EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Speech emotion recognition based on Graph-LSTM neural network

Yan Li[1], Yapeng Wang[1*], Xu Yang[1] and Sio-Kei Im[2]

Abstract

Currently, Graph Neural Networks have been extended to the field of speech signal processing. It is the more compact and flexible way to represent speech sequences by graphs. However, the structures of the relationships in recent studies are tend to be relatively uncomplicated. Moreover, the graph convolution module exhibits limitations that impede its adaptability to intricate application scenarios. In this study, we establish the speech-graph using feature similarity and introduce a novel architecture for graph neural network that leverages an LSTM aggregator and weighted pooling. The unweighted accuracy of 65.39% and the weighted accuracy of 71.83% are obtained on the IEMOCAP dataset, achieving the performance comparable to or better than existing graph baselines. This method can improve the interpretability of the model to some extent, and identify speech emotion features effectively.

Keywords Speech emotion recognition, Graph neural networks, Long short-term memory

1 Introduction

Speech emotion recognition (SER) is a branch of automatic emotion recognition and automatic speech recognition [1]. It recognizes the emotional state of speech by analyzing the acoustic features and linguistic content of the speech. It can currently be applied to multimodality generation tasks [2], assisted psychotherapy [3], video games [4] and telephone services [5]. The speech emotion recognition task is divided into two main phases: feature extraction and emotion classification. The speech signal is first processed based on time-domain and frequency-domain characteristics to quantize the raw speech. Subsequently, the processed data is fed into deep learning models for the purpose of emotion classification. The most popular models are convolutional neural network (CNN) [6], recurrent neural network (RNN) [7], long short-term memory network (LSTM) [8], as well as large-scale speech recognition models [9]. However, the voice state and emotional expression are variable at any time. It is still a great challenge to accurately identify the emotional state in short time.

Graph neural network (GNN) is an extension of convolutional networks on non-Euclidean data space, with the core idea being to construct good feature interpretability based on data association [10]. It has been successfully applied to computer vision and natural language processing tasks. Because speech is the combination of linear sequences, it is difficult to be converted into irregular non-Euclidean data. Therefore, the application of graph neural networks in speech signal processing is limited. In recent years, researchers have considered linear sequences as a special case of graph and applied graph convolution as encoder by transformations like line graphs, cycle graphs [11, 12], and complete graphs [13], building lightweight architectures with excellent performance. However, the relational structures of these compositions are single. The graph convolution is limited by the graph topology, which is not flexible, resulting in poor generalization ability in complex scenes.

This paper focuses on the task of the sentence-level speech emotion classification. To facilitate this task, individual frames are considered as nodes within the

*Correspondence:
Yapeng Wang
yapengwang@mpu.edu.mo
[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
[2] Macao Polytechnic University, Macao, China

framework. The backbone of the model is constructed using a cycle graph, while the feature similarity between speech frames is computed to determine the connections between nodes. Specifically, the K edges with the highest weights are selected to establish these connections. For complex topological graphs, we choose the Message Passing Neural Network (MPNN) based on spatial-domain convolution to design a more flexible classification model.

Our contributions are as follows.

1) The development of a more adaptable directed graph of speech by leveraging feature similarity allows for greater flexibility in representing speech.
2) The introduction of a graph neural network architecture based on an LSTM Aggregator employs a message passing mechanism to capture input dependencies and facilitates accurate recognition of speech emotions, particularly in graphs with higher complexity.
3) The proposal of a weighted graph pooling operation for graph-level classification tasks enables the extraction of global features. The experimental results show that the weighted pooling can effectively remove redundant information and lead to a more stable convergence trend.

2 Related work

2.1 SER based on deep learning

Currently, classifiers of SER can be categorized into two types, traditional classifiers and deep learning classifiers. Traditional classifiers include Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs) and Support Vector Machines (SVMs), etc. [14], which rely on a lot of preprocessing and precision feature engineering [15]. With the development of deep learning technology, the performance of SER has gained significant improvement. Some studies have combined Deep Neural Networks (DNNs) and traditional classifiers, e.g., [16] proposes a DNN-decision tree SVM model based on DNN, which can capture more distinctive emotion features compared with traditional SVM and DNN-SVM.

Most of recognition frameworks based on neural networks utilize CNNs, LSTMs, and combinitions [17, 18]. For example, [19] modifies the initial model with an incremental approach, and inputs multiple acoustic features to a 1D CNN, which improves the accuracy. [20] constructs a robust and effective recognition model based on key sequence fragments, combining CNN and BiLSTM. Attention mechanism is another key for recognizer based on deep learning to deal with hidden information. Attention-based DNN can mine unevenly distributed features in speech and emphasize saliently emotional information, which better adapts to changes in speech emotion [21]. By directing self-attention to deal with missing and hidden information, the more robust structure [22] obtains the satisfactory performance. Furthermore, the challenge of building SER systems based on neural networks lies in the poor generalization due to data mismatch. To address this problem, [23, 24] make significant progress on generalization by sharing feature representations among auxiliary tasks through multi-task learning. However, the traditional recognition system based on deep learning has the complex structure and weak interpretability of speech features. The graph has been introduced into speech tasks as a compact and efficient representation. And the superiority of GNNs in graph processing has received widespread attention.

2.2 SER based on GNNs

At present, the application of graph neural networks in the field of speech technology still has some limitations [25], but some scholars have verified the advantages of graph convolution in the field of speech technology and the possibility of being widely used through research, such as conversational speech recognition [26], sentence-level [27] / conversation-level speech emotion recognition [28], speech enhancement [29], and Q &A rewriting [30]. The methods of graph construction can be divided into sample point-based, frame-based, speech channel-based, and historical dialogue-based approaches, as shown in Fig. 1. In addition, graph neural network has good performance in low-resource speech emotion recognition, such as [31] using transduction integrated learning algorithms based on graph neural networks to accomplish the challenge of Portuguese speech emotion classification.

In current studies, researchers mostly use frame-based composition. Each frame is considered as one node. Additionally down-sampling is used to reduce the number of frames and simplify the structure. For example, the study [11] modeled the speech signal as a frame-based recurrent graph and constructed a lightweight and precise graph convolution architecture, achieving comparable performance with existing techniques. The studies [10, 12, 25] extend the context acceptance range by constructing neighbors within the specific times on the deep frame-level features obtained by recurrent neural networks. Similarly, the study [32] extends to dialogue speech emotion recognition by introducing CNN-BiLSTM to extract conversation features and constructing edges through a fixed past context window. These studies have a high dependence on the feature processing capability of sequence models, and the connections are relatively fixed. The study [33] proposes an ideal graph structure based on cosine similarity and constructs a graph convolutional network with better robustness.

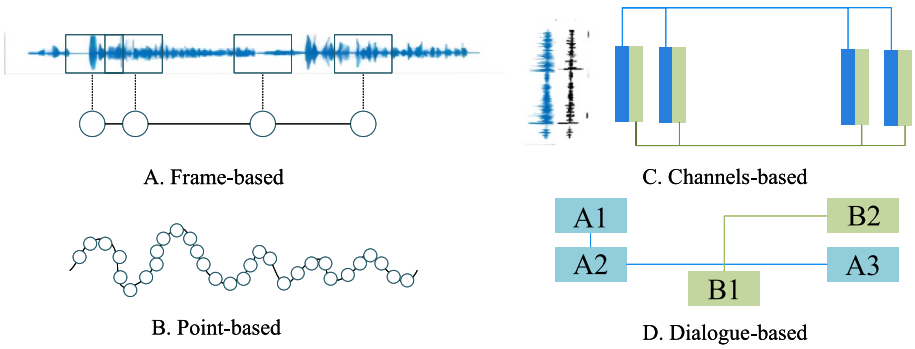

Fig. 1 It provides four examples of graph construction used in the above studies. The nodes of these graphs are frames, sample points, speech channels and dialogues

However, in practical applications, speech sequences are prone to problems of high feature similarity and feature instability. The threshold approach is not applicable to realistic scenarios.

To address the problems of inflexible graph structure and poor generalization ability in the above studies, this paper proposes a graph neural network based on LSTM aggregator and weighted pooling to transform the speech emotion recognition task into the graph classification task.

3 Proposed approach

In this part, we will discuss each component of Graph-LSTM neural network (GLNN) in detail.

3.1 Graph construction

Inspired by studies [11, 12], the speech signal is processed into frames, and each frame is considered as a node. To preserve feature integrity and build the scalable graph, the processing of downsampling and fixed-length cut is discarded. The speech with variable number of frames is transformed into the graph based on the temporal relationship and feature similarity. Thus, the speech graph is heterogeneous.

The graph dataset is represented as $G = (V, E)$. V is the set of nodes, and E is the set of edges. The feature matrix of nodes in the figure is represented as X, $X \in R^{n\times D}$, where n represents the number of nodes, and D represents the feature dimension, and x_i represents the feature vector of the i-th node. x, the feature vector of the node, is composed of a set of low-level descriptors extracted by openSMILE 3.0. The edges are constructed in two categories, one is the directed edges constructed by the temporal relationship. The one-way edges $\left\{v_i \rightarrow v_{i+1}\right\}_{i=1}^{n-1}$ are constructed only depending on the time, and finally the loop is established by $v_n \rightarrow v_1$. The directed cycle graph is used as backbone to improve the stability of the graph structure. The other category is the directed edges obtained from the feature similarity calculation. In order to reduce the computational complexity, the dot product similarity operation is used as follows:

$$X = \frac{X}{\|X\|^2}, weights = X \cdot X^T \tag{1}$$

$$edges = \left\{e_{ji} | j \in TopK(weights, k)\right\}_{i=1}^{n} \tag{2}$$

where $X \in R^{n\times D}$ is the feature matrix of nodes on the graph, and n represents the number of nodes, and D represents the feature dimension; X is standardized, and the dot product similarity between nodes is calculated to obtain the similarity weights. *edges* represent the set of constructed edges. j represents the index of the adjacent node of the i-th node selected by the TopK function. e_{ji} means the edge built between the i-th and the j-th nodes, pointing from the j-th node to the i-th node.

The heat map of the weights is shown in Fig. 2. According to the heat map, it is observed that the feature similarity between nodes is greatest in the region centered on the diagonal. And the feature similarity is higher in a small range of neighborhoods, which is consistent with the characteristics of speech temporal changes. In order to screen out redundant information and select the edges with the highest correlation, the TopK algorithm [34] is used to select the k nodes with the highest similarity to the target node v_i. By conducting experimental verification, the value of k is set to 10, resulting in improved stability of the model's convergence.

3.2 Graph-LSTM neural network

The structure of Graph-LSTM neural network (GLNN) is shown in Fig. 3. The architecture based on speech-graph consists of three graph convolution layers, a pooling layer and a classifier. In Fig. 3, A is the overall structure of GLNN; B is the structure of graph convolution,

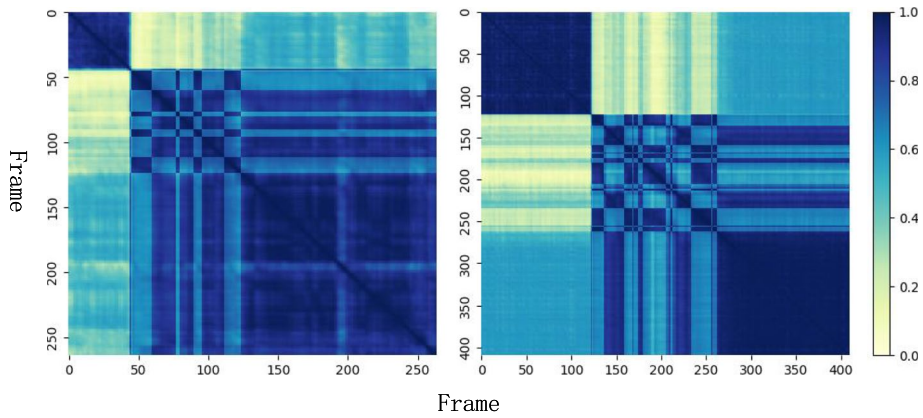

Fig. 2 Similarity weighting heat map

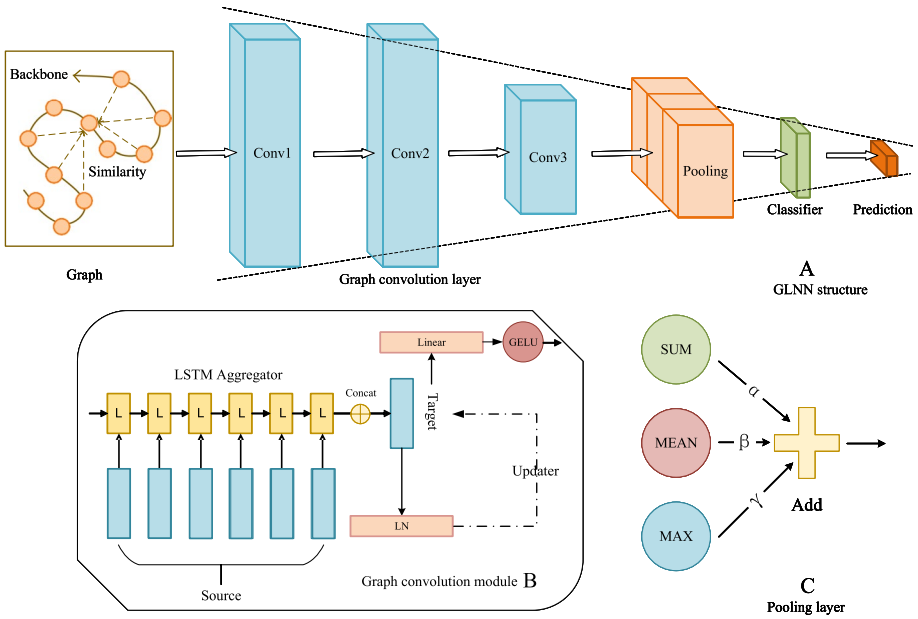

Fig. 3 The structure of GLNN. A is the overall structure of GLNN; B is the structure of graph convolution, consisting of LSTM aggregator and linear updater; C is the structure of weighted pooling layer. In addition, the solid line represents the backbone, and the dashed line represents the possible edges constructed by similarity in Graph of A

consisting of LSTM aggregator and linear updater; C is the structure of weighted pooling layer.

The model construction is based on the message passing network with two phases of forward passing, message aggregation and readout operation [35]. The convolution layers of Graph-LSTM model consist of aggregator and updater.

$$x_i^{'} = \varphi_\alpha(x_i) \tag{3}$$

$$x_{aggr} = \underset{j \in N(i)}{Aggregator_{LSTM}}\left(\oplus x_j\right) \tag{4}$$

$$x_{up} = \varphi_\beta\left(x_i^{'} \oplus x_{aggr}\right) + \gamma \tag{5}$$

where φ_α and φ_β represent the linear transformation; $N(i)$ represents the neighborhood of the target node; x_{aggr} represents the neighborhood features obtained by aggregation, and x_i represents the feature vector of the i-th node, and x_j represents the features vector of adjacent points.

Based on the graph structure of 3.1, considering the continuity and complexity of speech features, the simple aggregation operation [36] is no longer applicable to this application scenario. As a result, the LSTM aggregation operator [37] is chosen to accomplish inductive

representation learning of adjacent features. The flow is from the source nodes to the target nodes. The neighbors are combined into time series and fed into LSTM for inference, obtaining deep aggregated features.

The Graph-LSTM neural network consists of the 3-layers graph convolution module, which realizes the feature aggregation and update. Then it performs the read-out operation through the pooling layer to obtain graph-level features, which are input to the classifier.

3.3 Weighted pooling

The construction of the speech graph establishes connection relationships based on time sequence and similarity. There are a large number of overlapping regions and redundant features between neighbors. Conventional pooling operation is difficult to filter out representative features from dense connections, while the time sequence of speech needs to preserve the integrity of node features. Therefore, weighted pooling is constructed based on global pooling operations of sum, max and mean, which is calculated as shown in Eq. 6.

$$x_{pooling} = \alpha \cdot \max_{i=1}^{n}(x_i) + \beta \cdot mean_{i=1}^{n}(x_i) + \lambda \cdot sum_{i=1}^{n}(x_i) \quad (6)$$

where max, mean and sum represent the three types of global pooling operations; x_i represents the feature vector of the i-th node; $x_{pooling}$ represents the global feature vector; α, β and λ represent the weights of the three pooling operations respectively, which are set to {0.3, 0.3, 0.3} in the experiment. Through the weighted pooling operations, the feature integrity is retained while removing redundant information.

4 Experiments

4.1 Dataset and features

The dataset used for the study is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [38], containing 12 hours of audiovisual data. The data were collected from two-person situational dialogues which the actors performed in a scripted or improvised manner. The actors' facial expressions and hand movements were recorded simultaneously during the communication. The speech emotion recognition task in this study uses only speech data, for five binary dialogues divided into multiple sentences. IEMOCAP uses the multi-annotator to annotate these data with 11 emotion labels. For objective experimental analysis and performance comparison, we used four classes of data in our experiments, namely angry, happy, sad, and neutral, totaling 4490 utterances.

The extraction of audio features is done with the open source tool called openSMILE 3.0 [39]. openSMILE is a large-space audio feature interpreter that is widely used for sentiment computing tasks. Audio feature extraction can be achieved through command line and configuration files. The experiment uses the INTERSPEECH 2010 Paralinguistic Challenge feature set to extract a set of low-level descriptors (LLDs) consisting of mfcc, maxPos, amean, skewness, and smoothing using the corresponding first-order delta coefficients. The speech is framed by a fixed-size sliding window, with the frame length set to 25 ms and the shift set to 10 ms. In addition, spontaneous binary features are added to each frame, inspired by spontaneous learning [40]. As a result, 77-dimensional features are generated for speech frames.

4.2 Experimental setup

The dataset is divided into training set and test set by stratify to equalize the data categories, and the division ratio is 8:2. Training is performed using Adam optimizer. The learning rate is set to 1e-5. The decay weight is set to 1e-4, and the batch size is set to 8. All experiments are performed on NVIDIA Tesla V100 GPU. The model performance is evaluated using weighted accuracy (WA) and unweighted accuracy (UA) metrics.

4.3 Comparison model

We compare the proposed method with the sequence-based SER model and the graph-based SER model respectively.

4.3.1 SER models

We selected three SER models as baselines.

DCNN [41]: a 1-D convolutional neural network uses hybrid features as input and modifies the initial model using incremental methods to improve the classification accuracy. The model has good generalization.

ResNet34 [42]: a transfer learning method combines with acoustic spectrogram enhancement that can efficiently handle variable-length inputs using a pre-trained residual network. The method alleviates the over-fitting problem and improves the generalization ability of the model.

ADAN + SVM [43]: an adversarial data augmentation network generates augmented data and makes SVM classifiers outperform RNN classifiers in terms of local attention.

4.3.2 GNN baselines

Compact SER [11]: a lightweight graph convolutional network based on recurrent or linear graphs maintains comparable performance to existing techniques under reduced learning parameters.

PATCHY-SAN [44]: a general framework for extracting locally connected regions is based on convolutional

networks to learn the arbitrary graph, which is computationally efficient.

PATCHY-Diff [45]: a microscopic graph pooling module generates hierarchical representations to be combined with multiple graph neural networks in an end-to-end mode for graph classification tasks.

For the above methods, the four classes of data totaling 4490 utterances, angry, happy, sad and neutral, were used for analysis.

GA-GRU [25]: a speech emotion recognition framework applies graph attention mechanism to gating units, combining long time sequences and graph data to enhance feature saliency.

CoGCN [33]: a graph convolutional network is based on cosine similarity with good noise immunity.

LSTM-GIN [46]: a speech emotion recognition network based on LSTM and GIN applies Graph Isomorphism Network to extract global feature representations.

The above approaches merge the happy and excited categories when validating the model performance, and extend the happy category data to 1636 utterances, for a total of 5531 utterances.

4.4 Results and analysis

4.4.1 Performance comparison

Table 1 shows the results of GLNN compared with baselines. First, the basic architecture of GLNN, using global average pooling, obtains the WA of 68.15% on IEMOCAP, which exceeds the baseline methods. However, we found that the UA was only 59.16%, which was lower than ResNet34 [42] and compact SER [11]. The possible reason is category imbalance. The happy class with only 595 utterances is much lower than others, which might lead to lower UA values. For validation, the happy category is combined with the excited category, and the results are compared with the methods [25, 33, 46].

With more balanced categories, the difference between GLNN on WA and UA metrics reduces. Especially the UA has a significant improvement of 9.49%. It indicates that the number of training data in each categories has a large impact on the model performance, and GLNN is not accurate enough when training with the small and unbalanced dataset. Because the graph structure and graph convolution used by GLNN may lead to the problem of feature redundancy and unstable feature extraction for small training samples. To solve this problem, the weighted pooling layer is constructed.

After adopting the weighted pooling method, GLNN exhibits notable enhancement, achieving WA of 71.83% and UA of 65.39%. These results surpass the performance of the baseline models, indicating superior effectiveness. Furthermore, a reduction in the disparity between the two metrics is observed. In practical application, the category equalization problem is a common data problem. It is difficult to equalize the data. Therefore, it is more feasible to use weighted pooling to optimize the model performance and mitigate the oversmoothing problem.

4.4.2 Ablation

We set up three groups of ablation experiments to verify the rationality of the proposed method. Table 2 analyzes the effect of the number of layers of graph convolution and calculates the corresponding parameters. From the experimental results, it is clear that the best performance is obtained by the 3-layer convolution module, which has the large improvement compared with the 2-layer convolution. However, the model performance decreases by continuing to add the graph convolution layers.

Table 1 Comparison between SER baselines and proposed model

Model	UA (%)	WA (%)	Condition
DCNN 2020 [41]	-	64.3	4490 utterances
ResNet34 2021 [42]	61.61	66.02	
ADNN + SVM 2019 [43]	-	65.01	
Graph baselines			
PATCHY-SAN 2016 [11]	56.27	60.34	
PATCHY-Diff 2018 [11]	58.71	63.23	
Compact SER 2021 (cycle) [11]	**62.27**	65.29	
Ours (Mean pooling)	59.16	**68.15**	
Ours (Weighted pooling)	**65.39**	**71.83**	
LSTM-GIN 2022 [46]	65.53	64.65	5531 utterances
CoGCN 2022 [33]	63.67	62.64	
GA-GRU 2020 [25]	63.8	62.27	
Ours (Mean pooling)	**68.65**	**68.11**	

The Bold represents the best results. '-' means that the result is not recorded in the report

Table 2 Comparison between different layers

Layers	Params (KB)	UA (%)	WA (%)
2	361	55.93	63.81
3	409	**59.16**	**68.15**
4	591	58.62	67.82

Table 3 Comparison of number of K

K	UA (%)	WA (%)
5	57.97	58.55
10	59.16	**68.15**
15	**59.91**	68.04

Table 4 Comparison between different pooling methods

Pooling	Max	Sum	Mean	TopK	Weighted
UA (%)	57.97	58.55	59.16	61.88	**65.39**
WA (%)	66.15	67.71	68.15	68.49	**71.83**

Meanwhile, the complexity of the graph convolution is analyzed. The space complexity determines the number of parameters, and Table 2 records the parameter of the graph convolution. The parameter of the three-layer convolution is 409K with moderate training, which can fit well. And the time complexity is calculated from three links: feature mapping, feature aggregation and feature updating. n denotes the number of nodes. D denotes the original dimension of inputs, and D_1' denotes the mapping dimension, and D_2' denotes the feature dimension of outputs. Firstly, the features of all nodes are mapped with the time complexity of $O(n * D * D_1')$. Then the feature aggregation is performed by the LSTM aggregator with the complexity approximately equal to $O(n * D_1'^2)$. Finally, the feature updating is completed by the linear layer with the time complexity of $O(n * D_1' * D_2')$. In summary, the time complexity is $O(n(D_1' * D_2' + D_1'^2))$.

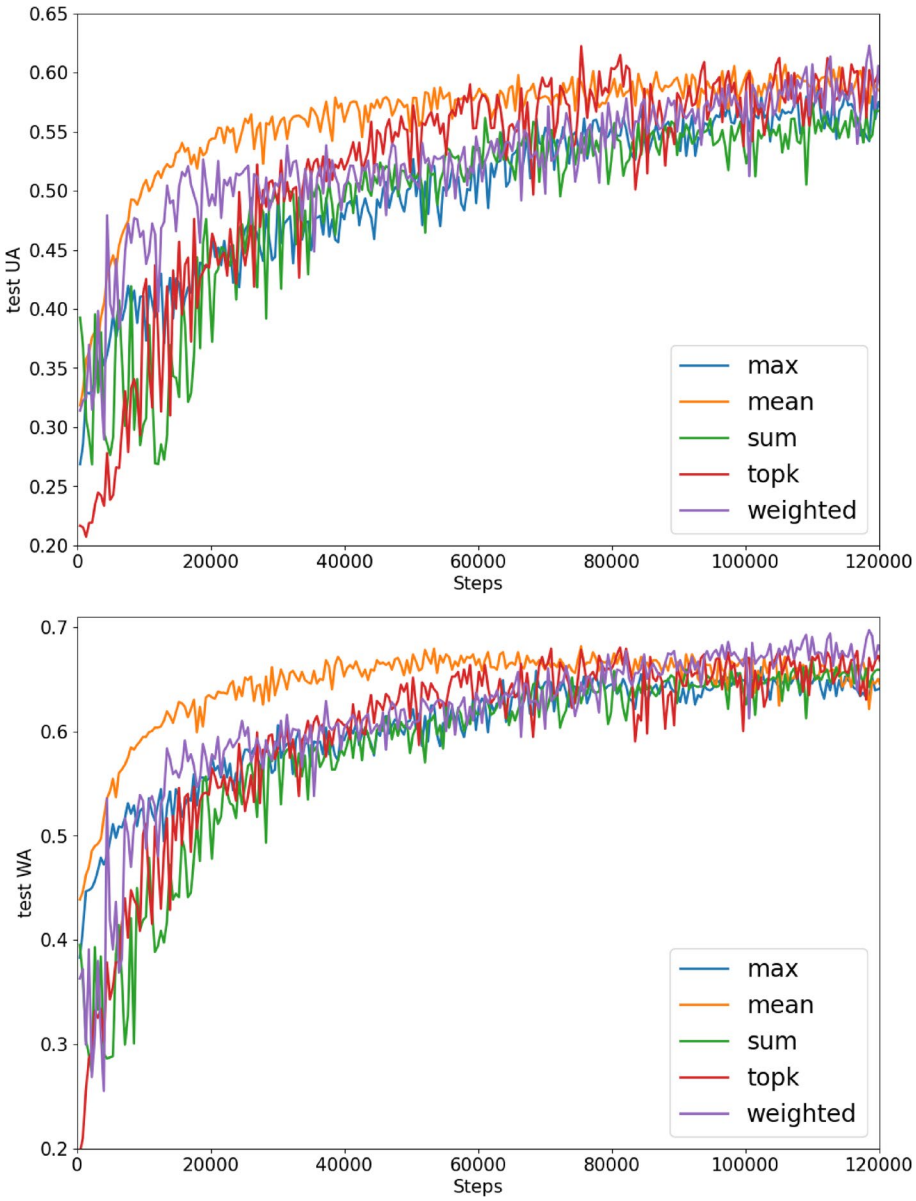

Fig. 4 The convergence curves of five pooling methods. The blue, orange, green, red and purple curves represent max-pooling, mean-pooling, sum-pooling, topk-pooling and weighted-pooling repectively. Two types of curves, WA curve and UA curve, are drawn separately

Table 3 analyzes the effect of k values when constructing edges by the TopK algorithm, i.e., the effect of the number of edges. The results in Table 3 show that the k value of 10 obtains the large improvement compared with the value of 5, with an improvement of 9.6% on WA. However, the gain of model performance is very small when the k value is taken as 15, indicating that the information obtained by adjacent nodes is saturated. Increasing the number of edges cannot bring extra information gain.

Table 4 compares the effects of different pooling methods on the accuracy. It should be noted that, in addition to three simple read-out operations of maximum, mean and summation, we also try to use topk pooling to filter out 50% of the nodes before performing mean pooling. From Table 4, the mean-pooling performs better than max-pooling and sum-pooling, but worse than topk-pooling. It indicates that filtering the nodes to remove redundant features helps to improve the performance. And weighted pooling maximally preserves the integrity of node features and effectively filters out representative features. Compared with other pooling methods, it has the better performance. Figure 4 shows the test curves of different pooling methods. As shown in Fig. 4, the weighted pooling can effectively mitigate oversmoothing and converge more stably.

5 Conclusion

In this paper, we explore a graph neural network based on LSTM aggregator and weighted pooling applied to speech emotion recognition task. The specific process is as follows. First, speech features are extracted by the openSMILE. Then, the connection relationship is selected for speech graph construction based on the feature similarity and TopK algorithm. Finally, a classification model is designed based on the message passing architecture to convert speech classification into a graph classification task. Our evaluation on the IEMOCAP dataset demonstrates superior performance compared to the baseline models. However, there are some shortcomings in the current stage, including 1) complex connections and a large number of redundant features in graph; 2) unstable processing and analysis of small datasets; 3) neglecting the speaker's information. The research focuses on adult speech, which is a lack of exploration of children's speech emotion recognition [47, 48].

In order to address the aforementioned challenges, we will adopt the following stategies in the next stage. 1) To address the issue of redundant features, we will consider more versatile approaches for graph construction to further reduce the requirement for data size and optimize the model framework. 2) Faced with the problem of data scarcity, the Transfer Learning strategy [49] is adopted to design a multi-task framework for speech recognition and emotion recognition, which improves the adaptability to small sample data through feature sharing. 3) To address the issue of differences in acoustic and linguistic features of speakers, a speaker converter is introduced to learn adaptive transformation, which enables the model to eliminate feature differences.

Abbreviations

SER	Speech Emotion Recognition
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory Network
GNN	Graph Neural Network
MPNN	Message Passing Neural Network
GLNN	Graph-LSTM neural network
IEMOCAP	The Interactive Emotional Dyadic Motion Capture
LLDs	Low-Level Descriptors
WA	The Weighted Accuracy
UA	The Unweighted Accuracy

Acknowledgements
Not applicable.

Authors' contributions
Yan Li and Yapeng Wang conceived the idea. Yan Li completed the experiment, analyzed the data, and wrote the initial draft. Yapeng Wang refined the idea, guided the research, and contributed to projection management, funding acquisition and manuscript revision. Xu Yang and SIO-KEI IM contributed to refining the idea, carrying out additional analyses and finalizing this paper.

Funding
This work is funded by Macao Polytechnic University under grant number RP/FCA-03/2022.

Availability of data and materials
The research dataset used is the Interactive Emotional Dyadic Motion Capture (IEMOCAP). To obtain the IEMOCAP data, we need to fill out an electronic release from https://sail.usc.edu/iemocap/iemocap_release.htm.

Code availability
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 9 June 2023 Accepted: 31 August 2023
Published online: 11 October 2023

References

1. J. de Lope, M. Graña., An ongoing review of speech emotion recognition. Neurocomputing. **528**, 1–11 (2023). https://doi.org/10.1016/j.neucom.2023.01.002

2. Y. Chen, J. Zhao, W.Q. Zhang., Expressive speech-driven facial animation with controllable emotions (2023). http://arxiv.org/abs/2301.02008
3. L.S.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber, N.B. Allen, Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans. Biomed. Eng. **78**(3), 574–586 (2011). https://doi.org/10.1109/TBME.2010.2091640
4. G. van Kleef, A. Cheshin, L. Koning, W. S.A., Emotional games: How coaches' emotional expressions shape players' emotions, inferences, and team performance. Psychol. Sport Exerc. **41**, 1–11 (2019). https://doi.org/10.1016/j.psychsport.2018.11.004
5. L.F. Parra-Gallego, J.R. Orozco-Arroyave., Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments. Digit. Signal Process. **120**, 103,286 (2022). arXiv:2108.11981
6. K. Wongpatikaseree, S. Singkul, N. Hnoohom, S. Yuenyong, Real-time end-to-end speech emotion recognition with cross-domain adaptation. Big Data Cogn. Comput. **6**(3), 79 (2022). https://doi.org/10.3390/bdcc6030079
7. C. Chen, P. Zhang, in *Interspeech*, CTA-RNN: Channel and temporal-wise attention RNN leveraging pre-trained ASR embeddings for speech emotion recognition (Korea, 2022), pp. 4730–4734. https://doi.org/10.48550/arXiv.2203.17023
8. A.H. Jo, K.C. Kwak, Speech emotion recognition based on two-stream deep learning model using korean audio information. Appl. Sci. **13**(4), 2167 (2023). https://doi.org/10.3390/app13042167
9. Sharma, Mayank., in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Multi-lingual multi-task speech emotion recognition using wav2vec 2.0 (IEEE, Singapore, 2022), pp. 6907–6911. https://doi.org/10.1109/ICASSP43922.2022.9747417
10. L. Wu, P. Cui, J. Pei, L. Zhao, X. Guo, in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Graph neural networks: foundation, frontiers and applications (Springer, Singapore, 2022), pp. 4840–4841. https://doi.org/10.1007/978-981-16-6054-2
11. A. Shirian, T. Guha, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Compact graph architecture for speech emotion recognition (IEEE, Canada, 2021), pp. 6284–6288. https://doi.org/10.1109/ICASSP39728.2021.9413876
12. Y. Hu, Y. Tang, H. Huang, L. He, A graph isomorphism network with weighted multiple aggregators for speech emotion recognition pp. 4705–4709 (2022). https://arxiv.org/abs/2207.00940
13. A. Shirian, S. Tripathi, T. Guha, Dynamic emotion modeling with learnable graphs and graph inception network. IEEE Trans. Multimed. **24**, 780–790 (2021). https://doi.org/10.1109/TMM.2021.3059169
14. T.M. Wani, T.S. Gunawan, S.A.A. Qadri, M. Kartiwi, E. Ambikairajah, A comprehensive review of speech emotion recognition systems. IEEE Access. **9**, 47795–47814 (2021). https://doi.org/10.1109/ACCESS.2021.3068045
15. B.J. Abbaschian, D. Sierra-Sosa, A. Elmaghraby, Deep learning techniques for speech emotion recognition, from databases to models. Sensors. **21**(4), 1249 (2021). https://doi.org/10.3390/s21041249
16. L. Sun, B. Zou, S. Fu, J. Chen, F. Wang, Speech emotion recognition based on DNN-decision tree SVM model. Speech Commun. **115**, 29–37 (2019). https://doi.org/10.1016/j.specom.2019.10.004
17. T. Anvarjon, Mustaqeem, S. Kwon, Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. Sensors. **20**(18), 5212 (2020). https://doi.org/10.3390/s20185212
18. A.A. Abdelhamid, E.S.M. El-Kenawy, B. Alotaibi, G.M. Amer, M.Y. Abdelkader, A. Ibrahim, M.M. Eid, Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm. IEEE Access. **10**, 49265–49284 (2022). https://doi.org/10.1109/ACCESS.2022.3172954
19. D. Issa, M.F. Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. Control. **59**, 101894 (2020)
20. M. Sajjad, S. Kwon, others., Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access. **8**, 79861–79875 (2020). https://doi.org/10.1109/ACCESS.2020.2990405
21. E. Lieskovská, M. Jakubec, R. Jarina, M. Chmulík, A review on speech emotion recognition using deep learning and attention mechanism. Electronics. **10**(10), 1163 (2021)
22. D. Li, J. Liu, Z. Yang, L. Sun, Z. Wang, Speech emotion recognition using recurrent neural networks with directional self-attention. Expert Syst. Appl. **173**, 114683 (2021). https://doi.org/10.1016/j.eswa.2021.114683
23. X. Cai, J. Yuan, R. Zheng, L. Huang, K. Church, in *Interspeech, Speech emotion recognition with multi-task learning,* vol. 2021, (ISCA, Czechia, 2021), p. 4508–4512
24. Y. Li, T. Zhao, T. Kawahara, others., in *Interspeech*, Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning (2019), pp. 2803–2807. https://doi.org/10.21437/Interspeech.2019-2594
25. B.H. Su, C.M. Chang, Y.S. Lin, C.C. Lee, in *Interspeech*, Improving speech emotion recognition using graph attentive bi-directional gated recurrent unit network. (China, 2020), pp. 506–510
26. S.H. Chiu, T.H. Lo, F.A. Chao, B. Chen, in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Cross-utterance reranking models with bert and graph convolutional networks for conversational speech recognition (Japan, IEEE, 2021), pp.1104–1110
27. A. Pentari, G. Kafentzis, M. Tsiknakis, in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Investigating graph-based features for speech emotion recognition (IEEE, 2022), pp. 01–05
28. Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, R. Li, in *Interspeech*, Conversational emotion recognition using self-attention mechanisms and graph neural networks (China, 2020), pp. 2347–2351
29. P. Tzirakis, A. Kumar, J. Donley, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Multi-channel speech enhancement using graph neural networks (IEEE, 2021), pp. 3415–3419. https://doi.org/10.1109/ICASSP39728.2021.9413955
30. S. Yuan, S. Gupta, X. Fan, D. Liu, Y. Liu, C. Guo, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Graph enhanced query rewriting for spoken language understanding system (IEEE, 2021), pp. 7997–8001. https://doi.org/10.1109/ICASSP39728.2021.9413840
31. E.L.S. Perin, E.T. Matsubara, In *Proceedings of the First Workshop on Automatic Speech Recognition for Spontaneous and Prepared Speech & Speech Emotion Recognition in Portuguese,* Transductive ensemble learning with graph neural network for speech emotion recognition (CEUR, 2022), p. 7
32. Y. Song, J. Liu, L. Wang, R. Yu, J. Dang, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Multi-stage graph representation learning for dialogue-level speech emotion recognition (IEEE, Singapore, 2022), pp. 6432–6436. https://doi.org/10.1109/ICASSP43922.2022.9746237
33. J. Kim, J. Kim, Representation learning with graph neural networks for speech emotion recognition. (2022). https://arxiv.org/abs/2208.09830
34. J. Duchi, S. Haque, R. Kuditipudi, A fast algorithm for adaptive private mean estimation (2023). https://arxiv.org/abs/2301.07078
35. K.T. Schütt, S. Chmiela, O.A. von Lilienfeld, A. Tkatchenko, K. Tsuda, K.R. Müller, Machine learning meets quantum physics. Lect. Notes Phys. (2020). https://doi.org/10.1007/978-3-030-40245-7
36. K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks? (2019). arXiv:1810.00826
37. W.L. Hamilton, R. Ying, J. Leskovec., Inductive representation learning on large graphs (2018). arXiv:1706.02216
38. C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database. **42**(4), 335–359 (2008). https://doi.org/10.1007/s10579-008-9076-6
39. F. Eyben, F. Weninger, F. Gross, B. Schuller, in *Proceedings of the 21st ACM International Conference on Multimedia*, Recent developments in opensmile, the munich open-source multimedia feature extractor (ACM, 2013), pp. 835–838. https://doi.org/10.1145/2502081.2502224
40. K. Mangalam, T. Guha., Learning spontaneity to improve emotion recognition in speech (2018). arXiv:1712.04753
41. D. Issa, M. Fatih Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. Control. **59**, 101894 (2020). https://doi.org/10.1016/j.bspc.2020.101894
42. S. Padi, S.O. Sadjadi, R.D. Sriram, D. Manocha, in *Proceedings of the 2021 International Conference on Multimodal Interaction*, Improved speech emotion recognition using transfer learning and spectrogram augmentation (ACM, Canada, 2021), pp. 645–652. https://doi.org/10.1145/3462244.3481003
43. L. Yi, M.W. Mak, in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Adversarial

data augmentation network for speech emotion recognition (2019), pp. 529–534. https://doi.org/10.1109/APSIPAASC47483.2019.9023347
44. M. Niepert, M. Ahmed, K. Kutzkov, in *International conference on machine learning*, Learning convolutional neural networks for graphs (PMLR, 2016), pp. 2014–2023
45. Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, J. Leskovec, Hierarchical graph representation learning with differentiable pooling. Adv. Neural Inf. Process. Syst. **31**, 4800–4810 (2018)
46. J. Liu, H. Wang, Graph isomorphism network for speech emotion recognition (2021), pp. 3405–3409. https://doi.org/10.21437/Interspeech.2021-1154
47. F. Albu, D. Hagiescu, L. Vladutu, M. Puica, in *Edulearn 15, 7th international conference on education and new learning technologies*, Neural network approaches for children's emotion recognition in intelligent learning applications (Barcelona, SPAIN, 2015)
48. V. Bhardwaj, M. Othman, V. Kukreja, Y. Belkhier, M. Bajaj, S.G. .B, A. Rehman, M. Shafiq, H. Hamam, Automatic speech recognition (ASR) system for children's: A systematic literature review. Appl. Sci. (2022). https://doi.org/10.3390/app12094419
49. L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. Albahri, B. Al-dabbagh, M. Fadhel, M. Manoufali, J. Zhang, A. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, Y. Gu, A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. J. Big Data. **10** (2023). https://doi.org/10.1186/s40537-023-00727-2

Publisher's Note