



# A Study on The Impact of Foundation Models on Automatic Depression Detection from Speech Signals

Bubai Maji<sup>1</sup>, Monorama Swain<sup>2</sup>, Shazia Nasreen<sup>1</sup>, Debabrata Majumdar<sup>1</sup>, Rajlakshmi Guha<sup>1</sup>,  
Aurobinda Routray<sup>1</sup>, Anders Sjøgaard<sup>2</sup>

<sup>1</sup>Indian Institute of Technology Kharagpur, India

<sup>2</sup>University of Copenhagen, Denmark

bubaim@kgpian.iitkgp.ac.in, rajg@cet.iitkgp.ac.in, aurobinda.routray@gmail.com

## Abstract

An automatic depression detection (ADD) system using spoken language offers the opportunity to develop practical, low-cost tools to detect symptoms early. However, limited data availability, privacy concerns, and transcription efforts pose significant challenges. Recent advancements in foundational models, capable of understanding and processing multimodal inputs, present opportunities for enhancing ADD systems. This study explores various speech foundation models to investigate their impact on ADD. We leverage Whisper and MMS for automatic transcription and integrate speech and text embeddings into a language model optimized with low-rank adaptation (LoRA). In addition, we examine the effects of fine-tuning strategies and prompt formats on model performance. We used English and Bengali datasets to demonstrate the potential of our method in ADD, even with moderate-quality transcriptions. The best speech and language foundation models outperform baseline models on both datasets.

**Index Terms:** depression detection, speech, foundation models, large language model

## 1. Introduction

Depression is a widespread mental health disorder affecting millions globally. It can negatively impact an individual's verbal communication ability [1]. At present, the diagnosis of depression relies on questionnaires and medical assessments, but their accuracy depends on participant cooperation and operator expertise [2]. An automatic depression detection (ADD) technique is highly desired for the diagnosis of depression. A growing body of research shows that both audio and text contain rich information on depressive symptoms. Therefore, several studies have focused on lexical [3, 4, 5], acoustic features [6, 7, 8], and their combination [9, 10, 11], to improve performance. However, existing models often struggle to integrate multimodal data effectively [11]. Furthermore, successfully deploying ADD systems in real-world applications faces challenges such as high annotation costs and limited model generalizability, particularly in resource-constrained settings.

A typical pre-processing step in developing an ADD system involves transcribing speech content. For instance, datasets like EATD [10] and DAIC-WOZ [12] are commonly used in depression detection tasks and rely on professionally transcribed audio dialogues, which are often obtained through commercial transcription services. However, this approach demands significant research and development (R&D) costs, as it typically involves training transcribers on guidelines and conducting quality checks. While crowdsourcing platforms such as Amazon's Mechanical Turk (MTurk) [13] have increased the efficiency of transcription and labeling tasks by enabling human workers to

perform such tasks at scale, they still require considerable labor hours, leading to significant financial costs. Moreover, when working with sensitive depression datasets, such as those enforced by the Institutional Ethical Committee (IEC) [14], often require that annotations be done in-house. This increases the demand for resources and effort.

The emergence of foundation models [15] has shown significant progress in speech recognition and language understanding, presenting new possibilities to improve data curation in depression detection through speech samples. For example, Whisper [16], an automatic speech recognition (ASR) model trained on thousands of hours of multilingual audio, offers exceptional zero-shot ASR performance, making it a powerful tool for the transcription of speech data in this domain. In addition to advances in ASR, large language models (LLMs) like GPT-4 [17] exhibit strong language understanding and generation abilities. Despite its success, the use of foundation models for ADD has not been studied. In this work, we investigate the impacts of different foundation models for speech-based ADD on the curation of depression datasets in transcribing.

Our main contributions are as follows:

- We first analyzed the impact of different foundation models on speech-based depression detection, evaluating their ability to capture depressive markers directly from raw audio.
- Next, we study Whisper and MMS as transcribing annotators, demonstrating that foundation models offer moderate-quality transcriptions but are useful for ADD tasks.
- We then investigated speech representation from foundation models with ASR transcripts and integrated them into an open-source LLM model, enabling joint analysis of acoustic and text content for accurate depression detection.
- Finally, we optimized the system using low-rank adaptation (LoRA) [18] and tested how using different types of prompts affected the performance of the LLM in analyzing acoustic and lexical information.

## 2. Related Work

### 2.1. Speech Foundation Models

Self-supervised learning has become a prominent method in speech modeling, offering pre-trained ASR models for encoding speech representations. Models such as wav2vec 2.0 [19] and AST [20] have demonstrated excellent performance in various tasks like emotion recognition [21] and mental health detection such as depression [22] and Parkinson's disease [23]. A noteworthy recent model in this area is Meta's Massively Multilingual Speech (MMS) [24] model, which has been pre-trained on 491K hours of speech data. In contrast, OpenAI's Whisper [16] employs a weakly supervised approach for tasks like lan-

guage identification, voice activity detection, and speech recognition, trained on 680K hours of labeled speech data. Additionally, research has shown that integrating ASR outputs can further improve the performance of these tasks [22, 25].

## 2.2. Large Language Models

Large Language Models have demonstrated a significant impact across various natural language reasoning tasks, including medical-related tasks [26]. Additionally, recent research indicates that decoder-only LLMs demonstrate superior generative abilities [27] compared to encoder-only models and encoder-decoder systems (e.g., BERT), particularly for tasks requiring extensive language generation. This makes models like GPT more suitable for tasks involving open-ended text generation. However, one limitation with models like GPT-4 or ChatGPT is the need to upload user data to remote servers, which poses substantial privacy risks, especially in sensitive applications such as healthcare. In light of these concerns, we opted to explore the use of open-source language models (see Table 1), which are capable of running on a single GPU, offering a more privacy-conscious alternative for local deployment. The application of LLMs to the task of depression detection, particularly through the integration of speech data, remains largely unexplored. This gap suggests a promising area for research that could benefit from the integration of speech and text to improve mental health assessment.

## 3. Experimental Corpus

### 3.1. Indic-Bengali Corpus

The Indic-Bengali depressed dataset [28] includes 58 Bengali-speaking participants aged 21 to 32 years. The participants were divided into two groups: 35 healthy subjects (25 male, 10 female) and 23 depressed subjects (12 male, 11 female). Depressed participants were evaluated at B.C. Roy Technology Hospital at IIT Kharagpur, India, while the healthy controls were college students from the same institute. All subjects provided informed consent before the experiment. Diagnosis of MDD as per DSM-V (F32.1) [29] was done by the psychiatrist. Furthermore, the subjects were screened by a clinical expert. Data were collected using unstructured Rorschach Inkblot Test (RIBT) cards, where participants described the images. The speech responses of each participant were recorded at a sampling rate of 44.1 kHz with 16 bit quantization.

### 3.2. DAIC-WOZ Corpus

The DAIC-WOZ dataset [12] is a widely used English benchmark for depression analysis. It contains clinical interviews of 189 participants, including speech recordings, texts, and facial features. However, only the acoustic recordings and transcriptions, representing audio and text, are chosen in this work. Each session is labeled with a PHQ-8 score, with participants scoring 10 or above classified as depressed. The dataset contains 22.5 hours of audio recordings sampled at 16 kHz. According to its description, 30 out of 107 interviews within the training set and 12 out of 35 interviews within the development set are classified as depressed. We reported the results based on the development set in line with previous studies [4, 10, 22].

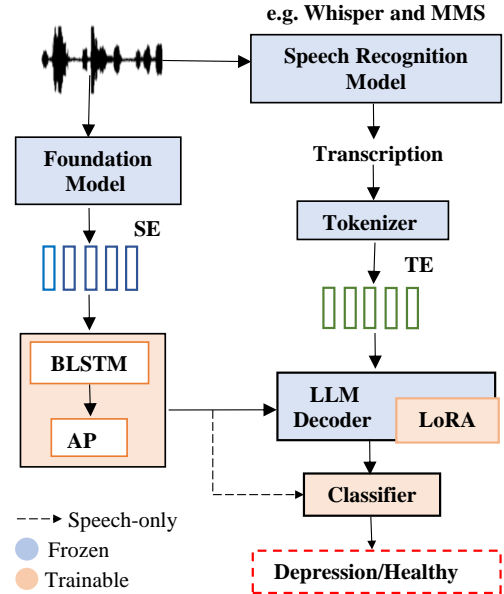


Figure 1: A schematic diagram of the proposed foundation model-based ADD framework. SE: Speech embeddings, TE: Text embeddings, AP: Attentive pooling.

## 4. Methodology

### 4.1. Foundation Model Assisted Transcribing and Embeddings Extraction

Our automatic transcription and embedding extraction from speech and text framework are shown in Fig. 1. Given a speech signal; speech foundation models directly extract speech embedding. Simultaneously, the ASR model generates transcriptions from the same speech, and the text foundation model is used to extract text embeddings. This study employs two widely used ASRs to generate transcription- Whisper-largeV2 and MMS-1B, both of which support English and Bengali. The details of all foundation models used in this study are listed in Table 1, including architecture, input, task details, etc.

### 4.2. Variations of Prompt Engineering in a LLM

Our prompt design incorporates various strategies, following the approach in [27], where instructing LLMs to annotate spoken utterances produces effective zero-shot performance. We apply three prompt strategies with LLMs to assess their impact on generating targeted labels for depression detection.

- No-prompt (P0): "Only the original transcription as input."
- Prompt 1 (P1): "Determine depression sign."
- Prompt 2 (P2): "You are a very good psychologist. The above is an interview on depression detection. Please determine whether there is a sign of depression."

Note: Prompts for the Bengali experiment were in Bengali.

### 4.3. Low-Rank Adaptation

Low-Rank Adaptation (LoRA) is a model adjustment technique grounded in reparametrization [18]. Rather than altering the original weights of a pre-trained model, LoRA incorporates trainable low-rank matrices into each Transformer layer.

Table 1: *Details of the foundation models used in this study; T.D: train data, ED: embeddings dimension, LSD: large-scale data*

Model	Architecture	Input	Model Size	TD.	Task	ED.
Whisper (tiny-medium)	Encoder-Decoder	Fbank	39M-1550M	680K	Speech Emb. Extraction	512-1024
Whisper-largeV2 [16]	Encoder-Decoder	Fbank	1550M	1M	Transcription	N/A
MMS-1B [24]	Encoder	Waveform	1B	1M	Transcription	N/A
Wav2vec2-Large [19]	Encoder	Waveform	311M	60K	Speech Emb. Extraction	1024
Wav2vec2-Base [19]	Encoder	Waveform	95M	60K	Speech Emb. Extraction	768
AST [20]	Encoder	Spectrogram	88M	1.8K	Speech Emb. Extraction	768
Qwen1.5-1.8B [30]	Decoder	Text	1.8B	LSD	Text Emb. Extraction	2048
BLOOM-1.7B [31]	Decoder	Text	1.7B	LSD	Text Emb. Extraction	2048
GPT-2 medium [32]	Decoder	Text	345M	LSD	Text Emb. Extraction	1024
mT5-base [33]	Encoder-Decoder	Text	580M	LSD	Text Emb. Extraction	768

This method effectively reduces the number of parameters that need to be trained for downstream tasks. It operates under the premise that pre-trained models have a low "intrinsic dimension," facilitating efficient learning within a constrained subspace [18]. This implies that weight updates during adaptation exhibit a low "intrinsic rank." Importantly, LoRA achieves this without adding any extra inference latency.

#### 4.4. Automatic Depression Detection Modeling

The full architecture of the ADD model is shown in Fig. 1. Our ADD model integrates both the speech and text backbones to extract their corresponding embeddings:

**Speech-only:** We first extracted speech embeddings from the foundation model, as discussed in Section 4.1. After extracting audio embeddings, two BiLSTM layers with 256 hidden units and dropout of 0.4 are used to capture temporal dependencies in the speech signal. This is followed by attentive pooling to aggregate the extracted acoustic information along the time dimension. Finally, a classifier consisting of a fully connected (FC) layer with a softmax function is used to classify.

**Text-only:** For the text modality, after generating text from the ASR model, we used AutoTokenizer to tokenize the text into input tokens. These tokens were passed into an LLM decoder, followed by the same FC layer for classification. We intend not to experiment with a larger version of Whisper and LLMs, as our setting requires prohibitively large GPU capacities.

**Fusion:** The extracted acoustic information is combined with the token representations of the LLM, guided by a specific prompt. These combined representations are passed through the LLM transformer decoder layers. Additionally, we also apply the LoRA module with default values of  $\alpha = 16$ , rank  $r = 64$ , and a dropout of 0.1 to optimize the model. At the same time, the LLM weights remain frozen during training. This approach enables more effective task-specific fine-tuning by addressing data limitations without updating the entire model.

## 5. Experiments And Results

### 5.1. Experimental Setup

During the data preprocessing stage, both datasets were down-sampled to 16 kHz. A data augmentation technique was employed to address the benchmark sample size limitation, using a sliding window of 3 seconds with a 50% overlap, which divided spoken dialogues into multiple segments as described in [7]. Following Shen et al.[10]’s approach to address imbalanced healthy and depressive samples of DAIC-WOZ, we selected ten questions from each participant’s responses to form one sample. We used the standard data split as provided in the dataset

description (see Section 3.2). In the Bengali dataset, consisting of continuous interview speech, we balanced the classes by randomly duplicating minority class samples. We used GroupK 3-fold cross-validation for Bengali data, ensuring that data from the same group did not appear in the training and evaluation sets. We set the batch size to 32, the learning rate to 0.00002, and the maximum training epoch to 30.

### 5.2. Results

To show the effectiveness of our methods, we first present the classification results of different speech foundation models of each dataset for the speech modality and prompt strategies for the text modality (see Tables 2-3). We used precision (Prec.), recall (Rec.), and F1-Score (F1), aiming to answer key research questions through these experiments.

#### RQ1: Does depression detection vary with different speech foundation models?

From the results presented in Tables 2-3, it can be seen that the Whisper models perform best across the datasets. Specifically, on English data, the Whisper-base model outperforms other versions, achieving an F1 of 0.689, and for Bengali data, Whisper-medium performs best with an F1 of 0.721. In comparison, Whisper (tiny-small), AST, and Wav2vec2 show lower performance. A likely explanation is that the Whisper-base and medium model provides a richer feature representation and capture subtle acoustic nuances better due to their larger architecture and more parameters. In addition, our best speech modality approach outperformed baseline methods [9, 34, 35]. We kept the best model from each dataset for the rest of the experiments.

#### RQ2: Does depression detection benefit from ASR transcription using the foundation model?

In this section, we compare the performance of ASR-generated text (Whisper-largeV2 and MMS-1B) using different LLMs, as shown in Fig. 2. Whisper-largeV2 has a word error rate (WER) of 14.5%, while MMS-1B has 29.12% on DAIC-WOZ reference text. Among LLMs, GPT-2 achieves the highest performance (F1 = 0.616) on Whisper-text due to its strong semantic understanding of English, allowing it to use accurate ASR text effectively. For MMS embeddings, GPT-2 achieves lower performance (F1 = 0.598) due to its higher transcription errors. For Bengali data, Qwen1.5 with Whisper-text performs best (F1 = 0.581) due to its robust multilingual capabilities and the ability to handle text representations effectively compared to other models such as mT5. Since the Bengali dataset does not include reference text, we primarily focused on how well the LLMs handle ASR-generated text from an unseen speech. Additionally, our method using ASR-generated text achieves competitive results compared to conventional systems [3, 9].

Table 2: Performance of different models using speech, text, and fusion features on the daic-woz; PM: proposed method

Inp.	Models	Prec.	Rec.	F1
Speech (S)	[34]	0.350	<b>1.00</b>	0.520
	[9]	0.710	0.560	0.630
	[22]	–	–	0.679
	[35]	–	–	0.613
	AST	0.634	0.541	0.584
	Whisper-tiny	0.680	0.597	0.636
	Whisper-base	<b>0.741</b>	0.644	<b>0.689</b>
	Whisper-small	0.684	<b>0.658</b>	0.671
	Whisper-medium	0.679	0.617	0.647
	Wav2vec2-base	0.577	0.609	0.593
Wav2vec2-large	0.603	0.621	0.612	
Text (T)	[9]	0.570	0.80	0.670
	[3]	–	–	0.640
	GPT-2 (P0) <sup>Ref</sup>	0.634	0.761	0.692
	GPT-2 (P1) <sup>Ref</sup>	<b>0.660</b>	<b>0.808</b>	<b>0.727</b>
	GPT-2 (P2) <sup>Ref</sup>	0.658	0.768	0.709
	GPT-2 (P0) <sup>W<sub>ASR</sub></sup>	0.593	0.641	0.616
	GPT-2 (P1) <sup>W<sub>ASR</sub></sup>	0.622	0.648	0.635
GPT-2 (P2) <sup>W<sub>ASR</sub></sup>	0.617	0.640	0.627	
S+T	[9]	0.710	0.830	0.770
	[22]	–	–	<b>0.829</b>
	PM (w/o Lora) <sup>Ref</sup>	0.758	0.842	0.798
	PM (w/ Lora) <sup>Ref</sup>	<b>0.795</b>	<b>0.859</b>	<b>0.826</b>
	PM (w/o Lora) <sup>W<sub>ASR</sub></sup>	0.703	0.754	0.728
PM (w/ Lora) <sup>W<sub>ASR</sub></sup>	0.737	0.786	0.761	

### RQ3: How do different prompt strategies affect the performance of LLMs in depression detection tasks?

Previous studies have demonstrated the efficacy of prompt engineering in improving the performance of LLMs [36]. We explored the impact of different prompt strategies on LLM performance for depression detection, comparing three distinct prompts (see Section 4.2). For English data, the results in Table 2 show that P1 outperformed P0 and P2 on both reference text and Whisper-text ( $W_{ASR}$ ). Its concise guidance helped the model focus on task-relevant information. However, the results from P2 suggest that overly detailed prompts may overwhelm the model can reduce the performance. For Bengali data (see Table 3), the P0 approach performed best with the best LLM. This could be due to limited exposure of LLMs to Bengali-specific instructions during pre-training, making structured prompts less effective. These findings suggest that selecting the right prompt can enhance model performance, and we kept the best prompt in the remaining experiments.

### RQ4: Can integrating speech embeddings into an LLM improve depression detection performance?

Finally, we investigate the impact of integrating speech embeddings with text embeddings into an LLM for depression detection, both with and without LoRA fine-tuning. The results presented in Tables 2-3 indicate that in both datasets, performance improved compared to single modalities (i.e., speech and text). This suggests that speech representations help LLMs extract relevant prosodic and acoustic cues along with linguistic patterns associated with depression. More importantly, the introduction of LoRA fine-tuning further enhances performance. In summary, the foundational ASR can improve performance

Table 3: Performance of different models using speech, text, and fusion features on the indic-bengali data

Inp.	Models	Prec.	Rec.	F1
Speech (S)	[9]	0.567	0.436	0.494
	[10]	0.626	<b>0.753</b>	0.684
	AST	0.670	0.738	0.703
	Whisper-tiny	0.607	0.576	0.592
	Whisper-base	0.676	0.626	0.650
	Whisper-small	0.707	0.681	0.694
	Whisper-medium	<b>0.719</b>	<b>0.723</b>	<b>0.721</b>
	Wav2vec2-base	0.652	0.707	0.678
Wav2vec2-large	0.687	0.711	0.699	
Text (T)	Qwen1.5 (P0) <sup>W<sub>ASR</sub></sup>	0.558	<b>0.605</b>	<b>0.581</b>
	Qwen1.5 (P1) <sup>W<sub>ASR</sub></sup>	<b>0.572</b>	0.514	0.565
	Qwen1.5 (P2) <sup>W<sub>ASR</sub></sup>	0.498	0.585	0.538
S+T	PM (w/o Lora) <sup>W<sub>ASR</sub></sup>	0.753	0.719	0.736
	PM (w/ Lora) <sup>W<sub>ASR</sub></sup>	<b>0.775</b>	<b>0.722</b>	<b>0.753</b>

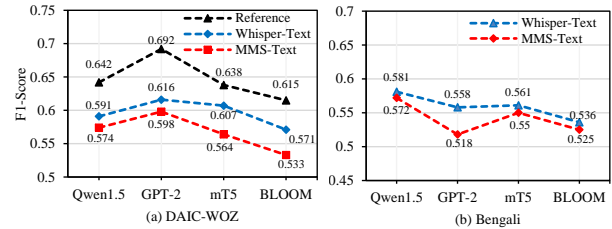


Figure 2: Performance comparison of different LLMs on both datasets using different ASR-generated text.

and make it suitable for integration into the ADD system.

## 6. Conclusions and Future Work

This paper explores the impact of foundation models for speech-based depression detection, approaching it as a language modeling problem. Our system learns jointly from a combination of acoustic and lexical embeddings. Experimental results demonstrate that the LLM backbone effectively integrates acoustic and lexical information across English and Bengali speech datasets. Furthermore, we observe that ASR-generated transcripts using foundation models enhance the performance of depression detection that typically relies solely on speech data. However, despite moderate WERs, our approach shows promising potential for real-world applications in depression detection. Lastly, our findings suggest that tailored prompt engineering within the LLM improved the focus on depression-specific cues and enhanced performance.

**Limitations and Future Work:** The main limitation of this study is that the model is trained on relatively small Bengali dataset, whereas the DAIC-WoZ is significantly larger. Testing our approach on other low-resource Indic languages, such as Odia and Assamese, with a larger scale and richer modalities, would help validate its effectiveness. Moreover, since emotional expressions are deeply affected by linguistic and cultural differences, exploring prosodic patterns between languages could provide further insight. To this end, we are integrating our model into a web-based application, which will be deployed in real-time scenarios across the public clinic sector.

## 7. References

- [1] H. Ellgring, *Non-verbal communication in depression*. Cambridge University Press, 2007.
- [2] U. Yadav, A. K. Sharma, and D. Patil, "Review of automated depression detection: Social posts, audio and video, open challenges and future direction," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 1, p. e7407, 2023.
- [3] E. Villatoro-Tello, G. Ramírez-de-la Rosa, D. Gática-Pérez, M. Magimai-Doss, and H. Jiménez-Salazar, "Approximating the mental lexicon from clinical interviews as a support tool for depression detection," in *Proceedings of the 2021 international conference on multimodal interaction*, 2021, pp. 557–566.
- [4] S. Burdisso, E. Villatoro-Tello, S. Madikeri, and P. Motlicek, "Node-weighted graph convolutional network for depression detection in transcribed clinical interviews," in *INTERSPEECH 2023*, 2023, pp. 3617–3621.
- [5] U. Yadav and A. K. Sharma, "A novel automated depression detection technique using text transcript," *International Journal of Imaging Systems and Tech.*, vol. 33, no. 1, pp. 108–122, 2023.
- [6] D. Wang, Y. Ding, Q. Zhao, P. Yang, S. Tan, and Y. Li, "Ecapadnn based depression detection from clinical speech," in *Inter-speech 2022*, 2022, pp. 3333–3337.
- [7] Q. Li, D. Wang, Y. Ren, Y. Gao, and Y. Li, "Fta-net: A frequency and time attention network for speech depression detection," in *INTERSPEECH 2023*, 2023, pp. 1723–1727.
- [8] B. Taşci, "Multilevel hybrid handcrafted feature extraction based depression recognition method using speech," *Journal of Affective Disorders*, vol. 364, pp. 9–19, 2024.
- [9] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Inter-speech*, 2018, pp. 1716–1720.
- [10] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6247–6251.
- [11] X. Zhang, B. Li, and G. Qi, "A novel multimodal depression diagnosis approach utilizing a new hybrid fusion method," *Biomedical Signal Processing and Control*, vol. 96, p. 106552, 2024.
- [12] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews," in *LREC*. Reykjavik, 2014, pp. 3123–3128.
- [13] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 5270–5273.
- [14] T. Feng, R. Hebbar, N. Mehlman, X. Shi, A. Kommineni, S. Narayanan *et al.*, "A review of speech-centric trustworthy machine learning: Privacy, safety, and fairness," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 3, 2023.
- [15] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.
- [16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [17] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
- [19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [20] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.
- [21] K. Liu, J. Wei, J. Zou, P. Wang, Y. Yang, and H. T. Shen, "Improving pre-trained model-based speech emotion recognition from a low-level speech feature perspective," *IEEE Transactions on Multimedia*, 2024.
- [22] W. Wu, C. Zhang, and P. C. Woodland, "Self-supervised representations in speech-based depression detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [23] D. Escobar-Grisales, T. Arias-Vergara, C. D. Ríos-Urrego, E. Nöth, A. M. García, and J. R. Orozco-Arroyave, "An automatic multimodal approach to analyze linguistic and acoustic cues on parkinson's disease patients," in *INTERSPEECH 2023*, 2023, pp. 1703–1707.
- [24] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.
- [25] Y. Li, P. Bell, and C. Lai, "Fusing asr outputs in joint training for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7362–7366.
- [26] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature med.*, vol. 29, no. 8, pp. 1930–1940, 2023.
- [27] D. Wagner, A. Churchill, S. Sigtia, P. Georgiou, M. Mirsamadi, A. Mishra, and E. Marchi, "A multimodal approach to device-directed speech detection with large language models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 451–10 455.
- [28] B. Maji, A. K. Roy, S. Nasreen, R. Guha, A. Routray, and D. Majumdar, "A novel technique for detecting depressive disorder: A speech database-based approach," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2023, pp. 1–4.
- [29] D. A. Regier, W. E. Narrow, E. A. Kuhl, and D. J. Kupfer, "The conceptual development of dsm-v," *American Journal of Psychiatry*, vol. 166, no. 6, pp. 645–650, 2009.
- [30] J. B. *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.
- [31] M. Al, "Bigscience large open-science open-access multilingual language model," *BigScience*, 2022.
- [32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, 2019. [Online]. Available: [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)
- [33] L. Xue, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.
- [34] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 35–42.
- [35] Y. Sun, Y. Zhou, X. Xu, J. Qi, F. Xu, Z. Ren, and B. W. Schuller, "Weakly-supervised depression detection in speech through self-learning based label correction," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.
- [36] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.