



An interpretable speech foundation model for depression detection by revealing prediction-relevant acoustic features from long speech

Qingkun Deng¹, Saturnino Luz², Sofia de la Fuente Garcia³

¹Division of Psychiatry, The University of Edinburgh, UK

²Usher Institute, The University of Edinburgh, UK

³Centre for Medical Informatics, Usher Institute, The University of Edinburgh, UK

qdeng2@ed.ac.uk

Abstract

Speech-based depression detection tools could aid early screening. Here, we propose an interpretable speech foundation model approach to enhance the clinical applicability of such tools. We introduce a speech-level Audio Spectrogram Transformer (AST) to detect depression using long-duration speech instead of short segments, along with a novel interpretation method that reveals prediction-relevant acoustic features for clinician interpretation. Our experiments show the proposed model outperforms a segment-level AST, highlighting the impact of segment-level labelling noise and the advantage of leveraging longer speech duration for more reliable depression detection. Through interpretation, we observe our model identifies reduced loudness and F0 as relevant depression signals, aligning with documented clinical findings. This interpretability supports a responsible AI approach for speech-based depression detection, rendering such tools more clinically applicable.

Index Terms: depression detection, speech foundation model, responsible AI, audio interpretation

1. Introduction

Depression is a common mental disorder, characterised by prolonged low mood, and loss of interest in activities, with an estimated 5% of adults suffering from it globally [1]. Recently, more research attention has been placed on developing automatic depression screening tools, using Deep Neural Networks (DNNs) to analyse patients' speech. These tools have the potential to monitor individuals' risk of depression at early stages and assist clinicians in providing rapid interventions. This paper addresses two issues that may hinder the clinical applicability of such clinical tools: predictions relying on short speech segments and a lack of model interpretability.

First, segmentation of data sequences is a common approach to avoid processing long sequences for DNN-based depression detection. For instance, [2] and [3] segmented audio data sequences along the temporal dimension, each labelled according to the subject's overall diagnosis. However, assuming that there are speech segments from depressed patients that contain *no* depression-relevant information (i.e. if depressive markers are *not* evenly spread across the whole speech), and are nonetheless labelled "depressed", this segment-level labelling approach can add noise to the model training and consequently reduce prediction accuracy in clinical practice.

To avoid labelling noise in social media analysis, [4] applied a post-level encoder to first encode social media posts from the same user into fixed-size embeddings. The sequence of embeddings from each user is then fed to a user-level encoder for a user-level depression classification given a single label. While segment-level labelling noise may be obvious in

social media analysis (i.e., text modality), where some posts do not explicitly contain depression-related utterances, its impact on speech-based depression detection (i.e., audio modality) demands further investigation.

Second, speech-based depression detection tools are hardly interpretable, hindering their clinical implementations in practice [5, 6]. In text-based depression detection (e.g., social media analysis), attention scores from transformer models have been used for model interpretation because they provide understandable weight distributions over input features [7]. For instance, [8] acquired attention scores from their Hierarchical Attention Network (HAN) to infer the importance of tokens in social media tweets for depression detection. Intuitively, input tokens at the sequence positions with high attention scores contribute more to depression detection.

However, this interpretation approach is insufficient because it ignores the computations that happened before the attention layers [9]. Specifically, in [8], before attention operations, each token has already been contextualised by every other token by a bidirectional Gated Recurrent Unit (biGRU). Therefore, attention scores derived from later layers do not map directly onto the input tokens, but rather onto their context-enriched representations, rendering the interpretation less precise. Additionally, raw attention scores often poorly correlate with gradient-based feature significance [10], questioning their validity for model interpretation. More importantly, this attention-based interpretation approach is not readily applicable to speech-based depression detection because unlike text, which is directly understandable to humans, raw audio signals require further processing to be human-interpretable.

To address these insufficiencies, we design an interpretable speech foundation model approach, leveraging pre-trained Audio Spectrogram Transformer (AST) [11], to make end-to-end depression detection using one's long speech instead of short segments. We compare the model to a segment-level AST to highlight the impact of segment-level labelling noise and the advantage of using longer speech duration for more reliable depression detection. Importantly, based on gradient-weighted attention maps [9], we introduce a novel frame-based interpretation method to extract human-interpretable acoustic features relevant to depression detection, making such speech foundation models interpretable for clinicians, and therefore more clinically applicable, for the first time.

2. Method

2.1. Data pre-processing

The D-vlog dataset [12] was used. To obtain audio segments consisting of natural sentences, we first obtained text data by applying the open-sourced Whisper model [13] to transcribe the

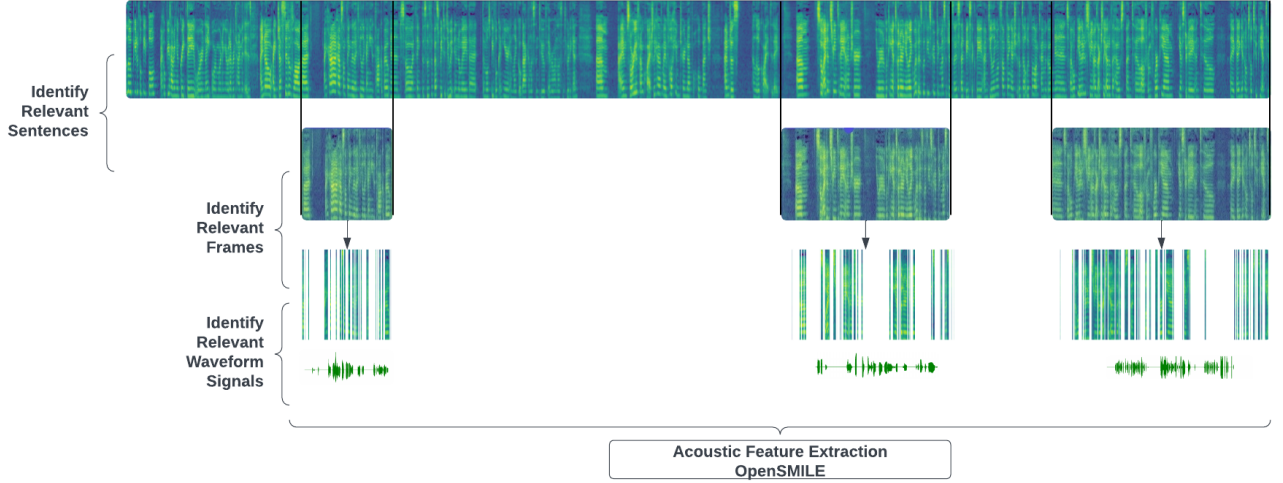


Figure 1: Workflow of the frame-based attention interpretation method. For demonstration purposes, the long spectrogram represents a long speech interval of ten sentences as an input for the proposed model. Here, the speech-level interpretation first identifies the most relevant five sentences, with indexes of 2, 5, 7, 8, and 9. Then, the sentence-level interpretation identifies the relevant frames, using a relevancy threshold of 0.3, for each sentence. Lastly, the waveform signals that temporally correspond to the relevant frames are identified, which are thereby processed by OpenSMILE for relevant acoustic feature extraction.

waveforms into texts with word-level time stamps. We then designed a sentence-level data segmentation approach to obtain text segments consisting of natural sentences, each with a number of words generally longer than seven.

Each text segment has word-level timestamps in millisecond units, indicating the start and end times of each sentence. These timestamps are utilised to retrieve the relative sentence-level waveforms to constitute the audio segments, with temporal lengths averaging approximately 5.53s (standard deviation, $sd = 13.53s$). Each waveform is transformed into a sequence of 128-dimensional log Mel filterbank features, utilizing a 25ms Hamming window with a step size of 10ms. The resulting Mel-spectrograms were uniformly padded or truncated to dimensions of 128×1024 .

2.2. Speech-level Audio Spectrogram Transformer

The proposed speech foundation model approach is based entirely on the attention mechanism [14]. First, the model consists of a sentence-level processing block, which encodes sentence-level spectrograms using a pre-trained frame-based AST. Instead of partitioning the spectrogram into patches of size 16×16 , the frame-based AST splits the spectrogram into frames of size 128×2 along the temporal dimension. We add attention masks to ignore padding. After the processing, the [cls] token from the audio data then represents the sentence-level audio data.

Second, the model consists of a speech-level processing block which is also a transformer encoder (six consecutive attention layers) that operates at the speech level. It receives the sequence of sentence representations produced from the sentence-level processing block for each participant and injects positional embeddings to consider the sequence order of each sentence. A speech-level [cls] token is prepended to the sequence to aggregate the sequence into a single representation which will be mapped onto a 2-dimensional space for the final binary classification.

Consider a given speech $S_i = \{s_{i1}, s_{i2}, \dots, s_{in}\}$ from a

participant P_i , where s_{ij} represents the Mel-spectrogram of the j^{th} sentence in the speech. The speech S_i is passed through the sentence-level processing block to be processed by the sentence-level encoders into a sequence of embeddings $E_i = \{e_{i1}, e_{i2}, \dots, e_{in}\}$, representing each sentence in the speech of participant P_i . The sequence E_i is then encoded by the speech-level processing block into a single representation r_i for binary classification.

2.3. Frame-based attention interpretation

Since our model has two processing blocks (i.e., sentence-level and speech-level), our frame-based interpretation method is hierarchical to first provide a speech-level interpretation, addressing the question “Which sentences within a given speech are most relevant to depression detection?”, and then a sentence-level interpretation, addressing the question: “Within the relevant sentences, which Mel-spectrogram frames are most relevant to depression detection?”. A visualisation of the interpretation procedure can be seen in Figure 1

For the speech-level interpretation, we derive attention scores from within the speech-level processing block. For each attention layer, we apply the approach introduced in [9] to weigh the relative importance of attention scores across attention heads to obtain a gradient-weighted attention map \bar{A} , set out in equation (1), where $\nabla A := \frac{\partial y_d}{\partial A}$ represents the gradients of the output for the depression class d with respect to the attention scores A . The Hadamard product \odot accounts for the relative importance of attention scores. We take the mean E_h across heads, with negative contributions removed.

$$\bar{A} = E_h[(\nabla A \odot A)^+] \quad (1)$$

We initialise a relevancy map with the identity matrix for the speech-level processing block as $R^{ss} = I^{s \times s}$, considering each sentence processing block and the speech-level [cls] token as initially “self-relevant”. Then, we apply equation (2), where XX can represent either ss or aa , to update R^{ss} with a forward pass across every self-attention layer within the block.

This provides a mechanism for continuously tracking relevancy between representations at deeper layers while updating the relevancy map.

$$R^{xx} \leftarrow R^{xx} + \bar{A} \cdot R^{xx} \quad (2)$$

After updating, we take the first row of the matrix R^{ss} , corresponding to the position of the [cls] token, which contains a relevancy score for each sentence representation. We interpret the sentences at the positions with the highest relevancy scores as most relevant to depression detection.

We then perform the sentence-level interpretation for the most relevant sentence representations. We first derive the attention scores with respect to these representations from within the sentence-level processing block. Next, we follow the same procedure as above to obtain a gradient-weighted attention map \bar{A} from each attention layer.

We initialise a relevancy map $R^{aa} = \mathbf{I}^{a \times a}$ to account for the self-attention interactions within the audio modality (i.e. between spectrogram frames). We then apply equation (2) to update R^{aa} across the attention layers within the sentence-level processing block with a forward pass.

After updating, we extract the first row from the relevancy map R^{aa} , corresponding to the position of the sentence-level [cls] token. This row $\mathbf{r} = [r_1, r_2, \dots, r_n]$ contains a relevancy score for every Mel-spectrogram frame n within the sentence. We then apply the Min-Max normalization process to this row, adjusting each element to have a value ranging between 0 and 1. The normalised relevancy scores are used to highlight the frames that are relevant to depression detection.

Lastly, we identify the waveform signals corresponding temporally to the relevant spectrogram frames, which are processed using OpenSMILE [15] to extract relevant acoustic features that are interpretable to human experts.

3. Experiments

3.1. Data

D-Vlog, the dataset used in our experiments, consists of 961 YouTube video vlogs labelled by trained annotators as either “depressed” or “normal”. The authors shared the YouTube video keys which were used to download the audio waveforms for our research purposes. However, some videos were made unavailable by their owners. Also, due to limited computational resources, we did not consider videos that are longer than 15 minutes. In total, we downloaded 698 waveforms, with 52.7% labelled as “depressed”. We stratified data splitting based on unique YouTube accounts to prevent data leaks. We also stratified based on class and gender to produce the train/development splits shown in Table 1, using an 8:2 ratio. One “depressed” sample with data processing issues was excluded from the development set.

Table 1: Number of Depression and Normal instances per set

	Depression	Normal
Train	294	261
Development	73	69

3.2. Segment-level model

To investigate if segment-level labelling noise affects audio modality for depression detection, we fine-tuned a frame-based

AST to make segment-level predictions (i.e. lacking the speech-level processing block compared to the proposed speech-level model). We used the first 42 sentences segmented from each participant’s speech, each labelled as either “depressed” or “normal” according to the participant’s overall label. Note that some participants have fewer than 42 sentences in their speech (mean = 39.41, sd = 11.42). To derive a speech-level probabilistic prediction for each participant, we calculate the predicted probability of depression as the ratio of sentences classified as “depressed” to the total number of speech sentences for that participant

3.3. Implementation details

One NVIDIA A100 GPU was used for conducting experiments. We trained the proposed speech-level model also using the first 42 sentences segmented from each participant’s speech. The sentence-level spectrograms from each participant were batched for the sentence-level processing block to process, which outputs one sequence of embeddings, with a minibatch size of 1, for the speech-level processing block to process. We applied gradient accumulation, whereby gradients are accumulated for 72 training steps before each parameter update.

For both the speech-level and segment-level models, we froze the first nine layers and randomly reinitialised the last three layers of the pre-trained AST. Both models were trained using an Adam optimizer [16] with a learning rate of 3×10^{-5} and a weight decay of 0.01. For a fair comparison, training was stopped at the onset of overfitting (epoch 9 for the speech-level model and epoch 8 for the segment-level model) as indicated by the cessation of further decrease in development loss. While we acknowledge that this approach may lead to overly optimistic estimates of performance for both models (lacking an independent test set due to limited sample size), we note that this work does *not* aim to benchmark performance on the D-Vlog dataset. Instead, we aim to provide robust evidence for the advantage of using a patient’s long speech over short segments for more reliable depression detection in clinical practice.

3.4. Model performance

Table 2 displays the AUC (Area Under the Receiver Operating Characteristics Curve) scores for the segment-level and speech-level models evaluated on the development set. The speech-level model demonstrates a statistically significant improvement in AUC score over the segment-level model ($p < 0.05$). The result suggests the presence of segment-level labelling noise in audio modality and the advantage of longer-duration speech analysis, which avoids using segment-level labelling, for more reliable depression detection in clinical practice.

Table 2: AUC performances for models, including DeLong’s test p -value for statistical comparison. AUC serves as a classification threshold-invariant measure of model discrimination ability (see a detailed explanation of AUC in [17]). The 95% confidence intervals (CI) for AUC scores were obtained through bootstrapping the development set (5,000 bootstrap samples), using the ConfidenceIntervals package (v0.0.3) [18].

Model	AUC (95% CI)	p-value
Segment-level	0.714 [−.085, +.082]	—
Speech-level	0.772 [−.080, +.074]	0.0127

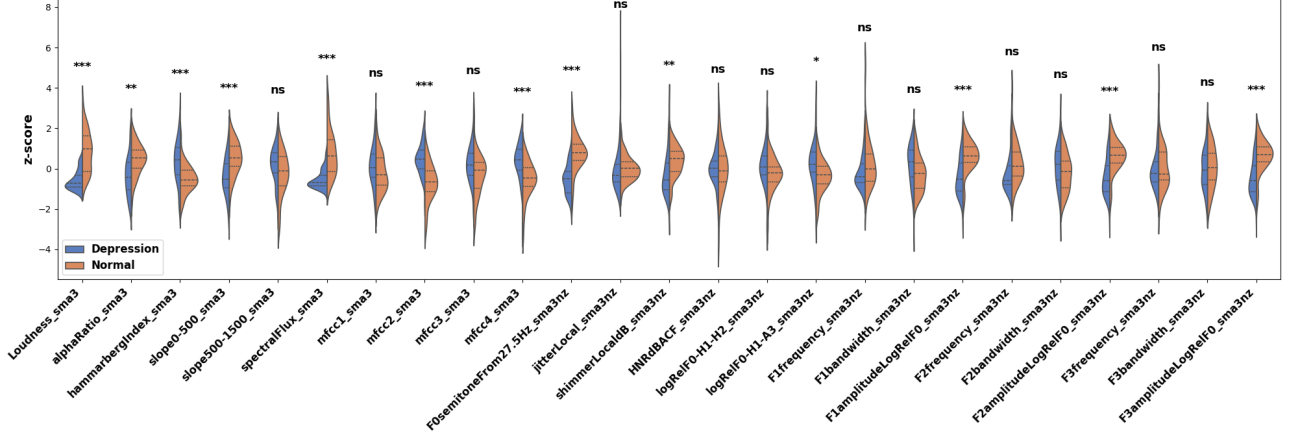


Figure 2: Violin plot of relevant acoustic feature value distributions (residualized for sex and standardized) between true positives ($n = 60$) and true negatives ($n = 41$). Relevant acoustic features were extracted from the waveform signals temporally corresponding to the spectrogram frames with relevancy scores higher than 0.3 from each sample’s five most relevant sentences. For true negatives, gradients of the output for the “normal” class regarding the attention scores were used to weigh the attention maps. Statistical significance was assessed using the Mann-Whitney U test, with Bonferroni correction applied for multiple comparisons. Significance levels were denoted as: *ns* for not significant; * for $p \leq 0.002$, indicating significant differences; ** for $p \leq 0.0004$; and *** for $p \leq 4 \times 10^{-5}$.

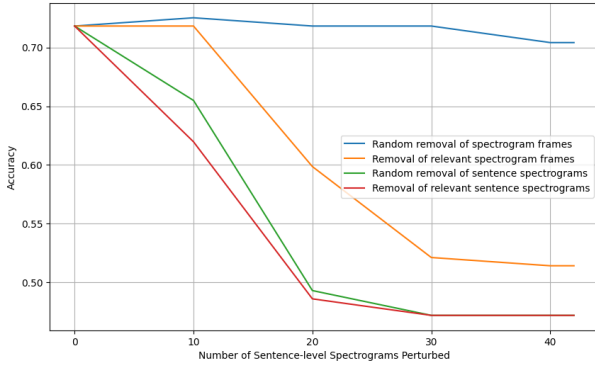


Figure 3: Perturbation test results for the proposed model. Accuracy was computed using a decision threshold of 0.527, reflecting the depression prevalence in the data. In the first test, sentence-level spectrograms were incrementally excluded in descending order of relevance, 10 at a time. In the second, only frames with a relevancy score above 0.3 within these sentences were removed. The results of these two tests were benchmarked against random exclusions: one with random sentence spectrogram exclusions, and another with random frames exclusions (30% within relevant sentence spectrograms). Accuracy never drops below 47.18%, likely due to the model’s bias towards positive predictions in the presence of a slight class imbalance.

3.5. Perturbation study

We conducted perturbation tests to explore the impact of omitting relevant sentence-level spectrograms or frames on our speech-level model’s performance. As shown in Figure 3, with detailed test descriptions provided in the figure caption, the selective removal of relevant frames led to a marked decline in model accuracy, substantially outpacing the accuracy degradation observed with random frame removal. The selective exclusion of relevant sentence-level spectrograms resulted in a further reduction in accuracy, which is marginally more pro-

nounced, especially for the first 10 sentences, than that observed with random removal of sentences. This observation implies differences between sentence-level spectrograms, or speech segments, in terms of signal importance for depression detection.

3.6. Relevant acoustic feature extraction

For the final step of the interpretation method, we locate the waveform signals temporally corresponding to the relevant spectrogram frames and extract a set of acoustics features from these signals using OpenSMILE (v2.5.0). We used the eGeMAPS (v02) feature set, which was designed to index affective changes in voice [19]. Figure 2 demonstrates the distribution differences in relevant acoustic feature values between true positive/negative samples. For a preliminary interpretation, we note that the model identifies reductions in loudness and fundamental frequency (F0) as relevant signals of depression, which aligns with the speech characteristics of depressed patients documented in clinical studies [20]. Future work is planned to validate the clinical relevance of our method using clinical speech datasets of depression (e.g., DAIC-WOZ [21]).

4. Conclusions

This work presents an interpretable speech foundation model approach that leverages pre-trained parameters, enabling scalability with more data, and uses patients’ long speech for more reliable depression detection in clinical practice. Crucially, it advances responsible AI by introducing a feasible audio interpretation method that reveals prediction-relevant acoustic features for clinician interpretation, which enhances early screening tools’ clinical applicability. With such interpretability, future works may systematically study the behaviour of speech-based depression detection systems. For instance, a system’s characteristics (e.g., reliance on certain acoustic features) in making false positives/negatives can be documented prior to their clinical implementations, which allows clinicians to critically assess the reliability of certain model predictions.

5. Acknowledgements

This work was supported by UKRI Grant No. 10102226 (University of Edinburgh) as part of the Horizon Europe Guarantee funding for participation in the INT-ACT project, funded by REA under the Horizon Europe Programme, grant 101132719.

6. References

- [1] World Health Organization. (2023) Depressive disorder (depression). <https://www.who.int/news-room/fact-sheets/detail/depression>. Accessed: August 15, 2023.
- [2] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, "Audio based depression detection using Convolutional Autoencoder," *Expert Systems with Applications*, vol. 189, p. 116076, Mar. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417421014147>
- [3] M. Muzammel, H. Salam, and A. Othmani, "End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis," *Computer Methods and Programs in Biomedicine*, vol. 211, p. 106433, Nov. 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169260721005071>
- [4] A.-M. Bucur, A. Cosma, L. P. Dinu, and P. Rosso, "An End-to-End Set Transformer for User-Level Classification of Depression and Gambling Disorder," Jul. 2022, arXiv:2207.00753 [cs]. [Online]. Available: <http://arxiv.org/abs/2207.00753>
- [5] M. Squires, X. Tao, S. Elangovan, R. Gururajan, X. Zhou, U. R. Acharya, and Y. Li, "Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment," *Brain Informatics*, vol. 10, no. 1, p. 10, Apr. 2023. [Online]. Available: <https://doi.org/10.1186/s40708-023-00188-6>
- [6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019, number: 5 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s42256-019-0048-x>
- [7] S. Wiegrefe and Y. Pinter, "Attention is not not Explanation," Sep. 2019, arXiv:1908.04626 [cs]. [Online]. Available: <http://arxiv.org/abs/1908.04626>
- [8] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol. 25, no. 1, pp. 281–304, Jan. 2022. [Online]. Available: <https://doi.org/10.1007/s11280-021-00992-2>
- [9] H. Chefer, S. Gur, and L. Wolf, "Generic Attention-Model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers," 2021, pp. 397–406. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Chefer_Generic_Attention-Model.Explainability_for_Interpreting_Bi-Modal_and_Encoder-Decoder_Transformers_ICCV_2021_paper.html
- [10] S. Jain and B. C. Wallace, "Attention is not Explanation," May 2019, arXiv:1902.10186 [cs]. [Online]. Available: <http://arxiv.org/abs/1902.10186>
- [11] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-Supervised Audio Spectrogram Transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10 699–10 709, Jun. 2022, number: 10. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/21315>
- [12] J. Yoon, C. Kang, S. Kim, and J. Han, "D-vlog: Multimodal Vlog Dataset for Depression Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 12 226–12 234, Jun. 2022, number: 11. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/21483>
- [13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 28 492–28 518, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v202/radford23a.html>
- [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- [15] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, Oct. 2010, pp. 1459–1462. [Online]. Available: <https://dl.acm.org/doi/10.1145/1873951.1874246>
- [16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Jan. 2017, arXiv:1412.6980 [cs]. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [17] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.
- [18] L. Ferrer and P. Riera, "Confidence intervals for evaluation in machine learning," <https://github.com/luferrer/ConfidenceIntervals>, 2023, [Computer software].
- [19] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016, conference Name: IEEE Transactions on Affective Computing. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7160715>
- [20] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, Jul. 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167639315000369>
- [21] 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.