Interspeech 2024
1-5 September 2024, Kos, Greece

Investigation of Layer-Wise Speech Representations in Self-Supervised Learning Models: A Cross-Lingual Study in Detecting Depression

Bubai Maji[1], *Rajlakshmi Guha*[1], *Aurobinda Routray*[2], *Shazia Nasreen*[1], *Debabrata Majumdar*[3]

[1]Rekhi Centre of Excellence for the Science of Happiness, IIT Kharagpur, India
[2]Department of Electrical Engineering, [3]B C Roy Technology Hospital, IIT Kharagpur, India
bubaim@kgpian.iitkgp.ac.in, rajg@cet.iitkgp.ac.in, aurobinda.routray@gmail.com

Abstract

Automated depression detection (ADD) from speech signals allows early identification and intervention, reducing costs to medical healthcare. However, most of the existing ADD studies are trained and evaluated on a single language corpus with a lack of sufficient training data. These limits the generalizability of models in other demographic groups in distinct languages. In this study, Semi-Supervised Learning (SSL) was applied to depression detection on two different language datasets. We evaluate the HuBERT and WavLM models in single-language, mixed-language, and cross-language scenarios to investigate the generalization to diverse populations at different recording environments. Moreover, we thoroughly analyzed layer-wise performance in the upstream model and pooling methods (i.e. max and mean pooling) in the downstream task. The results show that the WavLM features generalize better than the HuBERT features. Our best model surpasses previous works in the frozen upstream conditions.

Index Terms: self-supervised learning, depression detection, cross-lingual analysis

1. Introduction

Depression is one of the most common mental disorders and is primarily characterized by a persistent loss of interest, feeling hopeless, and low self-esteem [1]. According to the World Health Organization, in 2023, 5% of the world's population (approximately 368 million) suffer from depression [2]. Conventionally, diagnosis of depression is conducted by questionnaires [3]. However, many subjects might cause misdiagnosis [4]. Therefore, an automatic depression detection (ADD) system might help to detect depression in the early stage. To develop an ADD tool, growing research studies have shown that speech features are significantly used to distinguish depressed and healthy individuals [5, 6, 7].

Despite encouraging progress, two bottlenecks limit the performance of the existing studies. The first one is the insufficient annotated data is not enough to train a robust deep learning model [6, 8]. Second, in cross-corpus experiments, learning the discriminative depression information from speech is challenging due to variations in domain information [9, 10]. Moreover, the models that are trained and evaluated with the same corpus may not generalize well on unseen datasets [11]. Recently, self-supervised learning (SSL) models have been used in various speech-related tasks [12, 13, 14] to handle the data sparsity issue. An upstream model is typically trained on a vast speech data corpus through SSL to extract fundamental information. Additional layers may be added to build a downstream model, which is then trained on a smaller dataset for downstream tasks. It has been shown that the layer-wise representation in SSL models incorporated various aspects of speech information at different depths [15]. Even though the problem of data sparsity has been addressed in previous works on depression detection [16, 17], there is almost no study that investigates the cross-lingual databases for training and evaluation on depression detection tasks. Such cross-lingual depression detection experiments would ensure capturing the effect in different environments. Importantly, the use of SSL models as feature extractors has not been investigated in cross-lingual settings. Therefore, we investigate using the SSL model in an upstream task for speech representation (as shown in Figure 1) to observe their effect on generalizability in depression detection experiments. Our main contributions are:

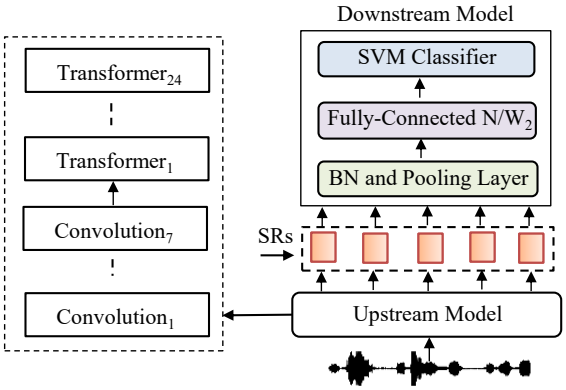

Figure 1: *The model architecture. BN: batch normalization, N/W: network, SRs: speech representations.*

- Empirically demonstrate the pre-trained SSL features in a downstream task with the support vector machine (SVM) classifier and analyze the performances in single and cross-lingual as well as mixed-lingual scenarios.
- We investigated layer-wise representations in various SSL models. We also compared different pooling methods (i.e., mean pooling and max pooling) in the downstream model.

2. Related Work

Prior studies of applying the SSL pre-trained models and the intermediate layer output of SSL models have proven effective for many downstream tasks, viz natural language processing [18], speech processing [19, 20], etc. Specifically, the SSL models, such as wav2vec 2.0 [12], HuBERT [13], and WavLM [14]. Fan et al. [18] utilized only wav2vec 2.0 for language identification and speaker recognition. The authors show that wav2vec 2.0 is capable of capturing information related to the speaker and

10.21437/Interspeech.2024-1737

language. In [19], the authors used HuBERT and WavLM for speech emotion recognition tasks and showed the effectiveness of both of the two models. Lebourdais et al. [20] also used the WavLM pre-trained feature to classify the overlapped speech and gender detection. This study shows that WavLM features effectively classify both speech and gender.

A few studies have examined the utilization of the SSL model in depression detection tasks. Zhang et al. [21] proposed a pretrained SSL model for speech-based depression recognition. Their model was trained to generate embedded spatial vectors from a spectrogram of an audio. Then a bi-directional long short-term memory (Bi-LSTM) was used for depression recognition. Wu et al. [22] compared three speech-based SSL models, such as wav2vec2.0, HuBERT, and WavLM. A layer-wise performance was also compared to the SSL models to analyze what type of information is generalized effectively in speech-based depression detection. Han et al. [23] also utilized SSL features and learned by a convolutional neural network (CNN) and LSTM network to classify depression severity. Besides in [24], the authors used latent representations extracted at different layers of the wav2vec 2.0 model to classify other pathological voices: Parkinson's disease, cleft lip and palate, laryngeal cancer, and oral squamous cell carcinoma. This type of approach allows the model to choose appropriate layer depths, which helps to reduce the computational cost. Conversely, our approach involved computing a single-layer representation from the upstream to the downstream task. Experimental findings showed that employing just a single layer proved to be just as effective as using a weighted sum of all layers.

Cross-lingual studies still have a major gap in the field of depression detection due to the great variation between source and target domain distributions. Mitra et al. [25] use the normalized spectral features for cross-dataset depression experiments. Recently, in [26], the authors performed a cross-corpus for classifying dementia in depression patients. However, given that it remains unclear whether the model can generalize across different languages, researchers need to focus on ADD for cross-lingual scenarios.

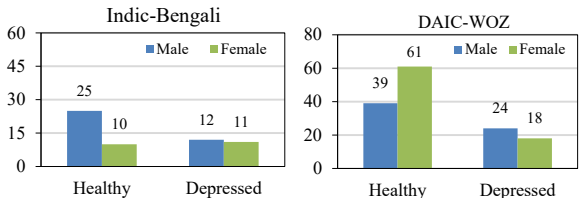

Figure 2: *Number of depression and healthy subjects from DAIC-WOZ and Indic-Bengali datasets.*

3. Experimental Corpus

3.1. Indic-Bengali Corpus

The Indic-Bengali depressed dataset we used is drawn from our previous study [27], consisting of 58 Bengali-speaking subjects aged between 21 and 32 years. Depressed participants were examined at B. C. Roy Technology Hospital, IIT Kharagpur, India, while the healthy subjects were college students from the same institute. All subjects provided informed consent before the experiment. Diagnosis of MDD as per DSM-V (F32.1) [28] was done by the psychiatrist. Furthermore, the healthy and depressed subjects were screened by the clinical expert. To collect the depressed data every subject was seated in front of a computer screen displaying a variety of unstructured Rorschach Inkblot cards (RIBT). They were instructed to vocalize and describe the images in the RIBT cards. The participants were not restricted to a time limit to complete the task. The initial and inquiry stages were conducted in line with Klopfer's data collection protocol. The responses from participants were recorded at a sampling rate of 44.1 kHz. The average length of the experiment was 40±4.6 minutes.

Ethical Approval: The design of the experiment was approved by the Institute's ethical committee under the approval number IIT/SRIC/DEAN/2023.

Data Availability: Due to the sensitive nature of the speech data, we cannot release our dataset to the public. Access to the data may be granted upon reasonable requests and will depend on obtaining local ethics approvals. For further information, please contact the author via email.

3.2. DAIC-WOZ Corpus

The Distress Analysis Interview Corpus Wizard of Oz (DAIC-WOZ) dataset [29] is an English benchmark depression dataset. The dataset comprises 189 clinical interviews between an interviewer and patients, with each session being annotated with a PHQ-8 score. The participant with a PHQ-8 score of 10 or greater is considered depressed. It includes a total of 22.5 hours of audio data from participants, recorded at a sampling rate of 16 kHz. According to the database description's data partition, 107 speakers were allocated for training and 35 speakers for development set. Following prior work [8, 16, 23], we also used the development set as evaluation or test data. Participant's details distribution from both databases is shown in Figure 2.

4. Methodology

4.1. Upstream models

In this study, we investigated the large versions of the HuBERT [13] and WavLM [14] speech models. These models are among the most commonly used SSL models, with the anticipation that their exploratory methods could extended to similar SSL models. These models include a CNN encoder followed by a series of transformer blocks, as shown in Figure 1. The convolutional encoder contains 7 consecutive one-dimensional convolution layers with 512 channels, kernel sizes (10,3,3,3,3,2,2) and strides (5,2,2,2,2,2,2). An input waveform is processed by the CNN encoder, which represents the initial (0th) layer, before being forwarded to the transformer encoder. The transformer block includes 24 transformer layers, an inner dimension 4096, 16 attention heads, and a model dimension of 1024. The resulting output sequence, enriched with both local and global information through the self-attention mechanism, is utilized for various downstream tasks. The upstream tasks of HuBERT are predictive, whereas WavLM is contrastive.

4.2. Downstream model

The downstream model used in this work includes a batch normalization (BN), a pooling layer, two fully connected (FC) networks, and an SVM classifier, as illustrated in Figure 1. A dropout layer was applied to the FC networks. The hidden dimensions of the FC layers and dropout rate were 256 and 0.3. We analyzed two different pooling methods: mean and max pooling. Mean pooling involves the use of a mean pooling layer applied to the time direction. Similarly, max pooling comprised

a max pooling layer used to the time axis. Finally, an SVM classifier was used to classify depression. SVM is a widely recognized and commonly utilized classifier for detection and regression tasks. In our implementation, we employed a linear kernel function and set the regularization parameter to 1. The value of gamma is determined using the formula $\gamma = \frac{1}{D \cdot \mathrm{Var}(X)}$, where D is the dimension of the features and $\mathrm{Var}(X)$ is the variance of the training samples.

5. Experiments

5.1. Experimental Setup

During data preprocessing, to match the sampling rate, both datasets are downsampled to 16 kHz. Due to the sample size limitation of the benchmarks, we use a data augmentation approach using a sliding window of 3 seconds and 50% overlap to divide the raw audio into multiple speech segments, as described in [30]. The upstream model was used as a feature extraction module and was fixed during the training of the downstream model. The evaluation was conducted in three parts: single-language, mixed-language, and cross-language evaluation. In the single-language evaluation (i.e., for both English and Bengali), the downstream model was trained and evaluated using only the individual database. The single-lingual evaluation allows us to distinguish the effectiveness of the SSL features where the recording conditions are the same for the training and testing stages. The second part of the experiments was a mixed-lingual evaluation where we trained the model by combining both language data and evaluated with separate languages. Therefore, the mixed-lingual evaluation helps the generalizability of the self-supervised features where unseen speakers alternate two or more languages. The last was a cross-lingual evaluation, where training and testing were done using different language databases (i.e., separately trained on English data and then evaluated using the Bengali data, and vice versa). The cross-lingual evaluation provides insights into the generalizability of the SSL features across different linguistic cultures.

5.2. Training Strategies

In the single-lingual experiment, for the DAIC-WOZ data, we used 107 subjects ($\sim$75%) for the training set and 35 subjects ($\sim$25%) for evaluation provided in the database description. For Bengali data, we kept the same split ratio (i.e., 75% training and 25% evaluation) to maintain consistency.

For the mixed-lingual experiment, we again use the same criteria, but this time, the respective training and evaluation data from both datasets are combined. For cross-lingual experiments, we use all data from one corpus as the training set and evaluate them on the other corpus (i.e., train in English and test in Bengali and vice versa). Note that there is no data overlap between the training and testing phases.

5.3. Evaluation Metrics

To assess the effectiveness of the ADD systems, this study focused on four widely recognized evaluation metrics: accuracy (A), precision (P), recall (R), and F1-score (F1).

5.4. Computing Resources and Code

All the experiments were implemented using the PyTorch framework and run with an NVIDIA Tesla P100 GPU with 16GB memory. The code is available at `https://github.com/bubaimaji/SSL-crosslingual`.

Table 1: *Performance metrics on the single-lingual setting. The best model with the pooling method is reported. In Model-N, where N is the number of layers, and Pool is defined the pooling method*

Model-N	Pool.	A (%)	P	R	F1
Bengali					
HuBERT-3	mean	83.87	0.827	0.822	0.824
HuBERT-1	max	82.31	0.831	0.821	0.826
WavLM-3	mean	**87.01**	0.863	0.852	0.857
WavLM-2	max	84.91	0.833	0.840	0.836
DAIC-English					
HuBERT-5	mean	80.37	0.732	0.863	0.794
HuBERT-6	max	79.10	0.684	0.832	0.750
WavLM-7	mean	**83.62**	0.794	0.914	**0.851**
WavLM-10	max	81.71	0.770	0.881	0.825
[22]	-	-	-	-	0.83
[23]	-	80	0.65	0.92	0.76

6. Results

In this section, we first report the results of the single-lingual experimental present in Section 6.1 followed by the experimental results of the mixed-lingual and cross-lingual in Section 6.2. Figures 3(a-e) show the accuracy with varying layer-wise depths when evaluating the SSL models. In these figures, the zeroth layer represents the CNN encoder.

6.1. Experiment 1: Single-Lingual Evaluation

The evaluation of layer-wise features extracted from the upstream model with varying pooling methods in the downstream model was first evaluated in a single-lingual setting. The classification accuracies are shown in Figures 3(a-b). Other metrics are displayed in Table 1, for the best model features. The results exhibit that the WavLM features perform better than the HuBERT features in both single-language experiments. For Bengali, WavLM-3 (where 3 is the layer number) with a mean pooling layer outperforms the best HuBERT-3 features by 3.14%. For English, again, the WavLM-7 (with mean pooling) outperforms the best HuBERT-5 features by an improvement of 3.25%. The best accuracies are 87.01% for Bengali and 83.62% for English. Moreover, the findings from the English language indicate that our best-performing model surpasses the previous studies where frozen upstream models were adopted [22, 23]. This outcome suggests that an upstream model can be highly effective when configured with optimal settings.

6.2. Experiment 2: Mixed and Cross-Lingual Evaluation

The classification accuracies for all layer-wise representations with varying pooling methods on the mixed-lingual and cross-lingual are shown in Figures 3(c-e). Table 2 depicts the other metrics for the best performed layer with the pooling method of the models. The performances for all scenarios generally degrade as expected. This is because different languages have unique phonetic and prosodic characteristics. This diversity makes it challenging for models to learn uniform indicators of depression across different linguistic and cultural contexts. In cross-lingual tasks, both HuBERT and WavLM perform relatively the same. A likely explanation is that these models can focus on the phonetic information and may face the same limitations in vocabulary mismatch between Bengali and English during cross-lingual tasks.

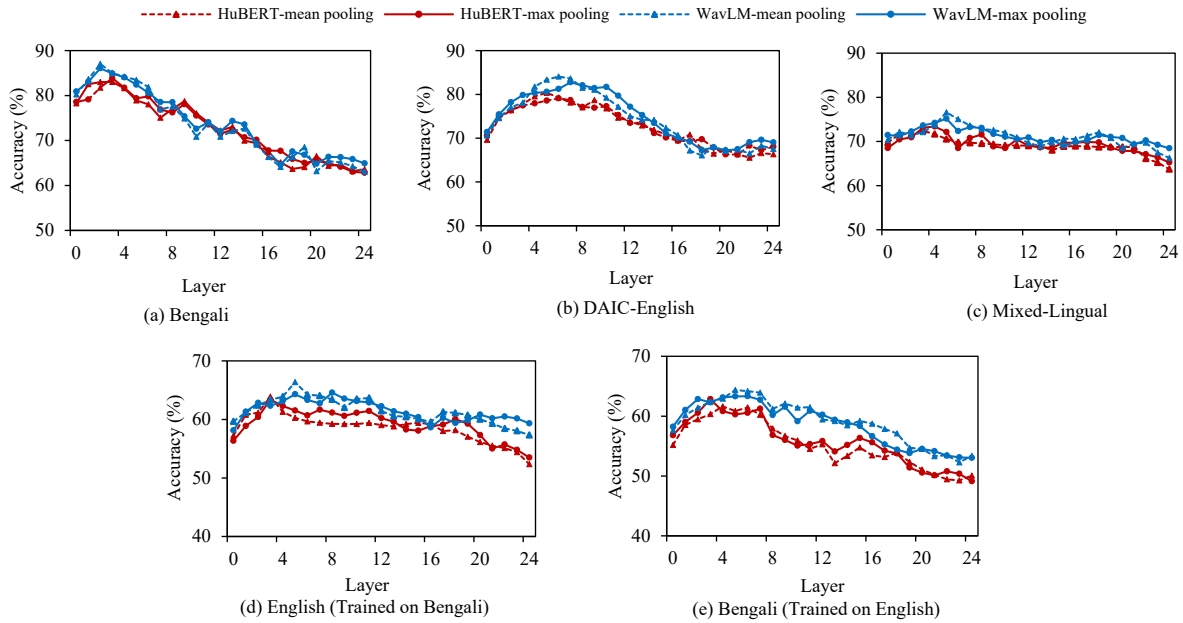

Figure 3: *Layer-wise analysis of depression detection accuracy (%) of HuBERT and WavLM with varying pooling methods.*

Table 2: *Performance metrics on the cross- and mixed-lingual setting. The best model with the pooling method is reported. In Model-N, where N is the number of layers, and Pool is defined as the pooling method*

Model-N	Pool.	A (%)	P	R	F1
English (Train on Bengali)					
HuBERT-3	mean	63.88	0.612	0.704	0.654
HuBERT-3	max	63.39	0.605	0.697	0.647
WavLM-5	mean	**66.37**	0.646	0.715	0.678
WavLM-8	max	64.57	0.648	0.692	0.670
Bengali (Train on English)					
HuBERT-4	mean	61.57	0.574	0.768	0.656
HuBERT-2	max	62.83	0.593	0.731	0.654
WavLM-5	mean	**63.29**	0.627	0.708	0.665
WavLM-5	max	62.81	0.631	0.693	0.660
Train and Test on Mixed-Lingual					
HuBERT-3	mean	72.42	0.748	0.685	0.715
HuBERT-4	max	73.38	0.762	0.691	0.724
WavLM-5	mean	**76.59**	0.771	0.754	0.762
WavLM-4	max	74.19	0.763	0.757	0.760

However, mixed-lingual achieve better performance than the cross-lingual settings. The fact that the systems perform better when tested on the other database may be attributed to the large volume of training data available in the mixed-lingual scenario. Compared to the single-database experiments in terms of accuracies, the best performances decreased by 23.72% for Bengali and by 17.25% for English.

7. Summary and Conclusions

In depression detection tasks, it is common to have many labeled utterances for the English language and lower availability of label utterances for low-resource languages. Based on this consideration, this study investigated the uses of SSL speech representations for depression detection in cross-lingual scenarios where paralinguistic information is crucial. In particular, we examined the effectiveness of the HuBERT and WavLM by varying layerwise speech representation in the upstream models. Moreover, we also analyzed the performances by varying the max and mean-pooling layers in the downstream task.

The following conclusions were drawn from these experiments: (1) The WavLM features outperformed almost all experiment scenarios for both databases. This demonstrates that the WavLM model has successfully learned to extract effective features suitable for depression detection. Moreover, the features from WavLM enable the detection system to generalize to new speakers due to the pre-training on extensive data. (2) In the mixed-lingual and cross-lingual scenarios, performance was significantly degraded compared to the single-lingual evaluation. For English, the cross-lingual scenario dropped by 17.25% compared to the best-performing single-lingual setting. Similarly, for Bengali, accuracy dropped by 23.72% when comparing the best-performing single-language setting. This indicates that relying solely on pre-trained SSL models might not be adequate for creating language-independent ADD systems. (3) The use of the mean pooling gives relatively higher performance than the max pooling layer in the downstream model. It was also observed that the final layer did not yield the best result for either corpus. These findings are consistent with the outcomes presented in [31]. Additionally, our best model outperformed the previous works using the frozen upstream model on the DAIC-English data.

The cross-lingual experiments with other modalities, such as video and text should be explored to confirm that combined embeddings can build a reliable mental diagnostic tool. Another future direction is to perform the fine-tuning upstream model. It could enhance the effectiveness of the ADD system.

8. References

[1] R. Peveler, A. Carson, and G. Rodin, "Depression in medical patients," *Bmj*, vol. 325, no. 7356, pp. 149–152, 2002.

[2] W. H. Organization *et al.*, "Depressive disorder (depression)," accessed: March 31, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

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

[4] D. M. Khan, N. Yahya, N. Kamel, and I. Faye, "Automated diagnosis of major depressive disorder using brain effective connectivity and 3d convolutional neural network," *IEEE Access*, vol. 9, pp. 8835–8846, 2021.

[5] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain and cognition*, vol. 56, no. 1, pp. 30–35, 2004.

[6] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[7] S. A. Almaghrabi, S. R. Clark, and M. Baumert, "Bio-acoustic features of depression: A review," *Biomedical Signal Processing and Control*, vol. 85, p. 105020, 2023.

[8] 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.

[9] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. F. Cohn, "Cross-cultural depression recognition from vocal biomarkers." in *INTERSPEECH*, 2016, pp. 1943–1947.

[10] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[11] S. Alghowinem, T. Gedeon, R. Goecke, J. F. Cohn, and G. Parker, "Interpretation of depression detection models via feature selection methods," *IEEE transactions on affective computing*, vol. 14, no. 1, pp. 133–152, 2020.

[12] 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.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, 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.

[14] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[15] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[16] V. Ravi, J. Wang, J. Flint, and A. Alwan, "Fraug: A frame rate based data augmentation method for depression detection from speech signals," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6267–6271.

[17] D. Wang, Y. Ding, Q. Zhao, P. Yang, S. Tan, and Y. Li, "Ecapa-tdnn based depression detection from clinical speech." in *INTERSPEECH*, 2022, pp. 3333–3337.

[18] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on Speaker Verification and Language Identification," in *Proc. INTERSPEECH 2021*, 2021, pp. 1509–1513.

[19] Y. Fang, X. Xing, X. Xu, and W. Zhang, "Exploring Downstream Transfer of Self-Supervised Features for Speech Emotion Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 3627–3631.

[20] M. Lebourdais, M. Tahon, A. LAURENT, and S. Meignier, "Overlapped speech and gender detection with WavLM pre-trained features," in *Proc. INTERSPEECH 2022*, 2022, pp. 5010–5014.

[21] P. Zhang, M. Wu, H. Dinkel, and K. Yu, "Depa: Self-supervised audio embedding for depression detection," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 135–143.

[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] Z. Han, Y. Shang, Z. Shao, J. Liu, G. Guo, T. Liu, H. Ding, and Q. Hu, "Spatial-temporal feature network for speech-based depression recognition," *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[24] D. Wagner, I. Baumann, F. Braun, S. P. Bayerl, E. Nöth, K. Riedhammer, and T. Bocklet, "Multi-class Detection of Pathological Speech with Latent Features: How does it perform on unseen data?" in *Proc. INTERSPEECH 2023*, 2023, pp. 2318–2322.

[25] V. Mitra, E. Shriberg, D. Vergyri, B. Knoth, and R. M. Salomon, "Cross-corpus depression prediction from speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4769–4773.

[26] F. Braun, S. P. Bayerl, P. A. Pérez-Toro, F. Hönig, H. Lehfeld, T. Hillemacher, E. Nöth, T. Bocklet, and K. Riedhammer, "Classifying Dementia in the Presence of Depression: A Cross-Corpus Study," in *Proc. INTERSPEECH 2023*, 2023, pp. 2308–2312.

[27] 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.

[28] 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.

[29] 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.

[30] Q. Li, D. Wang, Y. Ren, Y. Gao, and Y. Li, "FTA-net: A Frequency and Time Attention Network for Speech Depression Detection," in *Proc. INTERSPEECH 2023*, 2023, pp. 1723–1727.

[31] S. R. Kadiri, F. Javanmardi, and P. Alku, "Investigation of self-supervised pre-trained models for classification of voice quality from speech and neck surface accelerometer signals," *Computer Speech & Language*, vol. 83, p. 101550, 2023.