Learning Local to Global Feature Aggregation for Speech Emotion Recognition

Cheng Lu[†,1,2], Hailun Lian[†,1,3], Wenming Zheng[,1,2], Yuan Zong[*,1,2], Yan Zhao[1,3], Sunan Li[1,3]*

[1]Key Laboratory of Child Development and Learning Science of Ministry of Education,
Southeast University, Nanjing, China
[2]School of Biological Science and Medical Engineering, Southeast University, Nanjing, China
[3]School of Information Science and Engineering, Southeast University, Nanjing, China

{cheng.lu, lianhailun, wenming_zheng, xhzongyuan}@seu.edu.cn

Abstract

Transformer has emerged in speech emotion recognition (SER) at present. However, its equal patch division not only damages frequency information but also ignores local emotion correlations across frames, which are key cues to represent emotion. To handle the issue, we propose a Local to Global Feature Aggregation learning (LGFA) for SER, which can aggregate long-term emotion correlations at different scales both inside frames and segments with entire frequency information to enhance the emotion discrimination of utterance-level speech features. For this purpose, we nest a Frame Transformer inside a Segment Transformer. Firstly, Frame Transformer is designed to excavate local emotion correlations between frames for frame embeddings. Then, the frame embeddings and their corresponding segment features are aggregated as different-level complements to be fed into Segment Transformer for learning utterance-level global emotion features. Experimental results show that the performance of LGFA is superior to the state-of-the-art methods.

Index Terms: speech emotion recognition, Transformer, time-frequency feature, frame-level, segment-level

1. Introduction

Speech emotion recognition (SER) is a significant task of affective computing and has attracted wide attention in recent years [1], [2]. The key to addressing the SER is how to disentangle the emotion information hidden in speech from the confusion of diverse acoustic factors [3], [4], [5], e. g., background noise, language, speaker identity.

Actually, the emotional information is always discretely distributed in frames or segments of speech [6], [7] due to the presence of special frames or segments without emotional contexts, i .e., empty frames/segments. In other words, emotion information is always discretely distributed in some key frames or segments. Therefore, a practical approach is to capture long-range emotion dependencies from these key frames/segments [3], [4], [8], [9]. To this end, Recurrent Neural Networks (RNNs) [8], [9] are widely adopted for learning utterance-level emotion features from frame-level or segment-level features.

Although previous works based on RNNs, e. g., LSTM and Bi-LSTM, have achieved great success on SER, they still encounter some issues [10], e. g., high time and space complexity for computing cells and only modeling sequential long-term dependencies (from forward to backward, or reverse). With the emergence of Transformer [11], these issues have been handled effectively. In Transformer, the Multihead Self-Attention can describe the complete relationship between all speech frames/segments. Also, the time-space complexity could be effectively reduced by the matrix parallel calculation. Taking these advantages, the Speech Transformer models [12], [13] are promisingly developed from the Vision Transformer (ViT) [14].

However, Speech Transformer roughly divides the speech spectrogram into same "chunks" [15] (i. e., patches in ViT), leading to lossing local inter-frame relationships reflecting the fine-gained emotion distribution and corruption of frequency domain information. Since the frame-level and segment-level features contain the emotional information at different scales [16], [17], e.g., frames reflect the phoneme-level associations and segments respond to the word-level or phrase-level correlations, they should be aggregated complementarily to learn more emotion-discriminative speech features. Likewise, ViT also ignores the local structure information in image patches for computer vision. To handle the similar issue, Han et al. [18] proposed a Transformer in Transformer (TNT) to simultaneously learn inter-patch and intra-patch relationships.

Inspired by TNT [18], we propose a novel Local to Global Feature Aggregation learning (LGFA) method for SER. The LGFA nests a Frame Transformer inside a Segment Transformer to aggregate different-scale emotion dependencies for the speech emotion representation. The whole learning processing of LGFA is from frame-level to segment-level to utterance-level. Compared with other Speech Transformer-based methods, our LGFA is a novel and special Transformer-based model for SER and its advantages can be summarized as the following three folds:

1. *it aims to capture long-range emotion-related dependencies at different scales both inside frames and segments instead of the simple image patches adopted in Transformer.*

2. *it takes a frame and a segment as the input of Frame Transformer and Segment Transformer, respectively, instead of equally divided image patches. In this case, the frame and segment used in LGFA may contain the entire frequency domain information such that the frequency feature will not be damaged in the speech chunk division.*

3. *it also can be extended from the time domain to the frequency domain and time-frequency domain by different patch partition strategies. This extension can make full use of the time-frequency characteristic of speech signals to represent emotion information.*

2. Proposed Method

Considering the inter-frame time property of speech, LGFA feeds a Frame Transformer with frame features, then integrates frame embeddings and segment features as the segment-level aggregation features. This point is the main difference from TNT. Further, these aggregation features are regarded as the

input of a Segment Transformer to learn higher-level emotion correlations across segments. Consequently, we can obtain the global utterance-level features of speech emotions through joint training of the Frame and Segment Transformers. The overview of LGFA is shown in Figure. 1, in which the Frame Transformer takes the frame-level feature of speech as the input.

To this end, we firstly process the frame-level feature of speech. Given the log-Mel-spectrogram feature $\boldsymbol{x} \in \mathbb{R}^{F \times T \times C}$ of each emotional speech, the ith frame $\boldsymbol{x}_i \in \mathbb{R}^{F \times C}$ of the spectrogram $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^T$ is firstly encoded by a linear projection layer $FC(\cdot)$ as the ith frame embedding $\boldsymbol{x}_i' \in \mathbb{R}^{1 \times d_f}$, denoted as

$$\boldsymbol{x}_i' = FC(\boldsymbol{x}_i), \tag{1}$$

where F, T, and C represent the numbers of Mel-scaled frequency, time frame and channel, respectively. d_f is the dimension of frame embeddings. Then, to enhance inductive bias of Frame Transformer [14], we add a learnable position encoding $\boldsymbol{e}_i^f \in \mathbb{R}^{1 \times d_f}$ into $\boldsymbol{x}_i'$ as the input of Frame Transformer, which can be represented as

$$\boldsymbol{x}_i' \leftarrow \boldsymbol{x}_i' + \boldsymbol{e}_i^f, \tag{2}$$

where $\boldsymbol{e}^f = \{\boldsymbol{e}_i^f\}_{i=1}^T \in \mathbb{R}^{T \times d_f}$. In Frame Transformer, the sequence of speech frame embeddings $\boldsymbol{x}' = \{\boldsymbol{x}_i'\}_{i=1}^T$ is utilized to characterize local inter-frame correlations of emotions. Then, the frame-level encoding $\hat{\boldsymbol{x}}$ can be obtained by the frame embedding sequence $\boldsymbol{x}'$ through the following operations:

$$\boldsymbol{x}''^{,\ell} = MSA(LN(\boldsymbol{x}'^{,\ell-1})) + \boldsymbol{x}'^{,\ell-1}, \tag{3}$$

$$\hat{\boldsymbol{x}}^\ell = MLP(LN(\boldsymbol{x}''^{,\ell})) + \boldsymbol{x}''^{,\ell}, \tag{4}$$

where $\ell \in [1, ..., L]$ is the index of the stacked block, L is the number of blocks in Frame Transformer, and $\hat{\boldsymbol{x}}^\ell \in \mathbb{R}^{T \times d_f}$ is encoded by the ℓth block. Besides, in Equation (3) and (4), $MSA(\cdot)$, $MLP(\cdot)$ and $LN(\cdot)$ are the operations of Multi-head Self-Attention (MSA), MultiLayer Perceptron (MLP), and Layer Normalization (LN), respectively, according to [14], [18]. Notably, $\boldsymbol{x}'^{,0} = [\boldsymbol{x}_1', \boldsymbol{x}_2', ..., \boldsymbol{x}_T'] \in \mathbb{R}^{T \times d_f}$ in Equation (2) is the initial input of the frame embedding sequence $\boldsymbol{x}'$.

To aggregate the emotion-related dependencies at different scales, we further design a Segment Transformer to learn frame-level and segment-level correlations of speech emotion. Therefore, the input of Segment Transformer is the combination of the frame-level encoding $\hat{\boldsymbol{x}}$ and segment-level embedding $\boldsymbol{s}$. Specifically, the log-Mel-spectrogram feature $\boldsymbol{x}$ can be divided into a segment set, where each segment $\boldsymbol{s}_j \in \mathbb{R}^{k \times F \times C}$ consists of k frames, represented as $\boldsymbol{x} = \{\boldsymbol{s}_j\}_{j=1}^{T/k}$. Similar to the Frame Transformer, each segment $\boldsymbol{s}_j$ is firstly transformed to the segment embedding $\boldsymbol{s}_j' \in \mathbb{R}^{1 \times d_s}$ by a linear projection layer $FC(\cdot)$ in Segment Transformer. Besides, the kth frame-level encoding $\hat{\boldsymbol{x}}_j^s \in \mathbb{R}^{k \times d_f}$ corresponding to the jth segment are also used to aggregate into the segment embeddings after another linear projection $FC(\cdot)$, where $FC(\cdot)$ is to ensure dimension match for the addition of frame encoding and segment embedding. Then, the combination embedding $\boldsymbol{s}_j'' \in \mathbb{R}^{1 \times d_s}$ of frame-level encoding and segment embeddings is generated by

$$\boldsymbol{s}_j' = FC(Vec(\boldsymbol{s}_j)), \tag{5}$$

$$\hat{\boldsymbol{x}}_j^s = [\hat{\boldsymbol{x}}_{i+1}, \hat{\boldsymbol{x}}_{i+2}, .., \hat{\boldsymbol{x}}_{i+k}], \tag{6}$$

$$\boldsymbol{s}_j'' = \boldsymbol{s}_j' + FC(Vec(\hat{\boldsymbol{x}}_j^s)), \tag{7}$$
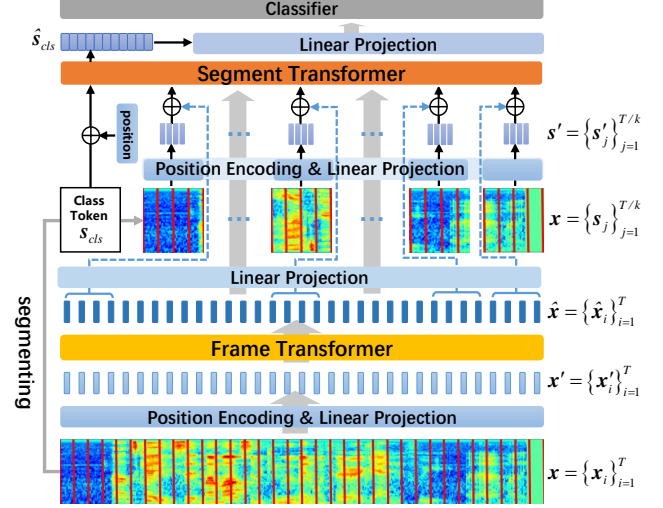

Figure 1: *Overview of Local to Global Feature Aggregation (LGFA) framework for SER. LGFA includes two Transformer-based networks, in which a Frame Transformer is nested inside a Segment Transformer.*

where $Vec(\cdot)$ is a vectorization operation to flatten the dimension of $\hat{\boldsymbol{x}}_j^s$ or $\boldsymbol{s}_j$ to $\mathbb{R}^{1 \times (k \times d_f)}$. Then, we also add a learnable class token $\boldsymbol{s}_{cls}$ into input sequence for the final emotion classification. Eventually, the segment-level embedding $\boldsymbol{s}'' \in \mathbb{R}^{(T/k+1) \times d_s}$ can be written to

$$\boldsymbol{s}'' = [\boldsymbol{s}_{cls}, \boldsymbol{s}_1'', \boldsymbol{s}_2'', ..., \boldsymbol{s}_{T/k}'']. \tag{8}$$

Similar to Frame Transformer, each segment-level embedding with frame-level aggregation is added the corresponding positions between segments to preserve time-sequence property of inductive bias on speech by a learnable position encoding $\boldsymbol{e}_j^s \in \mathbb{R}^{1 \times d_s}$, which can be denoted as

$$\boldsymbol{s}_j'' \leftarrow \boldsymbol{s}_j'' + \boldsymbol{e}_j^s, \tag{9}$$

where $\boldsymbol{e}^s = \{\boldsymbol{e}_j^s\}_{j=1}^{T/k} \in \mathbb{R}^{(T/k+1) \times d_s}$. The Segment Transformer also adopts L stacked standard transformer blocks to encode the aggregation embedding for the utterance-level representation of speech emotion, where the ℓth block transformations are formalized to

$$\bar{\boldsymbol{s}}^\ell = MSA(LN(\boldsymbol{s}''^{,\ell-1})) + \boldsymbol{s}''^{,\ell-1}, \tag{10}$$

$$\hat{\boldsymbol{s}}^\ell = MLP(LN(\bar{\boldsymbol{s}}^\ell)) + \bar{\boldsymbol{s}}^\ell, \tag{11}$$

where $\boldsymbol{s}''^{,0}$ is the initial segment embedding sequence in Equation (9).

With all the above operations, our proposed LGFA firstly models local emotion correlations within frames by Frame Transformer $I(\cdot)$, then aggregates the frame-level encoding $\hat{\boldsymbol{s}}$ and segment embeddings $\boldsymbol{s}''$ to capture global longer-dependencies for the utterance-level emotion representation $\hat{\boldsymbol{s}}$ through Segment Transformer $O(\cdot)$, which can be denoted as

$$\hat{\boldsymbol{s}} = O(\boldsymbol{s}''; I(\boldsymbol{x}')), \tag{12}$$

Furthermore, the class token $\hat{\boldsymbol{s}}_{cls}$ can be generated from $\hat{\boldsymbol{s}}$ to input the classifier for speech emotion prediction, represented as

$$y_{pred} = C(\hat{\boldsymbol{s}}_{cls}), \tag{13}$$

where y_{pred}, C, and $\hat{s}_{cls}$ are the predicted labels of emotions, classifier, and s_{cls} generated by LGFA, respectively. Note that the segment class token s_{cls}, frame position encoding e^f and segment position encoding e^s are all initialized as zeros in the letter.

3. Experiments

In the section, we will introduce the details of our implemented experiments, then discuss the comparison results of the proposed LGFA with state-of-the-art methods.

Database: To evaluate the performance of our proposed LGFA, two public emotional speech databases are selected to implement the experiments, i. e., the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [19] and the China Emotional Database (CASIA) [7]. In detail, IEMOCAP is an English multimodal database containing video, speech, and text scripts, which is recorded in 5 sessions (1 male and 1 female in each session) by inducing diverse emotions (*angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and others*) of 10 actors under improvised or scripted scenarios. CASIA is a Chinese Emotional Speech Database with 9 600 recording files under 6 emotions (*angry, fear, happy, neutral, sad, and surprise*). It is collected by inducing 4 actors (2 males and 2 females) to express 6 emotions under several fixed text contents. Note that we adopt 2 280 improvised samples and 4490 scripted+improvised samples with 4 emotions (*angry, happy, sad, and neutral*) in IEMOCAP, and 1 200 public released samples with 6 emotions in CASIA for experiments.

Experimental Settings: In our experiments, all speech sentences are re-sampled to 16 kHz for Short-Time Fourier Transform (STFT) using 20 ms Hamming window size with 50% frame overlapping. Then, they are divided into segments with 128 frames as experimental samples and pad 0 for the segment less than 128 frames. Finally, we obtain the log-Mel-spectrogram with the dimension of $\mathbb{R}^{64\times128\times1}$ for the input of our LGFA, where the number of Mel-filter is set as 64.

For the network of LGFA, the input sizes of Frame Transformer and Segment Transformer are assigned as $(64, 128, 1)$ and $(64, 8, 1)$. The number of stacked blocks L is 7. Furthermore, the projection dimensions and the head number of the Frame Transformer are set as 16 and 4, and they are assigned 256 and 4 in the Segment Transformer. The LGFA is implemented by PyTorch with NVIDIA A10 GPUs. And it is optimized by the AdamW Optimizer with a learning rate of 0.0001 and trained from scratch with a batch size of 64.

In addition, the Leave-One-Subject-Out (LOSO), i. e., k-fold cross-validation protocol (CV), is adopted for a fair comparison according to [3], [4], where k is the speaker number of dataset. Therefore, the speaker rate of training and testing data in IEMOCAP and CASIA are 9:1 and 3:1, respectively. Furthermore, since the IEMOCAP are class-imbalanced, the weighted average recall (WAR) and the unweighted average recall (UAR) [3], [4] are used to effectively evaluate the performance of the proposed method, where WAR is standard recognition accuracy while UAR is the class-wise accuracy.

Results and Analysis: We compare the performance of our proposed LGFA with several state-of-the-art methods on IEMOCAP, i. e., CNN+LSTM Model [8], DNN-HMM based model (DNN-HMM_SGMM-Ali.) [20], CNN model with spectrogram (model-2A(spectrogram)) [21], fusion model with different acoustic features (Model-3 (fusion) and Model-1 (dow.+ens.)) [22]. The above methods are all implemented on the improvised data (2280 samples). To further demonstrate

Table 1: *Experimental results on IEMOCAP, where the best results are highlighted in bold. The first five methods are implemented on the scripted+improvised data, and others are based on the improvised data.*

Comparison Methods	Accuarcy(%)	
	WAR	UAR
Bi-LSTM [23]	57.87	48.54
Greedy+Dro.+Att.+MLP [23]	56.33	49.96
ViT	63.57	56.62
TNT	63.14	56.18
LGFA (ours)	**64.12**	**57.78**
DNN-HMM_SGMM-Ali. [20]	62.28	58.02
CNN+LSTM [27]	64.50	61.70
Model-2A (spectrogram) [21]	71.30	61.60
Model-3 (fusion) [22]	72.34	58.31
ViT	70.22	58.58
TNT	70.61	59.72
LGFA (ours)	**73.29**	**62.63**

the performance of LGFA, we also compare the LGFA with other methods (i. e., Bi-LSTM and Greedy+Dro.+Att.+MLP) [23] on the scripted+improvised data (4490 samples). Moreover, we also choose other comparison methods on CASIA, i. e., LLDs with dimension reduction (LLD+DR) [24], DNNs with the extreme learning machine (DNN+ELM) [25], weighted spectral feature learning model (HuWSF) [26], and DCNN with discriminant temporal pyramid matching (DTPM) [16]. As homologous methods to LGFA, ViT [14] and TNT [18] were also used as comparasion methods. Note that the results of DTPM, ViT, and TNT are obtained through our own implementations with the released codes[1,2,3]. In addition, to evaluate the experimental performance more comprehensively, these selected comparison methods are based on two commonly used experimental protocols on IEMOCAP, i. e., 10-fold LOSO based on speakers and 5-fold LOSO based on sessions. For example, Bi-LSMT, Greedy+Dro.+Att.+MLP, CNN+LSTM, DNN-HMM_SGMM-Ali., ViT, TNT and our proposed LGFA are all based 10-fold CV, other methods are based on 5-fold CV.

The experimental results with WAR and UAR on IEMO-CAP are shown in Table 1, where ViT and TNT are implemented by the spectrogram size of 128×128 and the chunk size of 16×16 according to [18], [14]. From these results, it is obvious that the proposed LGFA achieves the competitive performance on both WAR and UAR. Specifically, based on the scripted+improvised data, our LGFA improves the accuracies (6.25% on WAR and 7.82% on UAR) than comparison methods. Based on the improvised data, LGFA is superior to RNN-based methods (i. e., CNN+LSTM), demonstrating the advantage of the Transformer-based methods in SER. Further, its results also outperform the ViT and TNT, which reveals LGFA effectively capture the long-range emotion dependencies inside frames and segments for better speech representation and is more suitable for the task of SER than ViT and TNT. Although our LGFA achieve the best performance, the UAR results are lower than the WAR ones on comparison methods because of the class-imbalance in IEMOCAP.

The results on CASIA, illustrated in Table 2, also reveal

---

[1]https://github.com/tzaiyang/SpeechEmoRec
[2]https://github.com/lucidrains/vit-pytorch
[3]https://github.com/huawei-noah/CV-Backbones/tree/master/tnt_pytorch

Table 2: *Experimental results on CASIA, where the best results are highlighted in bold.*

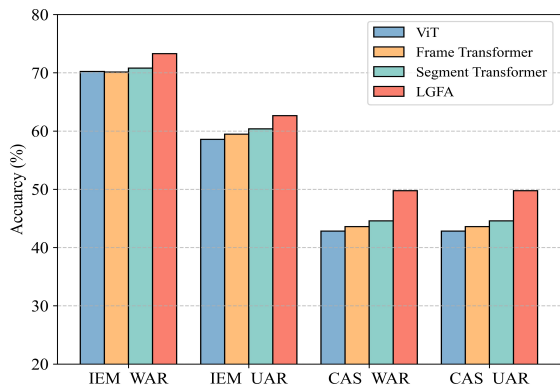| Comparison Methods	Accuarcy(%)	
	WAR	UAR
LLDs+DR [24]	39.50	39.50
DNN+ELM	41.17	41.17
HuWSF [26]	43.50	43.50
DTPM	45.42	45.42
ViT	42.83	42.83
TNT	46.58	46.58
LGFA (ours)	**49.75**	**49.75**

Figure 2: *Ablation study for LGFA, where IEM_WAR and CAS_WAR represent WAR results on IEMOCAP and CASIA, respectively. IEM_UAR and CAS_UAR are UAR results.*

the superiority of our LGFA (improving 3.17% on WAR and UAR). It is better than traditional methods (i.e., LLD+DR and HuWSF) and DNN-based approaches (i.e., DNN_ELM and DTPM). Similar to the results on IEMOCAP, our proposed LGFA proves its superiority on the SER again over ViT and TNT. Since CASIA is class-balanced, the results of WAR are equal to those of UAR.

Furthermore, to explore the effective components of LGFA, we implement extended experiments to analyze different architectures of our LGFA. Figure. 2 shows the results of ablation study, where ViT, Frame Transformer, and Segment Transformer are implemented by square chunks with the size of 16×16, frame chunks with the size of 64×1, segment chunks with the size of 64×8, respectively. The ablation results in Figure. 2 indicate that LGFA is superior in speech emotion representation over other architectures. Namely, our designed frame and segment aggregation learning is more suitable for SER than current Speech Transformers. Furthermore, the Segment Transformer outperforms the Frame Transformer, indicating larger chunks will promote the feature extraction of speech emotion for the Transformer.

Discussion on the extension of LGFA: In LGFA, to preserve the completeness of the frequency domain in the spectrogram, we divide the spectrogram feature as chunks only on the time domain. To further explore the effect of different chunk division strategies, we extend the chunk division of the proposed LGFA (i.e., LGFA_T in Table 3) from the time domain to the frequency and time-frequency domain (i.e., LGFA_F and LGFA_TF in Table 3). Compared with LGFA_T, LGFA_F takes

Table 3: *Extension experiments of LGFA on the time, frequency, and time-frequency domains, i.e., LGFA_T, LGFA_F, and LGFA_TF, where LGFA_T is the LGFA in Tables 1 and 2.*

Comparison Methods	IEMOCAP		CASIA	
	WAR	UAR	WAR	UAR
LGFA_T	**73.29**	**62.63**	49.75	49.75
LGFA_F	67.85	55.57	45.33	45.33
LGFA_TF	70.79	60.38	**50.17**	**50.17**

each frequency band as a frame and each frequency band group as a segment to learn the sentence-level emotion feature from the frequency domain. Thus, we can obtain the frequency-wise class token $\hat{s}_{cls}^{fre}$ of LGFA_F for the emotions prediction represented as $y_{pred} = C(\hat{s}_{cls}^{fre})$. Further, we will also complementarily combine the chunk division methods in the frequency and time domains to generate the fusion class token of LGFA_TF $\hat{s}_{cls}^{fu} = cat(\hat{s}_{cls}, \hat{s}_{cls}^{fre})$ to the emotion classifier $y_{pred} = C(\hat{s}_{cls}^{fu})$, where $cat(\cdot)$ is the concatenation operation on the feature dimension.

The experimental results of different chunk division strategies are shown in Table 3. From them, we observe that LGFA_T and LGFA_TF outperform LGFA_F, which may be due to the fact that speech emotion is closely related to the context within frames or segments. While in the frequency domain, not all emotions have obvious energy activations between frequency bands. Furthermore, the LGFA_TF outperforms LGFA_T on CASIA, while performs worse on IEMOCAP. The reason may be that chunk division in the frequency domain will not only complement the time-domain chunk division but may also integrate noise caused by the uncertain correlations on the frequency domain under emotions. The recording environment of CASIA contains less noise, while IEMOCAP is recorded in a open dialogue environment. Thus, the noise will affect frequency-domain correlations and impair the performance of the time-frequency fusion model. In other word, the frequency information should be screened to obtain this supplement.

4. Conclusions

We propose a novel Local to Global Feature Aggregation (LGFA) method for SER. LGFA integrates a Frame Transformer into a Segment Transformer to aggregate local emotion correlations at different scales both within frames and segments for the global utterance-level representation of emotional speech. Through the joint learning of two Transformers, we can obtain discriminative emotion features to learn speech emotion representation from frame-level to segment-level to sentence-level. Extensive experimental results on IEMOCAP and CASIA demonstrate the superiority of our proposed LGFA. Further, we will deeply explore the different chunk division strategies of LGFA for the better SER performance.

5. Acknowledgements

6. References

[1] B. W. Schuller, R. Picard, E. André, J. Gratch, and J. Tao, "Intelligent signal processing for affective computing," 2021.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 552–557.

[4] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5688–5691.

[5] C. Lu, Y. Zong, W. Zheng, Y. Li, C. Tang, and B. W. Schuller, "Domain invariant feature learning for speaker-independent speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2217–2230, 2022.

[6] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[7] J. T. F. L. M. Zhang and H. Jia, "Design of speech corpus for mandarin text to speech," in *The Blizzard Challenge 2008 workshop*, 2008.

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

[9] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu *et al.*, "Speech emotion recognition using capsule networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6695–6699.

[10] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] L. Tarantino, P. N. Garner, A. Lazaridis *et al.*, "Self-attention for speech emotion recognition." in *INTERSPEECH*, 2019, pp. 2578–2582.

[13] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, 2021.

[16] 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.

[17] G. Shen, R. Lai, R. Chen, Y. Zhang, K. Zhang, Q. Han, and H. Song, "WISE: Word-level interaction-based multimodal fusion for speech emotion recognition." in *INTERSPEECH*, 2020, pp. 369–373.

[18] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in Transformer," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[20] S. Mao, D. Tao, G. Zhang, P. Ching, and T. Lee, "Revisiting hidden markov models for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6715–6719.

[21] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding." in *INTERSPEECH*, 2018, pp. 3688–3692.

[22] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7189–7193.

[23] C.-W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition." in *INTERSPEECH*, 2016, pp. 1387–1391.

[24] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, 2018.

[25] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *INTERSPEECH 2014*, 2014.

[26] Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local Hu moments for speech emotion recognition," *Biomedical Signal Processing and Control*, vol. 18, pp. 80–90, 2015.

[27] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "CNN + LSTM architecture for speech emotion recognition with data augmentation," in *Workshop on Speech, Music and Mind 2018*. ISCA, 2018, pp. 21–25.