PROMOTING INDEPENDENCE OF DEPRESSION AND SPEAKER FEATURES FOR SPEAKER DISENTANGLEMENT IN SPEECH-BASED DEPRESSION DETECTION

Lishi ZUO [1]*, Man-Wai MAK*[1]*, and Youzhi TU*[1]

[1]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University

ABSTRACT

Recent studies have demonstrated the effectiveness of speaker disentanglement in mitigating the interference caused by speaker features in speech-based depression detection. However, the inherent entanglement between depression features and speaker features poses challenges to depression detection. In this study, we propose a mutual information-based speaker-invariant depression detector (MI-SIDD) that aims to promote independence between depression and speaker features to facilitate speaker disentanglement. Specifically, we disentangle the speaker features using a vanilla autoencoder with a well-tuned bottleneck layer and minimize the mutual information between depression and speaker features using a conditional mutual information constraint. Experimental results demonstrate the effectiveness of speaker disentanglement and the promotion of independence between depression and speaker features. Our MI-SIDD model achieves competitive performance compared to state-of-the-art methods on the DAIC-WOZ dataset.

***Index Terms*—** Speaker disentanglement, depression detection, mutual information, speaker embedding

1. INTRODUCTION

Depression [1] is a major public health concern, affecting hundreds of millions worldwide.[1] Because the speech production process could be affected by cognitive and physiological changes caused by depression [2–4], acoustic features offer potential for diagnosing depression. Particularly, deep learning-based methods have been promising in providing a reliable and scalable approach to speech-based depression detection in recent years [5–9]. However, developing accurate and robust depression detection models using deep learning requires a large and diverse dataset that is often hard to obtain due to ethical, legal, and practical constraints [10].

Small depression datasets can easily cause bias and overfit problems. For example, Bailey *et al.* [11] observed gender bias in their depression detector, primary because of the gender imbalance in the training data. In addition, the study in [9] shows that a depression detector may use speaker information for decision making. This phenomenon occurs when there is a limited number of speakers in the training set and the speaker labels and depression labels are correlated. Consequently, the detector learns speaker characteristics instead of depression characteristics. Previous research has shown that minimizing the influence of speaker information improves depression detection performance [6, 9], emphasizing the significance of reducing speakers' interference in depression detection.

An obvious solution is to disentangle speakers' information from depression information. However, a challenge in speaker disentanglement is the inherent entanglement between depression and speaker features. The entanglement makes separating speaker-related features difficult. One potential solution is to promote the independence of depression features and speaker features during speaker disentanglement.

Motivated by the above arguments, we propose a mutual information-based speaker-invariant depression detector (MI-SIDD) that aims to: (1) maximize depression information, (2) minimize speaker information, and (3) promote the independence of depression and speaker features. To be specific, inspired by [9, 12], we use a vanilla autoencoder that comprises an encoder, a decoder, and a projector to minimize speaker information. Like [9], we inject speaker features into the decoder to free the encoder from encoding speaker information. Importantly, inspired by [13], we use a conditional mutual information constraint to promote statistical independence between depression and speaker features in the bottleneck layer, which facilitates the MI-SIDD to learn representations that capture distinct depression information and to remove speaker features. Unlike [9], in which the model operates on utterance-based feature vectors, our model operates on frame-based feature vectors with minimum mutual information between speaker and depression features. To avoid confusion, we refer to the speaker-invariant depression detector in [9] as SIDD and the proposed detector as MI-SIDD.

In this work, we demonstrated that the disentanglement principle in SIDD [9] can be extended to the frame level. Our MI-SIDD achieves competitive performance compared to state-of-the-art methods on the Distress Analysis Interview Corpus – Wizard-of-Oz (DAIC-WOZ) dataset [14]. The

We acknowledge Weiwei Lin for providing the code of attentive pooling, which served as a reference for our work. This work was supported by the RGC of HKSAR, Grant No. 15210122.

[1] https://www.who.int/news-room/fact-sheets/detail/depression

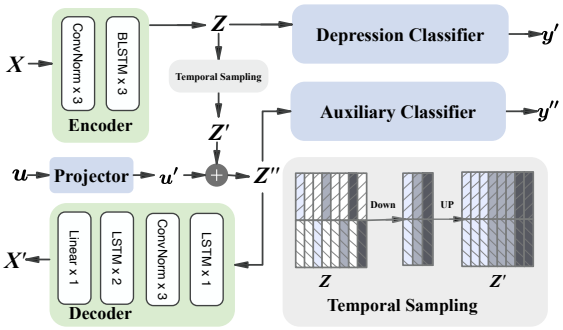

Fig. 1. Framework of MI-SIDD. The gray block represents temporal sampling, where the upper and lower shaded bars of $\boldsymbol{Z}$ correspond to the forward and backward passes of the bi-directional LSTM (BLSTM), respectively. Refer to Section 3.3 for the implementation details of each block.

experimental results demonstrate the effectiveness of minimizing speaker information and encouraging independence between depression and speaker features.

2. METHODOLOGY

Let $\{(\boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{u}_i)\}_{i=1}^N$ be a dataset of N speech segments for training a depression detector. $\boldsymbol{X}_i$ contains the frame-based features of the i-th segment, with T frames and D dimensions. $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$ represent the one-hot representation of the depression class and the speaker identity of the i-th segment, respectively. Also, denote $\tilde{Y}$ and $\tilde{U}$ as the random vectors representing the depression class and speaker identity, respectively. That is, $\boldsymbol{y}_i$'s and $\boldsymbol{u}_i$'s are the realizations of $\tilde{Y}$ and $\tilde{U}$, respectively.

The structure of MI-SIDD is shown in Figure 1. Our goal is to train an encoder that produces latent representation $\boldsymbol{Z}$ that possess the following characteristics: 1) maximum depression information; 2) minimum speaker information; and 3) minimum dependence between depression and speaker features. To this end, we minimize the following loss:

$$\mathcal{L}_{\text{MI-SIDD}} = -I_{\text{dep}}(\boldsymbol{Z}) + I_{\text{spk}}(\boldsymbol{Z}) + I(\tilde{Y}, \tilde{U}|\boldsymbol{Z}), \quad (1)$$

where $\boldsymbol{Z} \in \mathbb{R}^{D\times T}$ contains T frames of latent vectors of dimension D. The first and second terms in Eq. 1 denote depression and speaker information in $\boldsymbol{Z}$, respectively. The third term is the conditional mutual information between the random vectors $\tilde{Y}$ and $\tilde{U}$.

2.1. Conditional Mutual Information Constraint

To minimize the conditional mutual information $I(\tilde{Y}, \tilde{U}|\boldsymbol{Z})$, we utilize a mutual information constraint proposed in [13]. Specifically, $I(\tilde{Y}, \tilde{U}|\boldsymbol{Z})$ is estimated using two approximate distributions, $[q(\boldsymbol{y}|\boldsymbol{Z}) \simeq p(\boldsymbol{y}|\boldsymbol{Z})$ and $h(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{Z}) \simeq p(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{Z})]$, based on the rules of variational approximation:

$$\begin{aligned} &I(\tilde{Y}, \tilde{U}|\boldsymbol{Z}) = H(\tilde{Y}|\boldsymbol{Z}) - H(\tilde{Y}|\tilde{U}, \boldsymbol{Z}) \quad (2)\\ &= -\sup_q \mathbb{E}_{p(\boldsymbol{y},\boldsymbol{Z})}\{\log q(\boldsymbol{y}|\boldsymbol{Z})\} + \sup_h \mathbb{E}_{p(\boldsymbol{y},\boldsymbol{u},\boldsymbol{Z})}\{\log h(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{Z})\}, \end{aligned}$$

where $p(\cdot)$ denotes a probability distribution and $H(\cdot)$ denotes entropy. In practice, $q(\boldsymbol{y}|\boldsymbol{Z})$ and $h(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{Z})$ are implemented by neural networks, i.e., the depression classifier and the auxiliary classifier in Figure 1. Consequently, the loss function for minimizing $I(\tilde{Y}, \tilde{U}|\boldsymbol{Z})$ can be formulated as the difference between two classification losses (Eq. 9 of [13]):

$$\mathcal{L}_{\text{MI}} = \mathbb{E}_{p(\boldsymbol{X},\boldsymbol{y})}\mathcal{L}_{\text{CE}}(\boldsymbol{y}, \boldsymbol{y}') - \mathbb{E}_{p(\boldsymbol{X},\boldsymbol{y},\boldsymbol{u})}\mathcal{L}_{\text{CE}}(\boldsymbol{y}, \boldsymbol{y}''), \quad (3)$$

where $\mathcal{L}_{\text{CE}}$ denotes cross-entropy loss. $\boldsymbol{y}'$ and $\boldsymbol{y}''$ are the output logits of the depression classifier and the auxiliary classifier, respectively.

2.2. Speaker-invariant Autoencoder

AutoVC [12] was designed for voice conversion, and previous work has shown effectiveness in using voice conversion in assessing speech disorders [15]. Inspired by [12], we used a speaker-invariant autoencoder to minimize the speaker information in $\boldsymbol{Z}$. As shown in Figure 1, the speaker-invariant autoencoder consists of an encoder, a decoder, and a projector. The encoder produces speaker-invariant features $\boldsymbol{Z}$. These features are further processed via a temporal sampling procedure to obtain $\boldsymbol{Z}'$. Subsequently, the projector maps a one-hot representation $\boldsymbol{u}$ to a latent vector $\boldsymbol{u}'$, which is of the same dimension as $\boldsymbol{Z}'$. To align $\boldsymbol{u}'$ with $\boldsymbol{Z}'$, $\boldsymbol{u}'$ is replicated T times in the time dimension, denote as $\boldsymbol{U}'$. Speaker-dependent features $\boldsymbol{Z}''$ are then obtained by adding $\boldsymbol{U}'$ to $\boldsymbol{Z}'$. The decoder finally reconstructs the speaker-dependent depression features $\boldsymbol{X}'$ based on $\boldsymbol{Z}''$.

To ensure minimal speaker information in $\boldsymbol{Z}$, we directly present speaker information to the decoder via the projector, preventing the encoder from encoding speaker information in $\boldsymbol{X}$. However, $\boldsymbol{Z}$ needs to contain all necessary information to perfectly reconstruct $\boldsymbol{X}$. Thus, we restrict the amount of information that $\boldsymbol{Z}$ can carry (refer to as capacity) to ensure that $\boldsymbol{Z}$ will not simultaneously contain both speaker and non-speaker information. With limited capacity, the encoder will prioritize the extraction of non-speaker information, leading to speaker disentanglement:

$$I_{\text{mix}}(\boldsymbol{X}) \simeq I_{\text{dep}}(\boldsymbol{Z}) + I_{\text{spk}}(\boldsymbol{U}'). \quad (4)$$

In other words, the mixed information in $\boldsymbol{X}$ can now be separated into depression information in $\boldsymbol{Z}$ and speaker information in $\boldsymbol{U}'$.

We may control the capacity of $\boldsymbol{Z}$ by adjusting the size of the bottleneck layer and implementing a temporal sampling strategy.

Size of the bottleneck layer. The bottleneck factor v ($0 < v \leq 1$) controls the size M of the bottleneck layer of the autoencoder, i.e., $M = 2vD$. We empirically set $v = 1/7$ in the experiments.

Temporal sampling procedure. As illustrated in the gray block in Figure 1, the temporal sampling procedure reduces information by discarding frames in $\boldsymbol{Z}$. It has a downsampling process and an upsampling process. For downsampling, frames at indices $\{0, k, 2k, \ldots\}$ from the forward pass of

the bi-directional LSTM (BLSTM) and frames at indices $\{k-1, 2k-1, 3k-1, \ldots\}$ from the backward pass of the BLSTM are kept. For upsampling, each frame of the down-sampled $\boldsymbol{Z}$ is duplicated along the time axis to restore the original length of $\boldsymbol{Z}$. We set $k = 16$ in this work.

By adjusting the hyperparameters v and k, we can control the capacity of $\boldsymbol{Z}$. Increasing v and decreasing k result in a larger capacity. In the extreme, setting $v = k = 1$ gives the autoencoder its largest capacity, but it is undesirable because $\boldsymbol{Z}$ will contain too much speaker information. Conversely, if v is very small and k is very large, the autoencoder lacks sufficient capacity to reconstruct $\boldsymbol{X}$. Therefore, we minimize $I_{\text{spk}}(\boldsymbol{Z})$ in Eq. 1 by making v small and k large without compromising the capability of the autoencoder to reconstruct $\boldsymbol{X}$, which can be achieved by minimizing a reconstruction loss:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{p(\boldsymbol{X},\boldsymbol{u})} \sum_{t=1}^{T} ||\boldsymbol{x}'_t - \boldsymbol{x}_t||^2, \tag{5}$$

where $\boldsymbol{x}'_t$ and $\boldsymbol{x}_t$ are vectors in $\boldsymbol{X}'$ and $\boldsymbol{X}$, respectively.

2.3. Depression Classification

To encourage $\boldsymbol{Z}$ to contain as much depression information as possible, we want the distribution of the depression classifier's output $\boldsymbol{y}'$ to be as close to the class distribution $\boldsymbol{y}$ as possible. Therefore, we minimize the cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = \mathbb{E}_{p(\boldsymbol{X},\boldsymbol{y})} \mathcal{L}_{\text{CE}}(\boldsymbol{y}, \boldsymbol{y}'). \tag{6}$$

Consequently, by combining Eqs. 3, 5, and 6, the total loss $\mathcal{L}_{\text{MI-SIDD}}$ defined in Eq. 1 can be expressed as:

$$\mathcal{L}_{\text{MI-SIDD}} = \lambda \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{MI}}, \tag{7}$$

where λ is a hyperparameter.

3. EXPERIMENTAL SETUP

3.1. Dataset

The DAIC-WOZ dataset is a collection of clinical interviews from 189 participants [14]. The participants interacted with a virtual interviewer controlled by a researcher in a separate room. Each participant has only one recording, meaning the depression labels and speaker identities are correlated. The dataset has three subsets: a training set with 107 speakers, a development set with 35 speakers, and a test set with 47 speakers.

3.2. Data Pre-processing

We segmented the long audio recording of each speaker into 3.84-second segments. For each segment, 40-dimensional filterbank (FBank) vectors were extracted using a window of 64 ms and a frameshift of 32 ms. Also, frame-based features were extracted from a pre-trained wav2vec model [16].

We removed potential outliers from the 3.84-second segments in the dataset by flattening the frame-based features along the time axis and reducing their dimension to 10 using t-SNE. Subsequently, we employed a one-class SVM on the 10-dimensional vectors of each speaker to detect and remove the outliers.

3.3. Network Implementation

The implementation details of the MI-SIDD architecture are summarized in Table 1. The speaker-invariant autoencoder was modified from [12]. Refer to [12] for the implemetation details of the ConvNorm block. The hidden layers in the dense blocks use the `tanh` activation function. For the decoder and the projector, the output layer of the dense block is linear. For the classifiers, the outputs of the dense blocks are subject to a `softmax` function.

We used the attentive statistics pooling proposed in [17]. For each segment, the pooling layer produces a vector comprising the concatenation of the means, standard deviations, and the means of the first-order differences between consecutive frames from all attention heads.

Table 1. Implementation of the MI-SIDD. For the Dense blocks, $\textbf{dim}_{\textbf{in}}$, $\textbf{dim}_{\textbf{hidden}}$, and $\textbf{dim}_{\textbf{out}}$ refer to the number of nodes in the input layer, hidden layers, and output layer, respectively. For the ConvNorm blocks, these values refer to the number of feature maps.

	Blocks	# of layers	dim_{in}, $\text{dim}_{\text{hidden}}$, dim_{out}
Encoder	ConvNorm	3	D, 128, 128
	BLSTM	3	128, _ ,M
Decoder	LSTM	1	M, _ ,128
	ConvNorm	3	128, 128, 128
	LSTM	2	128, _ , 256
	Dense	1	256, _ , D
Projector	Dense	1	107, _ , M
Depression Classifier	Dense	3	M, 32, 16
	4-Head Attentive Pooling	1	16, 16, 4
	Dense	3	192, [32,16], 2
Auxiliary Classifier	Dense	3	M, 32, 16
	4-Head Attentive Pooling	1	16, 16, 4
	Dense	3	192, [32,16], 2

3.4. Network Optimization

The Adam optimizer was used to train the model for 15 epochs. Each batch contains one randomly sampled segment from each speaker in the training set, with a total of 40 batches per epoch. The learning rate was set to 10^{-4} and warmed up in the first 10% of iterations, followed by a cosine scheduling for the remaining iterations. λ was gradually increased from 0.5 to 1.0. In addition, class weighting was applied to address the class-imbalance problem.

When training the MI-SIDD using frame-based wav2vec features, we blocked the gradient propagation from the depression classifier to the encoder, preventing the classifier from using speaker information for decision-making [9]. Without this intervention, the encoder would produce $\boldsymbol{Z}$ displaying a significant increase in speaker separability (as calculated in [18]), indicating a strong bias towards speaker features. Conversely, when training the MI-SIDD using

FBank features, gradients are allowed to propagate to the encoder, and no increase in speaker separability was observed. Research in speaker verification has shown that speaker embeddings from wav2vec are more speaker discriminative than FBank features [19], which may explain why the MI-SIDD trained on wav2vec features is more susceptible to overfitting the speaker features than the one trained on FBank features.

3.5. Performance Metrics

We calculated the F1 scores for the depression class (F1(D)) and the non-depression class (F1(ND)), separately. We also determined the average of F1(D) and F1(ND), denote as F1-avg. Additionally, we reported the unweighted average recall (UAR) as another evaluation metric. We used a majority voting approach to make the final decision for each test speaker. Consistent with previous studies [5–8], we reported the performance on the development set.

4. RESULTS

4.1. Comparing with State-of-the-Art

Table 2 shows the performance of MI-SIDD and current state-of-the-art methods. Generally, methods with speaker disentanglement (MI-SIDD, [9], and [6]), outperform methods without speaker disentanglement [5, 7], indicating the efficacy of speaker disentanglement in depression detection. The superiority likely arises from the capability of these methods to suppress speaker features, mitigating the risk of overfitting the speaker features.

Methods employing a speaker-invariant autoencoder, such as MI-SIDD and SIDD [9], exhibit better performance compared to adversarial speaker disentanglement [6]. This result demonstrates the autoencoder's effectiveness, which encourages the encoder to learn more meaningful features while minimizing speaker information.

The SIDD under wav2vec features outperforms the MI-SIDD under FBank features, which may be due to wav2vec's better encoding of long-term information from extensive pretraining. In addition, data scarcity can also cause sub-optimal performance of MI-SIDD. However, MI-SIDD performs better than SIDD under the FBank features. Unlike SIDD, which performs speaker disentanglement at the segment level, MI-SIDD performs it at the frame level. Since embeddings from wav2vec features are more speaker discriminative than those from FBank features [19], it is easier for SIDD to disentangle speaker information from wav2vec features at the segment level. On the other hand, speaker information is still highly entangled with depression information at the segment level when FBank features are used. As a result, it is better to perform disentanglement at the frame level when FBank features are used.

4.2. Ablation Study

To assess the impact of individual modules within the MI-SIDD, we reconfigured the architecture in Figure 1 into three distinct models: an encoder–classifier (Enc–Cls), an autoencoder–classifier (AE–Cls), and a frame-based SIDD model (without mutual information minimization). The Enc–Cls comprises an encoder and a depression classifier only. The AE–Cls comprises a speaker-invariant autoencoder and a depression classifier (refer to the plain autoencoder in Figure 4 of [9]). Furthermore, we extended the AE–Cls by incorporating a projector to enable the disentanglement of speaker identities (refer to the SIDD in Figure 4 of [9]). To differentiate the extended AE–Cls from the SIDD [9], we refer to it as frame-based SIDD.

Table 2. Comparing with State-of-the-Art. F and S refer to frame-level and segment-level, respectively.

Method	Feature	Feature Type	F1-avg	F1(ND)	F1(D)
SpeechFormer [7]	wav2vec	F	0.69	-	-
Adv-Disentanglement [6]	wav2vec2.0	F	0.69	0.81	0.58
SIDD [9]	wav2vec	S	**0.81**	**0.87**	**0.76**
MI-SIDD	wav2vec	F	0.71	0.82	0.61
SpeechFormer [7]	Spectrogram	F	0.56	-	-
Depaudionet [5]	FBank	F	0.61	0.70	0.52
Adv-Disentanglement [6]	FBank	F	0.65	0.73	0.56
SIDD [9]	FBank	S	0.67	0.70	0.63
MI-SIDD	FBank	F	**0.76**	**0.81**	**0.71**

Table 3 shows the results of the ablation study on the FBank features, demonstrating consistent performance improvement with the inclusion of each component. Comparing the AE–Cls with the frame-based SIDD, we observe a notable 3% increase in F1-avg, indicating enhanced performance achieved by removing speaker information from the latent space. Moreover, the MI-SIDD surpasses the frame-based SIDD with an additional 3% improvement in F1-avg, showing the effectiveness of promoting independence between depression and speaker features.

We observe a significant performance improvement of 17% in F1-avg scores when comparing the AE–Cls with the Enc–Cls. We conjecture that the improvement is partly because $\mathcal{L}_{\text{recon}}$ (the second term in Eq. 7) encourages the encoder to learn more meaningful features that could describe the input segment rather than learning useless details of speakers, thus reducing the influence of speaker-related characteristics.

Table 3. Ablation Study using FBank Features.

Model	F1-avg	F1(ND)	F1(D)	UAR
Enc–Cls	0.53	0.72	0.33	0.53
AE–Cls	0.70	0.76	0.63	0.71
Frame-based SIDD	0.73	0.79	0.67	0.74
MI-SIDD	0.76	0.81	0.71	0.78

5. CONCLUSIONS

In this work, we proposed a conditional mutual information-based speaker-invariant depression detector (MI-SIDD), which encourages maximum depression information, minimum speaker information, and minimum dependence between depression and speaker features. The experimental results demonstrate the effectiveness of suppressing speaker information and encouraging the independence of depression and speaker features in depression detection.

6. REFERENCES

[1] Ashwani Arya and Tarun Kumar, "Depression: A review," *Journal of Chemical and Pharmaceutical Research*, vol. 3, pp. 444–53, 2011.

[2] 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.

[3] Klaus R. Scherer, "Vocal affect expression: A review and a model for future research.," *Psychological Bulletin*, vol. 99, no. 2, pp. 143, 1986.

[4] Gary Christopher and John MacDonald, "The impact of clinical depression on working memory," *Cognitive Neuropsychiatry*, vol. 10, no. 5, pp. 379–399, 2005.

[5] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang, "DepAudioNet: An efficient deep model for audio based depression classification," in *Proc. the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 35–42.

[6] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan, "A step towards preserving speakers' identity while detecting depression via speaker disentanglement," in *Proc. Interspeech*, 2022, pp. 3338–3342.

[7] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du, "SpeechFormer: A hierarchical efficient framework incorporating the characteristics of speech," *Proc. Interspeech*, pp. 346–350, 2022.

[8] Ying Shen, Huiyu Yang, and Lin Lin, "Automatic depression detection: An emotional audio-textual corpus and a GRU/BiLSTM-based model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6247–6251.

[9] Lishi Zuo and Man-Wai Mak, "Avoiding dominance of speaker features in speech-based depression detection," *Pattern Recognition Letters*, vol. 173, pp. 50–56, 2023.

[10] Tomasz Rutowski, Amir Harati, Elizabeth Shriberg, Yang Lu, Piotr Chlebek, and Ricardo Oliveira, "Toward corpus size requirements for training and evaluating depression risk models using spoken language," *Proc. Interspeech*, pp. 3343–3347, 2022.

[11] Andrew Bailey and Mark D Plumbley, "Gender bias in depression detection using audio features," in *European Signal Processing Conference (EUSIPCO)*, 2021, pp. 596–600.

[12] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *Proc. the 36th International Conference on Machine Learning (ICML)*, 2019, vol. 97, pp. 5210–5219.

[13] Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao, "Invariant information bottleneck for domain generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 7399–7407.

[14] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency, "The distress analysis interview corpus of human and computer interviews," in *Proc. the 9th International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 3123–3128.

[15] Tobias Weise, Philipp Klumpp, Andreas K. Maier, Elmar Nöth, Björn Heismann, Maria Schuster, and Seung Hee Yang, "Disentangled latent speech representation for automatic pathological intelligibility assessment," in *Proc. Interspeech*, 2022, pp. 4815–4819.

[16] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.

[17] Weiwei Lin, Man-Wai Mak, and Lu Yi, "Learning mixture representation for deep speaker embedding using attention.," in *Proc. Speaker Odyssey*, 2020, pp. 210–214.

[18] Lei Wang and Kap Luk Chan, "Learning kernel parameters by using class separability measure," *The 6th kernel machines workshop, in conjunction with Neural Information Processing Systems (NIPS)*, pp. 1–8, 2002.

[19] Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Proc. Interspeech*, 2021, pp. 1509–1513.