



# Emotion-Guided Graph Attention Networks for Speech-Based Depression Detection under Emotion-Inducing Tasks

Yuqiu Zhou<sup>1,2</sup>, Yongjie Zhou<sup>3</sup>, Yudong Yang<sup>1,2</sup>, Yang Liu<sup>1,2</sup>, Jun Huang<sup>1,2</sup>, Shuzhi Zhao<sup>1,2</sup>, Rongfeng Su<sup>1,2</sup>, Lan Wang<sup>1,2</sup>, Nan Yan<sup>1,2</sup>

<sup>1</sup>Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China

<sup>2</sup>Key Laboratory of Biomedical Imaging Science and System, Chinese Academy of Sciences, China

<sup>3</sup>Shenzhen Mental Health Center/Shenzhen Kangning Hospital, China

lan.wang@siat.ac.cn, nan.yan@siat.ac.cn

## Abstract

Depression affects emotional expression and perception. As a non-invasive and privacy-preserving method, speech is widely used for automatic depression detection. However, existing models often focus only on depressive features in speech, ignoring the differential emotion expression patterns across different emotion-inducing tasks. To address this, we propose an emotion-guided graph attention network (emoGAT) for depression detection. By collecting speech-text data from depressed individuals and healthy controls during emotion-inducing tasks, we construct graph embeddings using sentiment cues from both speech and text. Experimental results show our method reduces the standard deviation by 1.8% and improves accuracy by 4.36%. Graph attention visualization also reveals depression-specific characteristics, such as flattened prosody in neutral picture description tasks and cognitive biases toward negative information, offering deeper insights into emotional relational expressions.

**Index Terms:** emotion-inducing experiment, depression detection, multimodal learning, graph attention networks

## 1. Introduction

Adolescent depression is a growing public health concern worldwide, with rising prevalence rates, particularly in China, where estimates vary due to methodological differences. Some surveys report a 24.6% screening rate for depressive symptoms, emphasizing the need to distinguish between transient emotions and diagnosable depression. Current diagnostic practices rely on clinical interviews and self-report scales like the BDI and HDRS, but adolescents often show atypical symptoms (e.g., irritability, somatic complaints, academic decline) that differ from adult patterns, leading to misdiagnosis. Objective diagnosis is essential for avoiding biases in subjective assessments, enabling early screening, and improving treatment outcomes. However, biomarker research remains underdeveloped, lacking reliable physiological indicators.

Speech communication, as an easily accessible biomarker, is becoming a mainstream diagnostic marker for adolescent depression, particularly due to its non-invasive nature and capacity to capture nuanced psycholinguistic patterns associated with depressive symptomatology. This modality shows particular promise in clinical assessments and large-scale screening protocols, offering both ecological validity and cost-effectiveness compared to traditional neurobiological markers. In recent years, numerous studies have applied deep learning models to speech-based depression detection. Attention mechanisms [1] have played a crucial role in improving these models, leading to the development of Transformer-based and Graph Neural Network (GNN)-based architectures. For instance, Xu et al. uti-

lized Transformer-based models with speech-emotional expression cues to identify depressive speech patterns [2]. Sun et al. introduced a GCN-based model for depression detection by exploring the inter-class variability and intra-class audio consistency [3]. However, most existing methods primarily focus on dependencies within or between audio frames, neglecting the structural relationships of latent emotional cues across different emotion-inducing tasks.

Depression is a prevalent mental health disorder that significantly affects both psychological and physiological well-being. One of its key characteristics is altered emotional perception and expression, where individuals with depression often exhibit diminished positive affect and heightened sensitivity to negative stimuli compared to healthy individuals [4, 5, 6]. Psychological studies have explored emotion-inducing paradigms [7, 8, 9, 10] and analyzed behavioral differences between individuals with and without depression during tasks such as reading effective texts, describing affective pictures, and watching affective films. Among these modalities, speech-based data has gained significant attention due to its accessibility and ability to preserve patient privacy better than video-based approaches, making it a valuable data type for deep-driven depression detection [11, 12, 13, 14].

In this work, we recruited both depressed and healthy adolescents and guided them through multiple emotion-inducing tasks, including text reading and picture description. Speech data collected from these tasks was processed using automatic speech recognition (ASR) to obtain both audio and text modalities for depression detection. To explore the impact of different emotion-inducing tasks on depression classification, we employed a Graph Attention Network (GAT) [15] to aggregate utterance features across tasks, learning graph-based utterance representations for downstream classification. We also incorporated acoustic and semantic sentiment labels to construct adjacent matrices as graph masks, guiding GAT to capture intrinsic emotional relationships between tasks. Experiments demonstrate that integrating acoustic and semantic emotional information to construct cross-task utterance attention enhances GAT performance. To further interpret the model's learning process, we also visualized GAT's attention patterns for depressed and healthy subjects across different emotion-inducing tasks.

In summary, this study proposes a novel depression detection framework based on speech emotion-inducing tasks. It leverages both acoustic and semantic sentiment labels to guide graph attention learning. By visualizing attention differences between depressed and healthy subjects, we validate the effectiveness and interpretability of our approach.

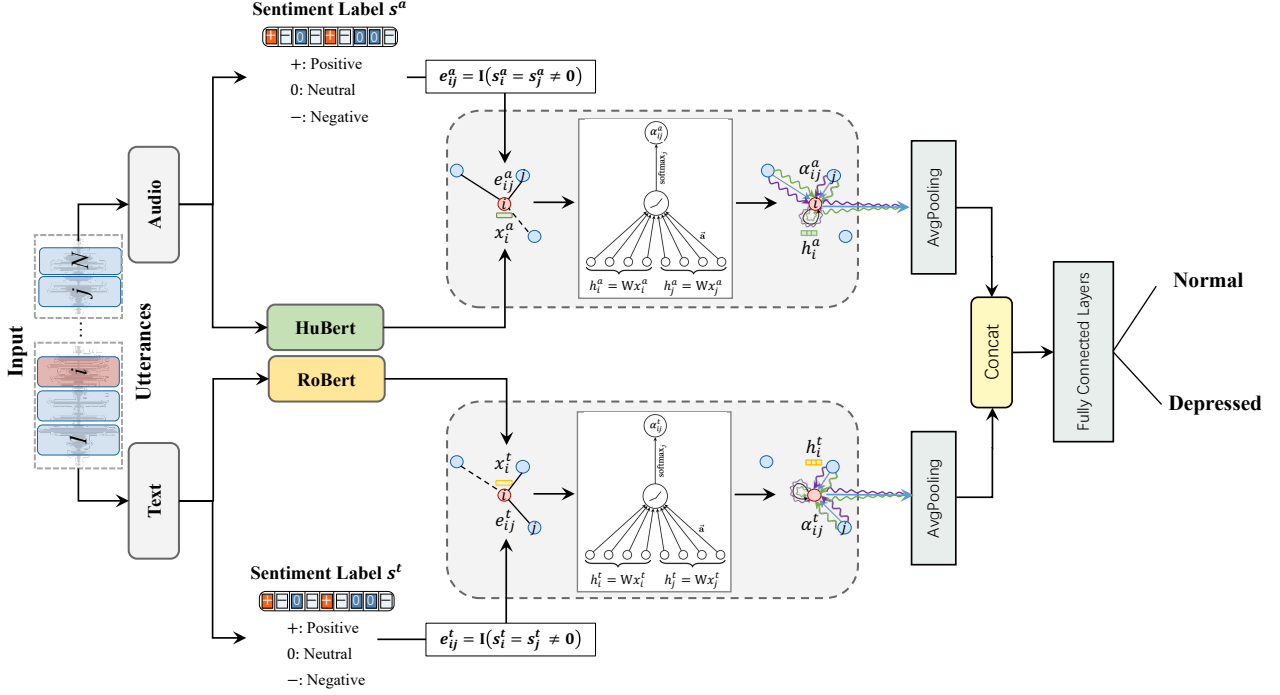


Figure 1: The overall architecture of emotion-guided graph attention networks for speech-based depression detection.

## 2. Methods

Figure 1 presents an overview of our proposed model, which consists of five main components: feature extraction, acoustic and semantic sentiment annotation, emotion-guided graph attention networks, multimodal fusion, and classification. Each component is described in detail below.

### 2.1. Feature Extraction

For audio data, we extract features using the pre-trained model HuBERT [16], utilizing embeddings from the 24th layer as acoustic representations. Corresponding semantic features are extracted using a pre-trained model RoBERTa [17].

### 2.2. Acoustic and Semantic Sentiment Annotation

Each utterance is assigned an acoustic sentiment label  $s_i^a \in \{+, 0, -\}$  based on HuBERT-extracted features with a pre-trained model to classify emotions as neutral (0), positive (+), or negative (-). For semantic sentiment analysis, we employ the Hownet-based Chinese sentiment analysis library `cnstent`<sup>1</sup> [18], using the `sentiment_calculate` function to compute sentiment scores. This function accounts for intensity adverbs and negation effects, computing both positive and negative scores. The final sentiment score is obtained by subtracting the negative score from the positive score. The semantic sentiment label  $s_i^t \in \{+, 0, -\}$  of each utterance is noted as positive (+) if the score is greater than zero, neutral (0) if equal to zero, and negative (-) if less than zero.

### 2.3. Emotion-Guided Graph Attention Networks

GAT utilize self-attention mechanisms similar to Transformers, where graph attention layers compute updated node embed-

dings by weighting neighboring node features based on learned attention coefficients. To enhance GAT’s performance for depression detection, we introduce both acoustic and semantic sentiment labels to construct adjacency matrices, ensuring attention is directed towards utterances with similar emotional characteristics. These coefficients determine the importance of connections between utterances, emphasizing those with similar emotional expressions. By aggregating features from utterances sharing the same non-neutral emotion, we (1) mitigate noise from irrelevant utterances and (2) capture intrinsic emotional relationships across different emotion-inducing tasks.

#### 2.3.1. graph attention layer

Each utterance represents a graph node. Each graph attention layer takes acoustic features  $\mathbf{x}^a$  or semantic features  $\mathbf{x}^t$  extracted from each utterance as input. For simplicity, we omit  $a$  and  $t$  as features  $\mathbf{x} = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N\}$  where  $\vec{x}_i \in \mathbb{R}^F$ . Each graph attention layer outputs aggregated graph embeddings  $\mathbf{h}' = \{\vec{h}'_1, \vec{h}'_2, \dots, \vec{h}'_N\}$ , where  $\vec{h}'_i \in \mathbb{R}^{F'}$ .

#### 2.3.2. Attention Score Calculation

Then we use multi-head self-attention to capture the interaction between any two utterances from emotion-inducing tasks.

$$a_{ij} = a(\mathbf{W}\vec{x}_i || \mathbf{W}\vec{x}_j) = a(\vec{h}_i || \vec{h}_j) \quad (1)$$

First, we use a linear mapping with shared parameters  $\mathbf{W}$  to transform the node features  $x_i$  into high-order features, concatenate the transformed features, and then use multi-head self-attention to calculate attention coefficients to each node for each head  $k$ , where  $a$  is a shared attention mechanism. Here,  $\mathbf{W} \in \mathbb{R}^{F' \times F}$  and  $a \in \mathbb{R}^{1 \times 2F}$  are two matrices with learnable parameters, and “||” denotes concatenation. To better normalize

<sup>1</sup><https://github.com/hiDaDeng/cntent>

Table 1: Classification result in utterance level (mean and standard deviation of 5-fold cross verification).

Modality	Model	Fusion	Accuracy	F1-Scores	Accuracy	Recall
Audio	Transformer[2]	-	65.67% (3.61%)	65.55% (3.59%)	66.33% (3.06%)	66.16% (3.06%)
Audio+Text	Transformer[2]	concatentation	73.47% (3.53%)	73.44% (3.51%)	73.90% (3.07%)	73.93% (2.98%)
Audio+Text	Transformer[2]	addition	73.81% (2.62%)	73.67% (2.70%)	73.94% (2.78%)	73.93% (2.54%)
Audio+Text	Transformer[2]	multiplication	73.54% (2.65%)	73.39% (2.74%)	73.73% (2.85%)	73.67% (2.63%)
Audio	emoGAT	-	74.99% (3.60%)	74.93% (3.62%)	75.27% (3.35%)	75.34% (3.19%)
<b>Audio+Text</b>	<b>emoGAT</b>	<b>concatentation</b>	<b>75.38% (3.55%)</b>	<b>75.35% (3.57%)</b>	<b>75.70% (3.23%)</b>	<b>75.80% (3.10%)</b>
Audio+Text	emoGAT	addition	75.49% (3.13%)	75.43% (3.20%)	75.58% (3.14%)	75.75% (2.95%)
Audio+Text	emoGAT	multiplication	75.39% (3.58%)	75.34% (3.59%)	75.66% (3.30%)	75.75% (3.15%)

Table 2: Classification result in subject level (mean and standard deviation of 5-fold cross verification).

Modality	Model	Fusion	Accuracy	F1-Scores	Accuracy	Recall
Audio	Transformer[2]	-	78.48% (8.47%)	78.40% (8.50%)	78.86% (8.36%)	78.50% (8.45%)
Audio+Text	Transformer[2]	concatentation	82.75% (7.21%)	82.59% (7.29%)	83.84% (6.89%)	82.79% (7.10%)
Audio+Text	Transformer[2]	addition	85.60% (6.75%)	85.55% (6.78%)	86.18% (7.11%)	85.64% (6.85%)
Audio+Text	Transformer[2]	multiplication	84.17% (8.38%)	84.14% (8.39%)	84.51% (8.52%)	84.21% (8.41%)
Audio	emoGAT	-	83.53% (8.14%)	83.46% (8.19%)	84.09% (8.03%)	83.59% (8.14%)
<b>Audio+Text</b>	<b>emoGAT</b>	<b>concatentation</b>	<b>87.11% (5.41%)</b>	<b>87.06% (5.45%)</b>	<b>87.65% (5.16%)</b>	<b>87.18% (5.40%)</b>
Audio+Text	emoGAT	addition	84.17% (7.15%)	84.08% (7.87%)	84.61% (7.87%)	84.15% (7.88%)
Audio+Text	emoGAT	multiplication	84.17% (8.38%)	84.14% (8.39%)	84.51% (8.52%)	84.21% (8.41%)

attention weights across all neighboring nodes  $j$ , we use softmax normalization and LeakyReLU as the activation function.

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a_{ij}))}{\sum_{l \in \mathcal{N}_i} \exp(\text{LeakyReLU}(a_{il}))} \quad (2)$$

where  $\alpha_{ij}$  denotes the attention weight between node  $i$  and node  $j$ .

### 2.3.3. Emotion-Guided Adjacent Matrix

If there is no extra information incorporated, the graph will just be fully connected, and the adjacent matrix will be an all-one matrix. To guide GAT to capture intrinsic emotional relationships across different emotion-inducing tasks, we introduce acoustic sentiment labels  $s^a$  and semantic sentiment labels  $s^t$  to construct the adjacent matrix as the attention mask:

$$\hat{\alpha}_{ij} = \begin{cases} \alpha_{ij}, & s_i = s_j \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where  $s_i$  denotes the sentiment label of node  $i$ . For simplicity, we omit  $a$  and  $t$ . Two nodes will only be connected if they have the same non-neutral sentiment label. Then the final graph embedding will be calculated by aggregation across neighboring nodes and average pooling over  $K$  heads:

$$\vec{h}_i = \frac{1}{K} \sum_{k=1}^K \sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \vec{h}_j^k \quad (4)$$

## 2.4. Multi-modal Fusion

We integrate the learned graph embeddings from speech and semantic modalities using three fusion strategies: concatenation, element-wise addition, and element-wise multiplication. These

strategies are evaluated to determine the optimal approach for maximizing classification performance.

## 2.5. Classification Layer

The fused graph embeddings are passed through a fully connected network for final classification. Cross-entropy loss is used for optimization, ensuring robust discrimination between depressed and healthy subjects.

# 3. Experiments and Analysis

## 3.1. Dataset and Experimental Settings

We recruited 70 individuals diagnosed with depression and 70 healthy controls to participate in emotion-inducing tasks while recording their speech using a lapel microphone. To mitigate the influence of gender-related differences, we maintained a 2:5 male-to-female ratio in both groups.

The emotion-inducing tasks included text reading (task 1) and picture description (task 2). In task 1, participants read three passages with negative, neutral, and positive content (task1\_neg, task1\_neu, task1\_pos). In task 2, participants described three pictures with corresponding emotional content (task2\_neg, task2\_neu, task2\_pos), expressing their feelings. Each participant produced six recordings, resulting in a dataset of 840 speech samples.

In our experiments, the graph embedding dimension in each Graph Attention layer was set to 1024. The training process spanned 50 epochs, incorporating an early stopping mechanism to halt training if the loss reduction fell below 0.1. We employed the Adam optimizer with a fixed learning rate of 0.000005, a batch size of 128 and a dropout rate of 0.2. Performance was evaluated using accuracy, F1-score, precision, and recall, based on subject-independent 5-fold cross-validation.

### 3.2. Data Preprocessing

First, we utilize the ASR toolkit Wenet<sup>2</sup>[19] to transcribe the collected speech data to obtain the accompanying text data. Silent segments are removed, and the speech is segmented into utterance-level clips based on the transcribed text. The final data will receive manual verification to ensure precise alignment between the audio and the corresponding text.

### 3.3. Results and Analysis

#### 3.3.1. Comparison Experiments

We compared the classification performance of GAT and Transformer-based models for depression detection. Integrating speech emotion cues to emoGAT led to significant performance improvements, surpassing Transformer-based models in all evaluation metrics. Aggregating utterances with similar emotional attributes enhanced GAT’s utterance-level classification accuracy, thereby improving subject-level classification.

Moreover, fusing audio and ASR text modalities by concatenation further boosted subject-level classification accuracy while enhancing the robustness of GAT (as indicated by a lower standard deviation). This demonstrates the benefits of leveraging multimodal information for depression detection.

#### 3.3.2. Ablation Experiments

In this section, we conduct ablation experiments to further investigate the impact of different components on the model’s performance. Based on the comparison experiments, the most effective model fused the Audio and Text modalities through concatenation and utilized a GAT architecture instead of the Transformer network. To evaluate the importance of specific features, we performed an ablation experiment where we removed the adjacency matrix constructed with the guidance of sentiment labels. Instead, we used a fully connected adjacency matrix. The results showed a significant reduction in both accuracy and robustness, demonstrating the crucial role of the emotion-guided adjacency matrix in enhancing model performance.

Table 3: Classification accuracy without emotion-guided adjacent matrix (mean and standard deviation of 5-fold cross verification).

Adjacent Matrix	utterance-level	subject-level
Emotion Guided	75.38% (3.55%)	87.11% (5.41%)
Full Connection	62.13% (9.25%)	68.65% (13.42%)

### 3.4. Visualization and Interpretation

To further explore the differential emotion expression patterns across different emotion-inducing tasks in recognizing depression-related features, we visualized the attention differences in GAT’s learned representations from both speech (Figure 2) and text (Figure 3). We computed utterance-wise attention scores based on activation values from different models (speech and text) and constructed attention matrices for depressed and healthy subjects. These matrices were grouped by task type, normalized, and analyzed for inter-task differences.

Figure 2 illustrates the attention differences across tasks and between depressed and healthy subjects. The size of each node and the thickness of each edge represent the significance

<sup>2</sup><https://github.com/wenet-e2e/wenet>

of attention differences for specific tasks. Solid edges indicate higher attention weights in depressed subjects compared to healthy controls, whereas dashed edges indicate the opposite.

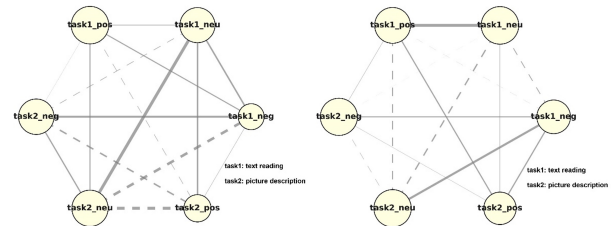


Figure 2: Attention differences in speech modality (left) and text modality (right)

Across both speech and text modalities, attention differences in neutral text reading (task1\_neu) and neutral picture description (task2\_neu) tasks, as well as their connections to other tasks, were more pronounced in depressed subjects compared to healthy controls.

In the speech modality (Figure 2, left), depressed subjects exhibited stronger connectivity between neutral picture description (task2\_neu) and neutral text reading (task1\_neu) (solid lines). In contrast, healthy subjects showed stronger connections between neutral picture description (task2\_neu) and both negative text reading (task1\_neg) and positive picture description (task2\_pos) (dashed lines). Given the emotional information embedded in our attention mechanism, we hypothesize that this difference arises because depressed individuals tend to maintain a monotonous speech tone with predominantly neutral affect, leading to stronger intra-neutral task connections. Conversely, healthy individuals exhibit more prosodic variations even in neutral tasks, resulting in stronger associations with emotionally charged tasks.

Unlike the speech modality, in the text modality (Figure 2, right), depressed subjects demonstrated a stronger connection between neutral picture description (task2\_neu) and negative text reading (task1\_neg). This suggests that depressed individuals may use more negative emotional expressions when describing neutral affective pictures. Previous studies have shown that individuals with depression exhibit negative cognitive biases, interpreting neutral stimuli more negatively and making fewer positive interpretations compared to healthy controls [20, 21, 22]. Our findings align with this theory, reinforcing the notion that depressed individuals process neutral stimuli differently across speech and text modalities.

## 4. Conclusion

This study explores the effectiveness of different network structures and modalities in distinguishing depressed patients from healthy controls using speech and text data. By integrating modality-specific emotional information with GAT, our approach improves depression detection accuracy and robustness, offering a novel way to examine emotion expression patterns across emotion-inducing tasks. However, the scarcity of suitable publicly available datasets remains a limitation. The neutral picture description task is key, showing that healthy controls have more emotional variation in speech, while depressed individuals use more negative words. These differences help the model distinguish between the two groups. Future work will refine multimodal integration and investigate advanced graph-based learning techniques for clinical applications.

## 5. Acknowledgements

This work is supported by the National Natural Science Foundation of China (U23B2018, NSFC 62271477), Shenzhen Science and Technology Program (JCYJ20220818101411025, JCYJ20220818101217037, JCYJ20220818102800001), Shenzhen Peacock Team Project (KQTD20200820113106007), and Shenzhen University-Huaqiang Project.

## 6. References

- [1] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
- [2] C. Xu, X. Wu, N. Li, X. Wang, X. Feng, R. Su, N. Yan, and L. Wang, "A transformer-based depression detection network leveraging speech emotional expression cues," in *International Conference on Social Robotics*. Springer, 2024, pp. 177–186.
- [3] C. Sun, M. Jiang, L. Gao, Y. Xin, and Y. Dong, "A novel study for depression detecting using audio signals based on graph neural network," *Biomedical Signal Processing and Control*, vol. 88, p. 105675, 2024.
- [4] J. Rottenberg, J. J. Gross, and I. H. Gotlib, "Emotion context insensitivity in major depressive disorder," *Journal of abnormal psychology*, vol. 114, no. 4, p. 627, 2005.
- [5] L. M. Bylsma, B. H. Morris, and J. Rottenberg, "A meta-analysis of emotional reactivity in major depressive disorder," *Clinical psychology review*, vol. 28, no. 4, pp. 676–691, 2008.
- [6] I. H. Gotlib, E. Krasnoperova, D. N. Yue, and J. Joormann, "Attentional biases for negative interpersonal stimuli in clinical depression," *Journal of abnormal psychology*, vol. 113, no. 1, p. 127, 2004.
- [7] H. Lyu, H. Huang, J. He, S. Zhu, W. Hong, J. Lai, T. Gao, J. Shao, J. Zhu, Y. Li *et al.*, "Task-state skin potential abnormalities can distinguish major depressive disorder and bipolar depression from healthy controls," *Translational Psychiatry*, vol. 14, no. 1, p. 110, 2024.
- [8] P. J. Lang, "The cognitive psychophysiology of emotion: Fear and anxiety," in *Anxiety and the anxiety disorders*. Routledge, 2019, pp. 131–170.
- [9] B. D. Dunn, T. Dalgleish, A. D. Lawrence, R. Cusack, and A. D. Ogilvie, "Categorical and dimensional reports of experienced affect to emotion-inducing pictures in depression," *Journal of abnormal psychology*, vol. 113, no. 4, p. 654, 2004.
- [10] A. B. Jin, L. H. Steding, and A. K. Webb, "Reduced emotional and cardiovascular reactivity to emotionally evocative stimuli in major depressive disorder," *International Journal of Psychophysiology*, vol. 97, no. 1, pp. 66–74, 2015.
- [11] S. Sahu and C. Y. Espy-Wilson, "Speech features for depression detection," in *INTERSPEECH*, 2016, pp. 1928–1932.
- [12] A. Y. Kim, E. H. Jang, S.-H. Lee, K.-Y. Choi, J. G. Park, and H.-C. Shin, "Automatic depression detection using smartphone-based text-dependent speech signals: deep convolutional neural network approach," *Journal of medical Internet research*, vol. 25, p. e34474, 2023.
- [13] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of biomedical informatics*, vol. 83, pp. 103–111, 2018.
- [14] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, p. 103107, 2022.
- [15] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48 550, 2017.
- [16] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [17] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.
- [18] X. Deng and P. Nan, "cntext: a python tool for text mining," 9 2022. [Online]. Available: <https://github.com/hiDaDeng/cntext>
- [19] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," *arXiv preprint arXiv:2102.01547*, 2021.
- [20] M. Milders, S. Bell, J. Platt, R. Serrano, and O. Runcie, "Stable expression recognition abnormalities in unipolar depression," *Psychiatry research*, vol. 179, no. 1, pp. 38–42, 2010.
- [21] J. Everaert, I. R. Podina, and E. H. Koster, "A comprehensive meta-analysis of interpretation biases in depression," *Clinical psychology review*, vol. 58, pp. 33–48, 2017.
- [22] F. Orchard, L. Pass, and S. Reynolds, "'it was all my fault': negative interpretation bias in depressed adolescents," *Journal of abnormal child psychology*, vol. 44, pp. 991–998, 2016.