WONDER - Waveform-based Optimal Neurological Depression Evaluation with Representations Leveraging Multilingual Self-Supervised Speech embeddings via Speaker Identity Invariant Training

Biman Najika Liyanage[1], Zhengwen Zhu[1], Jun Yang[1], Zongfeng Li[1], Akila Thalgahagoda[1], Harindu Ashan Sugathadasa[1] , Yun Hua Lin[2]*

Voice Health Tech, Beijing, China[1],
Institute of Mental Health, Peking
University Sixth Hospital, Beijing, China[2]

biman@voicehealthtech.com

Abstract

Recent progress in identifying depression through speech patterns has gained momentum due to the improved precision of foundation models. These models, trained on vast unlabelled speech datasets using self-supervised learning, extract powerful speech characteristics within their transformer encoder layers. This study introduces two novel approaches: one to tackle the challenge of limited data, and another to train the model in a way that minimizes speaker-specific biases. This ensures the model discerns speech features that are indicative of depression, irrespective of the speaker.

1 Introduction

Depression is a widespread and serious mental disorder that is often overlooked due to its similar signs and symptoms [1]. Timely diagnosis is essential for successful intervention. The non-invasive and continuous tracking of physiological and psychological markers through smartphone tech and wearable devices has recently garnered significant attention from the research community [2][3]. Consequently, there's a growing interest in fully automated depression screening tools. Notably, psychomotor retardation, a hallmark symptom of Major Depressive Disorder (MDD), is evident through unique verbal cues such as monotonic speech patterns, specific word choices, and uncommon speech interruptions [5][6][7]. This has led to a surge in the adoption of automated speech-based screening methods [4]. However, despite the depth of research, using conventional machine learning techniques for auto-detecting depression from speech is still problematic due to the scarcity of data [2].

Self-supervised learning (SSL)techniques aim to develop a versatile robust universal model that can be advantageous across numerous downstream tasks. In the recent years the progress of foundation models trained in a self-supervised way exhibits that the self-supervised representations related to acoustic word embeddings and learning with zero lexical resources [8].Such encoders in foundation models trained with SSL method contains information about both linguistics and paralinguistic parts of speech which can be applied to many areas especially in affective computing in SER (speech emotion recognition) [9]. Foundation models, like the Wav2Vec[10], huBERT[11], WavLM[12] excel in learning intricate speech features, largely due to their training on extensive amounts of unlabelled data, ranging from 60k to 94k hours. Such models not only effectively model speech content via masked speech prediction but also expand their versatility to non-ASR tasks by integrating speech denoising creating state of the art (SOTA) benchmarks[13].

Learning feature representations related to the speaker such as X-vectors, i-vectors , and other speaker embeddings have shown potential to increase the accuracy of detecting depression from speech [14][15][16].However Training a model in a speaker-invariant way involves ensuring that the model generalizes across various speakers rather than overfitting to specific vocal characteristics. By doing so, the model focuses on understanding

the content and nuances of the speech rather than the unique attributes of individual speakers. This approach ensures that the model is less prone to learning speaker-specific features, which could limit its applicability and accuracy when exposed to voices it hasn't encountered during training. In essence, a speaker-invariant training approach encourage the model to learn more subtle acoustic feature patterns related to psychomotor retardation, making the model more versatile and robust in real-world scenarios where diverse voices are encountered.

This paper we propose a robust approach for automatic depression detection that is both speaker and language independent. By analysing the latent speech representations from the initial encoder layers, we theorize that these representations are language-neutral due to their proximity to fundamental vocal features, To the best of our knowledge, this is the first work to establish vocal biomarkers associated with depression are inherently objective and remain unaffected by lexical nuances like accents and languages as the datasets used in training and evaluation consist of samples with different age groups and demographics. Self supervised representations of both English and Mandarin Chinese corpuses are fed to the model and trained in a speaker-invariant manner to minimize the effect on masking acoustic abnormalities caused by psychomotor retardation.

2 Proposed Method

In this section we present a unique approach for Speech-Based Depression Detection (SDD). Our method uses a two-pronged strategy: first, self-supervised learning processes are utilized to build feature representations with augmentations; second, these concatenated features are then applied to a training model that makes use of a combined loss-based adversarial learning framework. In a given training session with English and Mandarin depression speech corpus, this framework is intended to maximize the loss on speaker and language identification while concurrently minimizing the loss function related to the detection of Major Depressive Disorder (MDD).The approach ensures the model to learn paralinguistic features in speech with high corelation with MDD. The model structure is illustrated in Fig. 1 where the feature representations go through an data augmentation process and adversarial disentanglement of speaker and depression characteristics.

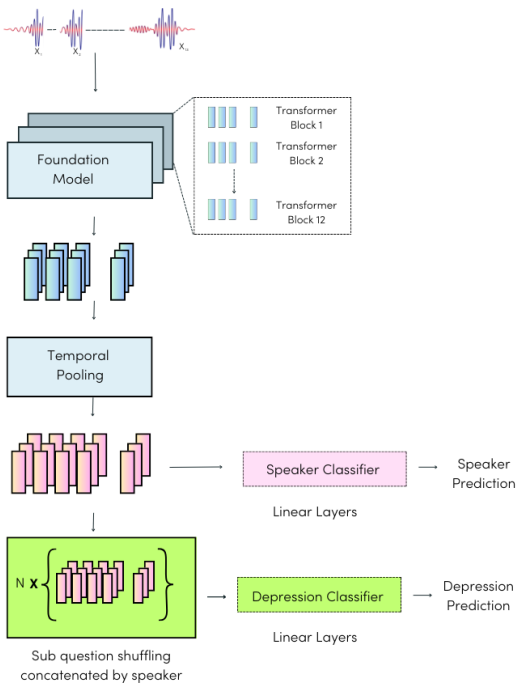

Figure 1: An illustration of the proposed system. The input raw audio segment is fed into foundation model and extracted feature representations are fed into a temporal pooling layer to have a common output vector dimension. The non augmented samples are fed to a speaker classification layer and the augmented features are fed to a depression classification layer where the total loss is a combination of a speaker loss and MDD loss

3 Dataset

3.1 DAIC-WOZ Dataset

Speech based depression detection systems are often benchmarked using the performance on The Distress analysis interview corpus wizard of Oz (DAICWOZ)[17], the dataset contains 189 participants. During the data collection process each participant completed patient health questioner (PHQ-8) [19] and assigned a depression score form a self rating index. The Dataset consist 100 speakers used for training and 30 speakers used for evaluation containing 58 hours of combined speech sampled at 16kHz.

3.2 Oizys Dataset

Oizys is a Mandarin Chinese speech corpus collected by Voice Health Tech annotated partnering with Peking University Sixth Hospital, The Fifth people's Hospital of Tangshan, and Weihai Mental Health Center etc [2].Unlike DAIC-WOZ dataset Oizys is a much diverse dataset containing 45,552 audio samples consists of 3343 unique speakers for training and evaluation as well as 161 speakers for independent validation. During the interview process the participants were diagnosed by a license psychiatrist following DSM-5 process and conclude if the participant was depressed or healthy, depression severity was estimated using Hamilton Depression Rating Scale(HAM-D) method. The participants were asked to use their own mobile device record answers for 13 uniquely designed questions complied according to the privacy and ethical standard capturing a wide range of acoustic properties in speech. Oizys is the only dataset in the world that has the scale and diversity on the voice protocols used to capture data containing Diadochokinetic tasks, sustained vowel, stroop test as well as open speech capturing data from ages ranging from 18-65 year old participants with an average age of 34 where 64.8% of whom are female and 35.2% of whom are male.

4 Data Processing and Augmentation

In this section we outline the procedures pertaining to data processing, feature extraction, and augmentation as utilized in our workflow. The model takes raw audio as input. The SDD model, as delineated in Figure 1 the audio representation features are extracted from the raw audio segments, During the initial stage, the system accepts a unique voice sample from a recording session denoted as X, that consists of 13 unique voice samples from the Oyisiz dataset . This sample is an aggregation of 13 individual voice responses, formulated as a result of a participant answering a set of questions. Hence, we represent the sample as $X = \{x^1, \dots, x^{13}\}$. The foundation model take in the raw audio segment, and give out an embedding matrix. Given F_l frame length and F_s frame stride and F_c denote the frame count .The shape of embedding matrix is $[Fc, D]$, the relationship between frame count for a given signal with a duration d $Fc = \left[\left(\frac{d-F_l}{F_s}\right) + 1\right]$. In our study, the embedding dimension D is 768 and the convolution stride F_l is $20ms$ and F_s is $320ms$ The raw audio after going through the foundation model will be feed in an adaptive average pooling layer to get the Fc to be 1 to let all the outputs to have a same dimensional output vector.

The Output features fed into a speaker invariant training pipeline uniform speaker with a combined loss function forcing the model to learn paralinguistic related features disentangling the model from learning speaker related features. Uniform speaker disentanglement (USD)[20] minimizes the prediction loss for the primary task of depression prediction and simultaneously maximizes the loss of the secondary task of speaker prediction.

$$L_{total} = L_d - \lambda(L_s) \quad (1)$$

In the calculated total loss combines the cross entropy loss for speaker prediction multiplied with λ where lambda is a hyperparameter governing the contribution of speaker loss, 1e-3 was selected as the initial hyper parameter value enforced by the similar literature [21]. In the formulation of the detailed equation, d denoted depression detection, s denoted speaker diarization, and CE denoted Cross Entropy Loss.

$$L_d = CE(\theta_d X, y_d) \quad (2)$$

L_d denotes the Cross Entropy Loss for depression detection, with θ_d representing the parameters of the depression detection model, X the input features, and y_d the true labels for depression. In the equation 2

$$L_s = CE(\theta_s X, y_s) \quad (3)$$

Similarly in the equation 3 L_s denotes the Cross Entropy Loss for speaker diarization, with θ_s representing the parameters of the speaker diarization, X the input features, and y_s the true labels for speaker recognition.

$$L_{total} = CE(\theta_d X, y_d) - CE(\theta_s X, y_s) (4)$$

$$\frac{\partial L_{total}}{\partial \theta_d} = \frac{\partial L_d - \lambda \partial L_s}{\partial \theta_d} (5)$$

The equation 5 computes the gradient of the total loss with respect to the parameters of the depression detection model.

$$\theta_d := \theta_d - \alpha \left(\frac{\partial L_d}{\partial \theta_d} - \lambda \frac{\partial L_s}{\partial \theta_d} \right) (6)$$

Finally, the parameters θ_d of the depression detection model are updated using a learning rate α and the calculated gradient from Equation 5. This update aims to minimize the loss for depression detection while maximizing the loss for speaker diarization, achieving the desired balance between the primary and secondary tasks. by leveraging this set of equations, the optimization step formulates a robust approach to simultaneously address the objectives of accurate depression prediction and effective speaker diarization. The inclusion of the trade-off hyperparameter λ offers flexibility in controlling the balance between the two competing objectives, catering to varying requirements and scenarios in practical applications.

We compel the model to prioritize depression-discriminative data related to paralinguistic features while disregarding certain speaker-specific details, thereby ensuring the model remains unaffected by variations in individual speaker characteristics

According to the Figure 1 The augmented Output features from the foundation model are reused for primary task of depression prediction. Oizys dataset consists of 13 questions with each question exploring different affective state being Neutral, Negative and Positive. Even though Depression is modeled as a binary classification problem, depression inherently contains samples with different severity levels. Different voice tasks for example read speech vs spontaneous speech detecting in the detection of depression, have varying impacts on diverse age groups [22].Therefore In order to balance the depressed and healthy sample count an alternative strategy is used where different permutations of samples from X with different affective states from the depressed class is used as an augmentation till the total sample size between the depressed and healthy class is equal.

The permuted Feature representation arrays from X are concatenated together and passed through to a depression detection classifier configuration of two transformer encoder layers, each operating in a 128-dimensional space and utilizing four attention heads to effectively manage sequence information, These layers are paired with a full-connected layer to facilitate the final classification output.

5 Experiments

The experiments were conducted using DAIC-WOZ dataset and Oizys datasets. In order to create a balanced representation of the dataset a sub data split was created Oizys_train, Oizys_ validation and Oizys_test. For the English dataset DAIC-WOZ with 189 samples containing train , dev and test set , train set was used in training while the dev set was used to evaluate the model final performance.
Detailed information on the training and validation set size is represented in the Table 1

Table 1: *Details about the datasets used for experiments.*

Dataset	Number of Speakers	Number of Samples
$Oizys_{train}$	1323	17191
$Oizys_{validation}$	333	4328
$Oizys_{test}$	161	2188
$DIAC-WOZ_{train}$	100	4450
$DIAC-WOZ_{dev}$	30	1286

The DAIC-WOZ dataset contains audio recordings of the participant and the interviewer's speech. The speech segments related to the participants were extracted using the time-stamps provided with the corpus. Hence speech segments per speaker represented as array. $X_{DW} = \{x^1, \ldots, x^n\}$ the speech segments varies from length and contains free form speech.

5.1 Experiments

5.1.1 DAIC-WOZ model

Model was trained using raw-audio features as input features, DAIC-WOZ dataset contains 50+ hours of audio samples across 189 participants however depressed participants represent 28% while healthy samples represent 72% in the total train set. Data processing and augmentation technique was applied to address the data imbalance issue. The speech segments per participant were randomly shuffled and new permutations were obtained as an augmentation strategy to balance the total depressed and healthy sample count. The permutations of the samples were generated preserving the order of the sequence to ensure the flow of speech.

The extracted foundation model features were channeled through an adaptive pooling layer to achieve a standardized, common as delineated in the speaker invariant training framework depicted in Figure 1. The output features concatenated by speaker and passed along to depression classifier While non concatenated features used in the speaker classifier yielded in optimal F1 scores.

5.1.2 Oizys model

The Oizys model was trained with similar architecture with slight variation in the data augmentation technique, Unlike the DAIC-WOZ dataset Oizys consist of 13 responses per speaker which have an affective state of expression being positive negative and neutral. Oizys_train consists of 8429 depressed samples and 8762 healthy samples, In order to preserve the original variation of the emotional responses the question order was preserved while premutation process. The Extracted Foundation models were later channeled through the speaker invariant training pipeline as depicted in Figure 1.

6 Results and Discussion

6.1 Evaluation Metric

Depression classification when modeled as a binary classification problem is often evaluated using the F1 score which is the harmonic mean of the precision and recall. However in the clinical setting additional metrics are often used to evaluate the algorithm performance. Given that Tp: True Positives , Tn : True Negatives , Fp: False positives and Fn: False negatives Sensitivity of the model is defined as the True positive rate or recall $Sensitivity = \frac{Tp}{Tp+Fn}$. Another important metric is the True negative rate $Specificity = \frac{Tn}{Tn+Fp}$ .In this context we can define Precession as $Precesion = \frac{Tp}{Tp+Fp}$. The AUC is a metric that evaluates the performance of a binary classification system as its discrimination threshold is varied. It represents the probability that a randomly selected positive instance will rank higher than a randomly selected negative instance. Mathematically, AUC is the integral of the ROC curve. The ROC curve plots Sensitivity (or True Positive Rate) against 1−Specificity (or False Positive Rate) for various threshold values. The AUC will be a value between 0 and 1, with 1 indicating a perfect classifier and 0.5 indicating a classifier that is no better than random chance.

6.2 Results

Results section is compromised into 3 sections where the baseline results are shown in the Table 2 . The next section elaborates the results obtained using the speaker invariant training pipeline. Furthermore, benchmark methods on depression classification evaluated as part of the literature review are summarized in the Table 3. And the effect on data augmentation for mixed model is presented in the section Table 4. Comparison of the methods are only performed on the $DIAC - WOZ_{dev}$ set using F1 Scores and additionally to F1 scores sensitivity, specificity and AUC is recorded for each experiment.

For creating the baseline model results were generated by channeling the self-supervised representations through a depression classifier model which consist of a fully connected layer. No

augmentations were applied to the raw data therefore it's evident to result in unbalanced performance in metrics

Comparing to the baseline model where only the self supervised representation are used for training results we can observe an overall a +10% improvement in the external independent test set denoted as $Oizys_{test}$ while Trained using the speaker invariant training pipeline.

In the Oizys_test dataset, we observed a marked enhancement in the performance metrics. The sensitivity exhibited a notable increase +0.10 , while the specificity showed a commendable rise of +0.14 . Additionally, the AUC score demonstrated a significant gain of +0.10. These findings underscore the efficacy of the modifications implemented in the model, emphasizing its potential in depression detection tasks.

6.1.1 Method Comparison

When elevating depression detection performance on various models and the SOTA baseline. The results are based on F1-Avg and F1(D) and F1(ND) for DAIC-WOZ dataset. The DepAudioNet model[23] considered to be the baseline model in MDD classification. The model is trained using Mel-spectrograms and archive 0.6 F1-Avg and also it is observed that F1(D) and F1(H) is imbalanced due to the nature of the dataset.

In the recent publications use of self supervised representations extracted from raw audio file have gained significant attention as they are more robust and have better generalization capabilities compared to traditional hand engineered features such as mel-spectrograms and formants[18].

In the pilot experiments comparing the performance on DAIC-WOZ with other implementations we have been able to gain +10% improvement in F1 scores when the concatenated self-supervised features are channeled through a speaker invariant training process. It's also significant that the SOTA methods struggles when maintaining a balanced representation of both F1(D) and F1(ND) metrices [23][24][25][26][27][28]. This is dues to the highly unbalanced nature of the dataset. This can be overcome augmentation techniques where the sample counts between depressed and non-depressed sample counts are balanced.

6.1.2 Mixed Language Dataset

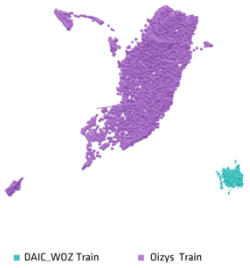

Figure 2: An illustration of the Self supervised representations for DAIC-WOZ and Oizys datasets, the figure shows the embedding space for the 2 corpuses share significant differences .

DAIC-WOZ and Oizys are fundamentally different corpuses. When both models are trained on the datasets individually shows excellent generalization ability while evaluating the evaluation metrics.

In order to further analyze the generalization ability in a mixed language setting Oyzis_train and DAIC-WOZ_train datasets we shuffled together and independently evaluated. The the self supervised representations extracted from the foundation models were fed into a program to visualize the embedding space. Embedding analysis is vital for understanding the behavior of classifiers. Embedding point clouds generated via Uniform Manifold Approximation and Projection(UMAP)[29] highlights the acoustic feature differences between the datasets highlighted in the Figure 2. Further experiments are required to evaluate and map the acoustic differences between the corpuses.

Table 2: *Depression detection performance on Oizys, DAIC-WOZ datasets. baselines is evaluated based on F1-AVG, F1-MAX, Sensitivity, Specificity. Baseline Models without augmentations.*

Model	Dataset Size	Sensitivity	Specificity	AUC	F1(D)	F1(ND)
$DAIC-WOZ_{dev}$	1286	0.65	0.5	0.61	0.55	0.56
$Oizys_{valvalidation}$	4328	0.79	0.71	0.83	0.76	0.74
$Oizys_{test}$	2188	0.67	0.66	0.72	0.66	0.65

Table 3: *Depression detection performance on Oizys, DAIC-WOZ with speaker invariant training pipeline, datasets are evaluated based on F1-AVG, F1-MAX, Sensitivity, Specificity.*

Model	Dataset Size	Sensitivity	Specificity	AUC	F1_avg	F1_max	Adv
$DAIC-WOZ_{dev}$	1286	0.83	0.89	0.88	0.83	0.86	$\lambda = 1e^{-3}$
$Oizys_{validation}$	4328	0.84	0.83	0.91	0.83		$\lambda = 1e^{-3}$
$Oizys_{test}$	2188	0.77	0.80	0.82	0.83		$\lambda = 1e^{-3}$

Table 3: *Performance in detecting depression across various models and SOTA baselines is evaluated based on F1-AVG, F1(ND), F1(D) , utilizing the DAIC-WoZ dataset. The SOTA baseline results are sourced from either reproduced values or directly reported from the relevant studies.*

Model	Input Features	Model Parameters	F1_avg	F1(D)	F1(ND)
DepAudioNet[23]	Mel-Spectrograms	280k	0.60	0.51	0.69
FVTC-CNN [24]	Formants	-	0.64	0.82	0.46
SpeechSimCLR[25]	Mel-Spectrograms	-	0.65	0.56	0.75
SpeechFormer[26]	Wav2vec	33M	0.69	-	-
CNN-LSTM [27]	Spk. Embd. + OpenSmile		0.68	0.51	0.86
WavLMPT[18]	Wavlm	95M	0.70	-	-
ECAPA-TDNN (E3)[28]	Raw-Audio	609k	0.73	0.63	0.83
VoiceHealthTech(ours)	Wavlm	95M	**0.83**	**0.84**	**0.88**

7 Conclusion

This paper approaches speech-based depression detection from invariant training using self-supervised representations to maximize the contribution of para-linguistic features in depression diagnosis. The model performance evaluated on English and Mandarin Chinese corpuses signifies the deep learning method used can be applied to fundamentally different language corpuses when detecting MDD using speech.The proposed approach trained in an adversarial manner surpasses the baseline model performance by +10% archiving SOTA on publicly available DAIC-WOZ dataset with only using raw audio acoustic speech form.

The data augmentation and concatenating the features by speaker represent more information regarding psychomotor retardation compared to random sampling approach applied in DepAudioNet [23] experiments.

8 Future Work

Data scarcity is one of the key issues when it comes to SDD, Future work would be focused on collecting data with a protocol similar to Oizys dataset to create common feature representations between different language corpuses.

References

1. World Health Organization. Depression and Other Common Mental Disorders: Global Health Estimates. Geneva: World Health Organization (2017).
2. Lin, Y., Liyanage, B. N., Sun, Y., Lu, T., Zhu, Z., Liao, Y., Wang, Q., Shi, C., & Yue, W. (2022, October 17). *A deep learning-based model for detecting depression in senior population*. Frontiers. https://doi.org/10.3389/fpsyt.2022.1016676
3. Levine, A. S. (2022, May 12). Mental Health Startup Uses Voice 'Biomarkers' To Detect Signs Of Depression And Anxiety. Forbes. https://www.forbes.com/sites/alexandralevine/2022/05/12/mental-health-startup-uses-voice-biomarkers-to-detect-depression-and-anxiety/
4. Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri, "A review of depression and suicide risk assessment using speech analysis," Speech Communication, vol. 71, pp. 10–49, 2015
5. S. Alghowinem et al., "Detecting depression: a comparison between spontaneous and read speech," in ICASSP. IEEE, 2013, pp. 7547–7551
6. F. Ringeval et al., "Avec 2019 workshop and challenge: stateof-mind, detecting depression with ai, and cross-cultural affect recognition," in Proceedings of the 9th AVEC, 2019.
7. D. M. Low et al., "Automated assessment of psychiatric disorders using speech: A systematic review," Laryngoscope Investigative Otolaryngology, vol. 5, no. 1, pp. 96–116, 2020.
8. Mohamed, A., Lee, H. Y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., Kirchhoff, K., Li, S. W., Livescu, K., Maaløe, L., Sainath, T. N., & Watanabe, S. (2022, May 21). *Self-Supervised Speech Representation Learning: A Review*. arXiv.org. https://doi.org/10.1109/JSTSP.2022.3207050
9. Ioannides, G., Owen, M., Fletcher, A., Rozgic, V., & Wang, C. (Year of Publication). Towards paralinguistic-only speech representations for end-to-end speech emotion recognition. Carnegie Mellon University; Amazon Alexa. (2023, August 24). https://assets.amazon.science/21/f3/1496cf78467399381a6e8bf0ae47/towards-paralinguistic-only-speech-representations-for-end-to-end-speech-emotion-recognition.pdf
10. Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2Vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS, Conference held virtually, 2020.
11. Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "HuBERT: Self-supervised speech representation learning by masked

prediction of hidden units," IEEE Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.

12. Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong 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
13. H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu, C.-I. Lai, J. Shi, X. Chang, P. Hall, H.-J. Chen, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities." Association for Computational Linguistics, 2022.
14. V. Ravi et al., "Fraug: A frame rate based data augmentation method for depression detection from speech signals," arXiv preprint arXiv:2202.05912, 2022.
15. Y. Di et al., "Using i-vectors from voice features to identify major depressive disorder," Journal of Affective Disorders, vol. 288, pp. 161–166, 2021.
16. S. H. Dumpala et al., "Significance of speaker embeddings and temporal context for depression detection," arXiv preprint arXiv:2107.13969, 2021.
17. D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stra- tou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L. P. Morency, "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proc. AAMAS*, Paris, 2014.
18. W. Wu, C. Zhang and P. C. Woodland, "Self-Supervised Representations in Speech-Based Depression Detection," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10094910.
19. K. Kroenke et al., "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.
20. V. Ravi et al., "A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement," in Proc. Interspeech, 2022, pp. 3338–3342.
21. Y. Yin et al., "Speaker-invariant adversarial domain adaptation for emotion recognition," in Proceedings of the 2020 International Conference on Multimodal Interaction, 2020, pp. 481–490.
22. S. Alghowinem et al., "Detecting depression: a comparison between spontaneous and read speech," in ICASSP. IEEE, 2013, pp. 7547–7551.
23. X. Ma et al., "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
24. Z. Huang et al., "Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments," in ICASSP. IEEE, 2020, pp. 6549–6553.
25. S. H. Dumpala et al., "Detecting depression with a temporal context of speaker embeddings," Proc. AAAI SAS, 2022.
26. W. Chen et al., "SpeechFormer: A Hierarchical Efficient Framework Incorporating the Characteristics of Speech," in Proc. Interspeech 2022, 2022, pp. 346–350.
27. D. Jiang et al., "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning," in Proc. Interspeech, 2021, pp. 1544–1548.
28. Wang, J., Ravi, V., Alwan, A. (2023) Non-uniform Speaker Disentanglement For Depression Detection From Raw Speech Signals. Proc. INTERSPEECH 2023, 2343-2347, doi: 10.21437/Interspeech.2023-2101
29. McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426* (2018).