INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Time-Frequency Representation Learning with Graph Convolutional Network for Dialogue-level Speech Emotion Recognition

Jiaxing Liu[1], *Yaodong Song*[1], *Longbiao Wang*[1,2,*], *Jianwu Dang*[1,2,3], *Ruiguo Yu*[1]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Tianjin University-Huiyan Technology Joint AI Lab, Tianjin University, China
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan

{jiaxingliu, songyaodong, longbiao_wang, rgyu}@tju.edu.cn, jdang@jaist.ac.jp

Abstract

With the development of speech emotion recognition (SER), dialogue-level SER (DSER) is more aligned with actual scenarios. In this paper, we propose a DSER approach that includes two stages of representation learning: intra-utterance representation learning and inter-utterance representation learning. In the intra-utterance representation learning stage, traditional convolutional neural network (CNN) has demonstrated great success. However, the basic design of a CNN restricts its ability to model the local and global information in the spectrogram. Therefore, we propose a novel local-global representation learning method for the intra-utterance stage. The local information is learned by a time-frequency convolutional neural network (TFCNN), which we published previously. Here, we propose a time-frequency capsule neural network (TFCap) to model global information that can extract more stable global time-frequency information directly from spectrograms. In the inter-utterance stage, a graph convolutional network (GCN) is introduced to explore the relations between utterances in a dialog. Our proposed methods were evaluated on the IEMOCAP database. The proposed time-frequency based method in the intra-utterance stage achieves an absolute increase of 9.35% compared to CNN. By integrating GCN in the inter-utterance stage, the proposed approach achieves an absolute increase of 4.05 % compared to the model in the previous stage.

Index Terms: dialogue level speech emotion recognition, capsule neural network, time-frequency

1. Introduction

Affective computing is a promising field of research that aims to endow intelligent systems to perform like humans and provide better service and information that people seek. Speech is the most commonly used communication method in the daily lives of people; therefore, speech emotion recognition (SER) has a very realistic and scientific research value [1]. Human speech contains a variety of emotion-related information, and therefore learning effective emotional representations is crucial to SER systems [2].

Emotional representation extraction methods can be categorized as traditional methods and deep learning methods. In traditional methods, such as mel-frequency cepstral coefficients (MFCC) [3], linear prediction cepstral coefficients (LPCC) [4], prosodic features [5, 6, 7], and the statistics of these segment features [8] perform well in automatic speech recognition (ASR) tasks, but these are not suitable for SER.

* Corresponding Author

Recently, with the rapid development of deep learning, numerous deep learning methods [9, 10, 11] have been introduced to SER. Among these deep learning models, convolutional neural network (CNN)-based models have achieved more competitive results. Satt et al. [12] proposed a famous model that used a CNN to learn emotional features from the spectrogram, and introduced bidirectional long short-term memory (BLSTM) to model the contextual information in an utterance. The CNN_BLSTM model has become the most widely adopted baseline model. However, this model ignored the special form of the spectrogram (time-frequency). In our previous work [13], three types of filters were used to capture the ignored time-frequency related information, which was called time-frequency CNN (TFCNN). However, TFCNN still faces the same problems as CNN, and the information is extracted from the local region.

In addition to the above-mentioned local problems of CNN and TFCNN, ubiquitous pooling layers usually inevitably drop some useful information. In response to the problems of CNN-based models, Sabour et al. [14] proposed a capsule network (CapsNet) that was quickly introduced in various research fields [15, 16, 17]. CapsNet contains vectors that represent the instantiation parameters of various global spatial information. The one-dimensional vectors and routing algorithm enable the CapsNets to model the global information well. However, it is unstable, and the learned representation is shallow. In [18], a densely connected capsule network (DenseCap) was proposed with a new routing algorithm that stacked the capsule layers to alleviate existing problems. Although the performance of DenseCap is better than that of CapsNet, the root of the problems—the poor robustness of the original one-dimensional structure in these works— remains unsolved.

We propose a time-frequency capsule neural network (TFCap) to solve the root of the problems existing in CapsNet and DenseCap. The TFCap is based on the newly designed matrix vectors (time-frequency) and a supporting routing algorithm. The time-frequency vectors explore more stable two-dimensional information to avoid mutation of one-dimensional vectors. Compared with one-dimensional vectors, time-frequency vectors have a wider range of length and direction, and maintain more global information. With the help of the proposed local-global time-frequency representation learning method, the proposed TFCNN_TFCap can learn more different and useful information than the traditional CNN model and CapsNet.

Due to the lack of inter-utterance contextual information, the current research on SER (intra-utterance stage) cannot service real-life applications, such as supporting dialogue sys-

http://dx.doi.org/10.21437/Interspeech.2021-2067

tems and generating more human-like speech . Therefore, we conduct further research on the dialogue-level (inter-utterance stage) SER to fit the needs of these real-life applications. Graph neural networks (GNNs) [19, 20] have received growing attention recently. In particular, the graph convolutional network (GCN) [21] was proposed by Kipf, which achieved state-of-the-art results in numerous benchmark datasets. The GCN was introduced in our work to learn the hidden relations between the utterances. The proposed model (TFCNN_TFCap)+GCN can model the contextual information in the intra-utterance and inter-utterance stages.

Our proposed dialogue-level SER (DSER) approach (TFCNN_TFCap)+GCN overcomes the limitations of the aforementioned methods. The contributions of our work can be summarized as follows: 1) TFCNN_TFCap using matrix vectors to model local-global time-frequency information in the intra-utterance stage is proposed. 2) A GCN is introduced to model the contextual information in the inter-utterance stage.

2. Dialogue-level SER Representation learning

2.1. The intra-utterance and inter-utterance stage in DSER

The proposed system as shown in Fig. 1 mainly consists two stages which are local-global time-frequency representation learning stage (intra-utterance stage) and dialogue-level contextual information learning stage (inter-utterance stage).The intra-utterance stage contains two steps: the first step is local-global time-frequency representation learning using TFCNN_TFCap. The TFCNN is the same as the model in [13]. The representations R_L and R_G, are learned in the first step. The second step is intra-utterance classification, using BLSTM. The second stage is the inter-utterance stage, which is based on the intra-utterance stage representation. The details of the proposed TFCap are presented in Fig. 2, and the details of the GCN model in the inter-utterance stage are shown in Fig. 3.

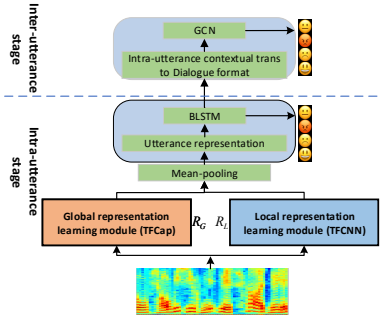


Figure 1: *Dialogue-level SER representation learning system.*

2.2. Global time-frequency representation learning

Compared with traditional CNNs, CapsNet uses a group of neurons, the length of which represents the existence probability and the orientation represents the instantiation parameters of various global spatial information. In other words, CapsNet outputs one-dimensional vectors instead of scalar values, which enables it to model global spatial information . The introduced one-dimensional vectors lead to an unstable original routing algorithm. We propose TFCap to solve the problems of the traditional CNN and CapsNet.

The (u_{ti}, u_{fi}) represents the i-th output of a capsule in the $l-1$ layer. The "prediction time-frequency vectors" $\hat{u}(T,F)$ is produced by

$$\hat{u}(T,F) = \sum_{n=0}^{N-1} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} (u_{nti}, u_{nfi}) \mathrm{K}(T-t, F-f) \quad (1)$$

$$\hat{u}(T,F) = \sum\nolimits_{N} (\hat{u}_{tj|i}, \hat{u}_{fj|i}) \quad (2)$$

In Eqs (1) and (2), N represents the input capsule number, T and F represent the input time and frequency dimensions, respectively, and K represents the kernel function. We use 2D-convolution functions to replace multiplying the weight s. The padding is the same in the convolution, and the output size remains the same as the input size.

$$(s_{tj}, s_{fj}) = \sum_{i} (c_{itj}\hat{u}_{tj|i}, c_{ifj}\hat{u}_{fj|i}) \quad (3)$$

In Eq. (3), (s_{tj}, s_{fj}) is the outer capsule. c_{itj} and c_{ifj} are coupling coefficients determined by Eqs (4) and (5).

$$c_{itj} = \frac{\exp(b_{itj})}{\sum_{k} \exp(b_{itk})} \quad (4)$$

$$c_{ifj} = \frac{\exp(b_{ifj})}{\sum_{k} \exp(b_{ifk})} \quad (5)$$

In this step, we get the output of N capsules:

$$S_N = \sum\nolimits_{N} (s_{tj}, s_{fj}) \quad (6)$$

At the same time, we can rewrite Eq. (6) as:

$$S_N = [s_1, \cdots, s_n \cdots s_N] \quad (7)$$

In Eq. (7), S_N can be treated as N matrix vectors, and the element s_n has a size $T \times F$. Matrix vectors contain two-dimensional information, and the structure is more stable. Based on the structural changes, we focus on N, the number of capsules.

In this step, we mainly have two concerns: 1) The information we obtain in each matrix vector represents different 'obvious features', which are treated equally. However, these components relate to emotion differently. 2) The sudden change is enhanced by the routing algorithm, and there is no mechanism to correct sudden change errors. Therefore, the generation of a suitable weight for each capsule matrix vector is important. Hence, we introduce capsule-wise attention in this step to address these concerns. The $n-th$ capsule in S is determined by

$$w_n = \frac{1}{T \times F} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} s_n(t,f) \quad (8)$$

Such capsule weights can be viewed as a collection of time–frequency descriptors, whose statistics contribute to express the whole capsule layer. In Fig. 2, $T_i \times F_i \times N_i$ is the input size of the primary capsule layer, and $T_o \times F_o \times N_o$ is the output size of the TFCap. T is the size of time, F is the size of the frequency, and N is the capsule number. $GP(\bullet)$ is the global average pooling, and $f(\bullet)$ is the 'Softmax' function. In this study, the input size is $4 \times 28 \times 64$, and the output size is $8 \times 32 \times 4$.

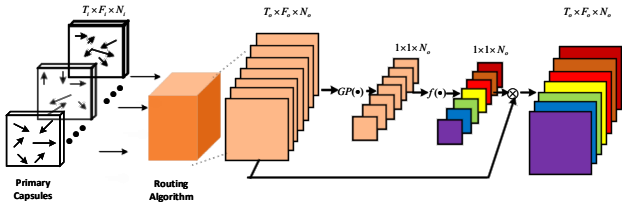


Figure 2: *The proposed model time-frequency capsule network (TFCap)*

$$\hat{S}_N = [w_1 s_1, \cdots w_j s_j, \cdots w_N s_N] = [\hat{s}_1, \cdots \hat{s}_j, \cdots \hat{s}_N] \quad (9)$$

The final step is nonlinear activation, where we introduce the similar activation with CapsNet, which is also called 'squash'. However, the 'squash' has two dimensions (time, frequency).

$$(v_{tj}, v_{fj}) = \frac{\|(\hat{s}_{tj}, \hat{s}_{fj})\|^2}{1 + \|(\hat{s}_{tj}, \hat{s}_{fj})\|^2} \frac{(\hat{s}_{tj}, \hat{s}_{fj})}{\|(\hat{s}_{tj}, \hat{s}_{fj})\|} \quad (10)$$

In the proposed TFCap, (c_{itj}, c_{ifj}), $(\hat{s}_{tj}, \hat{s}_{fj})$, and (v_{tj}, v_{fj}) are updated according to Eqs. (1)–(11). To increase the accuracy of (c_{itj}, c_{ifj}), (c_{itj}, c_{ifj}) is updated according to the following rule :

$$(c_{itj}, c_{ifj}) \leftarrow (c_{itj}, c_{ifj}) + (c_{itj}, c_{ifj})(v_{tj}, v_{fj}) \quad (11)$$

In summary, we reconstructed the data organization form of the capsule network and updated the new routing algorithm. The proposed TFCap can extract global time-frequency information directly from the spectrogram. The newly designed architecture is sufficiently stable to explore a more global representation.

2.3. Inter-utterance stage contextual representation learning

As shown in Fig. 3, we first used the intra-utterance contextual information learned by BLSTM in the previous stage. We then transform the intra-utterance contextual representation into a dialogue format. For example, a dialog consists of K utterances. The dialogue-level representation H_D is the input data.

$$H_D = [h_1, h_2, \cdots, h_k, \cdots h_K] \quad (12)$$

h_k represents intra-utterance representation. The representation of each utterance is treated as one node, and the relations between the utterances are the edges. The directed graph $G = (H_D, \varepsilon)$, and the edge is $(h_i, r, h_j) \in \varepsilon$. The hidden state in the t-th layer is

$$h_i^{(t)} = \sigma\left(\frac{G_i^T (h_i^{(t-1)} W + b)}{\sum G_i}\right) \quad (13)$$

$\sigma()$ is an activation function; we use $ReLU()$ in this study . W is the weight matrix and b is the bias. Finally, we obtain the inter-utterance representation $\hat{H}_D$.

3. Experiments and analysis

3.1. Experimental Setup

To verify the effectiveness of the proposed TFCap and dialogue-level SER, we set up three groups of experiments. The first group of experiments are visualization. The second group shows the classification results in the intra-utterance stage. The last group shows the classification result at the dialogue-level. The interactive emotional dyadic motion capture database (IEMOCAP) [22] is a database . We only used audio data, which had 5,531 utterances and sampled at 16KHz. The data consisted of four emotion categories: neutrality (29%), anger (20%), sadness (20%), and happiness (31%). The length of each segment containing effective emotional information is thus an open problem, and in this study, we use the same preprocessing method as Satt et al. [12]. The time of each segment is $265ms$, and the input spectrogram has the following $time \times frequency$: 32×128. We choose cross entropy as the cost function, Adamax as the optimizer, and ReLU as the activation. The batchsize was set to 128. There were five sessions in the IEMOCAP. Considering the form of data organization at the dialoguelevel, we set session 1 to session 4 as the training data, and session 5 as the testing data.

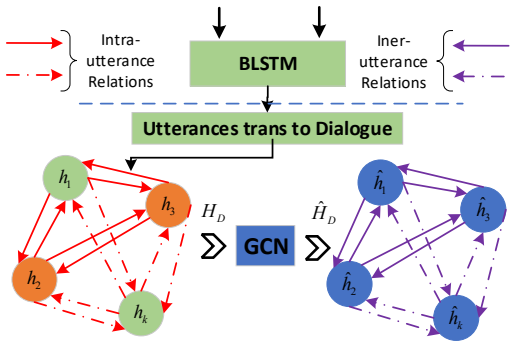


Figure 3: *Inter-utterance representation learning (GCN).*

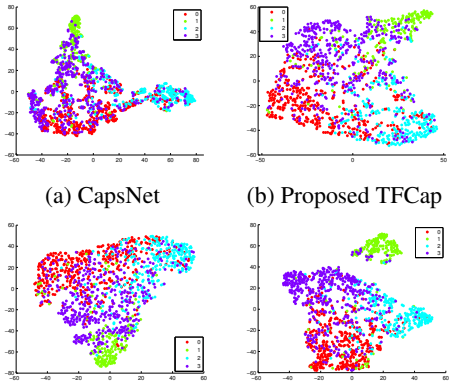
3.2. Experiment results and analysis

The evaluation criteria of the classification results are weighted accuracy (WA), unweighted accuracy (UA), and F1-score.

3.2.1. Visualization

To observe the representations extracted by CapsNet, the proposed TFCap, TFCNN [13], and proposed TFCNN_TFCap, t-distributed stochastic neighbor embedding (t-SNE) [23] was introduced to visualize the four emotional categories, as shown in Fig. 4.(0: Ne, 1: An, 2: Sa, 3: Ha)

We find that the distribution in Fig. 4(a) is different from the other three. The blue points (Sadness) were distributed throughout the range. The performance of Fig. 4(d) is the best, which combines the advantages of Fig. 4(b) and Fig. 4(c). In particular, the distribution of the purple points (happiness) have the



(a) CapsNet (b) Proposed TFCap

(c) TFCNN (d) TFCNN_TFCap

Figure 4: *The t-SNE visualizations of extracted representations*

best degree of aggregation.

3.2.2. *Classification results of inra- and inter- utterances*

To quantitatively evaluate the performance of the proposed model, the classification results of the four comparative experiments are provided in Table 1. The baseline model is a CNN [12]. For a fair comparison, all the experiments in the intra-utterance stage used BLSTM as the contextual information learning method, as shown in Table 1. Our proposed global model TFCap achieved 64.71% with absolute increments of 4.92% and 3.55% over CapsNet and DenseCap on WA respectively . The proposed local-global time-frequency representation learning method TFCNN_TFCap achieved 71.88% with absolute increments 9.35% and 8.06% over CNN and TFCNN on WA. The classification results in Table 1 prove two phenomena: the proposed TFCap is effective, and even has a performance similar to TFCNN; the local-global time-frequency representation learning frame is effective.

Table 1: *The results in intra-utterance stage*

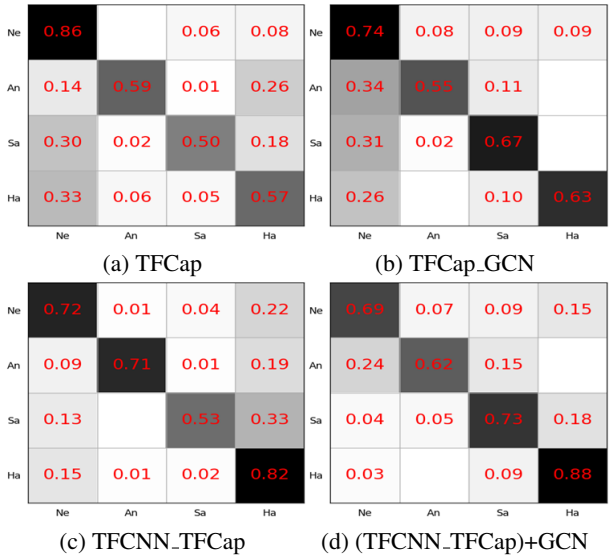
Model	WA(%)	UA(%)	F1(%)
CNN [9]	62.53	63.78	62.98
TFCNN [12]	63.82	65.53	64.06
CapsNet [2]	59.79	61.91	59.82
DenseCap [14]	61.16	61.46	61.44
TFCap	64.71	62.96	64.20
TFCNN_TFCap	71.88	69.60	72.25

The next group of experiments evaluates the effectiveness of the inter-utterance stage. Based on the previous experiments in Table 1, the results for the inter-utterance stage are shown in Table 2. In addition, two groups of confusion matrices for the proposed models in Table 2 are shown in Fig. 5.

From Table 2 and Fig 5, three phenomena were observed. The first is the increase in accuracy. The two experiments achieved 1.79% and 4.05% WA improvements, respectively. Second, the sensitivity to the four emotions is different. In IEMOCAP, the results in the intra-utterance stage have a higher sensitivity to neutrality and anger. On the contrary, the results in the inter-utterance stage have a higher sensitivity to sadness and happiness. Sadness and happiness are two two distinct atmospheres that are contained throughout the dialogue. More talking leads to better performance. Third, happiness is recognized as the most difficult. However, we obtained the best results for the proposed local-global time-frequency frame.

Table 2: *The ablation experiments in inter-utterance stage*

Model	WA(%)	UA(%)	F1(%)
TFCNN [12]	63.82	65.53	64.06
TFCap	64.71	62.96	64.20
TFCNN_TFCap	71.88	69.60	72.25
TFCNN+GCN	65.40	63.39	64.42
TFCap+GCN	66.50	64.85	65.60
(TFCNN_TFCap)+GCN	**75.93**	**72.73**	**72.92**



(a) TFCap (b) TFCap_GCN

(c) TFCNN_TFCap (d) (TFCNN_TFCap)+GCN

Figure 5: *Two groups of confusion matrices of proposed models*

4. Conclusions

Herein, we studied the local and global spatial representations from spectrograms combined with GCN for dialogue-level SER. The local representation is extracted from our previous work, TFCNN. The global representation was extracted using TFCap. TFCap is based on matrix vectors and a new routing algorithm. We also introduce GCN to model dialogue-level contextual information. The effectiveness of the proposed TFCNN+TFCap and GCN was verified through a series of comparative experiments on IEMOCAP. The proposed model achieved 71.88% in the intra-utterance stage and 75.93% in the in inter-utterance stage. In particular, the proposed TFCap model shows a significant improvement which can not only be used in SER, but also in many other speech tasks based on the spectrogram. The proposed architecture also provides a good direction for promoting research on SER .

5. Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grant 61771333 and the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

6. References

[1] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[2] L. Guo, L. Wang, and J. Dang, "A feature fusion method based on extreme learning machine for speech emotion recognition," *ICASSP 2018*, pp. 2666–2670, 2018.

[3] P. Zhou, X. Li, J. Li, and X. X. Jing, "Speech emotion recognition based on mixed mfcc," in *Applied Mechanics and Materials*, vol. 249, 2013, pp. 1252–1258.

[4] S. Lalitha, A. Mudupu, B. Nandyala, and R. Munagala, "Speech emotion recognition using dwt," in *2015 IEEE International Conference on Computational Intelligence and Computing Research*, 2015, pp. 1–4.

[5] K. Rao, S. Koolagudi, and R. Vempada, "Emotion recognition from speech using global and local prosodic features," *International Journal of Speech Technology*, vol. 16, pp. 143–160, 2013.

[6] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: Hsf-dnn, ms-cnn and lld-rnn," *Speech Communication*, vol. 120, pp. 11–19, 2020.

[7] B. T. Atmaja and M. Akagi, "The effect of silence feature in dimensional speech emotion recognition," *arXiv preprint arXiv:2003.01277*, 2020.

[8] A. C.N, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, pp. 155–177, 2012.

[9] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[10] J. Lee and I. Ivan, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. Interspeech*, 2015, pp. 1537–1540.

[11] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[12] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. Interspeech*, 2017, pp. 1089–1093.

[13] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, *Time-Frequency Deep Representation Learning for Speech Emotion Recognition Integrating Self-attention*, 2019.

[14] S. Sabour, N. Frosst, and G. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.

[15] K. Duarte, Y. Rawat, and M. Shah, "Videocapsulenet: A simplified network for action detection," in *Advances in Neural Information Processing Systems*, 2018, pp. 7610–7619.

[16] B. Zhang, X. Xu, M. Yang, X. Chen, and Y. Ye, "Cross-domain sentiment classification by capsule network with semantic rules," *IEEE Access*, vol. 6, pp. 1–1, 2018.

[17] Y. Min, M. Zhao, J. Ye, Z. Lei, Z. Z, and S. Zhang, "Investigating capsule networks with dynamic routing for text classification," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3110–3119, 2018.

[18] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7174–7178.

[19] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[20] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, no. 99, pp. 1–1, 2020.

[21] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[22] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, and et.al., "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, p. 335, 2008.

[23] L. V. D. Maaten, "Learning a parametric embedding by preserving local structure," *Journal of Machine Learning Research*, vol. 5, pp. 384–391, 2009.