IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2024EDL8034

Publicized:2024/07/22

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

Multimodal Speech Emotion Recognition Based on Large Language Model

Congcong FANG[†], **Yun JIN**[†*], **Guanlin CHEN**[†], **Yunfan ZHANG**[†], **Shidang LI**[†], **Yong MA**[††], ***Nonmembers***, ***and*** **Yue XIE**[†††], ***Member***

SUMMARY Currently, an increasing number of tasks in speech emotion recognition rely on the analysis of both speech and text features. However, there remains a paucity of research exploring the potential of leveraging large language models like GPT-3 to enhance emotion recognition. In this investigation, we harness the power of the GPT-3 model to extract semantic information from transcribed texts, generating text modal features with a dimensionality of 1536. Subsequently, we perform feature fusion, combining the 1536-dimensional text features with 1188-dimensional acoustic features to yield comprehensive multi-modal recognition outcomes. Our findings reveal that the proposed method achieves a weighted accuracy of 79.62% across the four emotion categories in IEMOCAP, underscoring the considerable enhancement in emotion recognition accuracy facilitated by integrating large language models.

key words: *Emotion recognition, GPT-3, Multimodal recognition*

1. Introduction

Affective computing, a concept pioneered by Professor Picard in 1997[1], comprises four key stages: signal acquisition, emotion recognition, emotion understanding and feedback, and emotion expression[2]. Emotion recognition encompasses diverse modalities such as speech, text, images, and videos. Each modality independently conveys emotions, and leveraging these modalities to extract emotion features for human emotion recognition has been a foundational approach in early emotion recognition technology. Initially, emotion recognition tasks predominantly centered on processing features from the speech modality to achieve emotion recognition[3][4][5]. However, with technological advancements and growing privacy concerns, the sensitive nature of speech has propelled the text modality to the forefront as a predominant and viable option. Researchers have pivoted their attention towards extracting emotion-related features from transcribed text using various tools, thus propelling the field of emotion recognition to new horizons[6][7][8][9].

Current pre-training methods in natural language processing still require fine-tuning for downstream tasks, demanding a large volume of task samples. In contrast, humans can readily tackle new language tasks with just a few samples. GPT[10] (Generative Pre-trained Transformer), a series of pre-trained language models rooted in the Transformer architecture developed by OpenAI, emerges as the most prevalent and commercially successful model in natural language processing. GPT-3, unveiled in 2020, signifies the third iteration in the GPT series. Boasting 175 billion parameters and adjustable weights, GPT-3 stands as the most sophisticated and potent version of the model to date.

In its model structure, GPT-3 maintains the model architecture of GPT while integrating the sparse attention module from the Sparse Transformer. Sparse attention diverges from conventional self-attention in that each token only engages in attention computation with a subset of other tokens, leading to a complexity of n*logn. To be precise, sparse attention sets the attention to 0 for all tokens except for those within a relative distance of k and multiples of k.

The advantages of employing sparse attention are twofold: Firstly, it diminishes the computational complexity of the attention layer, conserving memory and time. Consequently, it facilitates the processing of longer input sequences. Secondly, it demonstrates the trait of ”local tight correlation and remote sparse correlation,” prioritizing closely related contexts over distant ones.

Due to the unique structure of the attention mechanism, models can assign varying levels of importance to each word in a sentence based on the current task, making them well-suited for tasks such as emotion recognition and related endeavors[11]. GPT-3 has demonstrated remarkable efficacy in several domains[12]:

- **Zero-shot learning**: GPT-3 exhibits prowess in zero-shot learning, wherein language models can be applied to downstream tasks like translation and text summarization without necessitating additional task-specific data.
- **Encoding rich semantic knowledge**: GPT-3 adeptly encodes comprehensive semantic knowledge about the world, generating learned representations, typically fixed-size vectors, which are valuable for discriminative tasks.

Although GPT-3’s applications in emotion analysis tasks are relatively scarce, this paper utilizes text embeddings produced by the model for recognition purposes. Studies indicate that text embeddings derived from GPT-3 can reliably serve emotion analysis tasks, surpassing traditional emotion analysis methods and even rivaling fine-tuned models[13]. In essence, text embeddings based on the GPT-3 model present

[†]The author is with the School of Physics and Electronic Engineering, Jiangsu Normal University

[††]The author is with the School of Linguistic Sciences and Arts, Jiangsu Normal University

[†††]The author is with the School of Information and Communication Engineering, Nanjing Institute of Technology

[*]Presently, the author is with the Corresponding Author

a highly viable option for emotion analysis tasks and exhibit substantial potential for applications in this domain.

Building upon the superior performance of GPT-3, this paper delves into its potential to predict emotions from both speech and transcribed text, leveraging the extensive semantic knowledge embedded within the model. The efficacy of this approach is substantiated through multimodal fusion experiments, ultimately validating its effectiveness.

2. Methodology

The main method outlined in this paper is depicted in Figure 1, encompassing three primary components.

- The acoustic feature module (highlighted in blue) is tasked with providing acoustic features.
- The text feature module (highlighted in yellow) is responsible for extracting text features from the transcribed text.
- The feature fusion module (highlighted in green) merges the corresponding features obtained from the first two modules.

In Fig 1, we illustrate the two distinct procedures used for text modes. For training process, the training set includes transcriptions. Textual features are directly extracted from these transcripts during the training phase. For testing process, we simulate real-world scenarios by assuming the testing set only contains speech data, without accompanying transcripts. Therefore, an Automatic Speech Recognition (ASR) system is first employed to convert the speech into text. GPT-3 then extracts textual features from this generated text.

Subsequently, these fused features are inputted into a classifier for the final sentiment recognition. The ensuing subsections will furnish more comprehensive details regarding each module.

2.1 Text feature extraction

The fundamental approach of this paper revolves around acquiring text embeddings using GPT-3. To access the GPT-3 embeddings model, we utilize an endpoint provided in the OpenAI API, which is accessible to registered researchers. Specifically, in our research, we employ the "text-embeddings-ada-002" model to encode the transcribed text within the corpus.

These choices enable us to capture lexical, syntactic, and semantic attributes that are closely associated with emotions, particularly in lengthy statements. By leveraging these attributes, we aim to enhance the accuracy of emotion recognition tasks. The utilization of GPT-3 embeddings allows us to encapsulate rich linguistic information that can significantly contribute to the recognition and interpretation of emotions embedded within text data.

2.2 Feature fusion

To assess whether acoustic features and text embeddings

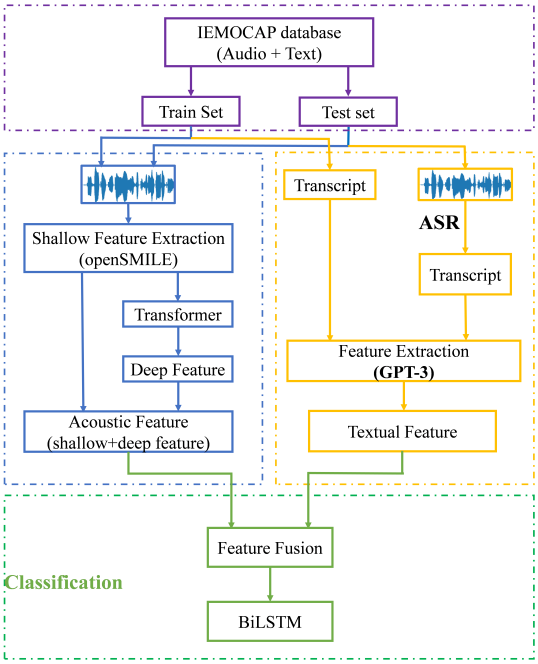


Fig. 1 Overview structure of the proposed method

can synergize to improve emotion classification, we merge acoustic features from voice audio data with GPT-3-based text embeddings to create multimodal emotion features. In our study, the acoustic features comprise 988-dimensional features extracted from the emobase configuration file using the openSMILE toolbox[14]. Subsequently, 200-dimensional depth features pertaining to emotion are extracted from the initial 988-dimensional features utilizing a Transformer encoder structure, resulting in a total of 1188 dimensions. In contrast, the text features, derived from GPT-3, possess a dimensionality of 1536 for transcribed text. The final fusion feature combining both modes totals 2724 dimensions.

Given the disparate dimensions of the acoustic and text features, we adopt a feature fusion approach. Specifically, we concatenate the acoustic features and text features, creating a singular high-dimensional feature vector as the final fused feature representation. This fused feature vector is then inputted into a BiLSTM network for recognition, ultimately yielding the final emotion recognition results.

By integrating information from both acoustic and text modalities, our method aims to capitalize on the complementary nature of these features, enhancing the model's ability to discern and interpret emotional content within the data. The fusion of acoustic and text features enables a more comprehensive representation of emotional cues, thereby improving the accuracy of emotion recognition tasks.

3. Experiments

3.1 Dataset

The experimental evaluations conducted in this paper uti-

lize the publicly available IEMOCAP emotion dataset [15]. This dataset encompasses a total of 10 emotion categories, including happiness, sadness, anger, neutral, excitement, surprise, disgust, frustration, fear, and other. However, for the purposes of classification in this study, four emotions have been selected. To maintain balanced data distribution, the emotions of happiness and excitement are merged into a single category. Consequently, the final experimental dataset comprises 5,531 utterances, with the following class distribution: anger: 19.9%, happiness/excitement: 29.5%, sadness: 19.5%, and neutral: 30.8%. This strategic amalgamation ensures a more equitable representation of emotional categories within the dataset, thereby enhancing the robustness and reliability of the experimental analyses conducted in the paper.

3.2 Parameter setting

In the multimodal fusion experiment, the original dataset is partitioned into 10 sections on average. Among these, 7 sections are allocated for training, 2 for testing, and 1 for validation. The BiLSTM network[16], consisting of both forward and backward LSTM layers, is employed. Additionally, a local attention mechanism is integrated to focus on segments of speech containing strong emotional information, mitigating the impact of uneven distribution of emotional features on the experimental outcomes.

Before the representation is passed to the final output layer, a dropout with a probability of 0.5 is applied to prevent overfitting. RMSprop is selected as the optimizer, while the ReLU function serves as the activation function before the fully connected layer. For the final output layer responsible for classification prediction, the softmax function is chosen as the activation function[17]. The classifier predominantly utilizes a bidirectional Long Short-Term Memory network (BiLSTM) comprising 200 neurons, with 100 nodes in the forward direction and 100 nodes in the backward direction. The training batch size is set to 64. Two commonly used evaluation metrics, Weighted Accuracy (WA) and Unweighted Accuracy (UA), are employed to assess the performance of the model. These metrics offer comprehensive insights into the model's effectiveness in emotion recognition tasks.

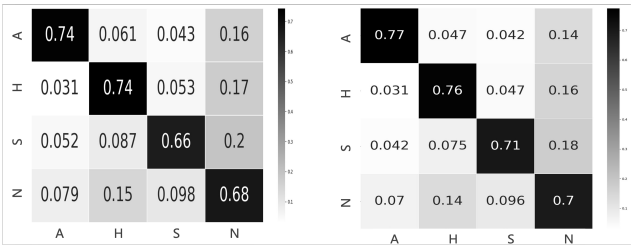
3.3 The text modality experiments

In the text modality, this paper predominantly employs two major language models to extract text features for emotion recognition. The first method involves utilizing the classical BERT model for extracting text features. The comparative experiment, which is the primary focus of this study, involves leveraging GPT-3 to extract text features for emotion recognition. The final recognition results for these two sets of experiments are presented in Table 1.

From Table 1, it's evident that using the same BiLSTM network, employing BERT to extract 768-dimensional text features results in a recognition accuracy with a WA of 68.78% and UA of 68.69%. Compared with that, utilizing GPT-3 to extract text features and inputting them into the same classifier can significantly improve the results. The WA is 72.84%, and UA is 72.90%, marking an enhancement of 4.15% and 4.21% in WA and UA, respectively, compared to the former method. The experimental results validate the effectiveness of the proposed approach using GPT-3 for text feature extraction in emotion recognition.

Table 1 The text modality experiments

Methos	WA(%)	UA(%)
BERT(768dim)	68.78	68.69
GPT-3(1536dim)	**72.84**	**72.90**



(a): BERT (b): GPT-3

A:angry H:happy S:sad N:neutral

Fig. 2 Confusion matrices for single text modality

The confusion matrices for the two experimental results presented in Table 2 are depicted in Figure 2. In Figure (a), the emotion recognition results are illustrated using text features extracted by BERT, while in Figure (b), the emotion recognition results using text features extracted by GPT-3 are displayed. From the graphs, it can be observed that the method employing GPT-3 for text feature extraction achieved accuracies of 77%, 76%, 71%, and 70% for the angry, happy, sad, and neutral categories, respectively. This represents an improvement of 3%, 2%, 5%, and 2%, respectively, compared to the method utilizing BERT for text feature extraction in the four emotion recognition categories. These findings highlight the overall superior performance of the proposed approach, leveraging the large-scale GPT-3 model for text feature extraction, over the use of BERT in handling text for the four emotion recognition tasks. Consequently, the effectiveness of the approach proposed in this paper is validated, emphasizing the significance of selecting appropriate language models for text feature extraction in emotion recognition tasks.

3.4 The multimodal experiments

In this section, the acoustic features and text features described in Section 2.2 are combined through feature fusion and inputted into the BiLSTM network for emotion recognition.The multimodal recognition results using BERT-extracted text features serve as the baseline system for this

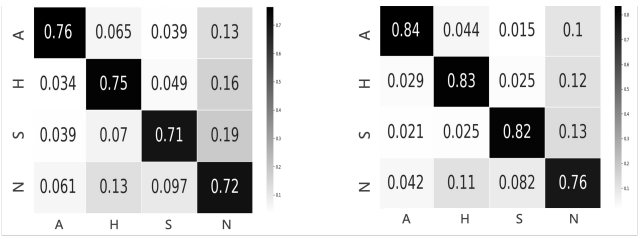
experiment. A comparison is made with the main recognition method proposed in this study and the recognition results obtained by other researchers in recent studies to validate the effectiveness of the system. The final experimental results are presented in Table 2.

Table 2 Experimental results of multimodal fusion

Methos	WA(%)	UA(%)
Fusing Pairwise Modalities for ERC[18]	69.57	69.34
BiGMF[19]	70.43	70.33
Key-Sparse Transformer[20]	74.3	75.3
Baseline system(1956dim)	71.01	71.69
Emobase+GPT-3(2724dim)(Ours)	**79.62**	**80.38**

The experiments detailed in Table 2 were exclusively conducted using the IEMOCAP dataset, showcasing that the model proposed in this paper surpasses other models in emotion recognition tasks. Under identical conditions, leveraging GPT-3 to extract text features significantly enhances the accuracy of multimodal emotion recognition. The Weighted Accuracy (WA) achieves 79.62%, and the Unweighted Accuracy (UA) reaches 80.38%. Comparing these results to the final recognition rates obtained using the BERT model to extract text features in the baseline system of our study, the WA increases by 8.61%, and the UA increases by 8.69%.

The forthcoming figure will display the confusion matrices for both the baseline system and the method proposed in this paper, offering deeper insights into the classification performance and the distribution of predicted classes.



(a): BERT+acoustic feature (b): GPT+acoustic feature
A:angry H:happy S:sad N:neutral

Fig. 3 Confusion matrices for multimodal experiments

Figure 3 illustrates the confusion matrix resulting from the fusion of text features and acoustic features extracted by BERT (Figure (a)), as well as the confusion matrix resulting from text features and acoustic features extracted by GPT-3 (Figure (b)).

Upon comparing the two, it becomes evident that in terms of recognition rates for the four emotion categories, the method employing GPT-3 for feature extraction outperforms the method using BERT for feature extraction. Specifically, the accuracies for the angry, happy, sad, and neutral categories using GPT-3 are 84%, 83%, 82%, and 76%, respectively. This represents an improvement of 8%, 8%, 11%, and 4%, respectively, compared to the accuracies obtained by utilizing BERT for feature extraction in the recognition of the four emotion categories.

These results provide further evidence of the effectiveness of the proposed approach, highlighting the superiority of GPT-3 in extracting text features for emotion recognition tasks. The enhanced performance achieved by leveraging GPT-3 underscores its capability to capture semantic nuances and improve the accuracy of emotion recognition systems when compared to other language models such as BERT.

4. Conclusion

In this study, we utilized the existing large language model GPT-3 to extract text features from transcription texts. By integrating these text features with the best acoustic features obtained from acoustic experiments, we achieved optimal results in multimodal experiments. The recognition accuracy significantly surpassed the baseline system in this paper, which employed the BERT model for text feature extraction. Moreover, it exhibited clear superiority over most existing bimodal recognition systems. These results provide evidence for the excellence of the proposed main methodology in this paper.

References

[1] Picard R W . Affective Computing[J]. technical report, 1997.
[2] Zhang YH, Lin XZ. Emotion can be calculated: A review of Emotion Computing [J]. Computer Science, 2008, 35(5):4.
[3] Liu S , Zhang M , Fang M , et al. Speech emotion recognition based on transfer learning from the FaceNet framework(a))[J]. The Journal of the Acoustical Society of America, 2021,149(2):1338-1345.
[4] Issa D , Demirci M F , Yazici A . Speech emotion recognition with deep convolutional neural networks[J]. Biomedical Signal Processing and Control, 2020, 59:101894.
[5] Batbaatar E , Li M , Ryu K H . Semantic-Emotion Neural Network for Emotion Recognition from Text[J]. IEEE Access, 2019, 7:111866-111878.
[6] PoRiA S, MAJuMDER N, HAzARikA D, et al. Multimodal sentiment analysis: addressing key issues and setting up the baselines [J]. IEEE Intelligent Systems, 2018, 33(6): 17-25.
[7] G. Sahu. Multimodal speech emotion recognition and ambiguity resolution. CoRR, abs/1904.06022, 2019.
[8] PENG Zixuan, LU Yu, PAN Shengfeng, et al. Efficient speech emotion recognition using multi-scale CNN and attention[C]//ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, ON, Canada. IEEE, 2021: 3020-3024.
[9] Li B, Dimitriadis D, Stolcke A, et al. Acoustic and Lexical Sentiment Analysis for Customer Service Calls[C]// Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 5876-5880.
[10] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.;2020 [cited 2022 Jul 14]. p. 1877–901.
[11] CHEN Mingyi, HE Xuanji, YANG Jing, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition[J]. IEEE Signal Processing Letters, 2018, 25(10): 1440-1444.
[12] Neelakantan A, Xu T, Puri R, et al. Text and code embeddings by

contrastive pre-training[J]. arXiv preprint arXiv:2201.10005, 2022.

[13] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. arXiv preprint arXiv:2303.04226,2023.

[14] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich ve rsatile and fast open-source audio feature extractor[C]// Proceedings of the 18th ACM international conference on Multimedia. 2010: 1459-1462.

[15] Busso C, Bulut M, Lee C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4):335-359.

[16] LI Yuanchao, ZHAO Tianyu, KAWAHARA T. Improved end-to-end speech emotion recognition using self- attention mechanism and multitask learning[C]//Interspeech 2019. ISCA: ISCA, 2019: 2803-2807.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[18] Chunxiao Fan, Jie Lin, Rui Mao, and Erik Cambria. Fusing pairwise modalities for emotion recognition in conversations. Information Fusion,106:102306, 2024.

[19] Lu N, Han Z, Han M, et al. Bi-stream graph learning based multimodal fusion for emotion recognition in conversation[J]. Information Fusion, 2024: 102272.

[20] W. Chen, X. Xing, X. Xu, J. Yang and J. Pang, "Key-Sparse Transformer for Multimodal Speech Emotion Recognition," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6897-6901.