Cross-Modal Attention Fusion of Electrocardiogram Emotion and Voiceprint for Depression Detection

Minghui Zhao[1], Lulu Zhao[1*], Hongxiang Gao[1], Keming Cao[1], Feifei Chen[1], Zhijun Xiao[1], Zhaoyang Cong[1], Jianqing Li[1], Chengyu Liu[1*]

[1] State Key Laboratory of Digital Medical Engineering, School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China

Abstract

Depression is one of the most important psychiatric disorders diseases. Early detection and intervention have a better impact on treatment outcomes. Depression symptoms typically manifest from ECG emotion and voiceprint. However, due to the heterogeneity, it is challenging to extract effective representations from various modalities. To address this issue, we propose a Lstm-based sinc network (Sinc-Lstm) to extract time-frequency domain features from both ECG and speech. Then, we employ a cross-modal attention mechanism to fuse these features based contrastive learning. Finally, we collect EEG and speech dataset for depression detection (ESDD). Comparative experiments were conducted on ECG emotion dataset (WESAD) and ESDD. Emotional classification accuracy on the WESAD dataset is 0.95, with an F1 score of 0.98. Depression classification accuracy on the ESDD dataset is 0.81, with an F1 score of 0.75. The results show that the crossed-attention mechanism along the temporal dimension effectively aggregates relevant features from various time periods, thereby realizing accurate and comprehensive depression assessment.

1. Introduction

Depression is one of the most important psychiatric disorders diseases. Early detection and intervention have a better impact on treatment outcomes. However, there is currently a lack of objective and effective detection methods. Currently, emotion recognition and multimodal fusion are widely applied in medical health. For depression detection tasks, effectively leveraging emotional features and integrating the depressive characteristics of different modalities can enhance the effectiveness and interpretability of depression detection.

Most emotion detection methods rely on behavioral recognition approaches like speech and facial recognition. However, these features can be easily masked through subjective control, making them insufficient for accurately identifying emotions [1]. In contrast, physiological signals, regulated by the autonomic nervous and endocrine systems, offer objectivity, authenticity, and reliability. Among these, electrocardiogram (ECG)-based strategies [2] are comfortable and convenient, suitable for mobile applications, wearable devices, and remote monitoring [3]. Users can effortlessly monitor their emotions in daily life. Speech, being more privacy-preserving than facial expressions or gestures, makes the effective fusion of ECG and speech signals particularly important.

Considerable research has been conducted in the field of ECG emotion recognition. Current approaches generally fall into three categories: the two-dimensional model (Valence-Arousal), the three-dimensional model (Valence-Arousal-Dominance), and discrete emotion classification. ECG signals contain rich physiological information, including time-domain characteristics, heart rate variability, and morphology features [4]. Classification algorithms used include KNN, SVM, random forests, and XGBoost. In deep learning, Behnam Behinaein introduced a neural network architecture combining convolutional layers and Transformer [5], while Ross Harper proposed emotion valence classification from unimodal ECG time series using a Bayesian framework [6]. Speech, a key medium for expressing emotions, shows distinct features in depression patients, such as reduced and sluggish speech, low mood, indecisiveness, and crying during interviews [7]. Exploring its integration methods is crucial [8].

Therefore, this study investigates depression recognition based on ECG emotion and voiceprint features. The main contributions are summarized as follows:

1. We propose a LSTM-based sinc [9] network (Sinc-Lstm) to extract time-frequency domain features from both ECG and speech signals, enabling the acquisition of aligned feature encodings from different modalities.

2. To integrate the complementary information from ECG and speech, a cross-modal attention mechanism is employed based on emotion and voiceprint pre-training,

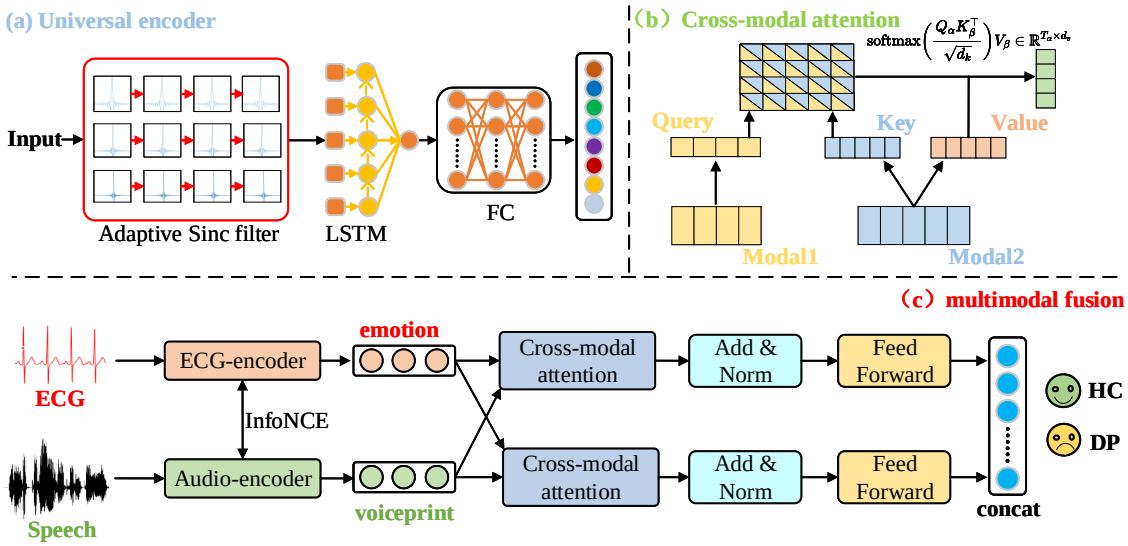

Figure 1. The architecture of cross-modal Attention fusion of ECG Emotion and voiceprint. (a) Unified ECG and voiceprint feature encoder. (b) Cross-modal attention mechanism. (c) Modality interaction block.

enhancing the interpretable and robustness of emotion recognition.

3. We collected datasets for depression detection involving both ECG and speech (ESDD). We conduct comprehensive experiments on public ECG emotion dataset (WESAD) and ESDD, obtaining superior or comparable results.

2. Proposed Method

To address the heterogeneity of ECG and speech, we employ the same architecture encoder, Sinc-Lstm, to extract time-frequency feature representations from both signals. For ECG signals, we conduct pre-training on emotion datasets, while for speech signals, we pre-train using voiceprint features. Subsequently, we conduct contrastive learning, utilizing intra-modal contrastive loss to extract shallow depression representations from each signal and align them within a unified feature space. Then, we employ a cross-modal attention mechanism to fuse these representations, enhancing the mutual constraints between different modal representations. The overall model architecture is illustrated in Figure 1.

In this study, we employ an LSTM-enhanced Sinc filter network, called Sinc-Lstm, which compels the first convolutional layer to discover more meaningful filters, as depicted in Figure 1 (a). SincNet is based on parameterized sinc functions to implement bandpass filters [9]. Unlike standard neural networks that learn all elements of each filter, our proposed method learns only the low cutoff frequency and high cutoff frequency directly from the data:

$$g[n, f_1, f_2] = 2f_2\text{sinc}(2\pi f_2 n) - 2f_1\text{sinc}(2\pi f_1 n), \quad (1)$$

The sinc function is defined as $\text{sinc}(x) = \sin(x)/x$. Subsequently, for subsequent features, integration is performed through recurrent neural networks:

$$\boldsymbol{x}_{\text{lstm}} = \text{LSTM}(\boldsymbol{x}_{\text{sinc}}), \quad (2)$$

where, LSTM refers to Long Short-Term Memory neural network. It can better capture important narrowband speaker characteristics in the speech domain, such as pitch and resonance peaks, as well as other audio features. Combining this filter with recurrent neural networks enables further extraction of ECG time-frequency domain features based on the proposed filter, leading to faster convergence and achieving natural inductive biases. Moreover, the learned filter parameters exhibit stronger interpretability, as the Sinc feature maps obtained in the first convolutional layer are understandable and match the manually extracted frequency domain features.

Figure 1 (b), (c)illustrates the self-supervised and fusion training process for multimodal signals. Initially, contrastive learning pre-training is conducted on the encoders for both ECG and speech, employing infoNCE to drive better alignment and understanding of features across modalities, thus establishing intrinsic connections between them. Subsequently, a cross-modal attention mechanism [10] is employed. This mechanism establishes precise and effective information transmission channels between signals of different modalities, enabling the model to dynamically focus on and extract key information segments based

on the causal relationships and dependencies between different modalities. In the cross-modal attention mechanism, for an input vector $\boldsymbol{x}_m \in \mathbb{R}^{t_s \times d_s}$, where $m \in (E, A)$ indicates ECG and audio,t_s, d_s are the dimension of time and frequency.

$$\text{Attention}(\boldsymbol{Q}_E, \boldsymbol{K}_A, \boldsymbol{V}_A) = \text{softmax}\left(\frac{\boldsymbol{Q}_E \boldsymbol{K}_A^T}{\sqrt{d_s}}\right)\boldsymbol{V}_A, \tag{3}$$

where $\boldsymbol{Q}_m$, $\boldsymbol{K}_m$, and $\boldsymbol{V}_m$ correspond to queries, keys, and values respectively, representing different modalities. This enables effective integration of depression-related information from different modalities while maintaining their consistency and complementarity. Ultimately, this approach facilitates depression recognition based on emotion and voiceprint features.

3. Experimental Evaluation

WESAD: A dataset for monitoring psychological stress and emotion based on wearable devices. This dataset records physiological and motion data from 15 subjects using wearable devices worn on the wrist and chest. For the purpose of this project, the ECG signals are utilized, sampled at 700Hz, with labels defined as follows: transient (0), baseline (1), stress (2), amusement (3), meditation (4). After data preprocessing and segmentation, the proportions of the corresponding ECG segments are 48.39%, 20.36%, 11.32%, 6.42%, and 13.50%. The emotion classification results are reported based on the segmented samples.

ESDD: We have collected a new ECG and speech depression dataset to facilitate research on depression detection. It comprises 18 depression and 22 normal controls (NC). All volunteers have signed informed consent forms and have assured the authenticity of the provided information. Each volunteer completed the PHQ-9 questionnaire, answering six questions, including reading aloud, positive, neutral, and negative questions, along with two TAT assessments.

Implementation Details: The sampling rate of the ECG is uniformly converted to 250 Hz. The sampling rate of the speech signals is uniformly converted to 16000 Hz, single-channel. The experiments are implemented using PyTorch 1.13.1 on an NVIDIA GeForce RTX 3090. To evaluate the model performance, metrics such as accuracy (ACC), precision (PRE), recall (REC), F1 score, and area under the curve (AUC) are utilized. The final individual predictions are obtained through segment voting. Specifically, F1 score, precision, and recall scores are averaged over the depression and NC categories, while the AUC score is computed considering depression as the positive class.

For ECG emotion features, we first conduct pre-training on the WESAD dataset, comparing the proposed Sinc-Lstm model with the AST model [11]. The effectiveness of the emotion recognition model is validated using a dependent subject approach though Hold-out Method. Different fixed window lengths and window functions of short-time fourier transform (STFT) are applied to the WESAD dataset to obtain corresponding ECG time-frequency spectrograms. These spectrograms are then input into the E-AST to obtain the final classification results. From Table 1, it can be observed that the model prediction accuracy is low when using undenoised time-frequency spectrograms. Moreover, when comparing different window lengths and window functions, it is found that window length has a greater impact, although the model F1 scores vary minimally. Subsequently, network comparison experiments are conducted, where raw data is cropped into 10-second segments as input to the model, resulting in higher accuracy and F1 scores, all attributed to the AST model with fixed STFT variations. Figure 2 shows the feature distribution of the Sinc filter, reduced in dimensionality using t-SNE. The Sinc-Lstm network exhibits better recognition performance. For emotion recognition based on ECG, longer ECG signals (approximately 10 seconds) are preferred. Moreover, compared to the results obtained using the fixed-resolution E-AST model, ECG emotion recognition tasks require the ability to effectively extract features at different frequency resolutions to achieve better results. Ultimately, comparative experiments were conducted on different cross-modal fusion approaches by integrating electrocardiogram emotion features with voiceprint features in signal processing. In the case of subject-independent analysis, we conduct experi-

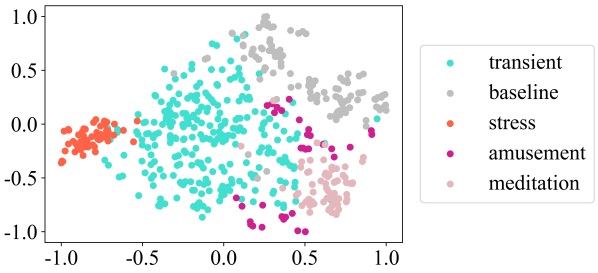

Figure 2. t-SNE visualization of Sinc-Lstm ECG emotion features.

Table 1. Emotional classification results of Sinc-Lstm and E-AST models on the WESAD dataset (%). **Bold** is the best.

Method	Window	Length	Padding	ACC	REC	PRE	F1
E-AST	Hanming(raw)	128	256	63.49	93.26	31.49	47.08
	Hanming	128	256	78.52	98.14	30.32	46.32
	Rectangular	128	256	77.23	97.79	30.34	46.31
	Hanming	256	256	73.31	90.36	31.90	47.15
Sinc-Lstm	×	×	×	**95.15**	**98.01**	**98.80**	**98.41**

ments using five-fold cross-validation. The ECG encoder utilizes weights from a Sinc filter pre-trained on the WESAD dataset to capture emotion and stress variation features in the experiments. The modal fusion was performed using three approaches: fully connected networks, graph attention networks, and cross-modal attention. The results are presented in Table 2. It is evident that utilizing only speech for depression detection yields an AUC of 0.6265 and an F1 score of 0.6154. However, the fusion of ECG and speech outperforms depression detection using solely speech, as indicated by various evaluation metrics. Furthermore, employing a cross-modal attention mechanism achieves the optimal AUC of 0.8527 and an F1 score of 0.7556. This validates that adopting a cross-modal attention along the temporal dimension of ECG and speech features effectively aggregates relevant features from different time periods, thereby realizing more accurate depression detection.

Table 2. Depression classification results using different fusion methods on the ESDD dataset (%). **Bold** is the best.

Method	Modal	ACC	F1	AUC	PRE	RECALL
Sinc-Lstm	Speech	72.50	61.54	62.65	67.67	62.05
FC	ECG Speech	75.50	67.04	84.37	66.63	68.14
GAT	ECG speech	76.00	70.57	83.78	71.35	73.00
Cross-modal attention	ECG speech	**81.00**	**75.56**	**85.27**	**79.28**	**75.85**

4. Conclusion

Based on the emotional features extracted from ECG signals and voiceprints, this paper proposes a cross-modal attention fusion method for depression detection. It employs the same architecture encoder, Sinc-Lstm, to extract the time-frequency feature representations of the signals. Subsequently, fusion is achieved through contrastive learning and cross-modal attention. Extensive experiments are conducted on the WESAD dataset for ECG dataset and a self-built ESDD dataset. The accuracy of emotional classification for ECG signals reaches 0.9515, with an F1 score of 0.9841, validating the effectiveness of the model in extracting emotional features. In the modal fusion experiments on the ESDD dataset, comparing the fusion methods of fully connected networks, GAT, and cross-modal attention, all yield optimal results. Ultimately, by integrating ECG and speech signals, the method obtains rich and comprehensive feature representations, enhancing the accuracy and interpretable of depression recognition.

5. Acknowledgement

This research was funded by the National Natural Science Foundation of China (62171123), and the National Key Research and Development Program of China (2023YFC3603600, 2022YFC2405600).

References

[1] Shoumy NJ, Ang LM, Seng KP, Rahaman D, Zia T. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. Journal of Network and Computer Applications 2020;149:102447. ISSN 1084-8045.

[2] Zhang S, Li Y, Wang X, Gao H, Li J, Liu C. Label decoupling strategy for 12-lead ecg classification. Knowledge Based Systems 2023;263:110298.

[3] Lopez E, Chiarantano E, Grassucci E, Comminiello D. Hypercomplex multimodal emotion recognition from eeg and peripheral physiological signals. In 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). 2023; 1–5.

[4] Hasnul MA, Aziz NAA, Alelyani S, Mohana M, Aziz AA. Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review. Sensors 2021; 21(15). ISSN 1424-8220.

[5] Behinaein B, Bhatti A, Rodenburg D, Hungler P, Etemad A. A transformer architecture for stress detection from ecg. In Proceedings of the 2021 ACM International Symposium on Wearable Computers, ISWC '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384629, 2021; 132–134.

[6] Harper R, Southern J. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. IEEE Transactions on Affective Computing 2022;13(2):985–991.

[7] Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Communication 2015; 71:10–49. ISSN 0167-6393.

[8] Zhang S, Li JQ, Fujita H, Li YW, Wang DB, Zhu TT, Zhang ML, Liu CY. Student loss: Towards the probability assumption in inaccurate supervision. IEEE Transactions on Pattern Analysis and Machine Intelligence 2024;.

[9] Ravanelli M, Bengio Y. Speaker recognition from raw waveform with sincnet. In Proc. IEEE Spoken Lang. Technol. Workshop. 2018; 1021–1028.

[10] Zong D, Ding C, Li B, Li J, Zheng K, Zhou Q. Acformer: An aligned and compact transformer for multimodal sentiment analysis. In Proceedings of the 31st ACM International Conference on Multimedia. 2023; 833–842.

[11] Gong Y, Lai CI, Chung YA, Glass J. Ssast: Self-supervised audio spectrogram transformer. In Proc. Conf. AAAI. Jun. 2022; 10699–10709.

Address for correspondence:

Lulu Zhao
Sipailou Road 2, Southeast University, Nanjing, China
Email: 101300153@seu.edu.cn
Chengyu Liu
Sipailou Road 2, Southeast University, Nanjing, China
Email: chengyu@seu.edu.cn