HHS Public Access

Author manuscript

CEUR Workshop Proc. Author manuscript; available in PMC 2024 April 22.

Published in final edited form as:

CEUR Workshop Proc. 2024 February ; 3649: 57–63.

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, University of California Los Angeles, California, USA 90095

[2]Department of Psychiatry and Biobehavioral Sciences, University of California Los Angeles, California, USA 90095

Abstract

The proposed method focuses on speaker disentanglement in the context of depression detection from speech signals. Previous approaches require patient/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 F1Score 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.

Keywords

Speaker disentanglement; Depression detection; Privacy; Healthcare AI; DAIC-WOZ

1. Introduction

Depression is anticipated to become the second leading cause of disability globally, revealing significant diagnostic accessibility gaps [1]. Recent advancements in speech-based automatic detection have proven invaluable in tackling the challenges posed by this formidable illness [2]. The evolution of speech-based depression detection encompasses diverse acoustic features [3, 4, 5], sophisticated backend modeling techniques [6, 7, 8], and innovative data augmentation frameworks [9, 10]. While the efficacy of depression detection

[*]Corresponding author. vijaysumaravi@ucla.edu (V. Ravi); wang7875@ucla.edu, (J. Wang); jflint@mednet.ucla.edu (J. Flint); alwan@ee.ucla.edu, (A. Alwan).

systems has seen notable improvements, safeguarding patient privacy remains a paramount concern in digital healthcare systems [11], particularly within the realm of mental health, where societal stigma persists as a formidable challenge [12].

Given the pivotal importance of privacy preservation in speech-based depression detection, numerous studies have attempted to address this issue. Approaches such as federated learning [13] and sine wave speech [14] 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 [15, 16], has demonstrated an enhancement in depression detection performance at the cost of a reduction in speaker classification accuracy. In the work by [17], non-uniform adversarial weights (NUSD) were identified as superior to vanilla adversarial methods in the context of raw audio signals. Additionally, in [18], 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 ad-versarial domain objective function [19]. 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.

Driven by the widespread adoption of unsupervised methods, of unsupervised learning approaches [20], this paper introduces a novel speaker disentanglement method to address the above-mentioned 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.

2. Proposed Method

In conventional speaker disentanglement methods [21, 22], 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 $L_{SPK-ADV}$. The total loss for the model training can be written as -

$$L_{total-ADV} = L_{MDD} - \alpha \cdot L_{SPK-ADV}, \tag{1}$$

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 -

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

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 [18], 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 model parameters. The proposed method is depicted in Figure 1.

Consider a depression classification model (θ_{MDD}) and a speaker classification model (θ_{SPK}). For a given speech input $X \in \mathbb{R}^{\mathbb{N}\times\mathbb{F}}$ (N is the batch size and F is the number of features) the latent embeddings of these models are:

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

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

H_{MDD_X} and $H_{SPK_X} \in \mathbb{R}^{\mathbb{N}\mathbb{X}\mathbb{D}}$ where D is embedding size. Next, we compute the predicted cosine similarity matrix between the two latent space embeddings by computing the cosine similarity between every pair of embeddings as follows -

$$Y_{pred_{(i,j)}} = \frac{H_{MDD_{X_i}} \cdot H_{SPK_{X_j}}}{\left\|H_{MDD_{X_i}}\right\| \cdot \left\|H_{SPK_{X_j}}\right\|} \tag{5}$$

where $1 <= i, j <= N$ and $Y_{pred} \in \mathbb{R}^{\mathbb{N}\mathbb{X}\mathbb{N}}$. The objective of the disentanglement process is to minimize the cosine similarity between the two embedding spaces by enforcing orthogonality between the depression and speaker latent spaces. To achieve this, we specifically set Y_{target} to 0, instead of −1. To enhance convergence during implementation, a small noise value, denoted as ϵ is incorporated [23].

$$Y_{target_{(i,j)}} = 0 + \epsilon, \tag{6}$$

$$\epsilon \in U(0, 1e-8) \tag{7}$$

We define the proposed speaker disentanglement loss function L_{USSD} as follows -

$$L_{USSD} = MSE(Y_{pred}, Y_{target}) \tag{8}$$

and the total loss as:

$$L_{total-USSD} = L_{MDD} + \alpha \cdot L_{USSD}, \tag{9}$$

Minimizing the loss function described in Eq. 9 compels the model to emphasize learning more discriminatory information related to depression while reducing its focus on speaker-

related distinctions. In contrast to ADV (Eq. 1), speaker disentanglement in the proposed method is achieved via loss minimization.

It is important to note that embeddings from θ_{SPK} can be extracted without the necessity of speaker labels, rendering the proposed speaker disentanglement method unsupervised. Moreover, only the parameters of θ_{MDD} require updating, as the θ_{SPK} model does not need finetuning and can remain a pre-trained model with frozen weights. Lastly, experiments where is set to zero, meaning the squared cosine similarity is directly minimized, yielded subpar performance compared to those with a non-zero ϵ. Consequently, results from experiments with $\epsilon = 0$ are not included in this paper.

3. Experimental Details

3.1. Dataset: DAIC-WoZ

The dataset [24], 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 [25, 26, 27, 28].

3.2. 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 [29], and the last hidden state of the Wav2Vec2 [30] 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 [31] 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 [26]. Since there is an imbalance in the dataset, similar to [25, 26], 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).

3.3 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 [32, 33]. 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 [34] (256-dimensional embedding) from the hugging face speechbrain library [35]. Note that the number of parameters reported for each experiment does not

include off-the-shelf speaker classification models that have not undergone re-training 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, the model weights and code repository will be publicly available here[1].

3.4. Evaluation Metrics

3.4.1. Depression Detection—As is common in the depression detection literature, to measure system performance, the F1 scores [36] 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.

3.4.2. Privacy Preservation—To assess the privacy-preserving capabilities of the models, we employ the De-Identification score (*DeID*[37]), a metric inspired by the voice privacy literature[38]. The *DeID* score calculation begins with a voice similarity matrix denoted as M_{AB}, computed for a set of N speakers. This matrix is derived from the log-likelihood ratio (LLR) of two segments—one from model A and the other from model B—considered to be from the same speaker. The LLR computation uses a Probabilistic Linear Discriminant Analysis (PLDA) model [39].

Subsequently, voice similarity matrices, M_{oo} and M_{od}, are calculated. M_{oo} utilizes embeddings solely from the baseline model (o), while M_{od} incorporates embeddings from both the baseline model (o) and the speaker-disentangled model (d).. The next step involves calculating the diagonal dominance $D_{diag}(M)$ for both M_{oo} and M_{od}. This measure is determined as the absolute difference between the average diagonal and off-diagonal elements in the matrices. The diagonal dominance value serves as an indicator of how identifiable individual speakers are within a given embedding space, ranging from 0 to 1.

When $D_{diag}(M_{oo})$ equals 1, speakers are completely identifiability in the original embedding space, whereas if $D_{diag}(M_{od})$ equals 0, speakers are unidentifiable after disentanglement. To measure how good the anonymization (disentanglement) process is, the *DeID* score is formulated as -

$$\text{DeID} = 1 - \frac{D_{diag}(M_{od})}{D_{diag}(M_{oo})} \tag{10}$$

DeID is expressed as a percentage, where 0% signifies poor anonymization, and 100% denotes fully successful anonymization. As *DeID* relies on voice similarity matrices constructed from embeddings pre and post-disentanglement, it is exclusively reported for the experiments involving speaker disentanglement.

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

4. Results and Discussion

4.1. 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% over the baseline was observed for ADV and USSD, respectively, for the 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 MelSpectrograms 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. This highlights the advantage of USSD over ADV in scenarios where speaker labels for the training set are unavailable.

4.2. 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 advantage of USSD over ADV in scenarios where speaker labels for the training set are unavailable .

4.3. 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*.

4.4. 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 F1Score of 0.89 (F1-Max) reported in [28], which involves a four-model ensemble, including parameter-heavy models like RoBERTa [40] and WavLM [41]. 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 [42], our results suggest that the information learned by speaker-disentanglement models can be complementary to content-related features.

5. 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 squared 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 [43] which are trained without supervision and 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

[1]. Mathers CD, Loncar D, Projections of global mortality and burden of disease from 2002 to 2030, PLoS Med. 3 (2006) e442. [PubMed: 17132052]

[2]. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF, A review of depression and suicide risk assessment using speech analysis, Speech Commun. 71 (2015) 10–49.

[3]. Afshan A, Guo J, Park SJ, Ravi V, Flint J, Alwan A, Effectiveness of Voice Quality Features in Detecting Depression, in: Proc. Interspeech 2018, 2018, pp. 1676–1680. doi:10.21437/Interspeech.2018-1399.

[4]. Dubagunta SP, Vlasenko B, Doss MM, Learning voice source related information for depression detection, in: ICASSP, IEEE, 2019, pp. 6525–6529.

[5]. Seneviratne N, Williamson JR, Lammert AC, Quatieri TF, Espy-Wilson C, Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression, in: Proc. Interspeech, 2020, pp. 4551–4555. doi:10.21437/Interspeech.2020-2758.

[6]. Harati A, Shriberg E, Rutowski T, Chlebek P, Lu Y, Oliveira R, Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus, in: ICASSP, IEEE, 2021, pp. 7273–7277.

[7]. Rejaibi E, Komaty A, Meriaudeau F, Agrebi S, Othmani A, Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech, Biomedical Signal Processing and Control 71 (2022) 103107.

[8]. Liu Z, Yu H, Li G, Chen Q, Ding Z, Feng L, Yao Z, Hu B, Ensemble learning with speaker embeddings in multiple speech task stimuli for depression detection, Frontiers in Neuroscience 17 (2023) 1141621.

[9]. Yang L, Jiang D, Sahli H, Feature augmenting networks for improving depression severity estimation from speech signals, IEEE Access 8 (2020) 24033–24045.

[10]. Ravi V, Wang J, Flint J, Alwan A, Fraug: A frame rate based data augmentation method for depression detection from speech signals, in: ICASSP, IEEE, 2022, pp. 6267–6271.

[11]. Lustgarten SD, Garrison YL, Sinnard MT, Flynn AW, Digital privacy in mental healthcare: current issues and recommendations for technology use, Current Opinion in Psychology 36 (2020) 25–31. [PubMed: 32361651]

[12]. Goldman LS, Nielsen NH, Champion HC, Council AMA on Scientific Affairs, Awareness, diagnosis, and treatment of depression, Journal of General Internal Medicine 14 (1999) 569–580. [PubMed: 10491249]

[13]. Bn S, Abdullah S, Privacy sensitive speech analysis using federated learning to assess depression, in: ICASSP, IEEE, 2022, pp. 6272–6276.

[14]. Dumpala SH, Uher R, Matwin S, Kiefte M, Oore S, Sine-wave speech and privacy-preserving depression detection, in: Proc. SMM21, Workshop on Speech, Music and Mind, volume 2021, 2021, pp. 11–15.

[15]. Ravi V, Wang J, Flint J, Alwan A, A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement, in: Proc. Interspeech, 2022, pp. 3338–3342. doi:10.21437/Interspeech.2022-10798.

[16]. Ravi V, Wang J, Flint J, Alwan A, Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement, Computer Speech & Language 86 (2024) 101605. URL: https://www.sciencedirect.com/science/article/pii/S0885230823001249. doi:10.1016/j.csl.2023.101605.

[17]. Wang J, Ravi V, Alwan A, Non-uniform Speaker Disentanglement For Depression Detection From Raw Speech Signals, in: Proc. Interspeech, 2023, pp. 2343–2347. doi:10.21437/Interspeech.2023-2101.

[18]. Zuo L, Mak M-W, Avoiding dominance of speaker features in speech-based depression detection, Pattern Recognition Letters 173 (2023) 50–56. URL: https://www.sciencedirect.com/science/article/pii/S0167865523002192. doi:10.1016/j.patrec.2023.07.016.

[19]. Xing Y, Song Q, Cheng G, On the algorithmic stability of adversarial training, NIPS 34 (2021) 26523–26535.

[20]. Yang S.-w., Chi P-H, Chuang Y-S, Lai C-IJ, Lakhotia K, Lin YY, Liu AT, Shi J, Chang X, Lin G-T, et al., Superb: Speech processing universal performance benchmark, arXiv preprint arXiv:2105.01051 (2021).

[21]. Gat I, Aronowitz H, Zhu W, Morais E, Hoory R, Speaker normalization for self-supervised speech emotion recognition, in: ICASSP, IEEE, 2022, pp. 7342–7346.

[22]. Li H, Tu M, Huang J, Narayanan S, Georgiou P, Speaker-invariant affective representation learning via adversarial training, in: ICASSP, IEEE, 2020, pp. 7144–7148.

[23]. Li L-Q, Xie K, Guo X-L, Wen C, He J-B, Emotion recognition from speech with stargan and dense-dcnn, IET Signal Processing 16 (2022) 62–79.

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

[25]. Ma X, Yang H, Chen Q, Huang D, Wang Y, Depaudionet: An efficient deep model for audio based depression classification, in: Proc. 6th Audio Visual Emotion Challenge, 2016, pp. 35–42.

[26]. Bailey A, Plumbley MD, Gender bias in depression detection using audio features, in: 29th EUSIPCO, IEEE, 2021, pp. 596–600.

[27]. Feng K, Chaspari T, Toward knowledge-driven speech-based models of depression: Leveraging spectrotemporal variations in speech vowels, in: IEEE-EMBS ICBHI, IEEE, 2022, pp. 01–07.

[28]. Wu W, Zhang C, Woodland PC, Self-supervised representations in speech-based depression detection, in: ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. doi:10.1109/ICASSP49357.2023.10094910.

[29]. Eyben F, Wöllmer M, Schuller B, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proc. 18th ACM-MM, 2010, pp. 1459–1462.

[30]. Baevski A, Zhou Y, Mohamed A, Auli M, wav2vec 2.0: A framework for self-supervised learning of speech representations, NIPS 33 (2020) 12449–12460.

[31]. Mikolov T, Chen K, Corrado G, Dean J, Efficient estimation of word representations in vector space,2013. arXiv:1301.3781.

[32]. Desplanques B, Thienpondt J, Demuynck K, ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, in: Proc. Interspeech, 2020, pp. 3830–3834. doi:10.21437/Interspeech.2020-2650.

[33]. Wang D, Ding Y, Zhao Q, Yang P, Tan S, Li Y, ECAPA-TDNN Based Depression Detection from Clinical Speech, in: Proc. Interspeech, 2022, pp. 3333–3337. doi:10.21437/Interspeech.2022-10051.

[34]. Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S, X-vectors: Robust dnn embeddings for speaker recognition, in: ICASSP, IEEE, 2018, pp. 5329–5333.

[35]. 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 RD, Bengio Y, SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624, arXiv:2106.04624.

[36]. Chinchor N, Muc-4 evaluation metrics in proc. of the fourth message understanding conference 22–29, 1992.

[37]. Noé P-G, Bonastre J-F, Matrouf D, Tomashenko N, Nautsch A, Evans N, Speech Pseudonymisation Assessment Using Voice Similarity Matrices, in: Proc. Interspeech 2020, 2020, pp. 1718–1722. doi:10.21437/Interspeech.2020-2720.

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

[39]. Kenny P, Bayesian speaker verification with, heavy tailed priors, Proc. Odyssey 2010 (2010).

[40]. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy N, Lewis M, Zettlemoyer L, Stoyanov V, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

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

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

[43]. Wang J, Ravi V, Flint J, Alwan A, Unsupervised Instance Discriminative Learning for Depression Detection from Speech Signals, in: Proc. Interspeech, 2022, pp. 2018–2022. doi:10.21437/Interspeech.2022-10814.

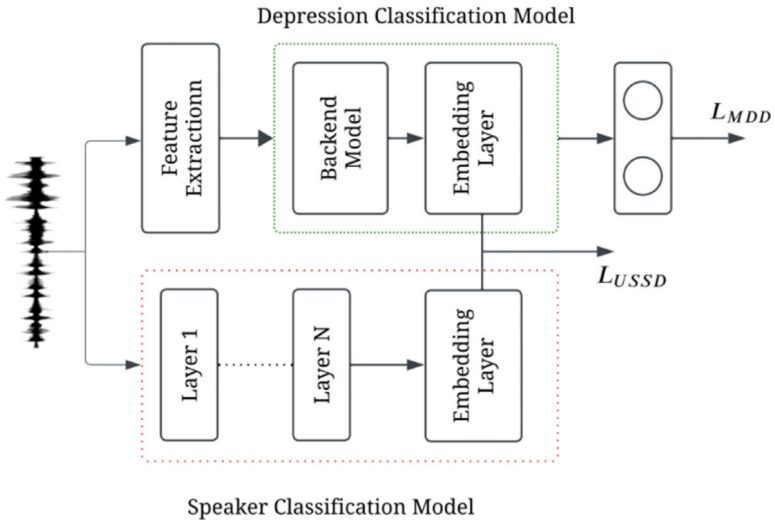

Figure 1:
The unsupervised speaker disentanglement method (USSD) aims to minimize cosine similarity between latent spaces of depression classification and speaker classification models.

Table 1

F1-scores using majority voting (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 parameter count for USSD does not include speaker ID models, as they are neither retrained nor fine-tuned. The best results are bold-faced.

Feature	Model	Disentanglement	Number of Parameters ↓	F1-AVG (MV) ↑	F1-ND ↑	F1-D ↑	*DeID* ↑
Mel-Spectrogram	CNN-LSTM	No	280 k	0.658	0.756	0.560	NA
		ADV	293k	0.694	0.773	0.615	14.01%
		USSD	280 k	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	445 k	0.669	0.792	0.546	NA
		ADV	459 k	0.709	0.809	0.609	55.83%
		USSD	445 k	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%

The symbols

'↑' *and* '↓' indicate a higher or lower value is better, respectively.

$^{+}$indicates improvements are not statistically significant.

Table 2

F1-AVG scores (MV) 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.

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%