Non-uniform Speaker Disentanglement For Depression Detection From Raw Speech Signals

Jinhan Wang[1], *Vijay Ravi*[1], *Abeer Alwan*[1]

[1]Dept . of Electrical and Computer Engineering, University of California, Los Angeles, USA

(wang7875@g, vijaysumaravi@g, alwan@ee).ucla.edu

Abstract

While speech-based depression detection methods that use speaker-identity features, such as speaker embeddings, are popular, they often compromise patient privacy. To address this issue, we propose a speaker disentanglement method that utilizes a non-uniform mechanism of adversarial SID loss maximization. This is achieved by varying the adversarial weight between different layers of a model during training. We find that a greater adversarial weight for the initial layers leads to performance improvement. Our approach using the ECAPA-TDNN model achieves an F1-score of 0.7349 (a 3.7% improvement over audio-only SOTA) on the DAIC-WoZ dataset, while simultaneously reducing the speaker-identification accuracy by 50%. Our findings suggest that identifying depression through speech signals can be accomplished without placing undue reliance on a speaker's identity, paving the way for privacy-preserving approaches of depression detection.

Index Terms: Depression-detection, Privacy, Healthcare AI, Computational Paralinguistics

1. Introduction

Speech signals have emerged as significant biomarkers of one's emotional and mental state [1, 2, 3, 4]. Several previous studies have successfully demonstrated the potential of using speech in developing automatic objective screening systems for mental health disorders, including serious illnesses such as Major Depressive Disorder (MDD) [5, 6, 7]. Various features and model architectures have been proposed in the past for the purpose of MDD diagnosis [8, 9, 10], each having its own distinct set of advantages and limitations. These include spectral [11, 12], prosodic [13], voice quality [14] and articulatory [15] features as well sophisticated modeling techniques such as data augmentation [16], model ensembles [17], transfer-learning [18] and self-supervised pre-training [19].

Similarly, speaker-identity-related features have been used in depression detection, with previous studies focusing on i-vectors [20], x-vectors [21, 22], or speaker embeddings [23]. Although these models result in good performance, the use of speaker-identity-related features raises privacy-preservation concerns. In the healthcare system, where there is often a stigma surrounding mental health, it is important to develop models that are less reliant on an individual's identity [24].

Although the field of privacy-preserving depression detection is relatively new, a few studies have attempted to endeavor in this direction. Among them, Federated learning [25] and sine-wave speech [26] are notable examples of such methods. Although these methods are promising, their application to low-resource depression detection from speech signals is still in its early stages, and results in significant performance loss [25].

More recently, [27, 28, 29] proposed to remove speaker-related information from speech signals using adversarial learning for speech emotion recognition. We refer to this approach as uniform speaker disentanglement (USD) where the whole model is trained with the same adversarial loss. Despite the promising results of USD in detecting depression, as reported in [30], the model has certain limitations that can impede its performance. One such limitation is the lack of consideration for the interactions between different layers of the model and the relationship between the tasks being performed and the intermediate representations. For example, recent research has shown that different layers of a model capture information differently [31]. It is, therefore, possible that some layers capture more depression information and less speaker information or vice versa, and applying speaker disentanglement to all the layers uniformly may result in sub-optimal performance.

In this paper, we hypothesize that speaker-related information encoded by different layers of a model is idiosyncratic, both in terms of quantity and quality, where some layers may encode more or fewer speaker characteristics than other layers, some of which may not be relevant for depression detection. Assigning a higher penalty to such layers during adversarial training can improve overall model performance. Hence, we propose a novel non-uniform speaker disentanglement method (NUSD) that regulates the proportion of speaker disentanglement applied to different model layers and show that NUSD outperforms USD.

We introduce a new model-input combination by training the ECAPA-TDNN model [32] with raw-audio speech signals as input. NUSD is implemented by adjusting the weighting of the adversarial loss between the two components of the model: the feature extraction (FE) and the feature processing (FP) sections. This method achieves audio-only state-of-the-art (AO-SOTA) performance on depression detection using DAIC-WoZ dataset [33] while simultaneously lowering speaker identification (SID) accuracy. We analyze the behavior of the model layers using a class separability framework, finding that a higher adversarial weight to the FE layers more effectively suppresses speaker information than USD, leading to a better encoding of depression information and performance improvement. To the best of our knowledge, our study is the first to suggest the use of a layer-behavior-based manipulation of loss, in that, we 1) propose differential weighting of the adversarial loss, and 2) utilize the functionality of the FE and FP layers to decide on weight distribution.

The paper is structured as follows: Section 2 presents the proposed NUSD method. Datasets, models, and experiments are described in Section 3. Results are discussed in Section 4, and Section 5 concludes the paper and outlines future directions.

This work was funded by NIH award number R01MH122569.

2. Speaker Disentanglement

Uniform speaker disentanglement (**USD**) [30] minimizes the prediction loss for the primary task and maximizes the loss of the auxiliary task. In the context of this paper, the primary task is depression detection, and the auxiliary task is SID. Consequently, the USD loss function is -

$$L_{USD} = L_{MDD} - \lambda(L_{SPK}) \tag{1}$$

where L_{MDD} is the depression-detection loss and λ controls how much of the SID loss, L_{SPK} contributes to the total loss, L_{USD}. Conventionally, L_{MDD} is Binary Cross Entropy loss, and L_{SPK} is multi-class Cross-Entropy loss.

A higher value of λ indicates a greater adversarial cost during training. This in turn scales the speaker-loss gradient of all the layers uniformly, by the same factor λ. Let the trainable parameters of a model be denoted as θ_{ALL}, then, the gradient of L_{SPK} in USD can be expressed as -

$$\frac{\partial L_{SPK}(USD)}{\partial \theta_{ALL}} = \frac{\partial(\lambda L_{SPK})}{\partial \theta_{ALL}} \tag{2}$$

During the optimizer's update step, the model's parameters are modified as follows:

$$\theta_{ALL} = \theta_{ALL} + \alpha(\frac{\partial L_{SPK}}{\partial \theta_{ALL}} - \frac{\partial L_{MDD}}{\partial \theta_{ALL}}) \tag{3}$$

where α is the learning rate. The negative term (positive sign) for the speaker gradient in Eq. 3 ensures that the model maximizes L_{SPK} while simultaneously optimizing L_{MDD} thereby partially disentangling speaker identity and depression status. Although USD has shown promising results in speaker disentanglement for depression detection [30], it can be further improved by providing better control over the proportion of adversarial disentanglement applied to different model layers.

To address this limitation, we propose a non-uniform speaker disentanglement (**NUSD**) approach. The idea is as follows - the loss gradients of the auxiliary task can be split into multiple components based on model layers and unlike USD, loss maximization can be applied differently to each component thereby allowing for varying levels of disentanglement to be applied to different layers.

As a preliminary study, we split the gradients into two components: the feature extraction component (FE) composed of the initial layers, and the feature processing component (FP) made up of the final layers as detailed in Section 3.3.1. The trainable parameters of these components can be represented as θ_{FE} and θ_{FP} for the FE & the FP layers, respectively. The speaker-loss gradients of FE and FP are non-uniformly scaled using different factors λ_1 and λ_2, respectively. Therefore, for NUSD, the gradient of L_{SPK} can be written as -

$$\frac{\partial L_{SPK}(NUSD)}{\partial \theta_{ALL}} = [\frac{\partial(\lambda_1 L_{SPK})}{\partial \theta_{FE}}, \frac{\partial(\lambda_2 L_{SPK})}{\partial \theta_{FP}}] \tag{4}$$

Comparing Eq. 4 with Eq. 2, it can be observed that NUSD helps us regulate adversarial disentanglement of different layers of the model differently by changing the ratio of λ_1 to λ_2 (denoted as β in later sections). For example, if $\beta < 1$ i.e., $\lambda_2 > \lambda_1$, then the FP layers are penalized more than the FE layers during adversarial training and vice-versa. Conversely, if $\beta = 1$, then NUSD is equivalent to USD.

3. Experimental Details

This section outlines experimental details, including the dataset, preprocessing steps, and the model architecture. Two models, ECAPA-TDNN [32] and DepAudioNet [34] were trained on a publicly available dataset to showcase our approach's effectiveness. ECAPA-TDNN is a SOTA model in SID [32] and emotion recognition [35, 36], while DepAudioNet is a common depression literature baseline [34].

3.1. Dataset and Input Features

3.1.1. Database

The DAIC-WoZ database [33] is a collection of audio-visual interviews in English featuring 189 participants, both male, and female, who underwent psychological distress evaluations. The dataset contains 107 speakers used for training and 35 speakers used for evaluation, consistent with the database description. Audio data from only the patients were extracted using the provided time labels. The validation set was utilized to report results, in line with previous literature.

3.1.2. Data Pre-processing and Input features

Models were trained using raw-audio features as input, with pre-processing steps implemented to address data imbalance [34, 37, 30, 19]. Prior to training, the training data were pre-processed with random cropping and sampling, where each utterance was randomly cropped to the length of the shortest utterance and segmented into multiple 3.84s segments (equivalent to 61440 raw-audio samples). To generate a balanced training subset, an equal number of depression and non-depression segments were randomly sampled without replacement. In each experiment, five models were trained using a randomly generated training subset, with the final prediction averaged across the five models. The raw-audio samples were normalized using mean-variance normalization [37].

Table 1: *Architecture details of the ECAPA-TDNN Model. [InC,OutC,K,S,P,D] are in-channels, out-channels, kernel, stride, padding and dilation, respectively.*

Layer Name	**InC,OutC,K,S,P,D**
Input Layer	[1,128,1024,512,0,1]
SE-Res2-1	[128,128,3,1,2,2]
SE-Res2-2	[128,128,3,1,3,3]
SE-Res2-3	[128,128,3,1,4,4]
Feature aggregation	-
Concat-Conv	[384,384,1,1,0,1]
AttentiveStatsPool	[384,768,-,-,-,-]
Embedding Layer	[768,128,-,-,-,-]
Speaker Prediction Layer	[128,107,-,-,-,-]
Depression Prediction Layer	[128,1,-,-,-,-]

3.2. Models

3.2.1. ECAPA-TDNN

In contrast to previous studies [38] that use spectrograms or MFCCs as inputs, the proposed ECAPA-TDNN model is trained using raw-audio signals. The model architecture was modified (see Table 1) to accommodate raw-audio speech signals as input and avoid overfitting the model to a small training dataset. Specifically, the kernel and stride of the input convolution layer, the number of channels in the intermediate layers, and the dimensions of the prediction layers were modified.

3.2.2. DepAudioNet

This model employs a CNN-LSTM architecture [34] with implementation based on [37]. Two 1-D Convolution layers followed by two unidirectional LSTM layers were used. Lastly, the MDD and speaker prediction layers were fully connected layers with output dimensions for speaker labels being 107.

3.3. Experiments

3.3.1. USD and NUSD

Both models share the same adversarial weight λ across all layers in the USD experiments. In contrast, in the NUSD experiments, the FE layers are weighted with λ_1 and the FP layers with λ_2. We consider the input layer and three SE-Res2 blocks of the ECAPA-TDNN model as FE, while the feature aggregation layer, concat-Conv layer, attention layer, fully connected embedding layer, and prediction layers are FP. Similarly, for DepAudioNet, the first 2 convolutional layers are FE with the two LSTM layers along with the prediction layers as FP. β and λ values are empirically chosen[1].

3.3.2. Speaker Identification Experiments

To investigate how speaker disentanglement affects speaker identity, we conduct a SID experiment by training a support vector classifier (SVC) using speaker embeddings (the embedding layer output from the ECAPA-TDNN model or the hidden representation of the last LSTM layer from the DepAudioNet model). During SVC training, embeddings are obtained from the baseline model without speaker disentanglement, while the evaluation embeddings are taken from the model with or without speaker disentanglement. Note that the SID branch of the model is discarded when extracting speaker embeddings.

3.3.3. Layer-wise GDV Analysis

Because the proposed method regulates the magnitude of adversarial disentanglement applied to different components of the models, we investigate the layer-wise behavior of the models with and without NUSD. This is accomplished with Generalized Discrimination Value (GDV) [39] analysis. Previously, GDV has been proposed as a metric to evaluate the separability of specific representations with respect to various classes and data labels. In this paper, we employ GDV to measure the speaker and MDD-separability of individual layers' outputs for the models in consideration. The prediction layers are excluded in this analysis and GDV values are sign-flipped, such that a higher GDV stands for better separability.

4. Results and Discussion

Results are shown in Table 2, where the best results from the literature are in the first part, and the baselines and proposed method, are in the second part. Methods are compared using speaker-level F1-scores for the depressed (F1-D), the non-depressed (F1-ND) classes, and their non-weighted (macro) average (F1-AVG). To measure the degree of speaker disentanglement, the accuracy of SID was reported if applicable.

4.1. Baseline Experiments

The DepAudioNet model ($D1$), trained on raw-audio, achieves an F1-Score of 0.6259, whereas the proposed ECAPA-TDNN model ($E1$), also trained on raw audio signals and without speaker disentanglement, achieves an F1-Score of 0.632, demonstrating a 1.12% improvement. Some previous studies have achieved better results than $D1$ and $E1$, for example the Vowel-based study [44] (0.7) and the SpeechFormer [43] (0.694). However, these studies have certain limitations. The Vowel-based study requires a trained vowel classification model, while SpeechFormer is a large model with 33M parameters. In contrast, the proposed raw audio-based ECAPA-TDNN model has only 609k parameters and does not need any auxiliary classifiers nor expensive self-supervised models making it simpler and more efficient.

4.2. Speaker Disentanglement

When USD is applied to the DepAudioNet Model ($D2$), its performance improves by 9.12% from 0.6259 to 0.6830 (λ = 3e$-$4). Furthermore, when the ECAPA-TDNN model is trained using USD ($E2$), it achieves an impressive F1-Score of 0.7086 (λ = 3e$-$3), outperforming $E1$ by 11.96%. Along with a significant increase in MDD classification performance, there is a decrease in SID accuracy of 11.2% (from 10.04% to 8.91%) and 77.8% (from 42.33% to 9.38%) for $D2$ and $E2$, respectively.

Next, we apply NUSD to the DepAudioNet and ECAPA-TDNN models and label the resulting best-performing models as $D3$ and $E3$, respectively. Model $D3$ achieves an F1 score of 0.7086 (λ_1 = 2e$-$3, λ_2 = 4e$-$4), an increase of 13.21% over $D1$ and 3.75% over $D2$, while only marginally reducing the SID accuracy to 8.05%. The overall best-performing model is $E3$, which achieves an F1 score of 0.7349 (λ_1 = 4e$-$5, λ_2 = 8e$-$6), outperforming other AO-SOTA models in the literature and surpassing the corresponding baseline models $E1$ and $E2$ by 16.12% and 3.7%, respectively while simultaneously reducing the SID accuracy to 4.68%. These results imply that applying NUSD can be an effective way to enhance depression classification performance while reducing SID performance.

The proposed $E3$ model surpasses AO-SOTA performance in depression detection without requiring additional training data, sophisticated pre-trained models, or complicated handcrafted features. Compared to some previously published AO-SOTA results, the method achieves an improvement of 4.98% vs. vowel-based [44] and 5.89% vs. SpeechFormer [43]. Although [23] has a better F1-ND when combining speaker embeddings with OpenSmile [45] features, our method achieves a better overall F1-AVG. Moreover, [23] uses a segment-level evaluation procedure, in contrast to using speaker-level as in this paper, and reports a lower F1-D of 0.43 and F1-ND of 0.82 when only speaker embeddings are used without feature-fusion.

In pilot experiments, we found that raw audio outperformed Mel-spectrograms, and ComparE16 [45]. Hence, we used the strongest baseline as the goal of the work is to provide a framework that can improve performance regardless of the features chosen and not to compare the performance based on features.

4.3. Ablation Experiments

In order to study the impact of the hyperparameter β on model performance, we conducted a series of experiments using different values of β ranging from 10 to 0.1. The F1-AVG scores were plotted as a function of β for both models (Figure 1). Our analysis revealed two key observations: Firstly, for both models, NUSD (β = 5) consistently outperformed USD (β = 1), indicating that a non-uniform manner of adversarial training can be beneficial for performance.

Secondly, we observed a trend in both DepAudioNet and

[1] Code repository: https://github.com/kingformatty/NUSD

Table 2: *Depression detection performance for various models and AO-SOTA baselines based on F1-AVG, F1(ND), F1(D), and Speaker ID accuracy using the DAIC-WoZ dataset. SOTA baseline results are either reproduced values or reported from the corresponding study. The symbol '$-$' indicates that those values were not reported in the corresponding study. The symbols '$\uparrow$' and '$\downarrow$' indicate a higher or lower value is better, respectively. Best results are highlighted in bold.*

Model Architecture	**Input Feature**	**Disentanglement Method**	**Model Parameters**	**F1-AVG** ↑	**F1(ND)** ↑	**F1(D)** ↑	**SID Accuracy** ↓
DepAudioNet [34]	Mel-Spectrogram	None	280k	0.6081	0.6977	0.5185	-
FVTC-CNN [40]	Formants	None	-	0.6400	0.4600	0.8200	-
Speech SimCLR [41]	Mel-Spectrogram	None	-	0.6578	0.7556	0.5600	-
CPC [42]	Mel-Spectrogram	None	-	0.6762	0.7317	0.6207	-
CNN-LSTM [23]	Spk. Embd. + OpenSmile	None	-	0.6850	**0.8600**	0.5100	-
SpeechFormer [43]	Wav2Vec	None	33M	0.6940	-	-	-
Vowel-based [44]	Mel-Spectrogram	None	-	0.7000	0.8400	0.5600	-
DepAudioNet [37] ($D1$)	Raw-Audio	None	445k	0.6259	0.7755	0.4762	10.04%
DepAudioNet [30] ($D2$)	Raw-Audio	USD	459k	0.6830	0.7826	0.5833	8.91%
DepAudioNet ($D3$)	Raw-Audio	NUSD	459k	0.7086	0.8085	0.6087	8.05%
ECAPA-TDNN ($E1$)	Raw-Audio	None	595k	0.6329	0.7273	0.5385	42.33%
ECAPA-TDNN ($E2$)	Raw-Audio	USD	609k	0.7086	0.8085	0.6087	9.38%
ECAPA-TDNN ($E3$)	Raw-Audio	NUSD	609k	**0.7349**	0.8333	**0.6364**	**4.68%**
Δ ($E3$ vs $E2$) in %	-	-	-	3.70	2.80	4.55	-50.11

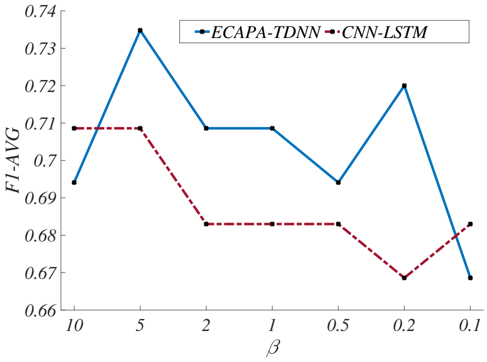

Figure 1: *A plot of F1-AVG versus NUSD β values for the ECAPA-TDNN and the DepAudioNet CNN-LSTM model. Best viewed in color.*

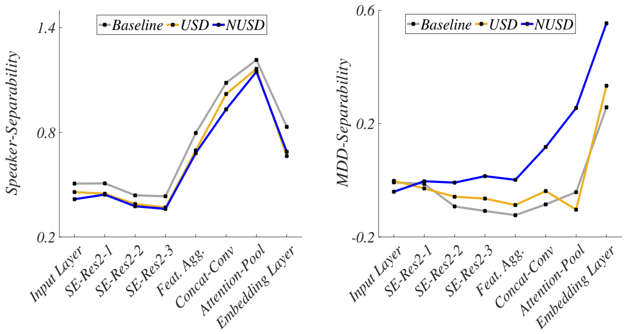

Figure 2: *A plot of the layer-wise speaker and MDD separability GDV scores of the ECAPA-TDNN model. Three models are analyzed - baseline without speaker disentanglement ($E1$), USD ($E2$), and NUSD ($E3$). X-axis represents the layers of the ECAPA-TDNN model. Best viewed in color.*

ECAPA-TDNN models wherein higher values of β produced better results up to $\beta = 5$. This finding suggests that assigning a higher penalty to the initial layers than to the final layers during adversarial training can improve model performance. One possibility is that assigning a higher weight to penalize FE layers in NUSD leads to more effective suppression of speaker-specific feature extraction that may not be too relevant to the primary task of depression detection, compared to assigning the same weight to both FE and FP layers as in USD. Although this is a consequential outcome and holds true for depression detection using the DAIC-WoZ dataset, further investigation is required to verify that the framework generalizes to other domains.

4.4. Layer-wise GDV Analysis

MDD and speaker separability of individual layers of the $E1$, $E2$, and $E3$ were analyzed using the GDV scores (Figure 2). Overall, these plots offer valuable insights into the behavior of the model and shed light on how the proposed method affects the separation of depression and speaker features within the model. Notably, for speaker-separability, NUSD had the lowest value among the three methods in all layers except in the embedding layer (0.664 for USD vs. 0.688 for NUSD) showing that NUSD was better at speaker disentanglement than USD in the FE layers and comparable to USD in the FP layers.

For depression separability, we observe that NUSD has a significantly better separability profile throughout the model compared to USD and the baseline counterparts. These findings support our hypothesis that speaker information encoded by different layers of the model is distinctive and non-uniform speaker disentanglement, which exploits these characteristics of model behavior, leads to better depression detection.

5. Conclusion

The proposed privacy-preserving approach of speech-based depression detection shows promising results by utilizing a non-uniform mechanism of adversarial SID loss maximization. The approach achieves an F1-Score of 0.7349 on the publicly available DAIC-WoZ dataset without any data augmentation, pre-training, or handcrafted features, outperforming other AO-SOTA methods while simultaneously reducing SID accuracy. These findings suggest that our approach leads to better model performance with improved speaker disentanglement. Future work will focus on analyzing the effects of the number of speakers in the database, exploring a more fine-grained, data-driven variant of NUSD, and extending the approach to other domains.

6. References

[1] N. Cummins *et al.*, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[2] A. Nilsonne, "Speech characteristics as indicators of depressive illness," *Acta Psychiatrica Scandinavica*, vol. 77, no. 3, pp. 253–263, 1988.

[3] N. J. Andreasen *et al.*, "Linguistic analysis of speech in affective disorders," *Archives of General Psychiatry*, vol. 33, no. 11, pp. 1361–1367, 1976.

[4] V. Ravi *et al.*, "Voice quality and between-frame entropy for sleepiness estimation," *Interspeech*, 2019.

[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: state-of-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] E. Rejaibi *et al.*, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, p. 103107, 2022.

[9] Y. Shen *et al.*, "Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model," in *ICASSP*. IEEE, 2022, pp. 6247–6251.

[10] K. Chlasta *et al.*, "Automated speech-based screening of depression using deep convolutional neural networks," *Procedia Computer Science*, vol. 164, pp. 618–628, 2019.

[11] M. H. Sanchez *et al.*, "Using prosodic and spectral features in detecting depression in elderly males," in *Interspeech*, 2011, pp. 3001–3004.

[12] S. P. Dubagunta *et al.*, "Learning voice source related information for depression detection," in *ICASSP*. IEEE, 2019, pp. 6525–6529.

[13] Y. Yang *et al.*, "Detecting depression severity from vocal prosody," *IEEE transactions on affective computing*, vol. 4, no. 2, pp. 142–150, 2012.

[14] A. Afshan *et al.*, "Effectiveness of voice quality features in detecting depression," *Interspeech*, 2018.

[15] N. Seneviratne and C. Espy-Wilson, "Multimodal depression classification using articulatory coordination features and hierarchical attention based text embeddings," in *ICASSP*. IEEE, 2022, pp. 6252–6256.

[16] L. Yang *et al.*, "Feature augmenting networks for improving depression severity estimation from speech signals," *IEEE Access*, vol. 8, pp. 24 033–24 045, 2020.

[17] A. Vázquez-Romero *et al.*, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy*, vol. 22, no. 6, p. 688, 2020.

[18] A. Harati *et al.*, "Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus," in *ICASSP*. IEEE, 2021, pp. 7273–7277.

[19] J. Wang *et al.*, "Unsupervised Instance Discriminative Learning for Depression Detection from Speech Signals," in *Proc. Interspeech*, 2022, pp. 2018–2022.

[20] 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.

[21] J. V. Egas-López *et al.*, "Automatic assessment of the degree of clinical depression from speech using x-vectors," in *ICASSP*. IEEE, 2022, pp. 8502–8506.

[22] V. Ravi *et al.*, "Fraug: A frame rate based data augmentation method for depression detection from speech signals," in *ICASSP*. IEEE, 2022, pp. 6267–6271.

[23] S. H. Dumpala *et al.*, "Detecting depression with a temporal context of speaker embeddings," *Proc. AAAI SAS*, 2022.

[24] S. D. Lustgarten *et al.*, "Digital privacy in mental healthcare: current issues and recommendations for technology use," *Current opinion in psychology*, vol. 36, pp. 25–31, 2020.

[25] B. Suhas *et al.*, "Privacy sensitive speech analysis using federated learning to assess depression," *ICASSP*, 2022.

[26] S. H. Dumpala *et al.*, "Sine-Wave Speech and Privacy-Preserving Depression Detection," in *Proc. SMM21, Workshop on Speech, Music and Mind 2021*, 2021, pp. 11–15.

[27] H. Li *et al.*, "Speaker-invariant affective representation learning via adversarial training," in *ICASSP*. IEEE, 2020, pp. 7144–7148.

[28] I. Gat *et al.*, "Speaker normalization for self-supervised speech emotion recognition," in *ICASSP*. IEEE, 2022, pp. 7342–7346.

[29] 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.

[30] V. Ravi *et al.*, "A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement," in *Proc. Interspeech*, 2022, pp. 3338–3342.

[31] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2022.

[32] B. Desplanques *et al.*, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[33] M. Valstar *et al.*, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on AVEC*, 2016, pp. 3–10.

[34] 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.

[35] M. Ravanelli *et al.*, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[36] E. Morais *et al.*, "Speech emotion recognition using self-supervised features," in *ICASSP*, 2022, pp. 6922–6926.

[37] A. Bailey *et al.*, "Gender bias in depression detection using audio features," in *2021 29th EUSIPCO*. IEEE, 2021, pp. 596–600.

[38] D. Wang *et al.*, "ECAPA-TDNN Based Depression Detection from Clinical Speech," in *Proc. Interspeech*, 2022, pp. 3333–3337.

[39] A. Schilling *et al.*, "Quantifying the separability of data classes in neural networks," *Neural Networks*, vol. 139, pp. 278–293, 2021.

[40] Z. Huang *et al.*, "Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments," in *ICASSP*. IEEE, 2020, pp. 6549–6553.

[41] D. Jiang *et al.*, "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning," in *Proc. Interspeech*, 2021, pp. 1544–1548.

[42] A. V. D. Oord *et al.*, "Representation learning with contrastive predictive coding," *Proc. of NIPS*, 2018.

[43] W. Chen *et al.*, "SpeechFormer: A Hierarchical Efficient Framework Incorporating the Characteristics of Speech," in *Proc. Interspeech 2022*, 2022, pp. 346–350.

[44] K. Feng *et al.*, "Toward knowledge-driven speech-based models of depression: Leveraging spectrotemporal variations in speech vowels," in *IEEE-EMBS ICBHI*. IEEE, 2022, pp. 01–07.

[45] F. Eyben *et al.*, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.