Optimizing Learnable Frequency-domain Filterbanks for Depression Detection via Speech Representation Disentanglement

Wenju Yang[1,2], LeYang Li[1], Yong Hao[1,2], Peng Cao[1,2,*], and Osmar R. Zaiane[3]

[1] College of Computer Science and Engineering, Northeastern University, Shenyang 110819, Liaoning, China
[2] Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education
caopeng@cse.neu.edu.cn
[3] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Alberta, Canada

Abstract. Speech has been adopted as a bioinformatic for detecting depression. Current speech-based depression detection (SDD) methods rely on raw signal and Mel-scale features as input. However, utilizing raw signal as input leads to high model complexity, while using Mel-scale features reduces model performance due to insufficient domain-specific adaptation. To solve these issues, we present a depression speech analysis (DSPA[4]) network, which contains a task-oriented learnable frequency-domain filterbanks (LFB) module for optimizing spectral feature generation via end-to-end tuning of filter parameters, and a spectro-temporal representation extraction (STE) module for identifying depression representations in the LFB learned features, while guiding filter parameters optimization. Furthermore, due to LFB exhibiting sensitivity to parameter initialization, a speech representation disentanglement strategy is designed to guide filters to focus on the emotional representations, where pre-trained parameters are used for initializing LFB. Our method yields F1-scores of 0.792, 0.927 and 0.702 on the DAIC-woz, CMDC and EATD-corpus datasets, respectively, with an average improvement of 8.9%, 1.2% and 8.7% over compared state-of-the-art methods. The results show that DSPA is effective in extracting depression-related features.

Keywords: Speech · Depression · Learnable filterbanks · Self-supervised learning · Disentangle representation.

1 Introduction

Depression is a prevalent psychiatric disorder and a primary contributor to the global burden of disease among young people, significantly increasing their risk of suicide and profoundly impacting their emotional well-being [3]. Recently, speech analysis offers a promising solution for depression diagnosis, since depressive speech is characterized by different acoustic features, such as lower pitch, slower

[4] https://github.com/IntelliDAL/Speech/tree/master/FD-learnableFilters

pitch change, slurring, long pauses and monotonous [8]. Developing an effective speech-based depression detection (SDD) method that holds the potential to advance depression screening is highly desirable.

Currently, deep learning methods [12] perform well in SDD tasks by using low-level spectral features as input or modeling directly on the raw signal [22]. For example, *vowel*-CNNs [9] and Wav2Vec [4] utilized multiple convolutional layers to reduce the signal dimensionality. It is worth noting that utilizing unconstrained large kernel convolutions for speech signal modeling will yield a large number of parameters [18]; Additionally, universal features such as Mel-filterbanks (MFbanks) and Mel-frequency cepstral coefficients (MFCCs) have been used as inputs to models like DepAudioNet [16] and Mfcc-LSTM [18]. However, these features are not tailored for SDD tasks, thereby failing to exhibit optimal domain adaptation [17,24]. Recently, several learnable time-domain (TD) filterbanks [22] with few parameters have been employed to extract spectral features by convolving the speech signal with the learnable filters (e.g., Gabor [23], Sinc [17] and Gammton [15] filters). These filters tune their shape and parameters in an end-to-end manner and achieve good domain adaptation. For example, Gabor filterbanks was used to mine differential pronunciation frequencies in depressed speech. [22]. Similarly, customized triangular and bell-shaped learnable filters were optimized for processing speech in speaker verification tasks [14].

Despite the advances in generating features with learnable TD filterbanks, two major challenges remain: 1) Although the learnable TD filters have few parameters, a wide time-window is necessary to achieve fine frequency resolution as required by the Heisenberg uncertainty principle[5]. Therefore, to retain the time information, these methods use a large convolution kernel with a smaller stride in the wavelet transform, typically setting the stride to 1. *This operation significantly increases the computational effort.* 2) DL-based SDD methods depend heavily on extensive and well-annotated datasets, as the data annotation requires significant effort from psychiatrists. As shown in Fig.1, the lack of annotated data poses an issue: *how to effectively initialize these learnable filters?*

Driven by these issues, an end-to-end **D**epression **SP**eech **A**nalysis (DSPA) network is designed to learn useful features and representations for speech data. DSPA assumes that variations in pronunciation are independent of the textual content in SDD tasks. These variations are localized within specific *frequency bands* during certain *time frames* in the spectral features. Therefore, an efficient **L**earnable **F**requency-domain filter**B**anks (LFB) is first used for learning depression features. It comprises a bank of learnable frequency-domain (FD) filters that adaptively identify key frequency bands from the speech signal with low computation cost and few parameters. In addition, a spectro-temporal representation extraction (STE) module is employed to extract key pronunciation frequency bands within certain time frames of the learned features (i.e., spec-

[5] It states that the product of signal's time window and frequency bandwidth must exceed a constant $(1/4\pi)$.

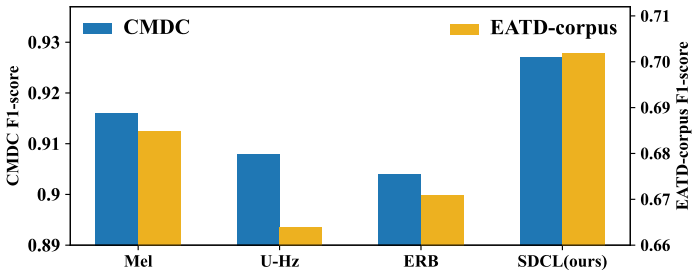


Fig. 1. Effect of different center frequency initial weights in our experiments, where U-Hz denotes uniform sampling in Hertz. An important observation is that classification results are influenced by variations in the initialization manner.

trogram patches). The combination of LFB and STE enables learnable filters to flexibly tune their frequency responses in an end-to-end manner.

Furthermore, a **C**ontrastive **L**earning-based **S**peech representations **D**isentanglement strategy, named SDCL, is introduced to disentangle useful representations from data internals via a self-supervised learning. Importantly, SDCL can guide LFB to obtain better initial weights from unlabeled data. SDCL ignores the different textual contents in speech segments and relies on two assumptions: (*i*) speech representations comprise both speaker representations (SRs) and emotional representations (ERs). SRs aim to identify speakers identity information, such as sex and age. ERs indicate instantaneous emotional variations in specific spectrum regions, such as changes in pitch, rhythm and energy; (*ii*) segments extracted from the same speech exhibit consistent emotional states over short intervals. Based on these assumptions, ERs and SRs from different segments are recombined to achieve the interaction and reconstruction of speech representations. SDCL exploits the potential general information in speech and captures the discriminative patterns in phoneme pronunciation beyond language textual contents. In the downstream task, the ERs-focused filter parameters of LFB will be extracted separately for weight initialization.

Our major contributions are summarized as follows:

1. The LFB module is proposed to automatically modulate the useful frequency feature in the speech for SDD tasks;
2. The SDCL strategy is introduced to learn better initial weights for learnable filterbanks by exploring emotional representation;
3. Our method is evaluated on three publicly available datasets, demonstrating promising performance over state-of-the-art methods.

2 Method

As shown in Fig. 2, DSPA is a joint optimization framework designed for SDD tasks. It involves two-stage: filterbank feature generation and depression representation extraction. First, to acquire depression-related features, a learnable

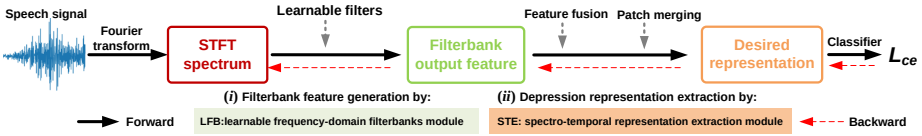


Fig. 2. Overall schematic of the DSPA model for the SDD tasks.

frequency-domain filterbank module (LFB) is designed to recalibrate the feature's importance within the spectrum. Then, a spectro-temporal representation extraction module (STE) is introduced to explore the depression-related patches and representations in the learned features.

2.1 Learnable frequency-domain filterbanks module (LFB)

As shown in Fig. 3, LFB uses a learnable FD filterbanks for retaining task-related acoustic information.

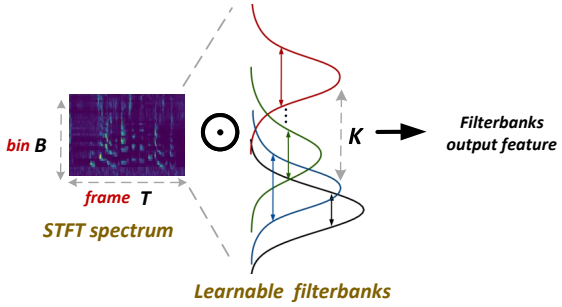


Fig. 3. Illustration of learnable frequency-domain filterbanks.

The convolution theorem provides both time-domain and frequency-domain approaches for signal analysis, it states that the Fourier transform $\mathcal{F}(\cdot)$ of the convolution of two signals is equivalent to the product of their Fourier transforms.

$$\mathcal{F}(\mathbf{x} * f^t) = f^r \odot \mathbf{S}, \tag{1}$$

where $\mathcal{F}(\cdot)$ indicates Fourier transform, $*$ and $\odot$ are convolution and point-wise multiplication operation. In our work, $\mathbf{x} \in \mathbb{R}_{L\times 1}$ and $\mathbf{S} \in \mathbb{R}_{B\times T}$ correspond to the signals and their short-time Fourier transform spectrum, respectively, and $f^t \in \mathbb{R}_{w\times 1}$ and $f^r \in \mathbb{R}_{1\times B}$ are impulse and frequency response of filters. Existing TD filtering methods parameterize the center and bandwidth of f^t and employ a convolutional layer with stride ω and kernel size l to approximate $\mathcal{F}(\cdot)$ [22][23]. However, obtaining the TD filterbanks features requires more computational cost, i.e., $(L-l+1)\cdot(\frac{L-l}{\omega}+1)$ convolutional operations, compared to one-shot

short-time Fourier transform (STFT) $\mathbf{x} \rightarrow \mathbf{S}$ and multiplication operation $\odot$. To solve it, in our work, $\mathbf{x}$ is first transformed into $\mathbf{S}$, then an efficient FD filtering is applied on $\mathbf{S}$ to obtain feature $\mathbf{X}$,

$$\mathbf{X}[k,:] = f_k^r \odot |\mathbf{S}|^2, k \in [1, K], \tag{2}$$

where k is the filter index. f_k^r indicates the Gaussian filter, $|\cdot|^2$ denotes absolute square operation, which aims to emphasize the amplitude information of speech,

$$f_k^r[b] = e^{-\frac{(b-c_k)^2}{2\sigma_k^2}}, b \in [c_k - \frac{B}{2K}, c_k + \frac{B}{2K}], \tag{3}$$

where $c_k \in [0, 1]$ and $\sigma_k \in [0, 0.1]$ are learnable center frequency and standard deviation, respectively. All filters are uniformly distributed across all bins with a length $\frac{B}{K}$, and the difference $\Delta c \geq 0$ between adjacent c_{k+1} and c_k. As a result, a learnable filterbanks $\mathbf{F} \in \mathbb{R}_{K \times B}$ is built by stacking a series of f_k^r. Given an STFT spectrum, the filterbanks outputs are produced as follows,

$$\mathbf{X} = \mathbf{F} \odot |\mathbf{S}|^2 \in \mathbb{R}_{K \times T}, \tag{4}$$

where each f_k^r serves as an aggregation function of the power spectrum, facilitating the retention or attenuation of pertinent information by adjusting its importance weight.

2.2 Spectro-temporal representation extraction module (STE)

As shown in Fig. 4, STE comprises two blocks:
(i) The multi-scale feature fusion block (MFB) for enhancing essential features and weakening irrelevant ones by fusing features under different receptive fields.
(ii) The spectro-temporal cross attention block (CAB) for assigning weights to different patches by comprehensively considering of time frames and frequency bands, thus preserving desired representations.

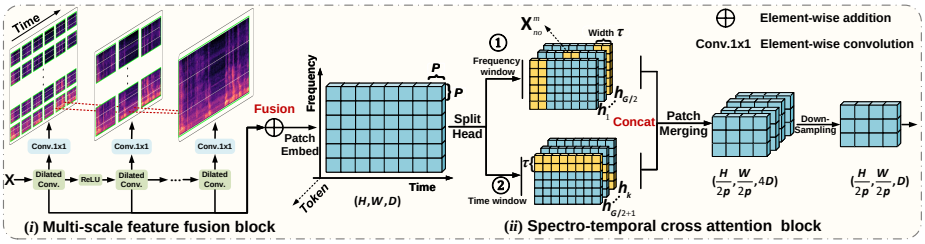


Fig. 4. Illustration of spectro-temporal representation extraction module.

As shown in Fig.4 (i), MFB utilizes dilated convolutions to extract multiscale features from $\mathbf{X}$, where the convolution operations are grouped and applied

individually along the filter axis. Here, η dilated convolutions are stacked, each one with a $2^{\eta-1}$ receptive field, and combined with a ReLU function and element-wise convolution. Then, the obtained multiscale features are fused using element-wise addition, yielding $\mathbf{X}_{fu}$.

As shown in Fig.4 (*ii*), CAB uses a patch-embed [7] convolution of kernel size P to split $\mathbf{X}_{fu}$ into different patch tokens sequentially, yielding $\mathbf{X}_{pe} \in \mathbb{R}^{H\times W\times D}$, where patches with same frequency bins in different time frames are arranged adjacently, H, W indicate the number of patches in vertical and horizontal directions, respectively, and D denotes the token size. Moreover, the patches are organized into different vertical or horizontal windows, and the self-attention is performed in parallel across frequency and time windows. Formally, $\mathbf{X}_{pe}$ is first linearly projected to G heads, and each head computes self-attention within either the frequency or time window. Then, the G heads are equally split into two parallel groups (each has $\frac{G}{2}$ heads). The first group of heads $[\mathbf{h}_1, \cdots, \mathbf{h}_{\frac{G}{2}}]$ executes frequency window self-attention, and the second one performs time window self-attention. The outputs of two parallel groups are concatenated together. For example, in Fig. 4-①, $\mathbf{X}_{pe}$ is partitioned into non-overlapping set $[\mathbf{X}_{no}^1, \cdots, \mathbf{X}_{no}^m, \cdots, \mathbf{X}_{no}^M]$ of equal window width τ. The frequency window self-attention output $\mathbf{Y}_g^m$ is defined as

$$\mathbf{Y}_g^m = softmax(\frac{\mathbf{X}_{no}^m \mathbf{W}_g^1 \cdot (\mathbf{X}_{no}^m \mathbf{W}_g^2)^{\mathrm{T}}}{\sqrt{d_g}}) \cdot \mathbf{X}_{no}^m \mathbf{W}_g^3, \tag{5}$$

where $\mathbf{X}_{no}^m \in \mathbb{R}^{(\tau\times W)\times D}$ and $M = \frac{H}{\tau}$. $\mathbf{W}_g^1, \mathbf{W}_g^2$ and $\mathbf{W}_g^3 \in \mathbb{R}^{D\times d_g}$ represent the projection matrices of queries, keys and values for the g_{th} head, respectively, and d_g is set as $\frac{D}{G}$. The time window self-attention can be similarly derived. Finally, the patch merging is used to merge adjacent patches as $(\frac{H}{2P} \times \frac{W}{2P}, 4D)$, and a fully connected (FC) layer is applied to select the desired representation to $(\frac{H}{2P} \times \frac{W}{2P}, D)$. As the STE is stacked, the representations of interest are retained, and a FC layer is used to complete the classification, and the loss function used the cross-entropy loss.

2.3 Speech feature disentanglement strategy (SDCL)

To better initialize the parameters of LFB, the self-supervised SDCL strategy is proposed to guide the filters to concentrate on emotional representations by leveraging unlabeled data. Specifically, our work design a pretext task to extract consistent ERs by supervising only the similarity of segment pairs. SDCL assumes that speech features comprise both speaker features ($\mathbf{X}_i^{sr}$) and emotional features ($\mathbf{X}_i^{er}$), $i \in \{m, n\}$. By disentangling SRs ($\mathbf{Z}_i^{sr}$) and ERs ($\mathbf{Z}_i^{er}$) from $\mathbf{X}_i^{sr}$ and $\mathbf{X}_i^{er}$ and subsequently exploiting ERs, the learnable filters are guided by SDCL to focus on emotional regions of interest. Importantly, $\mathbf{Z}_m^{sr}$ and $\mathbf{Z}_n^{sr}$ are extracted from distinct segments within the same speech and exhibit same speaker information, while $\mathbf{Z}_n^{er}$ and $\mathbf{Z}_m^{er}$ preserve identical emotional states. This motivates us to recombine $\mathbf{Z}_i^{sr}$ with $\mathbf{Z}_i^{er}$ to achieve the interaction and reconstruction of speech representations. SDCL involves three factors:

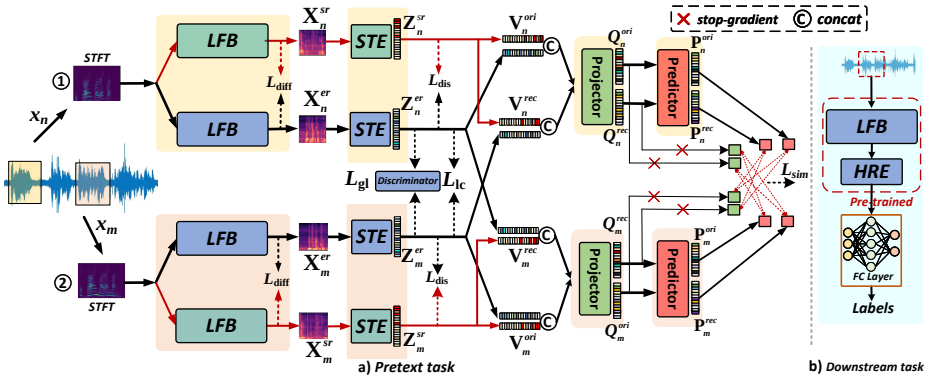


Fig. 5. Overview of the proposed speech representation disentanglement strategy (SDCL) and classification framework (*right*).

(*i*) **Speech sampling**: A straightforward data augmentation method is adopted to generate similar segment pairs by sampling non-overlapping segments $\mathbf{x}_m$ and $\mathbf{x}_n$ from the same speech.

(*ii*) **Representation disentangling**: As depicted in Fig. 5, SDCL consists of encoders, projectors, predictors and discriminators. First, the encoders consist of two LFBs (i.e., $\mathcal{E}_I$ and $\mathcal{E}_T$) and two STEs (i.e., $\mathcal{G}_I$ and $\mathcal{G}_T$), which share weights between two branches. Specifically, $\mathbf{x}_n, \mathbf{x}_m$ are fed into $\mathcal{E}_I$ and $\mathcal{E}_T$ to obtain features $\mathbf{X}_i^{sr}$ and $\mathbf{X}_i^{er}$, $i \in \{m, n\}$; the extractors, STE, explore SRs and ERs, instructing the learnable filters to focus on relevant regions to obtain $\mathbf{Z}_i^{er}$ and $\mathbf{Z}_i^{sr}$. The original and reconstructed representations $\mathbf{V}_i^{ori}$ and $\mathbf{V}_i^{rec}$ are achieved by concatenating $\mathbf{Z}_i^{er}$ and $\mathbf{Z}_i^{sr}$. The projector applies non-linear transformations to $\mathbf{V}_i^{rec}$ and $\mathbf{V}_i^{ori}$ to obtain $\mathbf{Q}_i^{rec}$ and $\mathbf{Q}_i^{ori}$. To enhance representation invariance, the predictor is utilized to ensure that one representation can be reconstructed and matched by another one extracted from distinct segments of the same speech, even after a non-linear perturbation, yielding $\mathbf{P}_i^{rec}$ and $\mathbf{P}_i^{ori}$. The negative cosine similarity $\mathcal{S}(\cdot)$ is chosen as the loss function to compute the similarity for segment pairs,

$$\begin{aligned} L_{\text{sim}} = &\frac{1}{2}(\mathcal{D}(\mathbf{P}_i^{ori}, sg(\mathbf{Q}_j^{ori})) + \mathcal{D}(\mathbf{P}_i^{rec}, sg(\mathbf{Q}_j^{rec}))) \\ &+ \frac{1}{2}(\mathcal{D}(\mathbf{P}_j^{ori}, sg(\mathbf{Q}_i^{rec})) + \mathcal{D}(\mathbf{P}_j^{rec}, sg(\mathbf{Q}_i^{ori}))), \end{aligned} \tag{6}$$

where $i, j \in \{m, n\}$, $i \neq j$, $sg(\cdot)$ denotes the stop-gradient.

(*iii*) **Loss function** To ensure successful disentanglement of the representations, several constraints are applied as follows:

$$L_{\text{total}} = L_{\text{sim}} + \alpha(L_{\text{diff}} + L_{\text{dis}}) + \beta L_{\text{lc}} + \gamma L_{\text{gl}}, \tag{7}$$

where α, β and γ are constraint terms that determine the contribution of each regularization to the overall loss.

1) The difference loss L_{diff} enforces the discrepancy of learned features to ensure that filters capture distinct regions,

$$L_{\text{diff}} = \|\mathbf{X}_i^{sr^\top} \cdot \mathbf{X}_i^{er}\|_{\mathcal{F}}^2, i \in \{m, n\}, \tag{8}$$

where $\|\cdot\|_{\mathcal{F}}^2$ is the squared Frobenius norm. Moreover, the discrimination loss L_{dis} guarantees the orthogonality of SRs and ERs as follows:

$$L_{\text{dis}} = \|\mathbf{Z}_i^{sr^\top} \cdot \mathbf{Z}_i^{er}\|_{\mathcal{F}}^2, i \in \{m, n\}, \tag{9}$$

2) The local similarity loss L_{lc} minimizes the discrepancy between ERs by aligning them in a shared subspace, utilizing the central moment discrepancy, as follows,

$$L_{\text{lc}} = \frac{\|\mathbb{E}(\mathbf{Z}_n^{er}) - \mathbb{E}(\mathbf{Z}_m^{er})\|_2}{\xi} + \sum_{o=2}^{\mathcal{O}} \frac{\|C_o(\mathbf{Z}_n^{er}) - C_o(\mathbf{Z}_m^{er})\|_2}{\xi^o}, \tag{10}$$

where ξ is an interval distance constant, $\mathbb{E}(\cdot)$ is empirical expectation, and $C_o(\cdot)$ is the o^{th} order sample central moment.

3) To ensure the global consistency of ERs across all samples, a discriminator is introduced to engage in a minimax adversarial game with LFB $\mathcal{E}_I$, which utilizes earth moving distance to reduce the discrepancy between ERs and the normal distribution $\mathbf{Z}^{\text{nor}}$. Under the Lipschitz continuity conditions,

$$L_{\text{gl}} = \mathbb{E}_{z \sim \mathbf{Z}_i^{er}}[\mathcal{D}(z)] - \mathbb{E}_{z^r \sim \mathbf{Z}^{\text{nor}}}[\mathcal{D}(z^r)], i \in \{m, n\}, \tag{11}$$

where $\mathcal{D}(\cdot)$ represents the discriminator function, and the loss added to generator LFB is $-\mathbb{E}_{z \sim \mathbf{Z}_i^{er}}[\mathcal{D}(z)]$.

3 Experiment

In this section, extensive experiments are conducted to evaluate the performance of the DSPA model on three datasets, aiming to investigate the following issues:

Q1. How does DSPA's performance compare to state-of-the-art methods?

Q2. Is the LFB module beneficial for SDD tasks?

Q3. What does the SDCL disentangled from speech?

3.1 Datasets

We perform extensive experiments on three publicly available datasets, DAIC-woz [10], CMDC [25] and EATD-corpus [20], with samples from each dataset as shown in Table 1. The DAIC-woz dataset comprises structured clinical interviews conducted by animated virtual interviewers targeting depression diagnosis and related psychological assessments.The CMDC dataset consists of semi-structured clinical interviews with predefined sets of questions designed to induce speech patterns related to depression.The EATD-corpus dataset contains self-reported

Table 1. Details of the three publicly available datasets.

Datasets	Criteria	Detail	Number	
			Normal control	Depression
DAIC-woz [10]	PHQ-8	Training set	77	30
		Development set	23	12
CMDC [25]	PHQ-9	Whole dataset	52	26
EATD-corpus [20]	SDS	Whole dataset	132	30

speech recordings across a wide range of affective states to present the speech features associated with depression.

Experimental setting: The component architectures are detailed in Table 2. All signals are resampled to 16 kHz, the segment length is $6s$, K=64, B=513, M=4, G=4, $D = 64$. STFT utilizes a 25 ms window and a 10 ms hop length. The cross-entropy and Adam optimizer are employed, with a batch size of 64 and an initial learning rate of $5e$-4. During pre-training, SDCL is trained on each dataset separately, and the learning rate set to $5e$-3. The hyperparameters α, β and γ are 1, 1 and 1, respectively. The SDCL module is trained on each dataset separately.

Table 2. Component structures, $\Rightarrow$ indicates connection direction.

Component	Layers
Projector	FC $\Rightarrow$ (BN+ReLU) $\Rightarrow$ FC $\Rightarrow$ (BN+ReLU) $\Rightarrow$ FC $\Rightarrow$ BN
Predictor	FC $\Rightarrow$ (BN+ReLU) $\Rightarrow$ FC
Discriminator	SN(FC) $\Rightarrow$ ReLU $\Rightarrow$ SN(FC) $\Rightarrow$ ReLU $\Rightarrow$ SN(FC)
Extractor	STE $\Rightarrow$ STE $\Rightarrow$ STE

* BN, batch normalization, SN, spectral normalization.

3.2 Comparison with the state-of-the-art methods (Q1)

To demonstrate the effectiveness of our DSPA model, it is compared with state-of-the-art methods (audio modality). The same experiment settings as the competing approaches are used to ensure competitiveness. Specifically, the results of the DAIC-woz dataset are tested on the development datasets. A 5-fold cross-validation is conducted on the CMDC datasets, and a 3-fold cross-validation is performed on the EATD-corpus datasets.

Table 3 shows that our method consistently achieves better F1 score than the comparison methods. 1) Particularly, the shallow methods (e.g., logistic, random forest (RF) and support vector machine (SVM)), perform poorly on all the datasets. 2) Similarly, our model outperforms the methods with Mel-spectral features. These results are consistent with our initial viewpoint that handcrafted spectral features underperform in comparison to learnable ones. 3) Experimental findings illustrate the effectiveness of our approach in leveraging the latent

Table 3. Comparison with the state-of-the-art methods, LLDs denote low-level descriptors: MFCCs, COMPARE and eGeMAPS.

Dataset	Methods	Inputs	F1 score	Precision	Recall
DAIC-woz	SVM [10]	LLDs	0.400	0.330	0.500
	DepAudioNet [16]	MFbanks	0.610	0.625	0.770
	DEPA(frame) [24]	STFT	0.640	0.640	0.640
	Mfcc-LSTM [18]	MFCCs	0.655	0.735	0.645
	STFN [11]	Signals	0.760	0.650	0.920
	NetVlad-GRU[20]	MFbanks	0.770	0.630	**1.000**
	DALF[22]	Signals	0.784	0.782	0.794
	DSPA	STFT	**0.792**	**0.785**	0.798
CMDC	SVM [25]	LLDs	0.910	0.920	0.910
	Logistic [25]	LLDs	0.840	0.850	0.840
	Naïve Bayes [25]	LLDs	0.890	0.890	0.890
	Bi-LSTM [13]	LLDs	0.910	**1.000**	0.830
	IIFDD [6]	LLDs	0.920	0.960	0.890
	DSPA	STFT	**0.927**	0.935	**0.920**
EATD-corpus	RF [20]	LLDs	0.500	0.480	0.530
	SVM [20]	LLDs	0.460	0.540	0.410
	NetVlad-GRU [20]	MFbanks	0.660	0.570	**0.780**
	Mm-LSTM [2]	MFbanks	0.490	0.440	0.560
	ABAFnet[21]	MFbanks	0.694	0.684	0.690
	DSPA	STFT	**0.702**	**0.681**	0.735

information hidden in the STFT spectrum, yielding high quality disease-related features. In contrast, other methods, except DALF, depend heavily on prior knowledge or intricate feature engineering, resulting in a limited capacity to extract useful representations. 4) The parameters of Mfcc-LSTM, DALF, DepAudioNet, Mm-lstm, NetVlad-GRU are 0.42 million (M), 1.89M, 0.55M, 1.10M, 0.92M, respectively. Our framework's parameters are 1.63M, which achieves a significant performance improvement (>5% on average) with a slight increase in computation cost. It demonstrates that our method achieves a balance between computational cost and performance.

3.3 Ablation studies (Q2)

The same STE extractor and classifier (FC) are employed for ablation studies to investigate the effectiveness of LFB and SDCL. Specifically, to evaluate the quality of features, pre-training of LFB is conducted to identify the region of interest in the STFT spectrum via SDCL, thus yielding LFB output features.

The results in Table 4 show that 1) on the DAIC-woz datasets, LFB improves the model's performance by 5.7% and 6.4% when compared to the MFbanks and MFCCs, respectively, which indicates that the LFB learned features are high-quality for the SDD tasks. 2) Due to the task-oriented adaptability of the learnable LFB, improved F1 scores are consistently achieved whenever it is utilized. 3) It demonstrates that distinct parameter initialization methods offer

Table 4. Ablation study on different inputs and initializations.

Input features	F1 score		
	DAIC-woz	**CMDC**	**EATD-corpus**
MFCCs	0.728	0.891	0.636
MFbanks	0.735	0.902	0.623
LFB **w/** Mel	0.764	0.916	0.685
LFB **w/** U-Hz	0.757	0.908	0.664
LFB **w/** ERB	0.747	0.904	0.671
LFB **w/** (SDCL **w/o** L_{diff})	0.740	0.894	0.633
LFB **w/** (SDCL **w/o** L_{lc})	0.754	0.905	0.668
LFB **w/** (SDCL **w/o** L_{gl})	0.771	0.903	0.672
LFB **w/** SDCL	**0.792**	**0.927**	**0.705**

different known priors for filters, and the self-supervised training of LFB via the SDCL strategy can effectively disentangle significant ERs, leading to superior prior guidance for filters. 4) Collaborative learning under multiple losses of L_{dis}, L_{diff}, L_{lc} and L_{gl} is essential for disintegrating speech representations.

3.4 Discussion

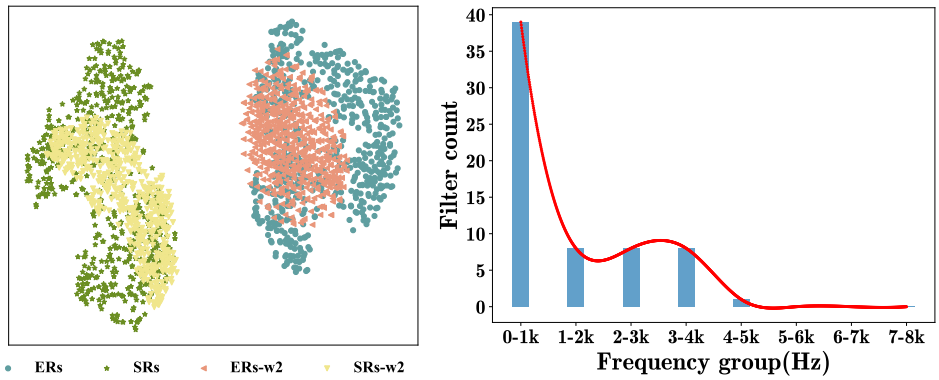
Several discussions are conducted to further analyze our method's performance. Additionally, the parameters of LFB are analyzed to understand what its focus.

A. How does the generalization ability of LFB? As depicted in Table 5, we investigate the generalization ability of LFB. The parameters of LFB are first optimized on one dataset, and then fixed or fine-tuned the LFB on another dataset to complete the classification. The accuracy decreases by 2.3% and 1.5% when using fixed LFB parameters on two datasets, respectively. However, the performance is improved after fine-tuning the LFB module. This is due to the inconsistent data distribution caused by the different corpus and questionnaire, but highlights the adaptability of LFB to variations in different scenarios.

Table 5. The generalization ability of LFB.

Datasets		**Module**		**F1 score**
Source	Target	LFB	Others	
CMDC	EATD-corpus	F	T	0.633
CMDC	EATD-corpus	T	T	0.687
EATD-corpus	CMDC	F	T	0.904
EATD-corpus	CMDC	T	T	0.912

* F indicates Fixed and T indicates fine-Tuning.



(a) Visualization of SRs and ERs. (b) Filter distribution in ERs.

Fig. 6. Analysis of the disentangled representations.

B. What does the SDCL disentangled from speech? (Q3) Building upon the effectiveness of SDCL, this part aims to prove that the ERs disentangled by SDCL are emotion-related. We select the Wav2Vec2.0 models that are fine-tuned in the speech emotion recognition and speaker identification tasks, respectively [5][19]. Both the Wav2Vec2.0 models and the pre-trained DSPA are used for representation extraction on the same test datasets of CMDC datasets. Finally, a t-SNE distributional analysis of ERs and SRs is conducted to investigate whether representations with distinct semantics are bounded.

As shown in Fig. 6 (a), representations with identical semantics are clustered in the overlapping space, while ERs and SRs are clearly separated. It is believe that when tasks are identical, the extracted representations lie in the same distribution, regardless of differences in network structure. These findings confirm that SDCL adheres to the pretext task during pre-training.

C. What do the learnable filters focus on? (Q3) A distributional analysis of ERs-focused filters is conducted to investigate what these filters learned. As shown in Fig. 6 (b), it can find that the filters predominantly cluster within the 0-1 kHz range. These findings are consistent with the observation that depression exhibits a correlation with the $F0$ and formant frequencies $F1$ [1]. Importantly, the first formant of the vowel phonemes /**e**/, /**i**/, /**o**/ and /**u**/ are in this frequency range. Based on our findings, a new text-corpus will be designed that contains more words constituted by mentioned vowel phonemes. Furthermore, these findings inspire us to construct precise spectral features by removing task-irrelevant regions.

4 Conclusion

In this paper, a depression speech analysis network (DSPA) is designed, which includes a specially designed learnable feature extraction component, LFB, tailored for SDD tasks, and the STE, a follow-on module for identify representations of interest in LFB learned features. With a well-designed self-supervised strategy, our SDCL strategy harnesses unlabeled data to disentangle robust emotional representations, providing a novel insight on initializing learnable filterbanks. Future work will be devoted to quantifying the features in the spectrum to enhance the interpretability of speech representations.

5 Acknowledge

This research was supported by the National Natural Science Foundation of China (No.62076059) , the Science and Technology Joint Project of Liaoning province (2023JH2/101700367) and the Fundamental Research Funds for the Central Universities (No. N2424010-7). Osmar Zaiane gratefully acknowledges the funding from Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canada CIFAR AI Chairs Program for Amii.

References

1. Afshan, A., Guo, J., Park, S.J., Ravi, V., Flint, J., Alwan, A.: Effectiveness of voice quality features in detecting depression. Interspeech 2018 (2018)
2. Al Hanai, T., Ghassemi, M.M., Glass, J.R.: Detecting depression with audio/text sequence modeling of interviews. In: Interspeech. pp. 1716–1720 (2018)
3. Altwaijri, Y.A., Al-Subaie, A.S., Al-Habeeb, A., Bilal, L., Al-Desouki, M., Aradati, M., King, A.J., Sampson, N.A., Kessler, R.C.: Lifetime prevalence and age-of-onset distributions of mental disorders in the saudi national mental health survey. International Journal of Methods in Psychiatric Research **29**(3), e1836 (2020)
4. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
5. Calabrés, E.H.: wav2vec2-lg-xlsr-en-speech-emotion-recognition (2024). https://doi.org/10.57967/hf/2045, https://huggingface.co/ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition
6. Chen, J., Hu, Y., Lai, Q., Wang, W., Chen, J., Liu, H., Srivastava, G., Bashir, A.K., Hu, X.: Iifdd: Intra and inter-modal fusion for depression detection with multi-modal information from internet of medical things. Information Fusion **102**, 102017 (2024)
7. Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., Dubnov, S.: Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 646–650. IEEE (2022)
8. Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection. Neural Networks **18**(4), 407–422 (2005)

9. Feng, K., Chaspari, T.: A knowledge-driven vowel-based approach of depression classification from speech using data augmentation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
10. Gratch, J., Artstein, R., Lucas, G.M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D.R., Rizzo, A.A., Morency, L.P.: The distress analysis interview corpus of human and computer interviews. In: LREC (2014)
11. Han, Z., Shang, Y., Shao, Z., Liu, J., Guo, G., Liu, T., Ding, H., Hu, Q.: Spatial-temporal feature network for speech-based depression recognition. IEEE Transactions on Cognitive and Developmental Systems (2023)
12. He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., Guo, C., Wang, H., Ding, S., Wang, Z., et al.: Deep learning for depression recognition with audiovisual cues: A review. Information Fusion **80**, 56–86 (2022)
13. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
14. Li, J., Tian, Y., Lee, T.: Learnable frequency filters for speech feature extraction in speaker verification. arXiv preprint arXiv:2206.07563 (2022)
15. López-Espejo, I., Tan, Z.H., Jensen, J.: Exploring filterbank learning for keyword spotting. In: 2020 28th European Signal Processing Conference (EUSIPCO). pp. 331–335. IEEE (2021)
16. Ma, X., Yang, H., Chen, Q., Huang, D., Wang, Y.: Depaudionet: An efficient deep model for audio based depression classification. In: Proceedings of the 6th international workshop on audio/visual emotion challenge. pp. 35–42 (2016)
17. Ravanelli, M., Bengio, Y.: Speaker recognition from raw waveform with sincnet. In: 2018 IEEE Spoken Language Technology Workshop (SLT). pp. 1021–1028. IEEE (2018)
18. 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**, 103107 (2022)
19. Saire2023: wav2vec2-base-finetuned-speaker-classification (2023), https://hf-mirror.com/Saire2023/wav2vec2-base-finetuned-Speaker-Classification
20. Shen, Y., Yang, H., Lin, L.: Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6247–6251. IEEE (2022)
21. Xu, X., Wang, Y., Wei, X., Wang, F., Zhang, X.: Attention-based acoustic feature fusion network for depression detection. Neurocomputing p. 128209 (2024)
22. Yang, W., Liu, J., Cao, P., Zhu, R., Wang, Y., Liu, J.K., Wang, F., Zhang, X.: Attention guided learnable time-domain filterbanks for speech depression detection. Neural Networks (2023)
23. Zeghidour, N., Teboul, O., Quitry, F.d.C., Tagliasacchi, M.: Leaf: A learnable frontend for audio classification. arXiv preprint arXiv:2101.08596 (2021)
24. Zhang, P., Wu, M., Dinkel, H.: Depa: Self-supervised audio embedding for depression detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 135–143 (2021)
25. Zou, B., Han, J., Wang, Y., Liu, R., Zhao, S., Feng, L., Lyu, X., Ma, H.: Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. IEEE Transactions on Affective Computing (2022)