A Privacy-Preserving Unsupervised Speaker Disentanglement Method for Depression Detection from Speech

Vijay Ravi[1], **Jinhan Wang**[1], **Jonathan Flint**[2], **Abeer Alwan**[1]

[1]Department of Electrical and Computer Engineering
[2] Department of Psychiatry and Biobehavioral Sciences
University of California Los Angeles, California, USA 90095
vijaysumaravi@ucla.edu, wang7875@ucla.edu, jflint@mednet.ucla.edu, alwan@ee.ucla.edu

Abstract

The proposed method focuses on speaker disentanglement in the context of depression detection from speech signals. Previous approaches require patient dataset speaker labels, encounter instability due to loss maximization, and introduce unnecessary parameters for adversarial domain prediction. In contrast, the proposed unsupervised approach reduces cosine similarity between latent spaces of depression and pre-trained speaker classification models. This method outperforms baseline models, matches or exceeds adversarial methods in performance, and does so without relying on speaker labels or introducing additional model parameters, leading to a reduction in model complexity. The higher the speaker de-identification score ($DeID$), the better the depression detection system is in masking a patient's identity thereby enhancing the privacy attributes of depression detection systems. On the DAIC-WOZ dataset with ComparE16 features and an LSTM-only model, our method achieves an F1-Score of 0.776 and a $DeID$ score of 92.87%, outperforming its adversarial counterpart which has an F1-Score of 0.762 and 68.37% $DeID$, respectively. Furthermore, we demonstrate that speaker-disentanglement methods are complementary to text-based approaches, and a score-level fusion with a Word2vec-based depression detection model further enhances the overall performance to an F1-Score of 0.830.

Introduction

Depression is anticipated to become the second leading cause of disability globally, revealing significant diagnostic accessibility gaps (Mathers and Loncar 2006). Recent advancements in speech-based automatic detection have proven invaluable in tackling the challenges posed by this formidable illness (Cummins et al. 2015). The evolution of speech-based MDD detection encompasses diverse acoustic features (Afshan et al. 2018; Dubagunta, Vlasenko, and Doss 2019; Seneviratne et al. 2020), sophisticated backend modeling techniques (Harati et al. 2021; Rejaibi et al. 2022; Liu et al. 2023), and innovative data augmentation frameworks (Yang, Jiang, and Sahli 2020; Ravi et al. 2022b). While the efficacy of depression detection systems has seen notable improvements, safeguarding patient privacy remains a paramount concern in digital healthcare systems (Lustgarten et al. 2020), particularly within the realm of mental health, where societal stigma persists as a formidable challenge (Goldman et al. 1999).

Given the pivotal importance of privacy preservation in speech-based depression detection, numerous previous studies have delved into this concern. Approaches such as federated learning (Bn and Abdullah 2022) and sine wave speech (Dumpala et al. 2021) have been explored to safeguard patient identity; however, these methods often incur a performance degradation in depression detection. More recently, adversarial learning (ADV), introduced in (Ravi et al. 2022a), has demonstrated an enhancement in depression detection performance at the cost of a reduction in speaker classification accuracy. In the work by (Wang, Ravi, and Alwan 2023), non-uniform adversarial weights (NUSD) were identified as superior to vanilla adversarial methods in the context of raw audio signals. Additionally, in (Zuo and Mak 2023), the utilization of reconstruction loss in conjunction with an autoencoder was found effective in achieving speaker disentanglement, consequently leading to improved depression detection performance.

Despite the notable progress achieved by the aforementioned studies in enhancing depression detection performance while reducing dependency on a patient's identity, there are significant drawbacks. Firstly, the training of these systems still necessitates speaker labels from patient datasets, posing a challenge to the privacy-preserving aspect of depression detection systems. Secondly, many prior methods rely on an adversarial loss maximization training procedure for speaker disentanglement. While effective in achieving good performance, it is acknowledged that loss maximization is inherently unstable due to the absence of upper bounds for the adversarial domain objective function (Xing, Song, and Cheng 2021). Thirdly, all the aforementioned methods introduce additional parameters, such as adversarial domain prediction layers or reconstruction decoders, to the model training framework, which are extraneous for the primary task. This inefficiency can be mitigated.

Motivated by the efficacy of unsupervised learning approaches (Yang et al. 2021), this paper introduces a novel speaker disentanglement method to address existing challenges. The proposed method focuses on reducing the cosine similarity between the latent spaces of a depression detection model and a speaker classification model. Operating at the embedding level, this approach eliminates the need

for speaker labels from the patient dataset. By reformulating the training process into a loss minimization framework, we overcome the issues of unboundedness associated with adversarial methods. Since the speaker classification models serve as embedding extractors and undergo neither retraining nor fine-tuning, our method achieves efficiency by not requiring domain prediction or reconstruction, resulting in fewer model parameters compared to previous approaches.

Extensive experiments are conducted to validate the efficacy of the proposed method, showcasing its superiority over baseline models (without speaker disentanglement) in terms of depression detection. Furthermore, the method demonstrates performance that is either better than or comparable to adversarial methods. Evaluation across multiple input features and backend models establishes the generalizability of the proposed framework to diverse architectures. The complementary nature of speaker disentanglement methods is highlighted through score-level fusion with text-based models, resulting in an enhanced overall performance when the models are combined.

Subsequent sections of this paper are: Section 2, which describes the proposed method, Section 3, which outlines experimental details, Section 4, which presents and discusses the results, and Section 5, which discusses future research directions.

Proposed Method

In conventional speaker disentanglement methods (Gat et al. 2022; Li et al. 2020), the loss function for the adversarial domain (speaker-prediction) is maximized. Consider the depression prediction loss L_{MDD} and the speaker prediction loss for the adversarial method $L_{SPK-ADV}$. Then the total loss for the model training can be written as -

$$L_{total-ADV} = L_{MDD} - \alpha \cdot L_{SPK-ADV}, \quad (1)$$

where α is a hyperparameter controlling the contribution of the adversarial loss to the main loss function where the negative sign indicates that the speaker prediction loss is maximized thereby forcing the model to learn more depression discriminatory features and less speaker discriminatory features. The speaker prediction loss $L_{SPK-ADV}$ is usually the Cross-Entropy loss defined as -

$$L_{SPK-ADV}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \cdot \log(\hat{y}_{ij}), \quad (2)$$

y is the ground-truth speaker label and $\hat{y}$ is the predicted speaker probabilities for N samples and C speakers.

As discussed earlier, this approach has three major issues: 1) this method requires the ground-truth speaker label y to achieve disentanglement, 2) the disentanglement of speaker identity is based on loss maximization ($-\alpha \cdot L_{SPK-ADV}$ which does not have an upper bound, resulting in degraded stability during training and 3) the speaker prediction branch in the model, to obtain $\hat{y}$, adds additional model parameters that are not useful for depression detection making this approach inefficient. In (Zuo and Mak 2023), along with speaker labels, feature reconstruction is used for speaker disentanglement which adds even more unnecessary parameters. In contrast, we propose an unsupervised method of speaker disentanglement that does not need any patient dataset speaker labels and neither involves loss maximization nor adds additional unwanted model parameters. The proposed method is depicted in Figure 1.

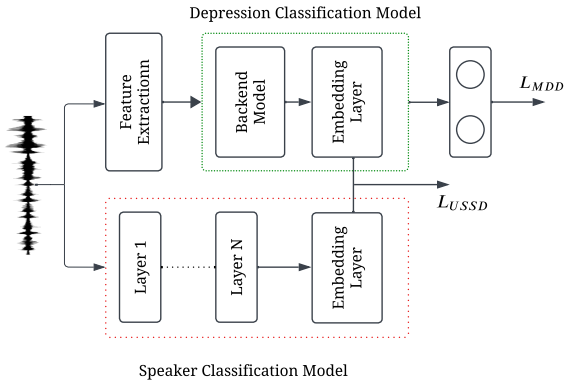

Figure 1: The unsupervised speaker disentanglement method (USSD) aims to minimize cosine similarity between latent spaces of depression classification and speaker classification models.

Consider a depression classification model (θ_{MDD}) and a speaker classification model (θ_{SPK}). For a given speech input X, the latent embeddings of these models are:

$$H_{MDD_X} = \theta_{MDD}(X) \quad (3)$$

$$H_{SPK_X} = \theta_{SPK}(X) \quad (4)$$

Then, we define the unsupervised speaker disentanglement loss function as - L_{USSD} as follows -

$$L_{USSD} = MSE(y_{pred}, y_{target}) \quad (5)$$

Here, we set y_{pred} as the cosine similarity between the two latent space embeddings, defined as:

$$y_{pred} = \frac{H_{MDD_X} \cdot H_{SPK_X}}{||H_{MDD_X}|| \cdot ||H_{SPK_X}||} \quad (6)$$

To ensure the orthogonality between the depression and speaker latent spaces, we set y_{target} to 0.

$$y_{target} = 0 + \epsilon, \quad (7)$$

ϵ is a small noise value added for better convergence (Li et al. 2022) and can be written as:

$$\epsilon = U(0, 1) * 1e - 8. \quad (8)$$

Minimizing the MSE loss of Eq. 5 ensures that the model learns more depression-discriminatory and less speaker-discriminatory information. The total loss is:

$$L_{total-USSD} = L_{MDD} + \alpha \cdot L_{USSD}, \quad (9)$$

Note that experiments in which the squared cosine similarity was minimized resulted in inferior performance.

Experimental Details

Dataset: DAIC-WoZ

The dataset (Valstar et al. 2016), comprises audio-visual interviews conducted in English with 189 participants experiencing psychological distress, including male and female speakers. For our experiments, 107 speakers were employed for training, while an additional 35 speakers were designated for evaluation purposes, aligning with the dataset specifications. The audio data only from the patients were extracted based on the provided time labels. For text-based experiments, the transcripts provided with the database were used. Results are reported using the validation set in line with previous research (Ma et al. 2016; Bailey and Plumbley 2021; Feng and Chaspari 2022; Wu, Zhang, and Woodland 2023).

Input Features

For the audio, four input features are evaluated to show that the proposed framework is independent of the acoustic features used. Mel-Spectrograms, raw-audio signals, ComparE16 features from the OpenSmile library (Eyben, Wöllmer, and Schuller 2010), and the last hidden state of the Wav2Vec2 (Baevski et al. 2020) model are used. Mel-Spectrograms are 40 and 80 dimensional, raw-audio features are 1-dimensional, ComparE16 features are 130-dimensional and Wav2vec2 features are 768 dimensional. For the text, a Word2vec model (Mikolov et al. 2013) is used to extract word-level embeddings from the transcripts of the patient's audio. The embeddings are 200 dimensional. Audio and text feature processing is based on publicly available code repository (Bailey and Plumbley 2021). Since there is an imbalance in the dataset, similar to (Ma et al. 2016; Bailey and Plumbley 2021), random cropping and segmentation are applied. To negate the bias effects of randomness, 5 models are trained with different random seeds, and performances are obtained via majority voting (MV).

Models

Similar to input features, multiple model architectures are designed for the audio modality to show that the proposed method generalizes to different model architectures. Mel-spectrogram features and Raw-Audio signals are used with two model configurations - CNN-LSTM and ECAPA-TDNN (Desplanques, Thienpondt, and Demuynck 2020; Wang et al. 2022a). The other two features, ComparE16 and Wav2vec2 are used with an LSTM-only configuration. For the speaker classification model, two pre-trained models are used - ECAPA-TDNN (128-dimensional embedding) and the X-Vector model (Snyder et al. 2018) (256-dimensional embedding) from the hugging face speechbrain library (Ravanelli et al. 2021). Note that the number of parameters reported for each experiment does not include off-the-shelf speaker classification models that have not undergone retraining or fine-tuning. For the text model, a simple CNN-LSTM framework was used. In the interest of space and since this paper does not propose any new neural network architecture but rather uses previously established models, we do not explain the model architecture in detail. However, model weights and code repository will be made publicly available here[1].

Evaluation Metrics

As is common in the depression detection literature, to measure system performance, the F1 scores (Chinchor 1992) for the two classes (Depressed: D and Non-Depressed: ND) F1-D and F1-ND as well as their macro-average, F1-AVG were reported. To evaluate the privacy-preserving capabilities of the models, the De-Identification score (Noé et al. 2020), inspired by the voice privacy literature, is used and measures how good the anonymization process is (Tomashenko et al. 2022). In the context of this paper, a better system has a higher F1-AVG as well as a higher $DeID$. Since $DeID$ is calculated using voice similarity matrices constructed using embeddings before and after disentanglement, it is only reported for the speaker-disentangled experiments.

Results and Discussion

Speaker Disentanglement versus Baseline

Table 1 shows enhanced depression detection performance (F1-AVG) across all experiments when applying speaker disentanglement, either in the form of ADV or USSD. On average, a notable improvement of 8.3% and 8.2% was observed for ADV and USSD, respectively, over six experiments. The highest improvement with ADV, 13.8%, occurred when utilizing Raw-Audio features with the ECAPA-TDNN model, while the lowest improvement, 5.3%, was observed with Mel-Spectrograms features and the ECAPA-TDNN model. In the case of USSD, the highest improvement was 11.7% with ComparE16 features and the LSTM-only model, and the lowest improvement was 3.8% with Mel-Spectrogram features and the CNN-LSTM model. These results collectively indicate that partially normalizing speaker identity-related information can significantly enhance depression detection performance.

USSD versus ADV

Comparing USSD to its adversarial counterpart, ADV, we observe that the proposed method outperforms ADV in 2 out of 6 experiments: Raw-Audio with CNN-LSTM (0.746 for USSD vs. 0.709 for ADV) and ComparE16 with LSTM-only (0.776 for USSD vs. 0.762 for ADV). Conversely, ADV exhibits better performance than USSD in 3 out of 6 experiments, with both methods yielding the same results in 1 out of 6 experiments. In the aggregate, ADV achieves the best overall results with an F1-Score of 0.79, whereas the corresponding USSD model achieves 0.773—a slight decrease of 2.15%, despite using 15k fewer parameters and not relying on speaker labels. Even without utilizing speaker labels or additional parameters for predicting speakers, USSD showcases comparable or superior performance to ADV. This highlights the potential advantage of USSD over ADV in scenarios where speaker labels for the training set are either unavailable or cannot be used.

[1]Model weights and code repository available at - https://github.com/vijaysumaravi/USSD-depression

Feature	Model	Disentanglement	Number of Parameters	F1-AVG (MV)	F1-ND	F1-D	$DeID$
Mel-Spectrogram	CNN-LSTM	No	280k	0.658	0.756	0.560	NA
		ADV	293k	0.694	0.773	0.615	14.01%
		USSD	280k	0.683	0.783	0.583	10.29%
	ECAPA-TDNN	No	515k	0.709	0.809	0.609	NA
		ADV	529k	0.746	0.826	0.667	3.69%
		USSD	515k	0.746	0.826	0.667	5.97%
Raw-Audio	CNN-LSTM	No	445k	0.669	0.792	0.546	NA
		ADV	459k	0.709	0.809	0.609	55.83%
		USSD	445k	0.746^{+}	0.826	0.667	45.35%
	ECAPA-TDNN	No	595k	0.694	0.773	0.615	NA
		ADV	609k	**0.790**	0.880	**0.700**	22.32%
		USSD	595k	0.773^{+}	0.851	0.696	19.90%
ComparE16	LSTM-only	No	1.15M	0.694	0.773	0.615	NA
		ADV	1.18M	0.762^{+}	0.857	0.667	68.37%
		USSD	1.15M	0.776	**0.885**	0.667	**92.87%**
Wav2vec2	LSTM-only	No	3.6M	0.683	0.783	0.583	NA
		ADV	3.7M	0.747	0.863	0.632	52.43%
		USSD	3.6M	0.720	0.840	0.600	58.65%

Table 1: F1-scores using MV and $DeID$, for speaker disentanglement through ADV and USSD using the DAIC-WOZ dataset. Recall that, unlike ADV, USSD does not use speaker labels for disentanglement. The best results are bold-faced. $^{+}$ indicates improvements are not statistically significant.

Privacy Preservation - $DeID$

Privacy is a crucial aspect of speech-based depression detection, and Table 1 demonstrates positive $DeID$ results for both USSD and ADV across all models. Notably, ComparE16 features with USSD achieve the highest $DeID$ at 92.87%. Despite a marginal depression detection performance drop in USSD compared to ADV, USSD excels in privacy preservation. An intriguing finding is that USSD's effectiveness is independent of the type or dimension of speaker embeddings used. Mel-spectrogram and Raw-Audio experiments employed ECAPA-TDNN speaker embeddings, while ComparE16 and Wav2Vec2 experiments used X-vector embeddings with dimension reduction. USSD's reliance on a pre-trained speaker classification model may contribute to leveraging pre-trained speaker embeddings, enhancing the masking of depression embeddings, and resulting in a higher $DeID$.

Text Fusion

Fusing speaker-disentangled audio models with Word2vec-based text models yields a notable improvement in depression F1-score, particularly for the top 2 audio-only models, as shown in Table 2. Specifically, when the ECAPA-TDNN model trained on Raw-Audio is combined with Word2vec, the depression detection F1-Score reaches 0.860. This result compares favorably to the state-of-the-art (SOTA) depression detection F1-Score of 0.89 (F1-Max) reported in (Wu, Zhang, and Woodland 2023), which involves a four-model ensemble, including parameter-heavy models like RoBERTa (Liu et al. 2019) and WavLM (Chen et al. 2022). In contrast, our approach utilizes only Raw-Audio/ECAPA-TDNN for audio classification and Word2vec/CNN-LSTM for text classification. Similar to ADV, the USSD model demonstrates a significant improvement in F1-Score when fused with text models. These findings underscore the complementarity of speaker-disentangled audio-based depression classification with text-based methods. Contrary to the assumption that speaker-disentanglement models shift focus from non-linguistic features to content-related features (Qian et al. 2022), our results suggest that the information learned by speaker-disentanglement models can be complementary to content-related features.

Audio-Model	Disent.	Audio-only	Word2vec Fusion (Text-only)	$DeID$ (Audio-only)
Raw-Audio ECAPA-TDNN	ADV	0.790	0.860 (0.762)	22.32%
ComparE16 LSTM-only	USSD	0.776	0.830 (0.762)	92.87%

Table 2: F1-AVG scores with and without score-level fusion with the Word2vec text model. Results are shown for the top 2 audio-only models together with their $DeID$s that illustrate the privacy-preserving feature of USSD.

Conclusion and Future Work

The proposed unsupervised method for speaker disentanglement in depression detection is a promising approach for improving model efficiency and privacy attributes. By reducing reliance on speaker labels and streamlining the model through the minimization of cosine similarity between latent spaces, we achieve superior performance compared to both baseline models and adversarial methods. A higher $DeID$ indicates better masking of speaker identity, contributing to the algorithm's enhanced privacy. The compatibility of speaker-disentanglement methods with text-based approaches further solidifies the versatility of the method. Future work will study dimension mismatch between speaker and depression embeddings, speaker-embedding extractors from SSL models such as Instance Discrimination Learning (Wang et al. 2022b) which are trained without supervision and are shown to capture significant speaker information, as well as understanding the nature of information learned through speaker disentanglement methods.

Acknowledgments

This work was funded by the National Institutes of Health under the award number R01MH122569- Combining Voice and Genetic Information to Detect Heterogeneity in Major Depressive Disorder.

References

Afshan, A.; Guo, J.; Park, S. J.; Ravi, V.; Flint, J.; and Alwan, A. 2018. Effectiveness of Voice Quality Features in Detecting Depression. In *Proc. Interspeech 2018*, 1676–1680.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NIPS*, 33: 12449–12460.

Bailey, A.; and Plumbley, M. D. 2021. Gender bias in depression detection using audio features. In *29th EUSIPCO*, 596–600. IEEE.

Bn, S.; and Abdullah, S. 2022. Privacy sensitive speech analysis using federated learning to assess depression. In *ICASSP*, 6272–6276. IEEE.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

Chinchor, N. 1992. MUC-4 evaluation metrics in Proc. of the Fourth Message Understanding Conference 22–29.

Cummins, N.; Scherer, S.; Krajewski, J.; Schnieder, S.; Epps, J.; and Quatieri, T. F. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.*, 71: 10–49.

Desplanques, B.; Thienpondt, J.; and Demuynck, K. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech*, 3830–3834.

Dubagunta, S. P.; Vlasenko, B.; and Doss, M. M. 2019. Learning voice source related information for depression detection. In *ICASSP*, 6525–6529. IEEE.

Dumpala, S. H.; Uher, R.; Matwin, S.; Kiefte, M.; and Oore, S. 2021. Sine-Wave Speech and Privacy-Preserving Depression Detection. In *Proc. SMM21, Workshop on Speech, Music and Mind*, volume 2021, 11–15.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. 18th ACM-MM*, 1459–1462.

Feng, K.; and Chaspari, T. 2022. Toward Knowledge-Driven Speech-Based Models of Depression: Leveraging Spectrotemporal Variations in Speech Vowels. In *IEEE-EMBS ICBHI*, 01–07. IEEE.

Gat, I.; Aronowitz, H.; Zhu, W.; Morais, E.; and Hoory, R. 2022. Speaker normalization for self-supervised speech emotion recognition. In *ICASSP*, 7342–7346. IEEE.

Goldman, L. S.; Nielsen, N. H.; Champion, H. C.; and Council on Scientific Affairs, A. M. A. 1999. Awareness, diagnosis, and treatment of depression. *Journal of General Internal Medicine*, 14(9): 569–580.

Harati, A.; Shriberg, E.; Rutowski, T.; Chlebek, P.; Lu, Y.; and Oliveira, R. 2021. Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus. In *ICASSP*, 7273–7277. IEEE.

Li, H.; Tu, M.; Huang, J.; Narayanan, S.; and Georgiou, P. 2020. Speaker-invariant affective representation learning via adversarial training. In *ICASSP*, 7144–7148. IEEE.

Li, L.-Q.; Xie, K.; Guo, X.-L.; Wen, C.; and He, J.-B. 2022. Emotion recognition from speech with StarGAN and Dense-DCNN. *IET Signal Processing*, 16(1): 62–79.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Yu, H.; Li, G.; Chen, Q.; Ding, Z.; Feng, L.; Yao, Z.; and Hu, B. 2023. Ensemble learning with speaker embeddings in multiple speech task stimuli for depression detection. *Frontiers in Neuroscience*, 17: 1141621.

Lustgarten, S. D.; Garrison, Y. L.; Sinnard, M. T.; and Flynn, A. W. 2020. Digital privacy in mental healthcare: current issues and recommendations for technology use. *Current Opinion in Psychology*, 36: 25–31.

Ma, X.; Yang, H.; Chen, Q.; Huang, D.; and Wang, Y. 2016. Depaudionet: An efficient deep model for audio based depression classification. In *Proc. 6th Audio Visual Emotion Challenge*, 35–42.

Mathers, C. D.; and Loncar, D. 2006. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med.*, 3(11): e442.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Noé, P.-G.; Bonastre, J.-F.; Matrouf, D.; Tomashenko, N.; Nautsch, A.; and Evans, N. 2020. Speech Pseudonymisation Assessment Using Voice Similarity Matrices. In *Proc. Interspeech 2020*, 1718–1722.

Qian, K.; Zhang, Y.; Gao, H.; Ni, J.; Lai, C.-I.; Cox, D.; Hasegawa-Johnson, M.; and Chang, S. 2022. Contentvec: An improved self-supervised speech representation by disentangling speakers. In *ICML*, 18003–18017. PMLR.

Ravanelli, M.; Parcollet, T.; Plantinga, P.; Rouhe, A.; Cornell, S.; Lugosch, L.; Subakan, C.; Dawalatabad, N.; Heba, A.; Zhong, J.; Chou, J.-C.; Yeh, S.-L.; Fu, S.-W.; Liao, C.-F.; Rastorgueva, E.; Grondin, F.; Aris, W.; Na, H.; Gao, Y.; Mori, R. D.; and Bengio, Y. 2021. SpeechBrain: A General-Purpose Speech Toolkit. ArXiv:2106.04624, arXiv:2106.04624.

Ravi, V.; Wang, J.; Flint, J.; and Alwan, A. 2022a. A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement. In *Proc. Interspeech*, 3338–3342.

Ravi, V.; Wang, J.; Flint, J.; and Alwan, A. 2022b. Fraug: A frame rate based data augmentation method for depression detection from speech signals. In *ICASSP*, 6267–6271. IEEE.

Rejaibi, E.; Komaty, A.; Meriaudeau, F.; Agrebi, S.; and Othmani, A. 2022. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71: 103107.

Seneviratne, N.; Williamson, J. R.; Lammert, A. C.; Quatieri, T. F.; and Espy-Wilson, C. 2020. Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression. In *Proc. Interspeech*, 4551–4555.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, 5329–5333. IEEE.

Tomashenko, N.; Wang, X.; Vincent, E.; Patino, J.; Srivastava, B. M. L.; Noé, P.-G.; Nautsch, A.; Evans, N.; Yamagishi, J.; O'Brien, B.; et al. 2022. The voiceprivacy 2020 challenge: Results and findings. *Computer Speech & Language*, 74: 101362.

Valstar, M.; Gratch, J.; Schuller, B.; Ringeval, F.; Lalanne, D.; Torres Torres, M.; Scherer, S.; Stratou, G.; Cowie, R.; and Pantic, M. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proc. 6th AVEC*, 3–10.

Wang, D.; Ding, Y.; Zhao, Q.; Yang, P.; Tan, S.; and Li, Y. 2022a. ECAPA-TDNN Based Depression Detection from Clinical Speech. In *Proc. Interspeech*, 3333–3337.

Wang, J.; Ravi, V.; and Alwan, A. 2023. Non-uniform Speaker Disentanglement For Depression Detection From Raw Speech Signals. *arXiv preprint arXiv:2306.01861*.

Wang, J.; Ravi, V.; Flint, J.; and Alwan, A. 2022b. Unsupervised Instance Discriminative Learning for Depression Detection from Speech Signals. In *Proc. Interspeech*, 2018–2022.

Wu, W.; Zhang, C.; and Woodland, P. C. 2023. Self-Supervised Representations in Speech-Based Depression Detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Xing, Y.; Song, Q.; and Cheng, G. 2021. On the algorithmic stability of adversarial training. *NIPS*, 34: 26523–26535.

Yang, L.; Jiang, D.; and Sahli, H. 2020. Feature augmenting networks for improving depression severity estimation from speech signals. *IEEE Access*, 8: 24033–24045.

Yang, S.-w.; Chi, P.-H.; Chuang, Y.-S.; Lai, C.-I. J.; Lakhotia, K.; Lin, Y. Y.; Liu, A. T.; Shi, J.; Chang, X.; Lin, G.-T.; et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Zuo, L.; and Mak, M.-W. 2023. Avoiding dominance of speaker features in speech-based depression detection. *Pattern Recognition Letters*, 173: 50–56.