Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 2, 2025

Self-Supervised Learning for Speech-Based Detection of Depressive States

Xinlin Li [1, 2], Changhe Fan [1, *] and Chengyue Su [2]

[1] Department of Psychiatry, The Affiliated Guangdong, Second Provincial General Hospital of Jinan University, Guangzhou Guangdong, 510317, China
[2] School of Physics and Optoelectronic Engineering, Guangdong University of Technology, Guangzhou Guangdong, 510006, China

*** Corresponding author:** Changhe Fan

Abstract: This study aims to enhance the accuracy of depression detection by leveraging representation learning from audio data. The data of depression speech sets are sparse and costly to annotate. Therefore, a self-supervised pre-training approach is employed to improve the performance, generalization capability, and training efficiency of downstream tasks. When processing unlabeled data, the pre-trained audio representations based on self-supervised learning may be interfered with by noisy data if there is a significant amount of noise or errors present. Consequently, it is necessary to effectively analyze long-distance sequence data to enhance anti-interference capabilities. However, traditional LSTM models have limitations in context extraction and robustness to input outliers. Thus, an improved method named CNN-BiLSTM is proposed in this paper. The network initializes the LSTM's embedding layer with pre-trained word vectors and extracts spatial and temporal features separately to ensure a full and complete expression of useful input information. Different weights are assigned based on the importance of the features to obtain fused features. Additionally, a random forest is used for classification to mitigate the risk of overfitting and to demonstrate good performance when processing high-dimensional data. Experimental results show that the proposed model exhibits good classification performance on the depression dataset, outperforming traditional methods and state-of-the-art investigations.

Keywords: Self-supervised Pre-training; CNN-BiLSTM; Depression Identification; Speech Detection.

1. Introduction

Due to the lack of effective measurable symptoms (physiological or psychological), the current assessment of depressive states is primarily diagnosed by clinical physicians through scoring. With the advancement of technology, ADE (Automatic Depression Diagnosis System) has been introduced to assist in diagnosis. Depression speech detection based on self-supervised pre-training is a method that utilizes deep learning technology to extract features from patients' speech to identify and diagnose depression. The core of this method lies in pre-training with a large amount of unlabeled data and then fine-tuning on specific downstream tasks to achieve efficient and accurate depression detection. Real-time assessment of the severity of depressive symptoms is of great significance for the diagnosis and treatment of patients with depression. In clinical practice, assessment methods mainly rely on psychological scales and doctor-patient interviews, which are time-consuming and labor-intensive. At the same time, the accuracy of the results largely depends on the subjective judgment of clinical physicians. With the development of artificial intelligence technology, an increasing number of machine learning methods diagnose depression through feature recognition. To address the issues with traditional depression detection methods, this paper starts from audio and selects the optimal path through the comparison of multiple deep learning models for depression auxiliary diagnosis effects.

2. DAIC-WOZ Dataset

Various deep learning models are trained using the speech data from the DAIC-WOZ dataset, where the voice signals are decomposed into individual frequencies and frequency amplitudes through Fourier transformation, converting the signals from the time domain to the frequency domain.

The DAIC-WOZ dataset (Distress Analysis Interview Corpus/Wizard-of-Oz set) is a corpus specifically designed for depression detection, comprising voice and text samples from 189 interviewees, with each participant provided two labels: a binary diagnosis of depression/health and the patient's eight-item Patient Health Questionnaire (PHQ-8) score. This dataset is part of the larger Disease Analysis Interview Corpus (DAIC), primarily used to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder (PTSD).

The audio portion consists of interviews conducted by an animated virtual interviewer named Ellie, with a recording frequency of 16kHz. Due to the presence of some noise and abrupt pauses in the original audio, irrelevant repetitive statements are removed during preprocessing, and features are re-extracted. Finally, the extracted speech feature dataset is divided into a training set of 70%, a validation set of 15%, and a test set of 15%.

3. Self-Supervised Pre-Training Models

A primary bottleneck in deep learning-based depression research is the limited availability and sparsity of data. In recent years, self-supervised learning has achieved success in pre-training voice signals and has been widely applied to data-sparse related tasks.

Self-supervised pre-training models can directly learn meaningful speech representations from raw audio waveforms, thereby reducing the workload of manual data annotation. This allows it to leverage a large amount of unmarked voice data for pre-training. In depression speech

detection, researchers have found significant differences in voice features between depressed patients and the normal population, such as lower pitch and slower speech rate, which can be effectively extracted and analyzed through self-supervised learning methods.

3.1. Feature Extractor

Wav2Vec2.0 is one of the most widely used self-supervised pre-training models in the field of voice signal processing. It learns effective encoding of voice signals by predicting the representations of masked parts in the audio sequence in an unsupervised setting and is widely used in downstream voice processing tasks such as speech recognition, speaker recognition, or emotion analysis.

Vq-wav2vec further improves upon wav2vec by adding a quantization module, including gumbel-softmax and K-means clustering quantization methods. This step aims to discretize the input, facilitating the use of the transformer's mask loss. The overall training process includes three steps:

1.Train vq-wav2vec.

2.Further pre-train BERT based on the discretized output of vq-wav2vec.

3.Use the BERT pre-training model as a feature extractor, and the extracted features as inputs for AM.

This paper mainly fine-tunes the vq-wav2vec model, which is used as a feature extractor for the audio files in the DAIC-WOZ dataset. The maximum voice length for vq-wav2vec is 10 seconds, and the sampling rate is fixed at 16kHz. Therefore, the original audio is segmented and preprocessed based on the start_time and stop_time in the TRANSCRIPT.csv file.

start_time	stop_time	speaker	value
411.950	413.320	Ellie	How have you been feeling lately?
414.090	417.140	Participant	Lately I've been feeling depressed.

Fig 1. Self-supervised extraction of speech information

3.2. Feature Representation

Utilizing the vq-wav2vec model, a self-supervised approach is employed to pre-train on a large number of unlabeled speech segments, followed by fine-tuning for emotion classification and English speech recognition tasks. The extracted embeddings are subjected to average pooling and dimensionality reduction along the temporal dimension, and the mean and standard deviation of the embeddings are calculated. This leverages the multi-layer feature representation capability of the self-supervised pre-training model to distinguish the characteristics of depressive state speech [4].

3.3. Hyperparameter Settings

During the self-supervised pre-training process, the model embedding dimension n is set to 1024, the self-attention layer mapping dimension m is set to 1024, the CNN-BiLSTM mapping dimension ne is 256, and the number of nodes in the fully connected layer d is 256. Adam optimizer is used for training, with a batch size of 64 and a maximum of 100 training epochs, and an initial learning rate of 0.004.

4. CNN-BiLSTM Depression State Detection

The LSTM module mitigates the issues of vanishing and exploding gradients through its gating mechanism, while the CNN-BiLSTM further improves the propagation of gradients by facilitating information flow in both forward and backward directions. This enables the model to leverage information from both past and future contexts at the current time step, thereby better capturing long-range dependencies within sequences. The proposed CNN-BiLSTM model, when confronted with noisy or incomplete data, utilizes its long-range dependency and context-aware capabilities to resist interference and make more accurate predictions by harnessing information from the entire sequence.

4.1. CNN-BiLSTM Model

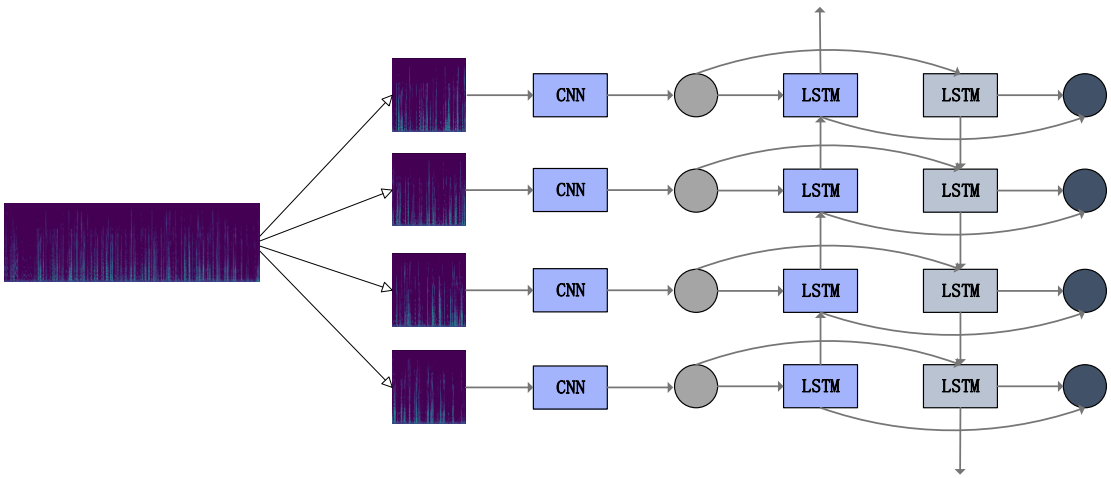

Fig 2. CNN-BiLSTM structure diagram

Each LSTM unit contains four gates (the forget gate f_t,the input gate i_t,the output gate o_t and the candidate state gate $\tilde{c}_t$):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

The LSTM unit generates its state based on the f_t and i_t of the gates:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

The hidden output in either direction.

$$h_t = o_t \odot \tanh(c_t)$$

In which, σ represents the activation function, $\tanh$ represents the hyperbolic tangent activation function, $\odot$ represents the element-wise product, W_f, W_i, W_o and W_c are the corresponding weight matrices, b_f, b_i, b_o and b_c are the bias terms, and $[h_{t-1}, x_t]$ is the concatenation of the previous time step's hidden state and the current input.

The CNN-BiLSTM model comprises two LSTM layers, one for forward propagation to obtain h_t and one for backward propagation to obtain $h_t^{backward}$, with the information from both eventually being merged by subsequent layers or the output layer of the model. Ultimately, the forward and backward hidden states can be concatenated to enhance pattern recognition capabilities and capture richer contextual information. The output of the BiLSTM at time step t is obtained as follows:

$$h_t^{BiLSTM} = [h_t, h_t^{backward}]$$

4.2. Experimental Setup

For the DAIC dataset, we employ a multi-task learning strategy to output binary classification and PHQ-8 scores. Data standardization is applied by calculating the global mean and variance on the training set and then using the mean and variance from the development set [5]. A dropout of 0.1 is taken after each LSTM layer to prevent overfitting. Results are reported based on the mean absolute error (MAE) and root mean square error (RMSE) for regression, and macro-average (classification) precision, recall, and their harmonic mean (F1) scores.

4.3. Experimental Results

As can be seen from the results in Table 1, the CNN-BiLSTM model shows an overall improvement in accuracy (Acc), precision (Pre), recall (Rec), and F1 scores compared to other models. This demonstrates that the CNN-BiLSTM model has superior performance in the analysis of depressive state speech for long sequences compared to other models.

Table 1. Comparison of Results Without Self-Supervised Pre-Training

Method	Acc	Pre	Rec	F1
RNN	0.634	0.595	0.654	0.623
GRU	0.703	0.639	**0.842**	0.727
LSTM	0.771	0.775	0.797	0.786
CNN-BiLSTM	**0.790**	**0.807**	0.815	**0.810**

The results presented in Table 2 indicate that after employing the method of self-supervised pre-training, there is a noticeable improvement in accuracy, precision, recall, and F1 scores for each model overall. Moreover, the CNN-BiLSTM model continues to outperform other models in terms of overall performance.

Table 2. Comparison of Results Using Self-Supervised Pre-Training

Method	Acc	Pre	Rec	F1
Self-supervised +RNN	0.651	0.633	0.654	0.743
Self-supervised +GRU	0.752	0.706	0.837	0.767
Self-supervised +LSTM	0.806	0.772	0.851	0.829
Self-supervised +CNN-BiLSTM	**0.863**	**0.851**	**0.868**	**0.860**

The results in Table 3 indicate an overall trend: all models perform slightly better on female samples than on male samples. As the complexity of the models increases (from RNN to CNN-BiLSTM), the performance difference between genders appears to decrease. Model-specific analysis: RNN: Performs significantly better on female samples than on male samples, with the largest performance difference. GRU: The gender difference remains noticeable but is reduced compared to the RNN.LSTM: The performance difference between genders is further reduced. CNN-BiLSTM: Shows the smallest gender difference, indicating that this model is more stable in processing speech features of different genders. Self-supervised pre-training + CNN-BiLSTM: Although there is still a slight gender difference, it is very small compared to other models. The slightly better performance on female samples in the experimental results of Table 3 may be related to the fact that females are more likely to express emotions, making speech features more pronounced. Complex models (such as CNN-BiLSTM) may be better at capturing speech features of different genders, thus resulting in smaller performance differences. Self-supervised pre-training further reduces gender differences, possibly because it can learn more universal speech feature representations.

Table 3. Comparison of Training Results by Gender Grouping

Method	Gender	Acc	Pre	Rec	F1
RNN	Male	0.620	0.585	0.640	0.611
	Female	0.648	0.605	0.668	0.635
GRU	Male	0.695	0.630	0.835	0.718
	Female	0.711	0.648	0.849	0.736
LSTM	Male	0.763	0.768	0.790	0.779
	Female	0.779	0.782	0.804	0.793
CNN-BiLSTM	Male	0.782	0.800	0.808	0.804
	Female	0.798	0.814	0.822	0.818
Self-supervised +CNN-BiLSTM	Male	0.855	0.844	0.861	0.852
	Female	0.871	0.858	0.875	0.866

5. Summary

This paper proposes an audio embedding pre-training method for the automatic detection of depression. Features are extracted using a self-supervised pre-training approach and then input into a CNN-BiLSTM network for depressive state detection. Data pre-trained on the DAIC dataset demonstrate that the results from the CNN-BiLSTM network with self-supervised embedding pre-training are significantly better compared to those without self-supervised pre-training and other methods that do not employ this technique. For similar tasks that require summarizing long sequences into a given single output label, such as most medical speech tasks, the use of a CNN-BiLSTM model with self-supervised pre-training can be a universal method.

Acknowledgments

Talents' plan Foundation of Guangdong Second Provincial General Hospital (2024C003)

Science and Technology Project (No. 202102010115)

Guangdong Yiyang Healthcare Charity Foundation (No. JZ2022001-3)

References

[1] Niizumi D, Takeuchi D, Ohishi Y, et al. BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation [Conference Proceedings] // 2021 INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN). 2021 [accessed 2022-01-07]. DOI:10. 1109/ IJCNN52387.2021.9534474.

[2] Zhang P, Wu M, Dinkel H, et al. DEPA: Self-Supervised Audio Embedding for Depression Detection [Conference Proceedings] // PROCEEDINGS OF THE 29TH ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, MM 2021. 2021: 135-143 [accessed 2021-01-01]. DOI:10.1145/3474085.3479236.

[3] Sun L, Lian Z, Liu B, et al. HiCMAE: Hierarchical Contrastive Masked Autoencoder for self-supervised Audio-Visual Emotion Recognition [Journal Article] // INFORMATION FUSION, 2024, 108 [accessed 2024-05-20]. DOI: 10.1016/j. inffus. 2024.102382.

[4] Gong X, Duan H, Yang Y, et al. Improving Audio Classification Method by Combining Self-Supervision with Knowledge Distillation [Journal Article] // ELECTRONICS, 2024, 13(1) [accessed 2024-01-29]. DOI:10.3390/ electronics 13010052.

[5] Liu A H, Glass J R, Gan C, et al. Method for self-supervised speech recognition through sparse subnetwork discovery in pre-trained speech self-supervised learning, involves pruning weights of lowest magnitude in new subnetwork regardless of network structure to satisfy target sparsity: US2023360642-A1 [Patent]. [2023-11-20].