://orcid.org/0000-0001-8829-037X
[g] https://orcid.org/0000-0002-6281-0901
[h] https://orcid.org/0000-0002-8254-8293

Hadžić, B., Ohse, J., Alkostantini, M. E., Peperkorn, N., Yorita, A., Weber, T., Kubota, N., Shiban, Y. and Rätsch, M.
Emotional Dynamics in Semi-Clinical Settings: Speech Emotion Recognition in Depression-Related Interviews.
DOI: 10.5220/0013415700003938

In *Proceedings of the 11th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2025)*, pages 104-113
ISBN: 978-989-758-743-6; ISSN: 2184-4984

"SER system" usually refers to a group of techniques that analyse and categorize speech signals in order to identify the emotions that are present in them (Akçay and Oğuz, 2020).

Speech emotion recognition in academic research traditionally was done manually with human raters or with self-report scales. But even humans when faced with such task, recognition rates are not higher than 90 % (Akçay and Oğuz, 2020), while self-report measures rely on retrospective insights and do not capture moment-to-moment fluctuations (Joormann and Stanton, 2016). Furthermore, such a process is highly subjective, context-dependent and bias-prone, without a valid methodology for further evaluation. Due to this methodological lack, the role and dynamics of emotions in various fields of application could not be explored properly by the academic community.

In the last decade, with the increase in computing power, deep learning has become a research hotspot when it comes to SER tasks. Compared with traditional machine learning methods, deep learning technology has the advantage of extracting high-level semantic features (Wang et al., 2021). Recent studies explored the numerous uses of deep learning models in SER tasks, particularly in fields where understanding and responding to human emotions is critical. Sentiment Analysis, Entertainment and Media and Mental Health Monitoring are some of the key application areas (Islam et al., 2024). Emotion2vec is likely one of the most effective solutions for speech recognition of emotions because of its specialized design and outstanding performance in capturing emotional elements across various languages and tasks (Atmaja and Sasou, 2022). Among others, Wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Van Niekerk et al., 2022) are more versatile and can be used for a broader range of speech tasks, including Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER), but may require more fine-tuning for optimal SER performance (Yang et al., 2024).

In our literature review process, we identified numerous earlier studies on SER systems, many focusing on the speech features and emotion categorization (Mohammed et al., 2024; Mustafa et al., 2018; Anusha et al., 2021) or depression prediction from the audio features (Prabhu et al., 2022; Rejaibi et al., 2022; Aloshban et al., 2022). But none of the studies we identified addressed the occurrence of emotions in semi-clinical or clinical settings, nor psychotherapy, counselling, or other mental health content-related conversations.

Such a literature gap is surprising, considering the long-standing recognition in clinical psychology of the strong correlation between specific emotions and depression. Even the Diagnostic and Statistical Manual of Mental Disorders (DSM V) states that depression cannot be diagnosed without some evidence of a persistent mood disturbance, such as excessive sadness and/or a greatly reduced experience of pleasure (APA, 2022). Therefore, a deeper exploration of this relationship is very much needed.

3 METHODOLOGY

3.1 Dataset

This study utilized the KID dataset introduced in our previous research on text-based depression detection (Danner et al., 2023; Hadzic et al., 2024a; Hadzic et al., 2024b; Ohse et al., 2024). The KID dataset is a set of semi-clinical depression-related interviews conducted at the Private University of Applied Sciences in Göttingen, Germany. Interviews were led by trained psychology students under the supervision of experienced mental health experts, with participants coming from the general population. The dataset includes a total of 217 interviews led in the German language.

The demographic breakdown of participants is as follows: $male = 67$, $female = 134$, $divers = 4$, $missing\ values = 12$, with an average age of $M = 32.67$, $SD = 11.78$, indicating a well-balanced dataset in terms of demographic characteristics. The average duration of the interviews is $M = 1050.21$, $SD = 497.20$ measured in seconds. These interviews followed the GRID-HAMD structure, a semi-structured interview protocol designed for depression-related assessments (Williams et al., 2008). Each participant involved in the interviews, also filled in PHQ-8 questionnaire to determine their ground truth depression scores. PHQ-8 scale is a concise and widely used psychometric tool that consists of eight items making it very time-efficient (Kroenke et al., 2009). Its brevity and strong psychometric validity make it very commonly used for research purposes.

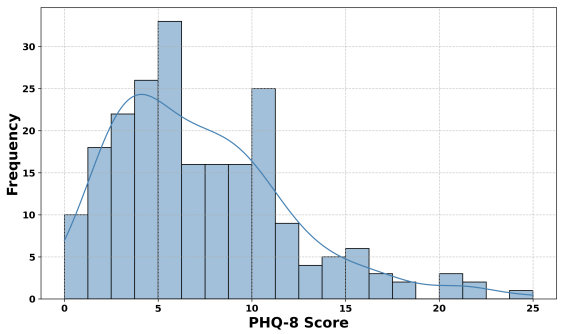

Figure 1: Distribution of PHQ-8 Depression Scores.

The descriptive statistics revealed that the mean PHQ-8 score among participants was $M = 7.27$, with a standard deviation of $SD = 4.72$, indicating a wide range of depression severity within the sample. The distribution of PHQ-8 scores is visualized in Figure 1.

3.2 Speech Emotion Recognition

The task of speech emotion recognition has been challenging researchers and scientists worldwide for the last few decades. In recent years, very promising open-source approaches are arising in the literature, such as HuBERT (Van Niekerk et al., 2022), Waw2vec 2.0 (Baevski et al., 2020) and emotion2vec (Ma et al., 2023) achieving impressive performance results. In this study, we employed emotion2vec+ large model due to its robust design and exceptional performance.

The **emotion2vec+ Model** is a series of foundational models for speech emotion recognition (SER), developed with an aim to overcome the effects of language and recording environments through data-driven methods to achieve universal, robust emotion recognition capabilities. The model was fine-tuned using filtered, large-scale pseudo-labelled data, resulting in a model size of approximately 300M parameters, specifically, it utilized approximately 40 000 hours of high-quality speech-emotion data, extracted from a pool of 160 000 hours, ensuring a rich and diverse dataset for training.

In the literature, the emotion2vec+ large model has established itself as a state-of-the-art solution in SER tasks, surpassing both traditional supervised learning methods and other widely used open-source models (Ma et al., 2023). Model was fine-tuned using academic datasets, including EmoBox a multilingual multi-corpus speech-emotion recognition toolkit designed to facilitate research in the field of speech-emotion recognition, enabling both intra-corpus and cross-corpus evaluations, making it a valuable resource for assessing the performance of speech-emotion recognition models (Ma et al., 2024).

Evaluations of the emotion2vec large model showcased that it achieves F_1 scores between .76 and .98 on standardised datasets across various languages (Ma et al., 2023). On the two commonly used German datasets, EmoDB and PAVOQUE, the model reached F_1 scores of .98 and .94.

The model is designed to predict a variety of emotions, such as anger, disgust, fear, happiness, neutrality, sadness and surprise. One further category is included, others/unknown, to capture feelings that do not fall into these predetermined categories.

3.3 Data Preprocessing

Speaker Diarization. The task of speaker diarization is the process of partitioning speech data into homogenous sections, with each segment corresponding to a specific speaker (Bredin, 2023). This was necessary, as our analysis of participants' emotions in clinical interviews required separating the interview audio files by speaker. Since only the interviewee's emotions were of interest, we preprocessed each file to identify the speaker and split the audio accordingly, using only the file containing the participant's responses for evaluation. To perform this speaker diarization, we utilized Pyannote (Bredin, 2023), an open-source toolkit built in Python. Pyannote employs pre-trained models and pipelines based on the PyTorch framework to achieve state-of-the-art performance by extracting distinctive vocal traits, speaking styles, and other speech-related characteristics from extensive labeled speech datasets (Plaquet and Bredin, 2023; Bredin, 2023).

After isolating only participant-related content, each recording was divided into shorter segments. Because each segment had a single estimate of emotional values, this enabled us to predict emotions throughout the interview at various points in time. This made it possible to assess the range of emotions that emerged throughout the interview.

For this, several segmentation approaches were tested, using sequences of 2, 2,5, and 3 seconds. While no significant differences were observed, the 2,5-second segmentation approach yielded slightly better results, though the improvement was not statistically significant, we noticed that 2,5-second segmentation led to the lowest number of noise predictions. Based on this observation, we have decided to proceed with 2,5-second segments for analysis. However, future studies should further explore this approach. The dataset was segmented into a total of 28 486 sequences, with an average of 131,27 segments (predicted emotions) per interview.

4 RESULTS

To explore the role of emotional dynamics in semi-clinical depression-related interviews we conducted various statistical analyses. First, we calculated basic demographic statistics, followed by t-testing to compare depressed and non-depressed samples on the level of each predicted emotion. Furthermore, we did multiple regression analysis to determine which of the variables measured by the system contribute to the prediction of depression scores. Lastly, we analysed the distribution of emotion prediction throughout the

interviews to examine changes over time.

4.1 Descriptive Statistics

Descriptive statistics were computed to better understand the distribution of the data. The average duration of the recordings was $M = 326.87$, $SD = 259.15$ measured in seconds after excluding the interviewer's contributions. The dataset was segmented into a total of 28 486 sequences, with an average of 131,27 predicted emotions per interview. Table 1 provides a visual representation of the distribution for each predicted emotion, PHQ-8 scores, and length of the interview. Among the identified emotions, *neutrality* was the most common, with a mean value of $M = 46.24$ *and* $SD = 14.35$. In contrast, the emotions *angry* ($M = 0.80$, $SD = 1.21$) and *surprise* ($M = 2.01$, $SD = 2.22$) had significantly lower mean values reflecting the rarity of their appearances among the predictions.

Table 1: Mean and Standard Deviations of Emotional and Temporal Variables.

Feature	Mean	SD
Interview Length	1050.21	497.21
Participant Length	326.87	259.15
Angry	0.80	1.21
Surprise	2.01	2.22
Fear	4.41	3.96
Disgust	6.72	6.13
Happy	8.05	6.38
Sad	16.03	12.04
Others	15.76	7.25
Neutral	46.23	14.35
PHQ-8	7.27	4.72

Table 2 presents Spearman coefficients of correlation among emotions and interview length with the PHQ-8 (depression) score. Introduced variable participant length stands for the duration of the interview with only participants' content isolated. The Spearman correlation analysis reveals significant associations between depression scores and several emotional variables. Emotions of sadness, fear, and disgust had positive correlations, while happiness and neutrality exhibited negative correlations with the PHQ-8 scores. Only the emotion of surprise had no significant correlation with the depression-related scores. Additionally, both interview length and participant length are found to be positively correlated with depression scores.

Table 2: Spearman Correlation Coefficients of PHQ-8 Scores with Emotional and Temporal Variables.

Variable	Spearman	p
Angry	−0.125	0.067
Sad	0.305	0.001
Happy	−0.191	0.005
Neutral	−0.193	0.004
Disgust	0.151	0.027
Fear	0.214	0.002
Surprise	−0.020	0.772
Interview length	0.447	0.001
Participant length	0.383	0.001

4.2 Statistical Analysis

In the first phase of our analysis, we divided participants into two groups depending on the PHQ-8 score they achieved. In the literature regarding PHQ-8 interpretation scores, as a threshold score between no indication of depression and mild-depression indication, a cut-off score of 10 is usually used (Kroenke et al., 2009). A total number of 60 participants in the sample achieved a score of 10 or higher, whereas 157 participants achieved a PHQ-8 score lower than 10. Such an imbalance between the two groups is also expected to be found in the general population. The distribution of emotion predictions and their comparison between depressed and non-depressed participants is shown in the Figure 2.

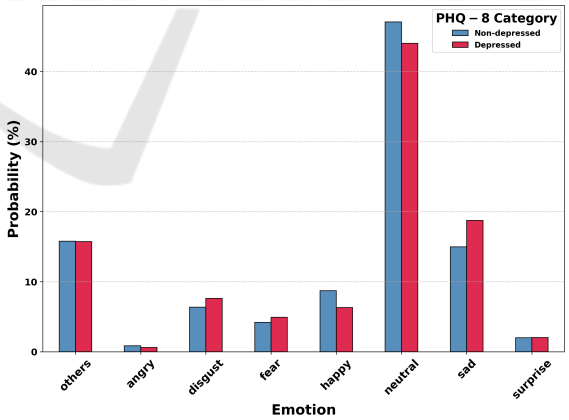

Figure 2: Emotional Probabilities Across PHQ-8 Depression Categories.

Comparison of Means: t-Test Analysis. To compare if there is statistical difference among depressed and non-depressed participants regarding the emotions being predicted in their interviews, we have calculated t-tests comparing these two groups on each of the seven predicted emotions. Before conducting the test,

Table 3: Multiple Linear Regression Results Predicting Depression Score (PHQ-8).

Variable	β	SE	t	p
(Constant)	6.527	5.516	1.183	0.238
Angry	−0.411	0.263	−1.563	0.120
Disgust	0.264	0.315	0.841	0.401
Fear	0.079	0.427	0.185	0.853
Happy	−1.352	0.431	0.185	0.002
Neutral	−2.536	1.158	−2.190	0.030^{**}
Sad	0.285	0.429	0.664	0.507
Surprise	0.496	0.274	1.813	0.071
Interview length	2.027	0.361	5.615	$<0.001^{***}$

Note: $R^2 = .237$, $F(8.208) = 8.08$, $p < .001$ * $p < .05$, ** $p < .01$, *** $p < .001$

we assessed whether all assumptions for parametric comparisons were met. To evaluate the normality of the distribution, we performed a Kolmogorov-Smirnov test.

The results indicated that the assumption of normal distribution was violated for five emotions: sad, happy, angry, disgust, and surprise. Only the distributions for neutral and others were not significant at $p < .05$, indicating normality.

Given the options of using non-parametric statistics or transforming the data, we opted for the latter, as further statistical analyses were intended later in the process. After performing a log transformation, a repeated Kolmogorov-Smirnov test confirmed that all emotions were normally distributed.

To assess the assumption of homogeneity of variance, we conducted Levene's test. The results indicated that the variances were equal across the groups $(p > .05)$ for each variable, satisfying the assumption for conducting parametric tests. After this confirmation, we proceeded with conducting t-tests.

The results of the t-tests revealed significant differences between high and low depression groups for two emotions. A significant difference was found in the level of happiness ($t = -2.037,\ p = .043$) with happy emotion being significantly more often predicted among non-depressed participants, while sad emotion ($t = 2.199,\ p = 0.029$) being predicted significantly less in the non-depressed group. For the other emotions, no significant differences were found: others ($t = -0.046,\ p = 0.963$), angry ($t = -0.781,\ p = 0.436$), disgust ($t = 1.607,\ p = 0.109$), fear ($t = 1.772,\ p = 0.078$), neutral ($t = 1.409,\ p = 0.160$) and surprise ($t = 0.734,\ p = 0.464$).

Multiple Linear Regression Analysis. To determine which emotion has the highest predictive power in explaining depression scores, we conducted a multiple regression analysis. In the predictive model, we included also the length of the participant speaking part in the interview. Before performing this analysis, it was essential to ensure that all assumptions for regression were met. These include the linear relationship between the variables, normality of residuals, not violating the assumption of homoscedasticity, low multicollinearity and independence of errors. Linearity was assessed using scatterplots of the independent variables against the dependent variable, and the assumption was found to be satisfactory. The normality of the residuals was achieved through the previously described log transformation, and the Kolmogorov-Smirnov test confirmed the normality of the residuals. The Breusch-Pagan test yielded a p-value of 0.228, indicating that the assumption of homoscedasticity was not violated. Furthermore, the Durbin-Watson statistic was 1.909, indicating that no significant autocorrelation of residuals is registered, as the value is close to 2, suggesting a satisfactory score regarding independence of errors. Lastly, the Variance Inflation Factor (VIF) indicated high multicollinearity due to the high correlation of variable others with the rest of the variables. After removing the variable others from the predicting model, all VIF scores were in the range of 1.11 to 1.98, indicating satisfactory multicollinearity between the variables in the model.

After confirming that all assumptions were met, we proceeded by including six emotions and the length of the interview, with only the participants' part included as an additional variable, to predict the depression score measured by the PHQ-8.

Results of conducted multiple regression analysis predicting depression scores (PHQ-8) from emotions and interview length have shown that the model explains 23.7 % of the variance in the depression scores ($R^2 = .237$). Such results indicated that the included model could predict a moderate portion of variance in depression scores. Among emotions, the only significant predictors were happy and neutral emotions.

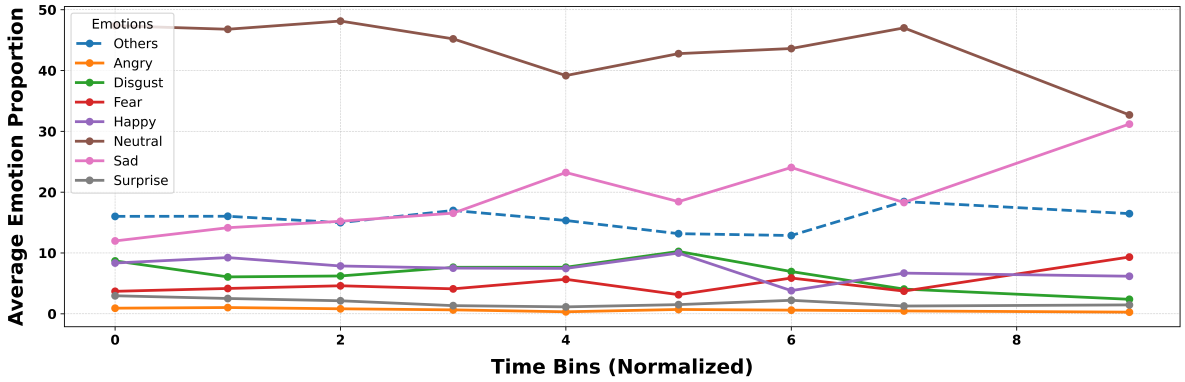

Figure 3: Emotion Trends over Time.

More precisely, higher levels of happiness were associated with a significant decrease in depression scores ($\beta = -1.352,\ p < .01$), while the same direction of effect is found in neutral emotion ($\beta = -2.536,\ p < .05$), meaning that increase of neutral emotions predicted is related to decrease in depression scores. Other emotions, such as anger ($\beta = -0.411,\ p > .05$), disgust ($\beta = 0.264,\ p > .05$), fear ($\beta = 0.079,\ p > .05$), sad ($\beta = 0.285,\ p > .05$), and surprise ($\beta = 0.496,\ p > .05$), did not show significant relationships with depression scores, suggesting that they do not have a strong predictive effect of depression scores in this model. On the other hand, Interview length was found to be a highly significant predictor of depression scores ($\beta = 2.027,\ p < .001$), indicating that longer interviews are associated with higher depression scores. Results are visually presented in the Table 3.

4.3 Dynamics of Emotions

As a final step of our analysis, we explored how the distribution of emotions varies throughout the interview. We divided interviews into eight time bins and in each time bin we calculated the proportion of specific emotions being predicted at each time bin. As the interviews followed the same guidelines, each session was standardised and the same questions were raised at the same time points in each interview.

Results are presented in the Figure 3. As the interview progressed, emotions remained relatively stable in the first three time bins, with minimal fluctuations in the emotional distribution. However, from the fourth time bin, greater variation in emotional expression was observed, indicating that the emotional dynamics of the interview changed as the interview progressed. In the final two time bins, there was a significant increase in the proportion of sad emotions, while the frequency of neutral emotions significantly decreased. Additionally, when comparing the beginning and the end of the interview, a growing trend in frequencies of variable fear was observed toward the end. When compared to the beginning, a notable decrease has also been found in the expression of disgust. Happiness, anger, surprise and others/noise proportions stayed quite stable during the whole process of the interview.

5 DISCUSSION

The goal of this study was to utilize a state-of-the-art Speech Emotion Recognition (SER) model to explore the dynamics of emotions in semi-clinical, depression-related interviews. On a set of recordings, where interviews averaged $17{,}5$ minutes in duration, we used a speaker diarization tool to isolate participants' spoken content and extract prosodic features for emotion predictions. Using only speaker-isolated data, we ran the SER model to predict emotions at $2{,}5$-second segments throughout the interviews.

The most frequently predicted emotions during the interviews were neutral and others/noise, which aligns with expectations for unstandardised, real-world applications. Among the basic emotions, sadness was the most frequently predicted. This was followed, at a significant gap, by happiness, disgust, and fear. Emotions such as anger and surprise occurred only rarely.

The dataset used in this study was collected in the context of depression screening, with participants' depression scores measured using the PHQ-8 scale. When the sample was divided into participants with and without indications of depression, t-tests were conducted to compare the groups based on the frequency of each emotion. Significant differences were observed for the emotions happy and sad. More precisely, non-depressed participants exhibited higher frequencies of happy emotions, while sadness was more frequently noted in the speech of depressed participants.

Results of multiple linear regression have shown

that the model with seven emotions (sad, happy, anger, disgust, fear, surprise and neutral) together with the variable interview length (only participant speech included) can explain 23.7 % of the variance in the depression (PHQ-8) score. Such a relatively small amount of variance explained by the model is also expected, as depression is a very complex variable influenced by numerous factors beyond emotional expressions. Among the emotions, only happy and neutral were significant predictors. Specifically, higher levels of happiness and neutral emotions were associated with lower depression scores, indicating that participants with higher depression scores exhibit more happy and neutral expressions. The variable length of the interview was proven to be the most significant predictor of the depression score in the model. A possible explanation for such results may be the fact that participants with depression symptoms elaborated more on their feelings and states, while non-depressed participants answered rather briefly on the occurrence of the depressive symptoms. The average duration of participants speaking in the depressed sample, measured in seconds amounted to $M = 290.45$, $SD = 240.20$, while in the non-depressed sample $M = 422.16$, $SD = 283.84$. While further analysis is needed for additional explanations, these results suggest that participants with higher depression scores are ready and open to discussing their mental health and symptoms in detail.

The observed shift in the emotional expression as the interview progressed, suggests that emotional expression is evolving and shifting during the duration of the interview. This may reflect participants' increasing comfort or discomfort as the interview continues, potentially influenced by factors such as the nature of the questions. An increase of negative emotions towards the end of the interview could potentially be attributed to participants' anticipation of the evaluation from the interviewer's side at the end of the interview. However, further in-depth analysis is needed to better understand these dynamics.

This study has as well several limitations that should be raised. While the KID dataset is well-structured, the generalizability of the findings could be limited due to the nature of the dataset, which consists primarily of semi-clinical interviews from a specific population. The findings might not fully reflect the emotional expressions or depression-related behaviours of other populations, such as those with clinical depression or from different cultural backgrounds.

Further limitation related to the dataset is that environmental surrounding and equipment was not standardised across the sample (e.g. different microphones), therefore the quality of recordings varied and could have potentially impacted the performance of the SER model as well. When it comes to the diarization tasks, although Pyannote provides state-of-the-art performance, no currently available diarization tool is perfect, and occasional misattributions may have occurred. In our diarization approach, we removed segments of the interview with no speech to reduce noise from non-emotional content. However, predictions for "others/noise" still occurred frequently.

One more limitation that we would like to point out is that we did not have a chance to test the performance of the emotion2vec+ model on our own, due to the lack of an annotated dataset that could be used. But in the literature, the model was tested on a huge variety of datasets and German database it achieved impressive precision levels between 94.7 % and 98.6 % (Ma et al., 2023). While the emotion2vec+ model is state-of-the-art, its performance in extreme cases, such as participants with severe depression or those who have difficulty articulating emotions, could be limited due to the lack of such data it was fine-tuned on.

In further studies, this approach could be enhanced by incorporating additional variables, such as demographic information, into the regression models to explore their influence on emotion prediction and depression correlation. A deeper focus on individual emotions could provide insights into their specific patterns and relevance to depression. Additionally, the interviewer's speech could be analysed as well to explore the emotional dynamics and its correlation between the interviewer and the participant.

Such an approach could as well be utilized to train machine learning models specifically designed for depression prediction, based on prosodic features of specific emotions.

Lastly, emotions could be categorized into positive and negative valence to assess the performance of a two-dimensional emotional model in the task of depression score prediction.

The practical implications of this approach are significant, particularly in research and clinical settings. In clinical settings, this method might one day contribute to the development of non-invasive, emotion-based automated screening tools that facilitate the early detection of mental health issues, such as depression. While the amount of variance in depression scores explained by the model is not sufficient for a standalone depression detection system, the system shows promise for supplementing other approaches with valuable information: While NLP-models evaluating interview transcripts have shown apt in inferring depression levels (Hadzic et al., 2024b; Ohse et al., 2024), they lack the capacity of evaluating emotive cues in prosody, which can readily be interpreted by human experts during the clinical interview. Integrating the

emotion prediction data from models such as emotion2vec+ might improve depression detection. Since a multichannel-approach by (Victor et al., 2019) has shown great promise for the detection of depression, an integration of emotion prediction with an NLP-based model warrants further investigation.

As emotional disturbances are a symptom of many mental disorders, emotion prediction holds great potential for automated early screening and symptom monitoring. Disorders in which the application of emotion prediction models appear particularly promising are for example anxiety disorders or bipolar disorders (Strakowski, 2012). Since the latter are characterized by episodes in which extreme mood states occur, emotion prediction and detection models might prove helpful as early warning systems (Cummins et al., 2020). For anxiety disorders, emotion prediction systems might be able to recognize heightened fear levels in patients during confrontations (e.g. in context of the Trier Social Stress Test (Kirschbaum et al., 1993)). The potential of emotion prediction could for example be used to infer mood states and provide users of mental health applications with just-in-time interventions fitting their current mood state (Teepe et al., 2021), or inform therapists conducting exposure therapy of a patients' current level of fear to optimize therapy (Boehnlein et al., 2020).

Transferred to the research-context, the detection of higher levels of fear from prosodic data could complement classic methods of fear detection like heart rate, skin conductance or cortisol levels (Hyde et al., 2019). This study's particular approach could furthermore be used to gather deeper insights into the emotional dynamics of clinical interviews and psychotherapy sessions. By analysing the dynamics of emotions between interviewers and participants, researchers could uncover patterns that may influence therapeutic outcomes or improve understanding of the communication process in mental health assessment processes.

There are various ethical considerations concerning the application of automated machine learning algorithms in clinical settings. An ethical and legal framework for all AI applications is provided within the EU AI Act (European Parliament and Council of the European Union, 2024). A specific resolution released by the German Chamber of Psychotherapists regarding the use of AI in psychotherapy and mental health diagnostics warns of the premature implementation of such applications, as machine learning does not follow any ethical guidelines and might harm patients and even society as a whole (Bundespsychotherapeutenkammer (BPtK), 2023). Said resolution clearly forbids the use of AI for diagnostic and psychotherapeutic processes, as these are only to be carried out by licensed professionals to avoid such harm. Therefore, AI might only be used in support systems for diagnosis and therapy, which in turn must be supervised by qualified experts (ibd.).

General ethical concerns regarding the use of AI in mental health settings involve risk of bias (Timmons et al., 2023), lack of transparency (e.g. the algorithm as a black box) (Fehr et al., 2024), data security (Rogan et al., 2024) and the question of accountability in case of harm (Smith, 2021). In the context of SER this means that more research is needed to validate the potential of the system for different groups of users, find sources for bias and ways to mitigate this bias. Within our study, a convenience sample from the German population was examined, yet whether the results obtained in this sample hold true in different populations requires more research. Regarding the lack of transparency, explainable AI approaches might inform on specific procedures learnt by the algorithm, while generally, the open-source nature of the model tested within this study already provides a layer of transparency. Data security was a priority due to the sensitive nature of clinical interview data. We strictly followed GDPR guidelines and good scientific practice. For a practical application of SER models, strict data protection guidelines must be defined. As the result of the SER models within our study did not inform any therapeutic processes or decisions, the potential for harm, as priorly checked by an ethics committee, was considerably low. Any application of SER models to practical settings, however, needs to be regulated with clear accountabilities and professional overview.

All in all, there appears to be some potential for the integration of state-of-the-art ML tools into clinical psychology and psychotherapy research, as well as for their integration into data-driven, personalised and scalable eHealth solutions. This approach therefore holds potential for significant contributions to both the theoretical and practical domains of mental health research.

6 CONCLUSION

In conclusion, this study underscores the potential of Speech Emotion Recognition models in clinical psychology and psychotherapy, providing deeper insights into the role emotions play in mental health assessment processes. By using state-of-the-art SER models, this approach enhances the understanding of emotional dynamics in semi-clinical depression-related interviews. Furthermore, it showcases the potential of machine learning and digital mental health tools for automated

early screening of mental health conditions, advancing eHealth solutions for mental health assessments.

ACKNOWLEDGMENTS

Work on this study was partially funded by VwV Invest BW – Innovation II funding program, with project number BW1 4056/03.

REFERENCES

Akçay, M. B. and Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76.

Aloshban, N., Esposito, A., and Vinciarelli, A. (2022). What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech. *Cognitive Computation*, 14(5):1585–1598.

Anusha, R., Subhashini, P., Jyothi, D., Harshitha, P., Sushma, J., and Mukesh, N. (2021). Speech emotion recognition using machine learning. In *2021 5th international conference on trends in electronics and informatics (ICOEI)*, pages 1608–1612. IEEE.

APA (2022). *Diagnostic and Statistical Manual of Mental Disorders (DSMV)*. American Psychiatric Association, 5th edition.

Atmaja, B. T. and Sasou, A. (2022). Evaluating self-supervised speech representations for speech emotion recognition. *IEEE Access*, 10:124396–124407.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Boehnlein, J., Altegoer, L., Muck, N. K., Roesmann, K., Redlich, R., Dannlowski, U., and Leehr, E. J. (2020). Factors influencing the success of exposure therapy for specific phobia: A systematic review. *Neuroscience & Biobehavioral Reviews*, 108:796–820.

Bredin, H. (2023). pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *24th INTERSPEECH Conference (INTERSPEECH 2023)*, pages 1983–1987. ISCA.

Bundespsychotherapeutenkammer (BPtK) (2023). Keine vorschnelle einführung von ki-anwendungen! Resolution verabschiedet vom 42. Deutscher Psychotherapeutentag, 5./6. Mai 2023, Frankfurt.

Cummins, N., Matcham, F., Klapper, J., and Schuller, B. (2020). Artificial intelligence to aid the detection of mood disorders. In *Artificial Intelligence in Precision Health*, pages 231–255. Elsevier.

Danner, M., Hadžić, B., Gerhardt, S., Ludwig, S., Uslu, I., Shao, P., Weber, T., Shiban, Y., and Ratsch, M. (2023). Advancing mental health diagnostics: Gpt-based method for depression detection. In *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, pages 1290–1296. IEEE.

European Parliament and Council of the European Union (2024). Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

Fehr, J., Citro, B., Malpani, R., Lippert, C., and Madai, V. I. (2024). A trustworthy ai reality-check: the lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, 6:1267290.

Hadzic, B., Mohammed, P., Danner, M., Ohse, J., Zhang, Y., Shiban, Y., and Rätsch, M. (2024a). Enhancing early depression detection with ai: a comparative use of nlp models. *SICE journal of control, measurement, and system integration*, 17(1):135–143.

Hadzic, B., Ohse, J., Danner, M., Peperkorn, N. L., Mohammed, P., Shiban, Y., and Rätsch, M. (2024b). Ai-supported diagnostic of depression using clinical interviews: A pilot study. In *VISIGRAPP (1): GRAPP, HUCAPP, IVAPP*, pages 500–507.

Hyde, J., Ryan, K. M., and Waters, A. M. (2019). Psychophysiological markers of fear and anxiety. *Current Psychiatry Reports*, 21:1–10.

Islam, M. S., Kabir, M. N., Ghani, N. A., Zamli, K. Z., Zulkifli, N. S. A., Rahman, M. M., and Moni, M. A. (2024). Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach. *Artificial Intelligence Review*, 57(3):62.

Joormann, J. and Stanton, C. H. (2016). Examining emotion regulation in depression: A review and future directions. *Behaviour research and therapy*, 86:35–49.

Kirschbaum, C., Pirke, K.-M., and Hellhammer, D. H. (1993). The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., and Mokdad, A. H. (2009). The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Ma, Z., Chen, M., Zhang, H., Zheng, Z., Chen, W., Li, X., Ye, J., Chen, X., and Hain, T. (2024). Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. *arXiv preprint arXiv:2406.07162*.

Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., and Chen, X. (2023). emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

Mohammed, P., Hadžić, B., Alkostantini, M. E., Kubota, N., Shiban, Y., and Rätsch, M. (2024). Hearing emotions: Fine-tuning speech emotion recognition models. In *Proceedings of the 5th Symposium on Pattern Recognition and Applications (SPRA 2024)*.

Mustafa, M. B., Yusoof, M. A., Don, Z. M., and Malekzadeh, M. (2018). Speech emotion recognition research: an analysis of research focus. *International Journal of Speech Technology*, 21:137–156.

Ohse, J., Hadžić, B., Mohammed, P., Peperkorn, N., Danner, M., Yorita, A., Kubota, N., Rätsch, M., and Shiban, Y.

(2024). Zero-shot strike: Testing the generalisation capabilities of out-of-the-box llm models for depression detection. *Computer Speech & Language*, 88:101663.

Plaquet, A. and Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. *arXiv preprint arXiv:2310.13025*.

Prabhu, S., Mittal, H., Varagani, R., Jha, S., and Singh, S. (2022). Harnessing emotions for depression detection. *Pattern Analysis and Applications*, 25(3):537–547.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., and Othmani, A. (2022). Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71:103107.

Rogan, J., Bucci, S., and Firth, J. (2024). Health care professionals' views on the use of passive sensing, ai, and machine learning in mental health care: Systematic review with meta-synthesis. *JMIR Mental Health*, 11:e49577.

Smith, H. (2021). Clinical ai: opacity, accountability, responsibility and liability. *Ai & Society*, 36(2):535–545.

Strakowski, S. (2012). Bipolar disorders in icd-11. *World Psychiatry*, 11(Suppl 1):31–6.

Teepe, G. W., Da Fonseca, A., Kleim, B., Jacobson, N. C., Salamanca Sanabria, A., Tudor Car, L., Fleisch, E., and Kowatsch, T. (2021). Just-in-time adaptive mechanisms of popular mobile apps for individuals with depression: systematic app search and literature review. *Journal of Medical Internet Research*, 23(9):e29412.

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., and Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science*, 18(5):1062–1096.

Van Niekerk, B., Carbonneau, M.-A., Zaïdi, J., Baas, M., Seuté, H., and Kamper, H. (2022). A comparison of discrete and soft speech units for improved voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6562–6566. IEEE.

Victor, E., Aghajan, Z. M., Sewart, A. R., and Christian, R. (2019). Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation. *Psychological assessment*, 31(8):1019.

Wang, H., Liu, Y., Zhen, X., and Tu, X. (2021). Depression speech recognition with a three-dimensional convolutional network. *Frontiers in human neuroscience*, 15:713823.

Williams, J. B., Kobak, K. A., Bech, P., Engelhardt, N., Evans, K., Lipsitz, J., Olin, J., Pearson, J., and Kalali, A. (2008). The grid-hamd: standardization of the hamilton depression rating scale. *International clinical psychopharmacology*, 23(3):120–129.

Yang, S.-w., Chang, H.-J., Huang, Z., Liu, A. T., Lai, C.-I., Wu, H., Shi, J., Chang, X., Tsai, H.-S., Huang, W.-C., et al. (2024). A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.