Neural Networks 165 (2023) 135–149

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Attention guided learnable time-domain filterbanks for speech depression detection

Wenju Yang [a,b], Jiankang Liu [a,b], Peng Cao [a,b,*], Rongxin Zhu [d], Yang Wang [d], Jian K. Liu [e], Fei Wang [d,**], Xizhe Zhang [c,***]

[a] College of Computer Science and Engineering, Northeastern University, Shenyang, 110819, Liaoning, China
[b] Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang, 110819, Liaoning, China
[c] School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, 211166, China
[d] Early Intervention Unit, Department of Psychiatry, Affiliated Nanjing Brain Hospital, Nanjing Medical University, Nanjing, 210096, China
[e] School of Computing, University of Leeds, Leeds, LS2 9JT, United Kingdom

ARTICLE INFO

Article history:
Received 25 December 2022
Received in revised form 13 May 2023
Accepted 20 May 2023
Available online 26 May 2023

Keywords:
Speech depression detection
Filterbanks
Time–frequency analysis
Interpretability
Affective computing

ABSTRACT

Depression, as a global mental health problem, is lacking effective screening methods that can help with early detection and treatment. This paper aims to facilitate the large-scale screening of depression by focusing on the speech depression detection (SDD) task. Currently, direct modeling on the raw signal yields a large number of parameters, and the existing deep learning-based SDD models mainly use the fixed Mel-scale spectral features as input. However, these features are not designed for depression detection, and the manual settings limit the exploration of fine-grained feature representations. In this paper, we learn the effective representations of the raw signals from an interpretable perspective. Specifically, we present a joint learning framework with attention-guided learnable time-domain filterbanks for depression classification (DALF), which collaborates with the depression filterbanks features learning (DFBL) module and multi-scale spectral attention learning (MSSA) module. DFBL is capable of producing biologically meaningful acoustic features by employing learnable time-domain filters, and MSSA is used to guide the learnable filters to better retain the useful frequency sub-bands. We collect a new dataset, the Neutral Reading-based Audio Corpus (NRAC), to facilitate the research in depression analysis, and we evaluate the performance of DALF on the NRAC and the public DAIC-woz datasets. The experimental results demonstrate that our method outperforms the state-of-the-art SDD methods with an F1 of 78.4% on the DAIC-woz dataset. In particular, DALF achieves F1 scores of 87.3% and 81.7% on two parts of the NRAC dataset. By analyzing the filter coefficients, we find that the most important frequency range identified by our method is 600–700Hz, which corresponds to the Mandarin vowels /***e***/ and /***ê***/ and can be considered as an effective biomarker for the SDD task. Taken together, our DALF model provides a promising approach to depression detection.

1. Introduction

Depression is a worldwide prevalent psychiatric condition and a primary contributor to the global burden of disease in young people (Gore et al., 2011). Depression often initially occurs during adolescence (Kessler, Avenevoli, & Merikangas, 2001), may continue or recur in adulthood (Lewinsohn, Rohde, Seeley, Klein, & Gotlib, 2003), and tends to become a lifelong chronic psychiatric disorder (Altwaijri et al., 2020). Nowadays, the methods of screening and detection of depression mainly rely on questionnaires and interviews supplemented by the clinical assessment of psychiatrists, such as the patient health questionnaire (PHQ) (Kroenke & Spitzer, 2002) and Hamilton depression scale (HAMD) (Hamilton, 1986). However, these assessments are easily biased by the experience of the interviewer, the quality of the question protocol, and the willingness of the patients (Cummins et al., 2015). Accordingly, an objective and convenient approach to depression detection is desirable. Speech is a non-invasive and easily accessible biological information, different emotional states are characterized by different acoustic features (Devillers, Vidrascu, & Lamel, 2005). Previous researches report that the speech of depressed patients typically exhibits whispering, monotonous, slow, slurred, prolonged pauses and

* Correspondence to: NO. 3-11, Wenhua Road, Heping District, Shenyang, China.
** Correspondence to: NO. 264, Guangzhou Street, Gulou District, Nanjing, China.
*** Correspondence to: NO. 101, Longmian Avenue, Jiangning District, Nanjing, China.
E-mail addresses: fei.wang@yale.edu (F. Wang), caopeng@cse.neu.edu.cn (P. Cao), zhangxizhe@njmu.edu.cn (X. Zhang).

https://doi.org/10.1016/j.neunet.2023.05.041
0893-6080/

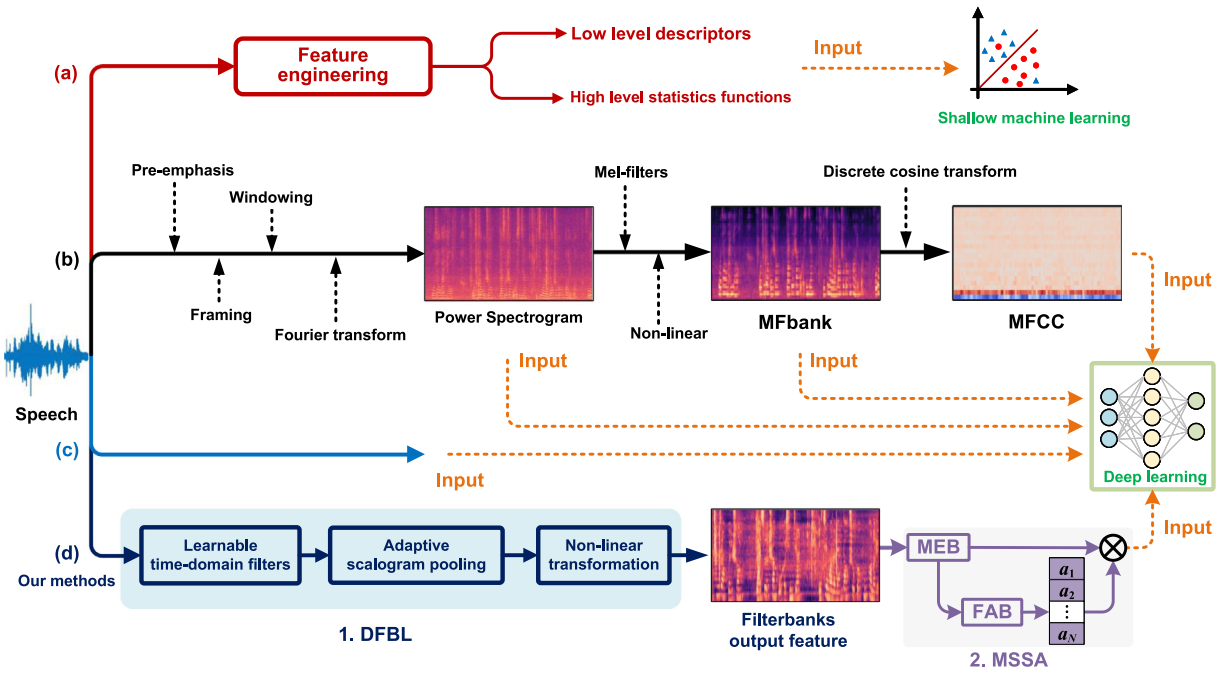


Fig. 1. Diagram of different SDD approaches.

stuttering (Williamson et al., 2016). It is essential to design an efficient speech depression detection (SDD) model that explores acoustic biomarkers crucial for differentiation and leverages them for depression analysis.

As shown in Fig. 1(a), in the early stage of the SDD research, the main efforts are to extract the low-level descriptors (LLDs), as well as their statistic feature sets by OpenSMILE (Eyben, Wöllmer, & Schuller, 2010) and COVAREP (Degottex, Kane, Drugman, Raitio, & Scherer, 2014) toolkits. For instance, the voice quality features including formants, jitter, and shimmer are considered to be most relevant to the SDD task (Ozdas, Shiavi, Silverman, Silverman, & Wilkes, 2004; Williamson et al., 2019). The prosodic features, such as loudness range, fundamental frequency (F0) and statistics features extracted from F0, have been proven to be associated with depression severity (France, Shiavi, Silverman, Silverman, & Wilkes, 2000; Jiang et al., 2018). Moreover, the Mel-frequency cepstral coefficients (MFCC) have proven their effectiveness in detecting depression compared to other cepstral features (Low, Maddage, Lech, & Allen, 2009). Given these extracted features, machine learning (ML), such as support vector machine (SVM) and random forests (RF), is employed to train predictive models (Valstar et al., 2014). However, these handcrafted features suffer from requirements of substantial prior knowledge and cannot provide more discriminative patterns hidden in the signals (Morales, 2018).

As shown in Fig. 1(b)–(c), researches on SDD using deep learning (DL) models have been witnessed a growing trend in recent years, which can be categorized into two groups: (1) Constructing DL models on the extracted low-level spectral features. (2) Constructing models directly on the raw speech signal. Approaches in the first category typically utilize spectral features as model inputs. As shown in Fig. 1(b), the power spectrograms (Shen, Yang, & Lin, 2022) are calculated by performing the short-time Fourier transform (STFT) on the signals. Then Mel-filterbanks (MFbanks) features (Zhang, Wu, Dinkel, & Yu, 2021) are obtained by applying the pre-defined Mel-filters and non-linear compression to the STFT spectrums. Lastly, the MFbanks features are converted into the Mel-frequency cepstral coefficients (MFCC) (Rejaibi, Komaty, Meriaudeau, Agrebi, & Othmani, 2022) after taking the discrete cosine transform operation. Each of these features can be considered individually as representative features of the SDD task, and can fed into the DL models, such as convolutional neural networks (CNNs) (Vázquez-Romero & Gallardo-Antolín, 2020) and long and short-term memory networks (LSTM) (Wei et al., 2022), to explore representative local patterns (He et al., 2022). The second category (Fig. 1(c)) involves modeling the speech signals directly without additional pre-processing of the data (Rejaibi et al., 2022; Zhang et al., 2021). In this case, it usually requires larger convolutional kernels and deeper model architectures to handle the high-dimensional data. Notably, previous works, including Sinc filter (Ravanelli & Bengio, 2018), Wavelet filter (Khan & Yener, 2018), Gabor filter (Zeghidour, Teboul, Quitry, & Tagliasacchi, 2021), and Gammatone filter (López-Espejo, Tan, & Jensen, 2021), focus on designing shape-specific learnable filters to replace the standard convolution kernels for signal decomposition. These learnable filters can extract relevant frequency features from the signals for various audio task. For example, learnable Gabor filters have been introduced to approximate the Mel-frequency spectral coefficients for improving the accuracy of the phone recognition task (Zeghidour, Usunier, Kokkinos, et al., 2018). Triangular and bell-shaped learnable filters are trained on the spectrograms and optimized specifically for the speaker verification task (Li, Tian, & Lee, 2022). Wavelet filters have been employed for spectral decomposition in automatic speech recognition task to eliminate the need for a large number of hyper-parameters during speech processing. Moreover, recent studies (Fu, Teng, White, Powell, & Schmidt, 2022) have focused on constraining the shape of frequency-domain filters that are multiplied with the spectrogram to obtain the learnable filterbanks for spoof speech. Unfortunately, no one has yet proposed such an approach towards the SDD task.

Despite the recent advances obtained by both category approaches, these DL methods for SDD still face several challenges. (1) Regarding feature learning, direct modeling on the raw signal without constraints yields a large number of parameters, which prevents the model from effectively exploiting the discriminable features at a fine-grained level (Tukuljac, Ricaud, Aspert, & Colbois, 2022). (2) The existing MFbanks and MFCC features are

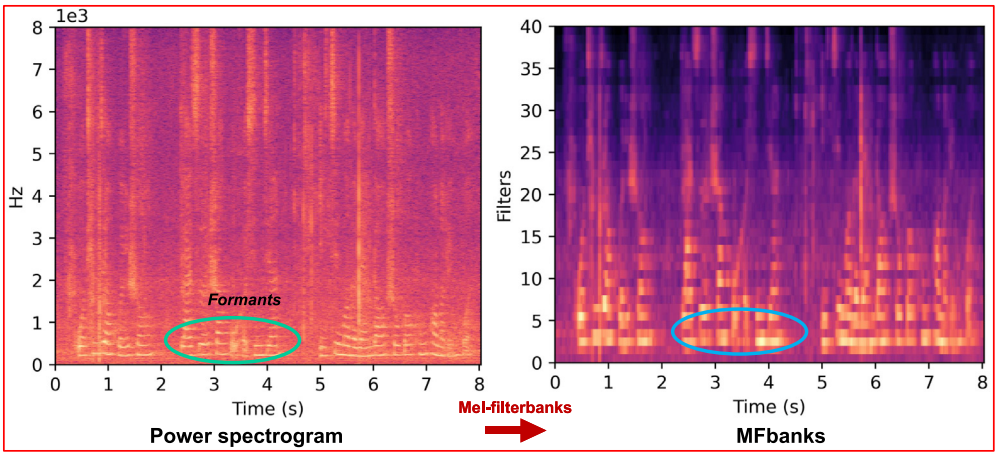


Fig. 2. Example of speech pre-processing results before and after Mel-filterbanks. The formants in the spectrogram, i.e., the high-energy frequency peaks, are blurred during the calculation of the MFbanks.

designed to match the human perceptual system and contain the rich prior factors relevant to the speech recognition task. Nevertheless, these universal acoustic features are not tailored for assessing depression disorders, thereby leading to being inappropriate for the SDD task. For example, as depicted in Fig. 2, the Mel-filterbanks hinder the extraction of crucial narrow-band features in speech, such as formants that are relevant to depression (Cummins et al., 2015; Ravanelli & Bengio, 2018). (3) Depression can cause changes in the vocal tract, shifting the speech energy from 500 Hz to 500–1000 Hz sub-bands as depression severity increases (Moore, Clements, Peifer, & Weisser, 2008; Ozdas et al., 2004). However, standard Mel-features typically employ fixed hyper-parameters, such as the filter numbers and the SFFT sampling points, which are insufficient for capturing the fine-grained frequency information relevant to SDD task (Li et al., 2022).

Driven by these important issues, we develop an end-to-end learning framework with Attention guided Learnable time-domain Filterbanks for Depression classification, called DALF. Our model, as shown in Fig. 1(d), consists of three components: a depression filterbanks features learning (DFBL) module, a multi-scale spectral attention learning (MSSA) module and a speech classification module.

(1) The DFBL module is designed to generate task-relevant spectral features, which involves three components: learnable time-domain filtering, adaptive scalogram pooling and parameterized nonlinear transformation. First, by utilizing time-domain band-pass filters with a small number of parameters, the shape-specific convolution kernels are capable of producing more discriminative acoustic features by selectively preserving critical frequencies. Additionally, the adaptive scalogram pooling can reduce the temporal resolution of the time-domain filtering components through fixed-stride depth-wise separable convolutions (DSC), while simultaneously smoothing the corresponding frequency scalograms of the specific filters. Finally, the parameterized nonlinear transformation compresses the dynamic ranges in the spectral features while preserving the important local patterns.

(2) The MSSA module is proposed to effectively leverage the multi-level resolution information and important channel local patterns from the output features of the DFBL module. It consists of two main blocks: the multi-scale features extraction block (MEB) and the frequency-aware attention block (FAB). On the one hand, MEB explores the multi-scale frequency information and captures long-range dependence along the filter axis. On the other hand, FAB captures the variable behaviors of the filters by integrating multi-scale features and learning the importance weight of each filter. It facilitates the concentration of important frequency sub-bands in depressed speech, as different frequency ranges contribute differently to the SDD task.

To the best of our knowledge, this work represents the first attempt to address the aforementioned challenges in SDD task. Our contributions are summarized as follows:

1. We propose a depression feature learning module, DFBL, which adaptively constructs unbiased filterbanks output features through a learnable approach with supervised learning scheme. Unlike traditional Mel-spectral features, our learnable filterbanks can automatically modulate the corresponding frequency regions that are relevant to the classification task. By analyzing the cumulative frequency response of the learnable filters, we can quantify the frequency regions of interest, enhancing the interpretability of our model. Furthermore, the DFBL module can easily replace the Mel-spectral features in existing deep SDD frameworks.
2. We develop an MSSA module that can re-calibrate channel feature responses. The MEB explores the different scale frequency features on each filter using dilated convolutions. Meanwhile, the FAB captures the dynamic behavior of the filters through multi-scale convolution. By stacking multiple MSSA layers, our model extracts critical information from fused multi-scale features and guides the filters to learn meaningful frequency sub-bands.
3. We collect a Chinese depression dataset called the Neutral Reading-based Audio Corpus (NRAC), which facilitates research in depression detection. Moreover, we thoroughly evaluate the DALF framework on the public dataset Distress Analysis Interview Corpus, Wizard of Oz (DAIC-woz) and NRAC dataset. Extensive experimental results demonstrate that our method not only outperforms state-of-the-art approaches in SDD task, but also excels at automatically identifying depression-related biomarkers in human speech.

2. Preliminaries

2.1. Problem definition

Mathematically, in our research, $\boldsymbol{x}_i \in \mathbb{R}^{L^{(i)}}, i \in \{1, 2, \ldots, I\}$ represents the raw speech of the ith subject with a length $L^{(i)}$, where I is the total number of subjects. The sliding window

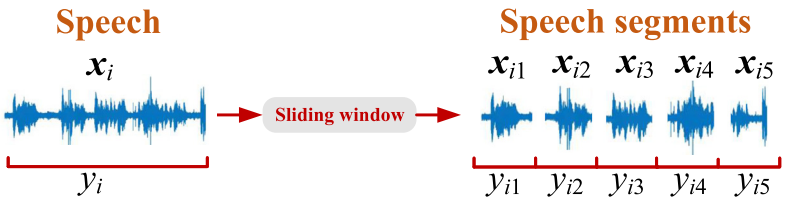


Fig. 3. Non-overlapping sliding window schematic.

method without overlapping, as shown in Fig. 3, is utilized to segment the $\boldsymbol{x}_i$ into a set $\{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ij}, \ldots, \boldsymbol{x}_{iJ}\}$. It is important to note that all segment labels $\{y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{iJ}\}$ of a specific subject are the same as the raw speech label $y_i \in \{0, 1\}$, J denotes the total number of segments. The length of $\boldsymbol{x}_{ij}$ is $l_{ij} < L^{(i)}$, and $y_{ij} \in \{0, 1\}$ represents the depression states for jth segment. Specifically, we define the SDD task into three stages: (1) The input $\boldsymbol{x}_{ij} \in \mathbb{R}^{1\times l_{ij}}$ is mapped into a two-dimensional feature $\boldsymbol{X}_{ij}$ by using the learnable time-domain filterbanks, pooling and nonlinear transformation $F_\psi\colon \mathbb{R}^{1\times l_{ij}} \to \mathbb{R}^{N\times M}$, where M denotes the filterbanks output feature length, N is the total number of filters, and ψ are the learnable parameters. (2) Then, the features metric $\boldsymbol{X}_{ij}$ will be tuned by the spectral attention learning module $\mathcal{G}_\theta(\cdot)$ with the parameters of θ. (3) Finally, both the parameters ψ and θ are optimized jointly with the classifier C_η with learnable parameters η to solve the supervised classification problems, which is written as:

$$\hat{\psi}, \hat{\theta}, \hat{\eta} = \arg\min_{\theta,\psi,\eta} \boldsymbol{E}_{(\boldsymbol{x},y)\in\mathbb{D}} \mathcal{L}\left(C_\eta(\mathcal{G}_\theta(F_\psi(\boldsymbol{x}_{ij}))), y_{ij}\right), \tag{1}$$

where $\mathcal{L}$ is the loss function, and $\mathbb{D}$ is the dataset. Our work aims to learn a better solution for the parameters ψ, θ and η via back-propagation.

3. Method

Our work attempts to learn representative biomarkers directly from raw speech signals, while providing a lightweight front-end for speech detection. In this section, we first describe the overview of the proposed framework. Then, we introduce the details of each module. Appendix A provides a list of the key symbols used in the paper along with their corresponding definitions.

3.1. Overview of the framework

We propose a framework that utilizes a depression feature learning module to directly process raw speech signals, while focusing on exploring the critical frequency regions in an end-to-end manner. The schematic of the DALF framework is shown in Fig. 4. It consists of a depression filterbanks features learning module, a multi-scale spectral attention module and a classification module.

3.2. Depression filterbank features learning module

According to the fact that speech signals are presented by blending time and frequency information together, we adopt our module to capture meaningful features from these two fundamental aspects. Specifically, as shown in Fig. 4(a), we propose a depression filterbanks features learning module (DFBL) to replace the fixed Mel-spectral features, DFBL is a task-oriented module that separates the input signal into multiple frequency sub-bands and contains three components:

(1) To accurately capture depression-related frequency components in speech, the learnable wavelet filters are proposed to perform time-domain filtering on the raw signals.

(2) To decrease the temporal resolution for each filter, an adaptive scalogram pooling is performed to downsample the outputs produced by the time-domain filters.

(3) To compress the dynamic range of the learned features while maintaining the local patterns, a parameterized nonlinear transformation is applied to activate the outputs from the learned filterbanks.

3.2.1. Time-domain filtering

The study (Sainath, Weiss, Wilson, Senior, & Vinyals, 2015) initially aimed to replicate the Mel-filterbanks structure using learnable gammatone filters. Similarly, the study (Zeghidour, Usunier, Kokkinos, et al., 2018) utilized the learnable Gabor filters to approximate the frequency response of the Mel-triangular filters. Both works focus on employing specific-shape filters to simulate the Mel-filterbanks, where filters deviate from their initial triangular form, but close to the Mel filterbanks, rather than getting rid of the Mel-prior limitations. Recently, research (Pu, Panagakis, & Pantic, 2021) demonstrated that tuning the location and shape of learnable wavelet filters is more effective for natural speech discrimination. Nevertheless, no systematic research on SDD task has been performed on learnable filterbanks in either time or frequency domain. Motivated by these observations, we employ a set of flexible time-domain (TD) filters to perform wavelet analysis on the original signals. Each filter can be tuned in shape and location by a finite number of trainable parameters. This adaptability enables the filters to automatically explore the critical frequency information hidden in the speech. The filter output is derived by convolving the impulse responses $\varphi_n \in \mathbb{C}^\kappa$ with the input segment waveform $\boldsymbol{x}_{ij} \in \mathbb{R}^{1\times l_{ij}}$ according to the Eq. (2),

$$\boldsymbol{X}_{ij}[n, :] = \boldsymbol{x}_{ij} * \varphi_n, \quad n = 1, \ldots, N, \tag{2}$$

where $*$ indicates the convolution operation, n is the index of the filter, $\boldsymbol{X}_{ij} \in \mathbb{R}^{N\times V}$ is a two-dimensional time–frequency feature of the jth segment, $V = l_{ij} - \kappa + 1$, κ is the filter length, and the horizontal and vertical axes are the time and the filter number, respectively.

Specifically, we choose complex Gabor filter (Noé, Parcollet, & Morchid, 2020; Zeghidour, Usunier, Kokkinos, et al., 2018) to better constrain the kernel coefficients so that the frequency response of the filter exhibits a regular shape within the desired frequency range, and we will discuss other learnable filters in the experiment. The impulse response φ_n of the Gabor filter is defined as the multiplication of a Gaussian kernel with a complex sinusoid,

$$\varphi_n(t) = \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{t^2}{2\sigma_n^2}} \cdot e^{i2\pi c_n t}, \quad n = 1, \ldots, N, \tag{3}$$

where t is the time step, and $t \in [-\frac{\kappa}{2}, \frac{\kappa}{2}]$, σ_n is the standard deviation, and c_n is the center frequency. In our work, N=64, κ=401, $c_n \in [0, 0.5]$, the c_n in Hertz is obtained by multiplying it with the samplerate 16000 Hz.

Moreover, as shown in Fig. 5(a), the complex Gabor filters are typically considered to be two out-of-phase filters that operate on the real and imaginary part of the complex function, respectively.

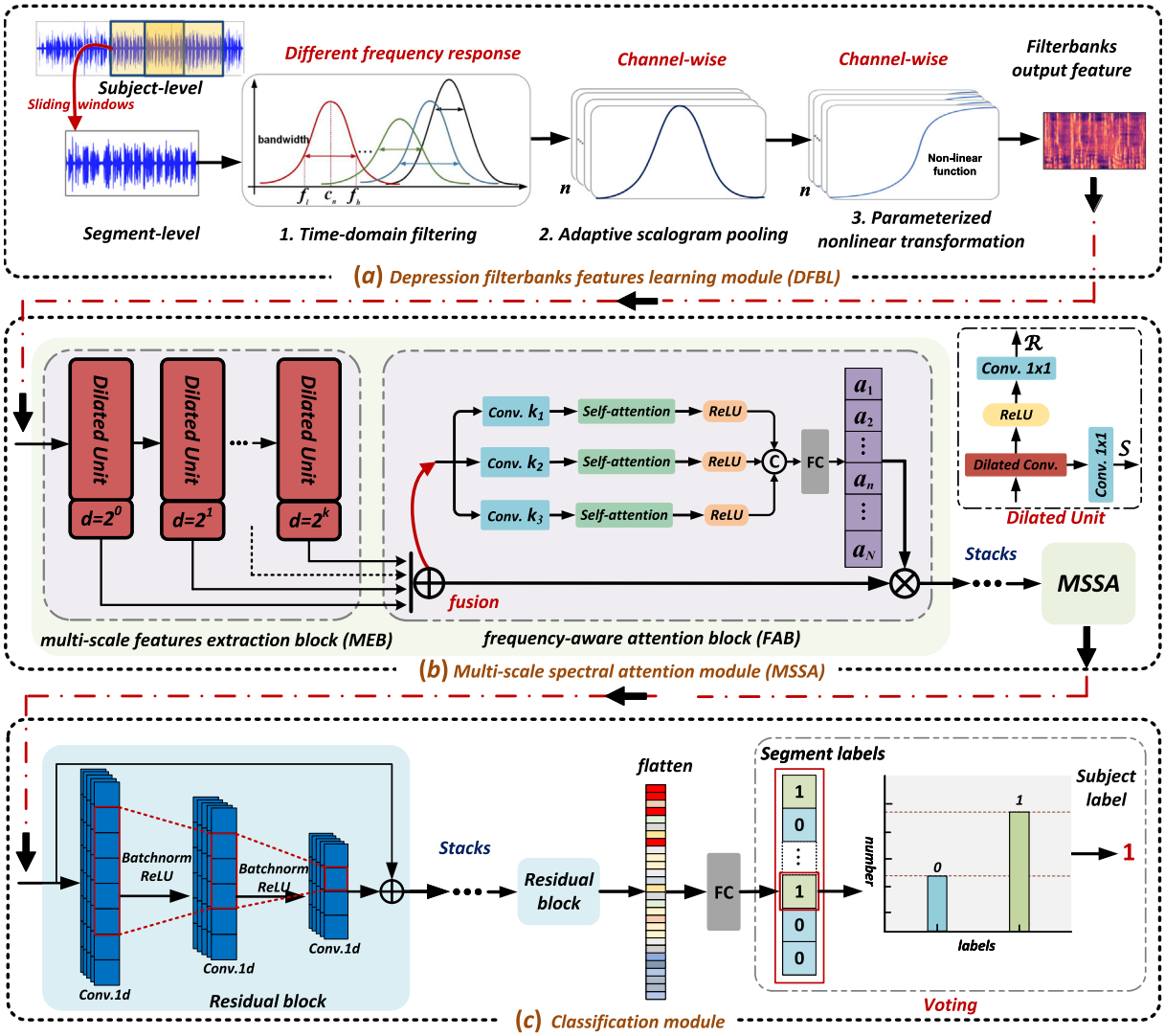


Fig. 4. The overall framework of the proposed approach consists of three stages: (1) A set of learnable time-domain filters is trained to acquire the time–frequency features for the original speech signals, and obtain the learned filterbanks features through the following scalogram pooling and nonlinear transformation. (2) Then, the MSSA module is proposed to focus on the extraction of discriminative acoustic features by both the multi-scale features extraction block (MEB) and frequency-aware attention block (FAB). (3) The classification module takes the learned spectral features as inputs and produces the outputs of the segment-level labels. Finally, the subject-level labels are determined by a majority voting of the segment-level predictions.

More specifically, the nth Gabor filter φ_n consists of a real $g_n^e(t)$ and an imaginary $g_n^o(t)$ part,

$$\varphi_n(t) = g_n^e(t) + i g_n^o(t), \tag{4}$$

and the $g_n^e(t)$ and $g_n^o(t)$ hold the filter as follows:

$$\begin{aligned} g_n^e(t) &= \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{t^2}{2\sigma_n^2}} \cdot \cos(2\pi c_n t), \\ g_n^o(t) &= \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{t^2}{2\sigma_n^2}} \cdot \sin(2\pi c_n t), \end{aligned} \tag{5}$$

we use both parts of the complex Gabor filters to decouple the speech signals. The two out-of-phase filters are helpful for instantaneously estimating frequency information and preserving the phase information (e.g. formants), which is critical for the SDD task (Noé et al., 2020). Additionally, the unmodulated positive responses of the Gabor filter can be obtained by calculating the square of the real and imaginary parts outputs, respectively. In our work, the squared ℓ_2-pooling is performed along the filter-wise axis to convert outputs into the magnitude (energy) features,

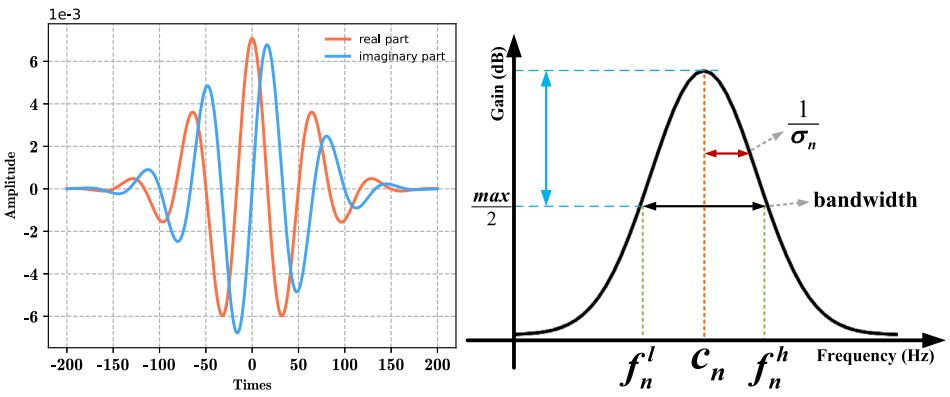
$$\boldsymbol{X}_{ij}[n,:] = \frac{|\boldsymbol{x}_{ij} * g_n^e|^2 + |\boldsymbol{x}_{ij} * g_n^o|^2}{2}, \quad n = 1, \ldots, N, \tag{6}$$

where matrix $\boldsymbol{X}_{ij}$ is considered as the scalogram of $\boldsymbol{x}_{ij}$. Following the Heisenberg uncertainty principle, a small convolution stride of 1 is utilized to finely explore the frequency feature and achieve a wide-band frequency output.

Particularly, to precisely control the filter passband and achieve an optimal trade-off between time and frequency localization, we parameterize the Gabor filter by considering its frequency response properties. Specifically, a Fourier transform is performed on the φ_n to obtain the magnitude frequency response H_n given by the equation,

$$H_n(f) = e^{-2\pi^2\sigma_n^2(f-c_n)^2}, \tag{7}$$

where f is the frequency, and c_n and σ_n are the center frequency and standard deviation in the time-domain, respectively. By manipulating H_n, we can dynamically explore critical frequency features. To enhance interpretability, as depicted in Fig. 5(b), H_n is controlled by only two trainable parameters: the low cutoff



(a) Impulse responses, including the real and imaginary parts.

(b) Frequency responses.

Fig. 5. Illustration of the Gabor filter.

frequency f_n^l and the high cutoff frequency f_n^h. Thus, both c_n and σ_n can be expressed in terms of f_n^l and f_n^h as

$$c_n = \frac{f_n^l + f_n^h}{2}, \quad \sigma_n = \frac{\mathcal{A}}{\pi \cdot (f_n^h - f_n^l)}, \tag{8}$$

where $\mathcal{A}$ is a constant set to $\sqrt{2\ln 2}$. Given f_n^l and f_n^h, the bandwidth ρ_n of the filter is determined as Eq. (9),

$$\rho_n = f_n^h - f_n^l, \tag{9}$$

which provides a more appropriate measure for quantifying the filter passband. Consequently, quasi-analytic Gabor filters with learnable parameters f_n^l and f_n^h are optimized in conjunction with the remaining module parameters, such that the filter positions and shapes can be fine-tuned to be optimal for the SDD task.

3.2.2. Adaptive scalogram pooling

The scalogram feature $\boldsymbol{X}_{ij}$ has a almost same length as the input speech signals, leading to an increase in computation cost. Therefore, the main purposes of the adaptive scalogram pooling are (1) to reduce $\boldsymbol{X}_{ij}$ to a lower temporal resolution, and (2) to smooth the spectral features and remove the noise components. Specifically, we downsample the features $\boldsymbol{X}_{ij}$ to $\boldsymbol{X}_{ij}^{pl} \in \mathbb{R}^{N\times M}$ using depth wise separable convolutions with fixed stride, where each group is associated with a learnable window smoothing function. It means that the scalogram contextual information of each channel is processed separately by its own function. Subsequently, it allows our model to adaptively adjust the effective response lengths of the windows to better aggregate the time-domain filters outputs. In this study, the Kaiser windows are parameterized for convolving the scalogram features. The Kaiser window is defined as:

$$\vartheta_n(t) = \frac{\mathcal{I}_0(\beta_n\sqrt{1-(\frac{2t}{\kappa})^2})}{\mathcal{I}_0(\beta_n)}, \tag{10}$$

where $\vartheta_n(\cdot)$ represents the value of the window at time $t \in [0, \kappa]$, $\mathcal{I}_0(\cdot)$ is the modified Bessel function of the first kind, and κ is the window length. Notably, β_n is a learnable parameter that controls the shape of the window. We can dynamically control the window shape with a single adjustable parameter β_n. Concretely, when β_n value is small, the Kaiser function preserves almost all information within the window, whereas it retains only a portion of the features when β_n is large enough. Here, the kernel size and convolution stride are set to 401 and 160, respectively.

3.2.3. Parameterized nonlinear transformation

In general, the acoustic features are compressed to replicate the human nonlinear perception of loudness. Several methods (Shen et al., 2022; Wei et al., 2022) use the fixed logarithmic compression function for full-band. However, the logarithmic function (1) depends on loudness and produces different features even if the underlying signal contents are the same, and (2) applies a large amount of dynamic frequency ranges to regions with low information. To address these issues, we use the per-channel energy normalization (Wang, Getreuer, Hughes, Lyon, & Saurous, 2017) method to normalize and compress the feature $\boldsymbol{X}_{ij}^{pl}$. Specifically, $\boldsymbol{X}_{ij}^{pl}(\gamma, n)$ is normalized along the channel by the exponential average of the past value $\boldsymbol{q}(\gamma, n)$,

$$\boldsymbol{X}_{ij}^{nor}(\gamma, n) = \frac{\boldsymbol{X}_{ij}^{pl}(\gamma, n)}{(\varepsilon + \boldsymbol{q}(\gamma, n))^{g_n}} \in \mathbb{R}^{N\times M}, \tag{11}$$

where $\gamma \in [0, M]$ denotes the time step, g_n is the degree of gain normalization, $\varepsilon = 1 \times 10^{-6}$, and the past value is

$$\boldsymbol{q}(\gamma, n) = (1-s)\cdot\boldsymbol{q}(\gamma-1, n) + s\cdot\boldsymbol{X}_{ij}^{pl}(\gamma, n), \tag{12}$$

where the smoothing coefficient s=0.025. Moreover, a compression operation is performed to reduce the dynamic range of the features. The compressed feature $\boldsymbol{X}_{ij}^{cs} \in \mathbb{R}^{N\times M}$ is written as follows

$$\boldsymbol{X}_{ij}^{cs} = \left(\boldsymbol{X}_{ij}^{nor}(\gamma, n) + o_n\right)^{e_n} - o_n^{e_n}, \tag{13}$$

where o_n and e_n are the offsets and exponents. We parameterized g_n, o_n and e_n and incorporate them as learnable parts of the DFBL module.

Finally, DFBL output $\boldsymbol{X}_{ij}^{cs}$ effectively emphasizes the frequency regions associated with depression classification.

3.3. Multi-scale spectral attention module

In fact, aggregating multi-scale speech context information and focusing on acoustic variations before and after a speech can provide a more comprehensive detection perspective (Troubat et al., 2021). Moreover, we consider that the learned multiple sub-bands should contribute differently to the SDD task, and using them directly as inputs may reduce features discriminability. Therefore, we introduce a multi-scale spectral attention (MSSA) module, as shown in Fig. 4(b) which consists of two components:

(1) To acquire multiscale features along the filter-wise axis, a multi-scale features extraction block (MEB) is proposed to extract information under different receptive fields.

(2) To capture the dynamic behavior of the learnable filters, a frequency-aware attention block (FAB) is proposed to obtain channel attention by different time-scale convolutions and assigning different weights to the filters.

3.3.1. Multi-scale feature extraction block

First, dilated convolution is employed to extract the features at various scales from the filterbanks output feature $\boldsymbol{X}_{ij}^{cs}$. Specifically, this operation is applied separately to each channel, allowing us to extract multi-scale contextual information associated with depression within each sub-band Yu and Koltun (2015). Given the nth sequence $\boldsymbol{\mathcal{X}}_n \in \mathbb{R}^{1\times M}$ of filterbanks feature $\boldsymbol{X}_{ij}^{cs}$ and a kernel $\xi \in \{0, \ldots, \kappa-1\}$, the d-dilated convolution $\mathcal{D}_d$ is defined as:

$$\mathcal{D}_d(t) = (\boldsymbol{\mathcal{X}}_n *_d \xi)(t) = \sum_{\mu=0}^{\kappa-1} \xi(\mu)\cdot \boldsymbol{\mathcal{X}}_{t-d\cdot\mu}, \tag{14}$$

where d is the dilation factor, t is the time step, κ is the kernel length, and $(t-d\cdot\mu)$ refers to the past time steps. In our work, as shown in Fig. 4(b), we stack K dilated units in our MSSA module, each one with a 2^{K-1} receptive field. Specifically, each dilated unit contains a dilated convolution operation, an activation function, and two 1×1 convolutions. By stacking the dilated unit, it allows our model to increase the receptive fields. Moreover, the dilated unit has two outputs, one output $\boldsymbol{\mathcal{R}}_{\mathrm{k}}$ is passed through the subsequent unit along the backbone network, and the other skip connection output $\boldsymbol{\mathcal{S}}_{\mathrm{k}}$ allows the block to explicitly capture the multi-scale features at several hierarchical levels. The formulas for the two outputs are written as:

$$\boldsymbol{\mathcal{R}}_{\mathrm{k}} = \boldsymbol{W}_{\mathrm{k}}^{r}(relu(\boldsymbol{W}_{\mathrm{k}}^{d}\cdot \boldsymbol{X}_{ij}^{cs})),\ \ \boldsymbol{\mathcal{S}}_{\mathrm{k}} = \boldsymbol{W}_{\mathrm{k}}^{s}(\boldsymbol{W}_{\mathrm{k}}^{d}\cdot \boldsymbol{X}_{ij}^{cs}), \tag{15}$$

where $\boldsymbol{W}_{\mathrm{k}}^{d}$ represents the dilated convolution weight at k layer, $relu(\cdot)$ is the rectified linear unit (ReLU), $\boldsymbol{W}_{\mathrm{k}}^{r}$ and $\boldsymbol{W}_{\mathrm{k}}^{s}$ are the trainable 1×1 convolution weights, respectively.

3.3.2. Frequency-aware attention block

Considering that the local spectral patterns of each filter are often different, directly leveraging the full-band frequencies as inputs cloud degrade the discriminative ability. Consequently, we present a frequency-aware attention block that models the importance of different channels and captures the dynamic behavior of filters. Specifically, the distinct learnable filters are treated as different channels and assigned different attention weights. First, we utilize the element-wise addition strategy to fuse the multi-scale features $\mathcal{S}_{\mathrm{k}}$ according to Eq. (16), thereby enhancing the essential features while weakening the irrelevant ones,

$$\boldsymbol{X}_{ij}^{fu} = \sum_{\mathrm{k}=1}^{K} \boldsymbol{\mathcal{S}}_{\mathrm{k}}, \tag{16}$$

where $\boldsymbol{X}_{ij}^{fu} \in \mathbb{R}^{N\times M}$ is the fused feature. Moreover, as shown in Fig. 4(b), three parallel one-dimensional convolutional attention modules (CAM) (Woo, Park, Lee, & Kweon, 2018) with kernel sizes of 3, 5 and 7 are used to effectively capture the important different time-scale features of $\boldsymbol{X}_{ij}^{fu}$. Then, the fully connected (FC) layers and softmax function are adopted to learn the important weights $\boldsymbol{a}^{att}$ from the concatenated time-scale features,

$$\boldsymbol{a}^{att} = [a_0, \ldots, a_n, \ldots, a_N]^{\mathrm{T}} \in \mathbb{R}^{N\times 1}, \tag{17}$$

where T indicates transposition, a_n represents the attention weight corresponding to the nth filter.

Finally, we obtain the weighted $\boldsymbol{X}_{ij}^{att} \in \mathbb{R}^{N\times M}$ with the learned weights $\boldsymbol{a}^{att}$, which is written as

$$\boldsymbol{X}_{ij}^{att} = \boldsymbol{X}_{ij}^{fu} \otimes \boldsymbol{a}^{att}, \tag{18}$$

where $\otimes$ is the Hadamard product operation. Accordingly, the model will prioritize frequency sub-bands that have a more pronounced impact on the SDD task. More importantly, by stacking several MSSA layers, our model effectively guides the learnable filters to capture the critical information from the fused multi-scale features, thus enhancing its sensitivity to depression features.

3.4. Classification module

As shown in Fig. 4(c), the classification module contains several residual blocks, and each residual block consists of three one-dimensional convolutional layers with kernel sizes of 7, 5 and 3. Batch normalization and ReLU functions are performed after each convolution layer. Finally, we flatten and feed the block output feature into the FC layers, and use the sigmoid function to map the outputs into probabilities belonging to each class. The cross-entropy loss is defined as follows

$$\mathcal{L}_{ce} = -\frac{1}{\mathcal{T}}\sum_{\tau=1}^{\mathcal{T}} \left[y_\tau \log(p_\tau) + (1-y_\tau)\log(1-p_\tau)\right], \tag{19}$$

where $\mathcal{T}$ is the segment sample total number, y is the true label, and p is the predicted probability. It is worth noting that the subject-level classification results $\bar{y}_i \in \{0, 1\}$ are derived by voting the segment-level prediction label sets, e.g.,

$$\bar{y}_i = voting(\{\bar{y}_{i1}, \bar{y}_{i2}, \ldots, \bar{y}_{ij}, \ldots, \bar{y}_{iJ}\}), \tag{20}$$

where $voting(\cdot)$ is the majority voting function, and $\bar{y}_{ij} \in \{0, 1\}$ represents the prediction label of jth segment for the subject i. The majority of labels in the segment label set are used to determine the final subject label $\bar{y}_i$.

4. Experiments and results

In this section, we conduct relevant experiments to evaluate the performance of the proposed DALF framework on two different datasets. The purpose of the experiments is to investigate the following research questions:

Q1. How does our DALF perform compared with the state-of-the-art methods?

Q2. Is DFBL more beneficial for the SDD task?

Q3. How do the filters behave when extracting depression-relevant filterbanks output features?

4.1. Datasets description

1. NRAC: We collect a dataset called the Neutral Reading-based Audio Corpus (NRAC), which encompasses participant recruitment, clinical observer-rating, self-rating and speech data collection. The objective is to systematically investigate the differences in speech variations between individuals diagnosed with depression and normal controls (NCs). Initially, participants are recruited from the inpatient and outpatient units of the Affiliated Brain Hospital of Nanjing Medical University, while NCs without depressive history are recruited through online advertisements. Specifically, the inclusion criteria are as follows: (a) $13\leq$Age≤ 24, regardless of sex. (b) Normal hearing. (c) Proficiency in Mandarin. Exclusion criteria include: (a) Organic mental disorders or other psychiatric conditions. (b) Medically diagnosed diseases that may affect vocalization. (c) Substance-induced mental disorders. Currently, the dataset comprises 155 depressions and 110 NCs samples.

Furthermore, the severity of depression in participants is assessed using two scales: the Hamilton depression rating scale 17-items (HAMD-17) (Ma et al., 2021) and the patient health

Table 1
Detail of NRAC and corresponding depression severity.

Group	Detail	NC	Depression		
			Mild	Moderate	Severe
PHQ-9	Scores	0–4	5–9	10–19	$\geq$ 20
	Number	34	22	56	30
HAMD-17	Scores	0–7	8–16	17–23	$\geq$ 24
	Number	86	13	62	80

questionnaire 9-items (PHQ-9) (Kroenke & Spitzer, 2002). HAMD-17 is an observer-rating scale used by clinicians, while PHQ-9 is a self-rating completed by the participants themselves. According to these two scales, participants are categorized into two groups:

Group PHQ-9 : Participants in this group evaluate their depression severity using the PHQ-9, and speech recordings are labeled with scores ranging from 0 to 27. The scale scores correspond to the following levels of depression severity: normal control (0–4), mild depression (5–9), moderate depression (10–19) and severe depression ($\geq$20).

Group HAMD-17: Participants in this group assess their depression severity using the HAMD-17, and speech recordings are labeled by scores ranging from 0 to 52. The severity levels based on the scale scores are as follows: normal control (0–7), mild depression (8–16), moderate depression (17–23), and severe depression ($\geq$24).

During the data collection phase, all participants are positioned in a peaceful and enclosed environment. Clear instructions are provided to the participants, emphasizing the importance of maintaining a natural and relaxed state throughout the recording session. For capturing the audio, a standardized recording pen is utilized for capturing the audio. Participants are asked to read a specific text "Let life be beautiful like summer flowers". The speech lengths range from 2 to 4 min, and the speech files are recorded at a sampling rate of 44.1 kHz, PCM at 16-bits.

Table 1 summarizes the basic information of the NRAC datasets, and Appendix B provide the demographic details. All participants are informed about the study and provided written informed consent, which is signed by themselves and their legal guardians, following the regulations set by the Medical Ethics Committee of Nanjing Brain Hospital (MECNJBH). The study received official approval from MECNJBH under the approval number 2021-KY108-01.

2. DAIC-woz: As the eight-item Patient Health Questionnaire (PHQ-8) has been shown to be as effective as the PHQ-9 for screening depression (Shin, Lee, Han, Yoon, & Han, 2019), we also utilize the Distress Analysis Interview Corpus, Wizard of Oz (DAIC-woz) (Gratch et al., 2014) in our study. The DAIC-woz is a publicly available dataset in English that includes a series of clinical interviews to diagnose depression, anxiety, post-traumatic stress and other mental disorders. The length of each audio file ranges 7-33 min with a fixed samplerate of 16000 Hz. Each participant has been given the PHQ-8 score based on their answers to the questionnaires. Each record is labeled by the PHQ-8 score and PHQ-8 binary label, where the binary score is determined by a threshold of 10, dividing the patients into depression and NCs groups. As shown in Table 2, the train and development set of the DAIC-woz includes 107 and 35 samples, respectively.

4.2. Pre-processing and setup

We convert the samplerate of the speech signal to 16000 Hz by FFmpeg (Tomar, 2006), and all experiments are implemented on NVIDIA A100 GPU using PyTorch.

Table 2
Detail of DAIC-woz.

DAIC-woz	Detail	NC	Depression
	Scores	$\leq$ 10	$>$ 10
Number	Train	77	30
	Development	23	12

Table 3
The hyper-parameter settings of our model, where length indicates the kernel size.

Names	Input	Output	Length	Stride	Group	Layer	Block
TD filters	1	64	401	1	1	1	–
Pooling	64	64	401	160	64	1	
Dilated conv.	64	64	2^d	1	64	6	3
CAM	64	64	[3, 5, 7]	1	64	6	
Classifier	64	128	[3, 5, 7]	1	1	3	3

Table 4
Experimental results on NRAC, with the best results in bold.

Scales	Tasks	Measures	F1	Accuracy	Precision	Recall
PHQ	DC	NC vs. Mild	0.839±0.07	0.839±0.07	0.797±0.13	0.810±0.10
		NC vs. Moderate	0.786±0.08	0.789±0.08	0.815±0.05	0.858±0.09
		NC vs. Severe	0.841±0.07	0.844±0.07	0.831±0.11	0.867±0.13
		NC vs. Depressive	**0.873±0.05**	**0.873±0.05**	**0.932±0.05**	**0.906±0.06**
		4-class classification	0.477±0.06	0.499±0.05	0.485±0.03	0.467±0.29
	SC	Mild vs. Moderate	0.616±0.04	0.616±0.05	0.616±0.05	0.627±0.05
		Mild vs. Severe	0.576±0.03	0.648±0.06	0.669±0.05	0.833±0.02
		Moderate vs. Severe	0.647±0.08	0.586±0.11	0.500±0.06	0.500±0.06
HAMD	DC	NC vs. Mild	0.794±0.10	0.789±0.13	0.233±0.20	0.367±0.37
		NC vs. Moderate	0.773±0.02	0.776±0.02	0.757±0.13	0.705±0.11
		NC vs. Severe	**0.849±0.05**	**0.847±0.05**	0.872±0.04	0.812±0.13
		NC vs. Depressive	0.817±0.03	.819±0.03	**0.893±0.06**	0.819±0.04
		4-class classification	0.417±0.04	0.550±0.02	0.435±0.06	0.600±0.05
	SC	Mild vs. Moderate	0.639±0.12	0.640±0.15	0.777±0.07	0.771±0.16
		Mild vs. Severe	0.755±0.04	0.773±0.09	0.871±0.04	**0.875±0.14**
		Moderate vs. Severe	0.572±0.06	0.578±0.06	0.636±0.06	0.588±0.15

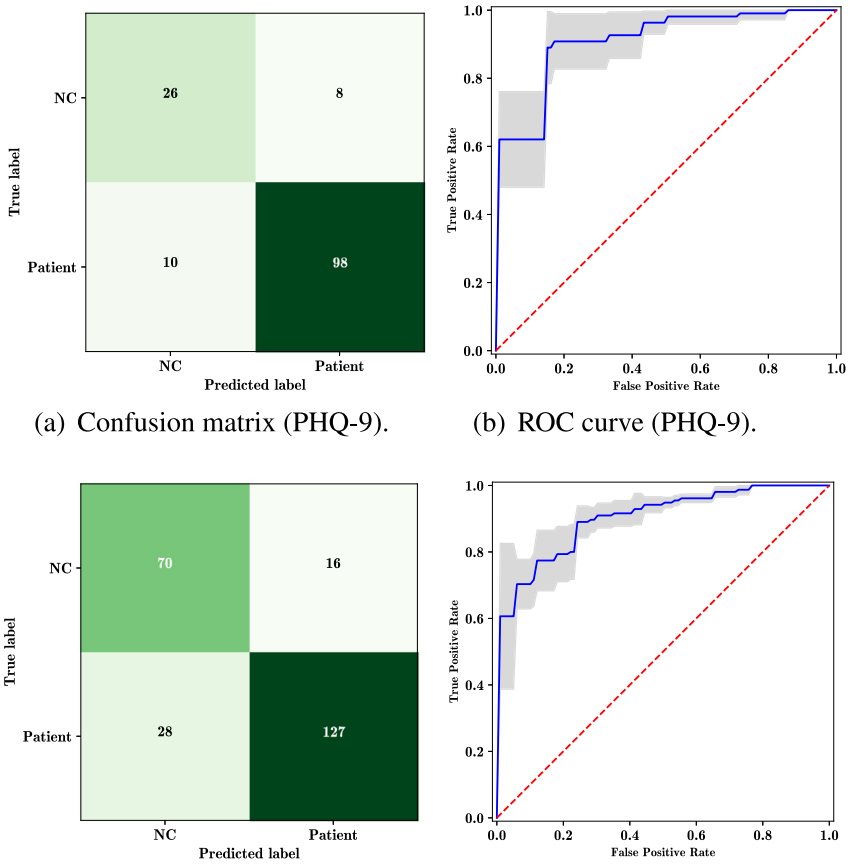
Parameters setting: The hyper-parameters setting of each module is shown in Table 3. For the optimization, we use the Adam optimizer with a weight decay of 0.001, the batch size is set to 32, and the initial learning rate is 0.0005.

Evaluation Metrics: For DAIC-woz dataset, we follow the same random sampling procedure as the DepAudioNet (Ma, Yang, Chen, Huang, & Wang, 2016), and the results are reported based on the development set. For each task in NRAC, all datasets are stratified into 5-fold to preserve the percentage of samples in each class. Additionally, we select 10% of the training data as the validation dataset for hyperparameter tuning purposes. To evaluate the performance of the model, we consider accuracy, recall, F1-score and precision as the chosen evaluation metrics. We report the means and standard deviations of all results, and use the paired t-test at a 5% level of significance to explore whether there are statistically significant differences between the indicators of DALF and other methods.

4.3. Performance evaluation on the NRAC dataset

To intuitively explore the performance of DALF on the SDD task, we first perform a comprehensive validation on the NRAC dataset with PHQ-9 and HAMD-17 scales. Specifically, we cast depression detection as a classification task. The purpose of the depression severity classification (SC) is to explore the discriminative patterns through speeches. Moreover, the diagnosis-based classification (DC) task focuses on detecting the differences in speech between NCs and patients. Table 4 reports the results of DALF, there are some observations worth highlighting:

1. With regards to the depression and NCs classification in PHQ-9 and HAMD-17 groups, DALF demonstrates strong performance with F1 scores of 87.3% and 81.7%, respectively. Moreover, Fig. 6(a) and (c) depict the confusion matrices for the NCs and depression classification, while the corresponding ROC curves are



(a) Confusion matrix (PHQ-9). (b) ROC curve (PHQ-9).

(c) Confusion matrix (HAMD-17). (d) ROC curve (HAMD-17).

Fig. 6. Confusion matrix and ROC for classification of depression and NCs in PHQ and HAMD groups, respectively.

shown in Fig. 6(b) and (d). These results highlight the excellent classification capabilities of the DALF model in the SDD task.

2. When comparing the classification results of discriminating NCs from patients with mild, moderate, and severe symptoms in the PHQ-9 groups, a notable observation is the consistently excellent classification performance regardless of the severity of depression. These results demonstrate that patients with varying symptoms exhibit significant differences in speech when compared to NCs. Furthermore, the analysis of both groups reveals that the model performance becomes significantly improves as the symptom is severe, indicating that the acoustic features associated with depression are more pronounced in severe cases.

3. When observing the results of depression severity classification, including the classification tasks of NCs vs. each severity class and the classification tasks of different severity classes, we can find that the classification performance among mild, moderate and severe symptoms performs poorly, especially when symptoms are close together. This is attributed to the fact that depressive speech often exhibits common features and similar presentations, resulting in weak distinctions among different symptom severities.

4. Considering the weaker performance in classifying mild and NCs within the HAMD-17 group, it can be inferred that the PHQ-9 scale provides an equitable reflection of depression speech. In other words, speech variations are evident regardless of the scores. Moreover, we find a weaker performance in the HAMD-17 group. The PHQ-9 scale, grounded in the actual feeling of the patient, is well-suited for preliminary screening.

4.4. Comparisons with the state-of-the-art models

4.4.1. Performance analysis

To further demonstrate the effectiveness of our DALF. On the one hand, we compare our approach with several state-of-the-art methods using the public DAIC-woz dataset. In particular, we conduct a comprehensive comparison of DALF model with the traditional methods and the deep learning methods. On the other hand, several comparable methods are tested on the PHQ-9 group of the NRAC dataset. For all the compared approaches, we employ the same setting as the original papers to make sure that they are competitive in the comparison. All Mel-spectral features (N_{fb}=64, N_{mfcc}=40) are performed at the 16 kHz samplerate using the 25 ms window length and the 10 ms hop length.

Traditional methods:

In this section, we select two common ML classification methods: support vector machine (SVM) and random forest (RF) as baselines. Then we use the openSMILE (Eyben et al., 2010) toolkit to extract the acoustic features including low-level MFCCs, COMPARE and eGeMAPS, and input all concatenated feature representations into the models.

Deep learning models:

We compare our model with the following models:

DepAudioNet[1] (Ma et al., 2016) is a SDD framework that uses one-dimensional CNN and LSTM to capture the middle-term and long-term correlations to produce a more comprehensive speech representation.

[1] https://github.com/adbailey1/DepAudioNet_reproduction

Table 5
Comparison between DALF and other methods on different datasets, where RS denotes raw speech signal. The best indicators are boldfaced and those with † are statistically significant with p value < 0.05.

Data	Model	Features	Accuracy	F1	Precision	Recall
DAIC	SVM	LLDs	–	0.400	0.330	0.500
	RF	LLDs	–	0.570	0.500	0.570
	DepAudioNet	MFbanks	–	0.610	0.625	0.770
	CNN-AE	RS	–	0.710	0.705	0.710
	En-CNN	MFbanks	0.740	0.725	0.720	0.760
	DEPA	MFbanks	–	0.610	0.610	0.610
	DEPA	STFT	–	0.640	0.640	0.640
	Mfcc-LSTM	MFCC	0.763	0.655	0.735	0.645
	Vlad-GRU	MFbanks	–	0.770	0.630	**1.000**
	ConvbiLSTM	MFbanks	–	0.610	0.560	0.660
	DALF	RS	**0.786**	**0.784**	**0.772**	0.794
NRAC	RF	LLDs	$0.688 \pm 0.02^{\dagger}$	$0.638 \pm 0.07^{\dagger}$	$0.675 \pm 0.04^{\dagger}$	$0.726 \pm 0.06^{\dagger}$
	SVM	LLDs	$0.738 \pm 0.11^{\dagger}$	$0.748 \pm 0.10^{\dagger}$	$0.763 \pm 0.03^{\dagger}$	0.739 ± 0.17
	DepAudioNet	MFbanks	$0.802 \pm 0.03^{\dagger}$	0.768 ± 0.09	0.875 ± 0.09	0.889 ± 0.10
	Vlad-GRU	MFbanks	$0.676 \pm 0.10^{\dagger}$	$0.629 \pm 0.05^{\dagger}$	$0.782 \pm 0.05^{\dagger}$	$0.818 \pm 0.21^{\dagger}$
	ConvbiLSTM	MFbanks	$0.788 \pm 0.09^{\dagger}$	0.795 ± 0.09	0.919 ± 0.04	$0.794 \pm 0.15^{\dagger}$
	DALF	RS	**0.873±0.05**	**0.873±0.05**	**0.932±0.05**	**0.906±0.06**

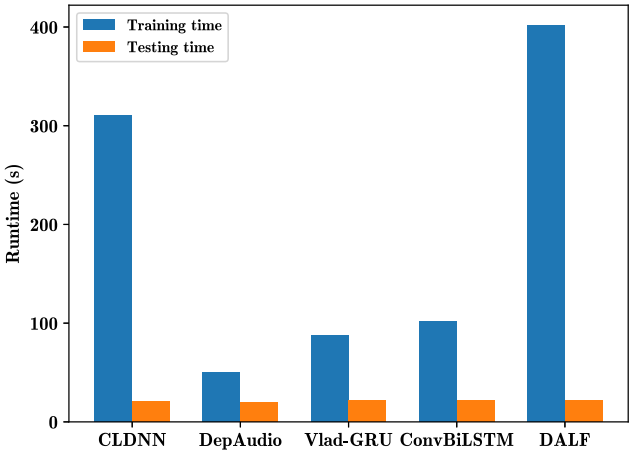


Fig. 7. Running time of the models on NRAC dataset.

DEPA (Zhang et al., 2021) is a self-supervised audio embedding pre-training model for depression detection, in which a self-supervised encoder–decoder model is trained to predict and reconstruct the center spectrogram.

Mfcc-LSTM (Rejaibi et al., 2022) is a successive model based on the LSTM that utilizes the MFCC features to detect depression and assess its severity levels.

CNN-AE (Sardari, Nakisa, Rastgoo, & Eklund, 2022) is an end-to-end depression detection model that uses the CNN auto-encoder to learn the highly relevant and discriminative features from the raw audio signals.

En-CNN (Vázquez-Romero & Gallardo-Antolín, 2020) takes advantage of an ensemble learning strategy to integrate voting mechanisms with 1D-CNN architecture to improve the performance of depression classification.

Vlad-GRU[2] (Shen et al., 2022) uses the NetVLAD to produce the same-dimensional speech embedding from extracted Mel spectrograms, and the Gate Recurrent Unit (GRU) is used to automatically summarizes representations of depressed speech.

ConvBiLSTM[3] (Wei et al., 2022) captures short and long-term temporal as well as spectral features by leveraging a hierarchical CNN and Bi-LSTM structure to achieve end-to-end depression estimation.

Table 5 shows that our model consistently has better performance than the baselines. Particularly, we have the following observations.

1. Compared with state-of-the-art, DALF achieves better performance on the DAIC-woz. Moreover, compared with the traditional methods such as handcraft features combined with RF and SVM, DALF observes a substantial improvement of 21.4% and 38.4% in terms of F1, respectively. Similarly, the performance of our DALF also outperforms other deep learning methods that only use the audio modality data. For example, when compared to CNN-based model (DepAudioNet, CNN-AE and En-CNN), DALF achieves an average F1 improvement of 10.20%. Additionally, in comparison to the RNN-based methods (DEPA, Mfcc-LSTM, Vlad-GRU and ConvbiLSTM). DALF achieves an average F1 improvement of 12.7%. These results highlight the effectiveness of the DALF framework.

2. On the NRAC dataset, as shown in Table 5, our model consistently achieves the best results in terms of accuracy, F1 and other metrics when compared to other state-of-the-art models. This improvement demonstrates that DALF framework clearly outperforms the state-of-the-art methods. It is due to the fact that the DFBL module captures more relevant spectral features associated with depression in high-dimensional signals.

3. An interesting observation is that the models typically achieve better classification results on the NRAC than on the DAIC-woz dataset. This can be attributed to the fact that the DAIC-woz dataset is collected based on conversation scenarios with fewer training samples and more imbalanced class distribution. In contrast, the NRAC dataset uses neutral reading for speech collection, with low speech content heterogeneity and more balanced samples. As a result, we believe that more clinically valuable findings can be achieved from the NRAC dataset.

4.4.2. Complexity analysis

The main efforts of our work are to extract filterbanks features from the raw signals using the DFBL module. Accordingly, we analyze the complexity of DFBL. First, we compare the parameter number $\mathcal{O}_s$ of DFBL with other methods that learn features on the raw signal. Specifically, the parameters of DFBL are as follows:

$$\mathcal{O}_s^1 = 4N + 3, \tag{21}$$

where N is the filter number. Additionally, DCNN (He & Cao, 2018) feeds the speech signals into two convolutional layers to generate features of the same dimension as MFbanks features. The parameter number is written as

$$\mathcal{O}_s^2 = \kappa \cdot (N^2 + N). \tag{22}$$

Accordingly, when the kernel size κ=401, N=64, we can observe that the whole parameters of DFBL are over 6440 times less than unconstrained CNNs alternatives. Moreover, we compare the runtime $\mathcal{O}_t$ of DALF with the other methods on the NRAC PHQ-9 group. Fig. 7 illustrates the runtime of training (50 epochs) and testing on the same device.

In conclusion, DALF exhibits a longer runtime during the training phase compared to models utilizing Mel-spectral features. This discrepancy arises from the fact that the spectral-based model bypasses the feature extraction process from the raw signals. Remarkably, when processing the same test dataset and conducting tests, the runtimes of all models are essentially similar. Furthermore, a comparison with the end-to-end CLDNNs (Sainath et al., 2015) reveals that feature extraction from the original signal is typically time-consuming. Nevertheless, the small number of parameters of DFBL is easy to control, allowing the model to better extract fine-grained and discriminative acoustic features. Moreover, the trade-off between runtime and accuracy is reasonable, as it allows for achieving higher accuracy while maintaining a reasonable computational cost.

[2] https://github.com/speechandlanguageprocessing/ICASSP2022-Depression
[3] https://github.com/pingcheng-wei/depressionestimation

Table 6
Performance of the DFBL module.

Method	F1	Accuracy	Precision	Recall
CNNs	0.778±0.06	0.782±0.06	0.849±0.04	0.871±0.08
MFCC	0.803±0.07	0.796±0.08	0.909±0.05	0.814±0.09
MFbanks	0.833±0.08	0.831±0.08	0.826±0.05	0.850±0.12
DFBL	**0.873±0.05**	**0.873±0.05**	**0.932±0.05**	**0.906±0.06**

Table 7
Generalization ability of the DFBL module. F indicates the fixed and T is the Fine-tuning.

Datasets		**Module**		F1	Accuracy
Source	Target	DFBL	Others		
DAIC-woz	NRAC	F	T	0.833±0.02	0.845±0.03
DAIC-woz	NRAC	T	T	0.887±0.08	0.887±0.07
NRAC	DAIC-woz	F	T	0.772	0.760
NRAC	DAIC-woz	T	T	0.786	0.787

4.5. Ablation analysis

We conduct several ablation studies to investigate how these modules (DFBL, MSSA) affect the classification performance in the PHQ-9 group of NRAC. Moreover, a series of experiments are conducted to analyze possible factors that affecting the DALF performance.

4.5.1. Effectiveness of the DFBL module

First, to further elaborate on the effectiveness of the depression filterbanks features learning module, DFBL is compared with the common acoustic features, such as MFbanks and MFCC, which are all used as the original input features of the model. Moreover, we also replace DFBL with the standard CNNs to demonstrate the effectiveness of the pre-processing on raw speech signals. Specifically, all MFCC and MFbanks features are performed at the 16 kHz samplerate using the 25 ms window length and the 10 ms hop length. The kernel and stride length of the convolution is set to 401 and 160. Then, we investigate the generalization ability of the DFBL module. We optimize the DFBL module parameters on one dataset, and then tune the classifier or all of the module parameters on another dataset.

As shown in Table 6, we find that the model utilizes the DFBL outperforms the Mel-scale features and large kernels convolution. These results reveal several meaningful points:

1. The MFCC-based model performs the worst in terms of accuracy. Our findings seem to yield conclusive evidence that the useful acoustic information for the SDD task is discarded during the MFCC feature extraction procedure. In contrast, more redundant acoustic information is retained in the CNN-based model used directly on the raw signals, both of which are detrimental to the performance improvement of the SDD task.

2. The DFBL module achieves the best classification performance, which validates the significance of end-to-end feature learning again. Through joint learning with the downstream classification task, DFBL overcomes the limitations of prior knowledge and effectively explore depression-relevant spectral features.

3. Regarding generalizability, as shown in Table 7, the accuracy decreases by 4.0% and 1.4%, respectively, when fixing the parameters of DFBL trained on other datasets. However, significant performance improvement is observed after pre-training the DFBL module. This reason is that the filters are insufficiently trained due to the limited training dataset. It serves as a reminder that pre-training of filter parameters by designing appropriate pretext tasks should be considered in future work.

Table 8
Effectiveness of each component in MSSA.

Components	F1	Accuracy	Precision	Recall
w/o MEB	0.822±0.08	0.816±0.09	0.923±0.04	0.832±0.13
w/o FAB	0.860±0.05	0.859±0.05	0.930±0.05	0.888±0.08
DALF	**0.873±0.05**	**0.873±0.05**	**0.932±0.05**	**0.906±0.06**

Table 9
Influence of different time-domain filters.

Kernels	F1	Accuracy	Precision	Recall
1D-Conv filter	0.848±0.05	0.852±0.05	0.898±0.03	0.916±0.08
Sinc filter	**0.877±0.05**	0.866±0.06	0.909±0.08	0.852±0.06
Wavelet filter	0.831±0.06	0.831±0.06	0.920±0.07	0.861±0.08
Gamma-tone filter	0.827±0.05	0.831±0.04	0.895±0.05	0.888±0.06
Gabor filter	0.873±0.05	**0.873±0.05**	**0.932±0.05**	**0.906±0.06**

4.5.2. Role of the MSSA module

In this section, experiments are conducted to further evaluate the effectiveness of each component in the MSSA module. Specifically, the MEB and FAB components are removed individually to verify their relative contributions. Experimental results are reported in Table 8.

1. From Table 8, we remove the MEB module from DALF and only use the FAB module to exploit the filterbanks features and guide the behaviors of filters. It can be observed that the model results decrease by 5.1% and 7.4% in terms of F1 and recall, respectively, which demonstrates that MEB has a significant impact on performance by fusing the multiple hierarchical spectral features from the filter-wise perspective.

2. It is obvious that removing FAB from DALF leads to a decrease of 1.8% in recall. The reason for performance degradation is mainly due to the fact that FAB is designed to assist filters in capturing critical channels within the spectral structure. When the FAB is removed, the filters tend to favor non-representative frequency regions, which is detrimental to the SDD task.

4.5.3. Influence of different time-domain filters

To investigate the role of different parameterized TD learnable filters in DFBL. We employ different time-domain filter functions, and all these functions are normalized in the continuous domain.

Sinc filters (Ravanelli & Bengio, 2018) have frequency responses that approximate the band-pass rectangular windows, the functions $h_{sc}(\cdot)$ can be written as

$$h_{sc}(t) = 2f_n^h \cdot sinc(2\pi f_n^h t) - 2f_n^l \cdot sinc(2\pi f_n^l t), \tag{23}$$

where $t \in [\frac{-\kappa}{2}, \frac{\kappa}{2}]$, f_n^l and f_n^h are the low and high cutoff frequencies, center frequencies $c_n^r = (f_n^l - f_n^h)/2$.

Gamma-tone filters (Zeghidour, Usunier, Synnaeve, Collobert, & Dupoux, 2018) are given by gamma probability distribution functions multiplied by sinusoidal tones,

$$h_{tone}(t) = at^{r_n-1} \cdot e^{-2\pi b_n t} \cdot \cos(2\pi c_n t), \tag{24}$$

where $t \in [0, \kappa]$, the learnable parameters are the bandwidth b_n and center frequencies c_n of the filters, r_n are the order of the filters, a is a constant to control the amplitude.

Wavelet filters (Khan & Yener, 2018) are defined by the wavelet functions. Specifically, we select the Ricker wavelet with a single parameter,

$$h_w(t) = \frac{2}{\pi^{\frac{1}{4}}\sqrt{3s_n}}\left(\frac{t^2}{s_n^2} - 1\right) \cdot e^{-\frac{t^2}{s_n}}, \tag{25}$$

where $t \in [\frac{-\kappa}{2}, \frac{\kappa}{2}]$ and s_n is the learnable scaling parameter.

Table 9 presents the results of DALF with different learnable filters. The Gabor filters realize better trade-offs in time and frequency resolution through the relative phases of the real and

Table 10
Influence of different segment lengths.

Overlap	Length	F1	Accuracy	Precision	Recall
0	5	0.835±0.08	0.830±0.08	0.925±0.05	0.841±0.12
	6	**0.873±0.05**	**0.873±0.05**	**0.932±0.05**	0.906±0.06
	7	0.852±0.06	0.852±0.06	0.911±0.04	0.898±0.10
	8	0.851±0.02	0.852±0.03	0.911±0.03	0.897±0.07
50%	5	0.833±0.04	0.831±0.04	0.917±0.04	0.860±0.09
	6	0.866±0.06	0.866±0.06	0.920±0.04	0.907±0.08
	7	0.859±0.06	0.859±0.06	0.920±0.04	0.898±0.09
	8	0.850±0.04	0.852±0.04	0.904±0.04	**0.908±0.08**

Table 11
Performance of different pooling manners.

Manners	F1	Accuracy	Precision	Recall
Max-pooling	0.822±0.06	0.838±0.05	0.853±0.03	0.899±0.05
Avg-pooling	0.816±0.01	0.831±0.01	0.852±0.02	0.909±0.04
DSC	0.845±0.04	0.845±0.05	0.912±0.04	0.889±0.09
DSC w/ Gaussian	0.859±0.04	0.866±0.04	0.899±0.05	0.935±0.06
DSC w/ Hanning	0.833±0.02	0.845±0.03	0.867±0.03	**0.944±0.07**
DSC w/ Kaiser	**0.873±0.05**	**0.873±0.05**	**0.932±0.05**	0.906±0.06

imaginary parts, and thus perform better in the SDD task. Moreover, the Since and Gabor filters outperform the standard 1D convolution by 2.9% and 2.5% in terms of F1, respectively. These observations show that the results can be significantly improved when the filter coefficient is effectively constrained. In particular, the learnable filters achieve a natural inductive bias that exploits the knowledge with respect to the filter shape, while retaining the flexibility to adapt to the speech.

4.5.4. Influence of different segment lengths

Actually, different segment lengths represent different levels of acoustic information. We use different lengths of segment samples to verify the influence of sample length. In this part, each speech signal is split into segments of lengths l_{ij} with the option of 0 or 50% overlapping.

As shown in Table 10, the worst model results are observed when the segment has a small length and overlapping. This can be attributed to the increase in noise samples and the lack of sufficient speech information. Conversely, large segments reduce the number of training samples leading to model overfitting. It is worth noting that larger segments contain more information that needs to be processed. However, once the segments contain sufficient representative acoustic features, further increasing their length becomes redundant and offers no additional benefit.

4.5.5. Influence of different pooling manners

To investigate the effectiveness of different downsampling methods in our work. We compared our method with several approaches such as avg-pooling (Balestriero, Cosentino, Glotin, & Baraniuk, 2018) and max-pooling (Noé et al., 2020). Moreover, we employ different window functions to tune the convolution kernel coefficients of depth-wise separable convolutions (DSC) to achieve different smoothing performances, including the Gaussian window (Zeghidour et al., 2021) and Hanning window. The results are shown in Table 11. Several observations from these results are worth highlighting:

1. The results show that the performance improves when convolution with stride is used instead of max-pooling and avg-pooling. We argue that DSC has a dynamic adjustment behavior in retaining different frequency sub-bands for each filter, whereas max-pooling and avg-pooling not only over-attenuate the retained frequency regions during dimensionality reduction, but also lose important feature information.

2. Compared with Hanning and Gaussian windows, we also find that the performance of Kaiser window leads to an average 2.7% improvement in F1, as well as a significant improvement compared to the fixed pooling approaches. This finding suggests that the adaptive Kaiser window has a better shape adjustment capability for preserving and smoothing the feature region of corresponding channels.

Table 12
Impact of different classification modules.

Inputs	Classifiers	F1	Accuracy	Precision	Recall
MFbanks	ResNet	0.843±0.02	0.838±0.02	0.932±0.03	0.852±0.04
	TE	0.818±0.07	0.817±0.07	0.927±0.06	0.832±0.12
	LSTM+ResNet	0.824±0.09	0.817±0.10	0.937±0.04	0.813±0.30
	CNN+LSTM[a]	0.652±0.06	0.647±0.07	0.632±0.14	0.804±0.08
DFBL	ResNet	**0.873±0.05**	**0.873±0.05**	0.932±0.04	**0.906±0.06**
	TE	0.772±0.09	0.760±0.01	**0.947±0.07**	0.730±0.09
	LSTM+ResNet	0.845±0.04	0.845±0.05	0.912±0.04	0.889±0.09
	CNN+LSTM[a]	0.738±0.11	0.748±0.10	0.903±0.03	0.739±0.17

[a]The structure is the same as DepaudioNet.

4.5.6. Impact of different classification modules

In this section, we focus on answering whether a unified classification module exists for the SDD task, regardless of inputs. Specifically, the classification module attempt to extract useful information from the spectral features in both temporal and spatial dimensions, while completing the guidance of the DFBL module. Some baseline models we selected are as follows:

Residual neural network (ResNet): contains 3 residual blocks, each block consists of 3 one-dimensional convolutional layers with kernel sizes of 7, 5, and 3, hidden size set to 64. The batch normalization and ReLU functions are performed after each convolution layer, and the average-pooling and FC layer is performed after the last layer.

Long short-term memory network (LSTM): contains 2 LSTM layers, each bias of the LSTM set to false.

Transformer encoder (TE): contains 3 transformer layers, each layer consists of 16 multi-head attention. The batch normalization, feed-forward layer and GELU functions are performed after each layer, hidden channel size and dropout rate are set to 64 and 0.1, respectively. Finally, the FC layer is performed after the last transformer layer.

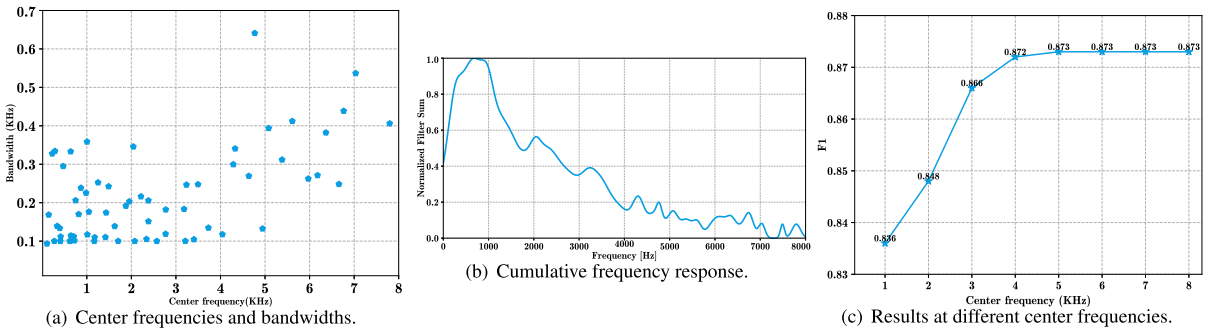
Table 12 presents the performance of the classification modules using two different input features. It can be observed that the ResNet module achieves better results with both features. However, the incorporation of the sequential models decrease the performance, indicating their limited capabilities in capturing local feature patterns. Additionally, the poor performance of the transformer encoder can be attributed to the limited dataset size and insufficient training.

4.6. Interpretability analysis

First, to explore the behavior of the learnable filters and to know which sub-bands are critical in the SDD task, we analyze the parameters learned by Gabor filters. Specifically, we illustrate c_n^r and ρ_n in Fig. 8(a) to visualize what the filters learn during training. Moreover, since the magnitude frequency response H_n is usually considered as the frequency decay function of the nth filter, a value close to 1 indicates that the corresponding frequency region is retained more by the filter. Accordingly, we calculate the cumulative frequency response Q_c as follows

$$Q_c = \sum_{n=1}^{N} H_n, \tag{26}$$

where Q_c is obtained by summing up the findings $\mathcal{H}_n$ of all learnable filters. This approach allows us to visually highlight the frequency sub-bands covered by all filters, and the normalized



(a) Center frequencies and bandwidths. (b) Cumulative frequency response. (c) Results at different center frequencies.

Fig. 8. Analysis of filters parameters.

results are shown in Fig. 8(b). Finally, as shown in Fig. 8(c), we constrain the filter center frequencies below [1, 2, 3, 4, 5, 6, 7, 8] kHz, respectively, and then re-validate the DALF model on the same dataset, the horizontal and vertical axes are the center frequencies and F1 results. We make several observations:

1. As in Fig. 8(a), we can observe that the learned filters are focusing on different sub-bands. All filters adaptively adjust their learnable parameters, and the bandwidth of learned filters increases with the center frequency increasing. The frequency response regions of all learned Gabor filters cover almost the full-band of the speech.

2. As in Fig. 8(b) and 8(c), the cumulative frequency responses of trained filters pay more attention on the lower frequencies below 4000 Hz, and we find two main peaks in the lower frequency region around 600–700 Hz and 2050–2100 Hz. Furthermore, we observe that the F1-score improves with increasing frequency range when the filter center frequencies are located in the range of 0–4k Hz, while the influence on the model becomes weaker for the center frequencies over 4k Hz.

In conclusion, these findings collectively suggest that:

1. The learnable Gabor filters achieve the feature extraction by flexibly optimizing two cutoff frequency parameters, which is more selective than other fixed feature construction methods. Each filter focuses on the different parts of the frequency full-band. We believe this is one of the reasons that explain the performance of DFBL module, as it focuses on selecting the more appropriate frequency ranges to represent the speech samples.

2. Interestingly, the learned Gabors filters emphasize the frequency ranges around 600–700 Hz and 2050–2100 Hz, corresponding to the first formants of the Mandarin vowel {/***e***/, /***ê***/} and the second formants of Mandarin vowels /***ê***/ (Howie & Howie, 1976; Liu & Ng, 2009). This adaptation to human speech shows that DALF is able to learn the contents that are important for the SDD task. In other words, the center frequencies of the learnable filter are concentrated in the specific formants range of Chinese vowels, especially in the 600–700 Hz frequency range. We believe this is a biomarker that assists in distinguishing depressive patients and NCs. Moreover, the learned filters have smaller bandwidths in the critical frequency region, suggesting that the narrow-band information is more appropriate for the SDD task.

4.6.1. Limitation of our work

All of the experiments conducted in this study have demonstrated the potential of speech as biomarkers for depression detection and confirmed the effectiveness of the DALF model in the SDD task. However, there are still two limitations that need to be addressed. (1) The small size of the datasets is likely to result in insufficient optimization of the filter parameters. To address this limitation, future studies should recruit more volunteers for speech collection or pre-train the DFBL module on larger emotion recognition (SER) (Lei, Zhu, & Wang, 2022) datasets. (2) In order to achieve a wide-band frequency output, the Gabor filter must be narrow in the time-domain. However, the frequency response of the filter away from the center frequency is close to 0. To alleviate this issue, the convolution stride is set to a small value, which inevitably leads to an increase in computational cost.

5. Conclusion and future work

In this work, we propose an attention-guided learnable time-domain filterbanks representation network for speech depression detection. DALF employs a lightweight architecture, DFBL, to compute filterbanks features at the initial layer of the model. DFBL is fine-tuned with the guidance of the objective function to obtain task-relevant spectral features. Compared with fixed Mel-features inputs, the learned filterbanks features consistently outperform other feature acoustics in our experiments. Moreover, we propose an MSSA module to guide the filters to learn the useful frequency sub-bands, so that our model not only focuses on those timestamps and frequency ranges that are discriminative, but also emphasizes the informative channels and suppresses the less useful ones. As a result, our DALF model achieves competitive performance on both the DAIC-woz and NRAC depression datasets compared to previous works. Furthermore, by analyzing the frequency responses of filters, we discover that the Mandarin vowel /***e***/, around 600–700 HZ, serves as a representative biomarker of depressed speech. In future work, we plan to extend work to the multi-modal of depression and to the multi-view (Tian et al., 2019) of speech. Additionally, we will explore a self-supervised learning strategy to optimize the learnable filterbanks parameters under limited annotated data.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in or the review of the manuscript entitled, Attention Guided Learnable Time-domain Filterbanks for Speech Depression Detection. We declare that this research was carried out following the principles of the Declaration of Helsinki.

Data availability

Data will be made available on request.

Acknowledgments

This study is funded by the National Key Research and Development Program (2022YFC2405603 to Xizhe Zhang), National Natural Science Foundation of China (62076059 to Peng Cao, 62176129 to Xizhe Zhang), Science Project of Liaoning Province, China (2021-MS-105 to Peng Cao), National Science Fund for Distinguished Young Scholars (81725005 to Fei Wang), the National Natural Science Foundation Regional Innovation and Development Joint Fund (U20A6005 to Fei Wang), Jiangsu Provincial Key Research and Development Program, China (BE2021617 to Fei Wang).

Appendix A. Major symbols and definitions

As shown in Table A.1, we provide a list of key symbols along with their corresponding definitions.

Appendix B. Demographics of the NRAC

Descriptive analyses are employed to summarize socio-demographic variables. To assess the normality of the age variable, the Kolmogorov–Smirnov test is applied, and the means and standard deviations are reported. Subsequently, the differences between the NCs and depression groups are compared using the t-test. Sex differences between the two groups are examined using the chi-square test. $p < 0.05$ is considered statistically significant. Table B.1 presents the demographics results of the NRAC dataset. It is evident that a significant age difference exists between the depression patients and NCs, which motivates us to conduct two subsequent experiments:

1. Retest (R-t): Ten speeches are collected from 10 NCs of age [13, 14, 16, 16, 17, 17, 18, 18, 19, 19] and HAMD-17 score $\leq$7.

Table A.1
The major symbols and definitions of this paper.

Symbol	Definition
$\boldsymbol{x}_i$, $\boldsymbol{x}_{ij}$	$\boldsymbol{x}_{ij}$ is the jth segment of the speech $\boldsymbol{x}_i$
y_i, y_{ij}	y_{ij} and y_i are segment and subject labels
$\boldsymbol{X}_{ij}$	scalogram features matrix of segment $\boldsymbol{x}_{ij}$
$\boldsymbol{a}^{att}$	attention weight with elements of a_n
$\boldsymbol{W}$	convolution weight matrix
$\mathcal{S}$, $\mathcal{R}$	skip and backbone connection output features
$*$	convolution operation
$\rightarrow$	mapping
$\otimes$	Hadamard product
$\mathbb{C}$	complex number field
$\mathbb{R}$	real number field
$\mathbb{D}$	datasets
κ	Kappa denotes convolution kernel length
V	length of the scalogram
K	number of the dilated unit
M, N	length of filterbanks features and number of the filters
I, J	number of total subjects and segments
$l, L^{(\cdot)}$	length of segment and speech
f^l, f^h	low and high cutoff frequencies
c, σ	time-domain center frequency and standard deviation
ρ	filter bandwidth
β	learnable parameter of the Kaiser windows $\vartheta(t)$
Q_c	cumulative frequency response
$F_\psi(\cdot)$	filterbanks features learning function with parameter ψ
$\mathcal{G}_\theta(\cdot)$	spectral attention learning function with parameter θ
$C_\eta(\cdot)$	classification function with parameter η
$\varphi(\cdot)$	impulse response
$H(\cdot)$	frequency response
$\mathcal{D}_d(\cdot)$	dilated convolution with factor d and function $\xi(\cdot)$
$g^e(\cdot)$, $g^o(\cdot)$	real and imaginary parts of the complex function
$\mathcal{L}(\cdot)$	loss function

Table B.1
Detail demographics of the NRAC.

Groups	Detail	NC	Depression			p
			Mild	Moderate	Severe	
PHQ	Male	13 (38.24%)	4 (18.18%)	15 (26.79%)	6 (20.00%)	0.083
	Female	21 (61.76%)	18 (81.82%)	41 (73.21%)	24 (80.00%)	
	Age	22.93±3.45	15.73±1.28	15.75±2.19	15.03±1.99	< 0.001
HAMD	Male	27 (31.40%)	4 (30.77%)	20 (32.26%)	15 (18.75%)	0.298
	Female	59 (68.60%)	9 (69.23%)	42 (67.74%)	65 (81.25%)	
	Age	22.62±1.94	16.08±1.55	15.18±1.44	15.63±2.29	< 0.001

Table B.2
Experiments on the impact of age.

Scales	Experiments	Accuracy	F1
HAMD-17	R-t	0.86±0.05	0.925±0.03
HAMD-17	A-c	0.573±0.02	0.583±0.07
PHQ-9	A-c	0.587±0.03	0.562±0.06

These samples are then validated using 5 trained DALF models to assess performance.

2. Age-classification (A-c): The NCs group is further categorized into two groups based on the average age of 22. Individuals with 13 < age $\leq$ 22 are assigned to group 0, and 25 > age > 22 are assigned to group 1. Subsequently, age classification is conducted utilizing these two groups.

Table B.2 shows the results of our experiments. Accordingly, we can conclude that significant differences in covariate age do not affect our work, which is attributed to the fact that adolescents ($\geq$14 years) have reached adult levels of physical development of the vocal cords and speech development (Demirci, Köse, Aydinli, İncebay, & Yilmaz, 2021).

References

Altwaijri, Y. A., Al-Subaie, A. S., Al-Habeeb, A., Bilal, L., Al-Desouki, M., Aradati, M., et al. (2020). 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), Article e1836.

Balestriero, R., Cosentino, R., Glotin, H., & Baraniuk, R. (2018). Spline filters for end-to-end deep learning. In *International conference on machine learning* (pp. 364–373). PMLR.

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

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 IEEE international conference on acoustics, speech and signal processing* (pp. 960–964). IEEE.

Demirci, A. N., Köse, A., Aydinli, F. E., İncebay, Ö., & Yilmaz, T. (2021). Investigating the cepstral acoustic characteristics of voice in healthy children. *International Journal of Pediatric Otorhinolaryngology*, *148*, Article 110815.

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, *18*(4), 407–422.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on multimedia* (pp. 1459–1462).

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, *47*(7), 829–837.

Fu, Q., Teng, Z., White, J., Powell, M. E., & Schmidt, D. C. (2022). Fastaudio: A learnable audio front-end for spoof speech detection. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 3693–3697). IEEE.

Gore, F. M., Bloem, P. J., Patton, G. C., Ferguson, J., Joseph, V., Coffey, C., et al. (2011). Global burden of disease in young people aged 10–24 years: a systematic analysis. *The Lancet*, *377*(9783), 2093–2102.

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., et al. (2014). The distress analysis interview corpus of human and computer interviews. In *LREC*.

Hamilton, M. (1986). The hamilton rating scale for depression. In *Assessment of depression* (pp. 143–152). Springer.

He, L., & Cao, C. (2018). Automated depression analysis using convolutional neural networks from speech. *Journal of Biomedical Informatics*, *83*, 103–111.
He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., et al. (2022). Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, *80*, 56–86.
Howie, J. M., & Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones, vol. 18*. Cambridge University Press.
Jiang, H., Hu, B., Liu, Z., Wang, G., Zhang, L., Li, X., et al. (2018). Detecting depression using an ensemble logistic regression model based on multiple speech features. *Computational and Mathematical Methods in Medicine*, *2018*.
Kessler, R. C., Avenevoli, S., & Merikangas, K. R. (2001). Mood disorders in children and adolescents: an epidemiologic perspective. *Biological Psychiatry*, *49*(12), 1002–1014.
Khan, H., & Yener, B. (2018). Learning filter widths of spectral decompositions with wavelets. *Advances in Neural Information Processing Systems*, *31*.
Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric Annals*, *32*(9), 509–515.
Lei, J., Zhu, X., & Wang, Y. (2022). BAT: Block and token self-attention for speech emotion recognition. *Neural Networks*, *156*, 67–80.
Lewinsohn, P. M., Rohde, P., Seeley, J. R., Klein, D. N., & Gotlib, I. H. (2003). Psychosocial functioning of young adults who have experienced and recovered from major depressive disorder during adolescence. *Journal of Abnormal Psychology*, *112*(3), 353.
Li, J., Tian, Y., & Lee, T. (2022). Learnable frequency filters for speech feature extraction in speaker verification. arXiv preprint arXiv:2206.07563.
Liu, H., & Ng, M. L. (2009). Formant characteristics of vowels produced by Mandarin esophageal speakers. *Journal of Voice*, *23*(2), 255–260.
López-Espejo, I., Tan, Z. H., & Jensen, J. (2021). Exploring filterbank learning for keyword spotting. In *2020 28th European signal processing conference* (pp. 331–335). IEEE.
Low, L. S. A., Maddage, N. C., Lech, M., & Allen, N. (2009). Mel frequency cepstral feature and Gaussian mixtures for modeling clinical depression in adolescents. In *2009 8th IEEE International Conference on Cognitive Informatics* (pp. 346–350). IEEE.
Ma, X., Yang, H., Chen, Q., Huang, D., & Wang, Y. (2016). 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).
Ma, S., Yang, J., Yang, B., Kang, L., Wang, P., Zhang, N., et al. (2021). The patient health questionnaire-9 vs. the hamilton rating scale for depression in assessing major depressive disorder. *Frontiers in Psychiatry*, *12*, Article 747139.
Moore, E., II, Clements, M. A., Peifer, J. W., & Weisser, L. (2008). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Transactions on Biomedical Engineering*, *55*(1), 96–107.
Morales, M. R. (2018). *Multimodal depression detection: An investigation of features and fusion techniques for automated systems*. City University of New York.
Noé, P. G., Parcollet, T., & Morchid, M. (2020). Cgcnn: Complex gabor convolutional neural network on raw speech. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7724–7728). IEEE.
Ozdas, A., Shiavi, R., Silverman, S., Silverman, M., & Wilkes, D. (2004). Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, *51*(9), 1530–1540.
Pu, J., Panagakis, Y., & Pantic, M. (2021). Learning separable time-frequency filterbanks for audio classification. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing* (pp. 3000–3004). IEEE.
Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop* (pp. 1021–1028). IEEE.
Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2022). MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, *71*, _eprint: 1909.07208.
Sainath, T., Weiss, R. J., Wilson, K., Senior, A. W., & Vinyals, O. (2015). Learning the speech front-end with raw waveform CLDNNs.
Sardari, S., Nakisa, B., Rastgoo, M. N., & Eklund, P. (2022). Audio based depression detection using Convolutional Autoencoder. *Expert Systems with Applications*, *189*(April 2021), Article 116076.
Shen, Y., Yang, H., & Lin, L. (2022). Automatic depression detection: An emotional audio-textual corpus and a GRU/BiLSTM-based model. _eprint: 2202.08210.
Shin, C., Lee, S. H., Han, K. M., Yoon, H. K., & Han, C. (2019). Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: analysis of psychiatric outpatient data. *Psychiatry Investigation*, *16*(4), 300.
Tian, X., Deng, Z., Ying, W., Choi, K. S., Wu, D., Qin, B., et al. (2019). Deep multi-view feature learning for EEG-based epileptic seizure detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(10), 1962–1972.
Tomar, S. (2006). Converting video formats with ffmpeg. *Linux Journal*, *2006*(146), 10.
Troubat, R., Barone, P., Leman, S., Desmidt, T., Cressant, A., Atanasova, B., et al. (2021). Neuroinflammation and depression: A review. *European Journal of Neuroscience*, *53*(1), 151–171.
Tukuljac, H. P., Ricaud, B., Aspert, N., & Colbois, L. (2022). Learnable filter-banks for CNN-based audio applications. In *Proceedings of the northern lights deep learning workshop, vol. 3*.
Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., et al. (2014). Avec 2014: 3D dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (pp. 3–10).
Vázquez-Romero, A., & Gallardo-Antolín, A. (2020). Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, *22*(6).
Wang, Y., Getreuer, P., Hughes, T., Lyon, R. F., & Saurous, R. A. (2017). Trainable frontend for robust and far-field keyword spotting. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 5670–5674). IEEE.
Wei, P. C., Peng, K., Roitberg, A., Yang, K., Zhang, J., & Stiefelhagen, R. (2022). Multi-modal depression estimation based on sub-attentional fusion. arXiv preprint arXiv:2207.06180.
Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., et al. (2016). Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 11–18).
Williamson, J. R., Young, D., Nierenberg, A. A., Niemi, J., Helfer, B. S., & Quatieri, T. F. (2019). Tracking depression severity from audio and video based on speech articulatory coordination. *Computer Speech and Language*, *55*, 40–56.
Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision* (pp. 3–19).
Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.
Zeghidour, N., Teboul, O., Quitry, F. d. C., & Tagliasacchi, M. (2021). LEAF: A learnable frontend for audio classification. arXiv preprint arXiv:2101.08596.
Zeghidour, N., Usunier, N., Kokkinos, I., Schaiz, T., Synnaeve, G., & Dupoux, E. (2018). Learning filterbanks from raw speech for phone recognition. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 5509–5513). IEEE.
Zeghidour, N., Usunier, N., Synnaeve, G., Collobert, R., & Dupoux, E. (2018). End-to-end speech recognition from the raw waveform. arXiv preprint arXiv:1806.07098.
Zhang, P., Wu, M., Dinkel, H., & Yu, K. (2021). Depa: Self-supervised audio embedding for depression detection. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 135–143).