INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Stable Speech Emotion Recognition with Head-k-Pooling Loss

Chaoyue Ding[1*], *Jiakui Li*[1*], *Daoming Zong*[1*], *Baoxiang Li*[1†], *Tianhao Zhang*[1,2], *Qunyan Zhou*[1]

[1]SenseTime Research
[2]University of Science and Technology Beijing
{dingchaoyue, lijiakui, zongdaoming, libaoxiang, zhangtianhao1}@sensetime.com

Abstract

Speech emotion recognition (SER) aims to detect the emotion of the speaker involved in a given utterance. Most existing SER methods focus on local speech features by stacking convolutions and training all segments of an utterance with an utterance-level label. Two deficiencies exist in these methods: *i*) learning only local speech features may be insufficient for SER due to the ambiguity of emotions; *ii*) consistent supervision of each segment may lead to label error propagation, as the true emotions of some segments may not match the utterance label. To solve the two issues, we first devise a global-local fusion network to model both long- and short-range relations in speech. Second, we tailor a novel head-k-pooling loss for SER tasks, which dynamically assigns labels for each segment and selectively performs loss calculation across segments. We test our method on the IEMOCAP and the newly collected ST-EMO dataset, and the results show its superiority and stability.

Index Terms: speech emotion recognition, self-attention, temporal convolution, negative sampling

1. Introduction

Emotion, as one of the basic paralinguistic information, conveys the user's intention and status, which helps the speech interaction system to improve user experience. Speech emotion recognition (SER), aiming to identify human emotions from speech, has been an active research field for decades [1, 2, 3]. It is widely used in numerous applications, such as intelligent robots, automated call centers, and distance education [4, 5].

Recently, deep learning methods have garnered increasing attention due to their powerful representation capabilities [6, 7, 8]. Particularly, convolutional neural network (CNN)-based methods have seen substantial improvements in SER over traditional methods, as the inductive biases inherent to CNNs, such as spatial locality and translation equivariance, are believed to be helpful for learning sound speech representations [9, 10]. Earlier works use fixed-scale convolutional kernels to extract emotion-related representations, but ignores the variation in the expression of emotions at different scales [11, 12, 13]. Later, work on multi-scale CNNs was proposed to address this issue, such as Light-SERNet [9] and deep-CNN [14]. They proposed to aggregate the information of multi-scale features extracted by multiple branches of convolutional kernels with varying sizes. Despite their modest success, these methods still suffer from either *emphasizing too much on local relationship modeling* or *ignoring segment-level label mismatch*.

For the first point, notice that most existing CNN-based methods can directly capture the local relationship between adjacent feature frames, but lack the ability to model the long-term dependencies between temporally distant frames, which may also be valuable for SER recognition [15, 16]. Recent approaches also explore the self-attention mechanism [17] to capture long-term relationships in speech for emotion detection [4, 18, 19]. Since self-attention can relate any two temporal spans of the feature input, it comes naturally to integrate self-attention and CNNs to model global and local temporal dependencies for SER. For the second point, given an utterance, current SER methods usually divide it into multiple segments using fixed windows and hop lengths, then use the same label to supervise each segment for training. However, the true emotions of these segments may not match the original utterance label. This is because the degree of human emotional intensity is time-varying. Emotional variability can be represented by two main dimensions: valence (a continuum from negative to positive) and arousal (a continuum that varies from low to high) [20]. Depending on the speaker's individual pronunciation habits (*e.g.*, variation in intonation and pitch), the valence and arousal value of an utterance may not be maintained at the same level over time. This means, as shown in Fig. 1, a Chinese Mandarin utterance labeled as `happy` may also contain some `neutral` segments. With this in mind, training all segments with the original utterance label will introduce noise data and

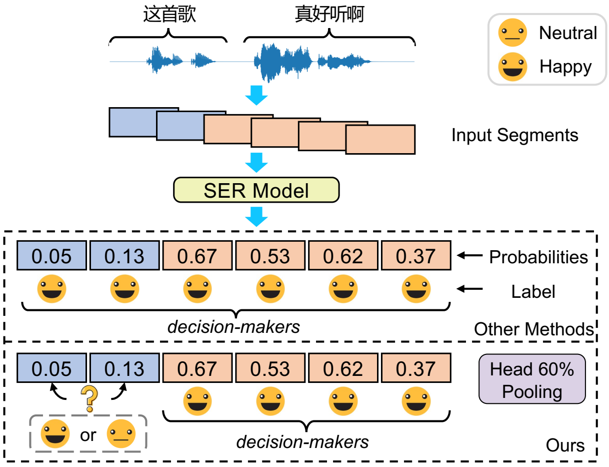

Figure 1: *Illustration of the head-k-pooling loss. In the example utterance "这首歌 (this song) 真好听啊 (is so nice)", the tone rises at the end of the utterance, so the emotional states of the front segments are ambiguous, possibly neutral. Therefore, instead of using all segments, we pick the top k% confident segments as decision-makers and only use them to compute loss. The gradients are set to zero for those uncertain segments.*

[*] Equal contribution
[†] Corresponding author

10.21437/Interspeech.2023-80

may lead to model performance degradation.

In this work, we propose a **g**lobal-**l**ocal **r**elation **f**usion network, named GLRF, for speech emotion recognition. The core component of GLRF is its inherent global-local relation-aware block (GLRA), where one branch specializes in local context modeling (by temporal convolution) while another branch specializes in long-distance relationship modeling (by multi-head self-attention). Besides, to reduce error propagation caused by misallocation of segment labels, we design a novel head-k-pooling loss for GLRF training, which enables the model to pick the top $k\%$ confident segments to teach itself. More specifically, we automatically select the top $k\%$ segments as *decision-makers* according to the posterior probability of the current emotion category output by the model, and these *decision-makers* are trained using the same label of utterance. The remaining segments are treated as *assistants*. Under the current emotion category, the gradients for those *assistants* are set to zero, as it is uncertain whether these *assistants* have the same emotion category as the *decision-makers*, and they may present a `neutral` affective state. When performing negative sampling, these *assistants* are only judged to be negative samples of categories that differ greatly from the valence and arousal value of the current emotion category. As an example shown in Fig. 1, for an assistant segment of `happy`, we just assume that its class is unlikely to be `fear`, `sad`, or `angry` and do not regard it as a negative sample of the `neutral` class. To summarize, our main contributions are as follows:

- We propose a GLRF, a novel architecture for speech emotion recognition, which can efficiently capture both long-term and short-term temporal dependencies in speech.
- We customize a head-k-pooling loss to enhance the training of GLRF. This loss helps the model select the top $k\%$ of segments with the highest confidence based on posterior probabilities, reducing errors caused by inaccurate segment label assignments and minimizing the propagation of errors in the training process.
- We introduced ST-EMO, the largest (153 hours) Chinese Mandarin speech emotion dataset for advancing SER research. In particular, ST-EMO covers some in-car driving scenarios.

2. Method

Model Pipeline. Fig. 2 illustrates the overall architecture of the proposed GLRF. Given a long-form utterance, we first split it into several 2s segments by a sliding window of 1.8s overlap. We then extract the mel-frequency cepstral coefficients (MFCCs) features per segment as input. During training, we send each segment to GLRF and obtain the segment-level confidence scores for all emotion categories by learning multiple binary classifiers for each class *vs.* all other classes (*i.e.*, one-*vs*-rest). We then select segments with the top $k\%$ confidence and train them with the same label as the corresponding utterance via multiple binary cross-entropy losses. Below we detail the core ingredient of GLRF, namely the global-local relation-aware (GLRA) block, and head-k-pooling loss tailored for SER.

2.1. Global-Local Relation-Aware Block

Global-Local Relation Module. Formally, GLRF takes a set of MFCCs features $\mathbf{X} \in \mathbb{R}^{\mathsf{N}\times\mathsf{F}\times\mathsf{T}}$ as input, where N is the number of segments, T is the number of frames per segment, and F denotes the feature dimension. To begin with, $\mathbf{X}$ is first temporal convolved by a 1D convolution with a kernel size of 3, followed by batch normalization and ReLU activation, yielding the feature map of $\mathbf{F}_0 \in \mathbb{R}^{\mathsf{N}\times\mathsf{C}_0\times\mathsf{T}}$. $\mathbf{F}_0$ is then sent into the GLRA block, which can be decomposed into global and local relation-aware branches (see Fig. 2). The global branch model global contexts by a normal multi-head self-attention module, while the local branch captures local contexts by a composition of temporal convolutions and linear layers. Both branches use residual connections. Such a dual-stream approach places attention and convolution modules in parallel, encouraging them to have different views on the speech from global and local perspectives, so that the architecture can benefit from specialization and achieve higher efficiency.

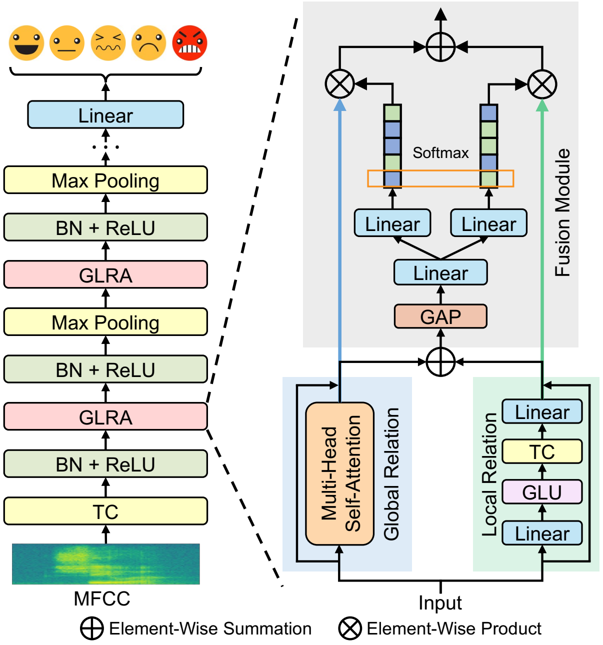

Figure 2: *Illustration of the architecture design of GLRF as well as the global-local relation-aware (GLRA) block.*

Global-Local Fusion Module. Later, we fuse information from two branches via gates. To determine the contribution of the information flow, gates need to integrate information from two branches. This can be achieved through an element-wise summation. Then we apply global average pooling to generate channel-wise statistics as $\mathbf{s}$. Further, we compress $\mathbf{s}$ into a compact feature $\mathbf{z}$ via a simple fully-connected (FC) layer for efficiency. Next, $\mathbf{z}$ is distributed to calculate the attention scores of each information flow. Finally, we multiply the normalized attention scores by the results from two branches. After a GLRA block, we get $\mathbf{F}_1 \in \mathbb{R}^{\mathsf{N}\times\mathsf{C}_1\times\mathsf{T}}$. We stack multiple GLRA blocks and perform max-pooling between two GLRA blocks to halve the size of the temporal dimension T. The output of the final GLRA block is fed into an FC layer for scoring each segment.

2.2. Head-k-Pooling Loss

It is common practice to segment an original utterance into multiple segments for data augmentation or to simulate streaming speech input. However, due to differences in individual pronunciation habits and variability in the distribution of arousal values, segments in an utterance may not have a consistent emotional state. For instance, in the utterance “这首歌真好听啊”, the upscaling is concentrated at the end of the utterance, so the emotional classes of the front segments are ambiguous, and they may be neutral. Therefore, instead of using all segments, we choose to use those most certain segments, referred to as *decision-makers*, to guide the model. Formally, let N denote the number of segmented segments for an utterance. C is the cardinality of emotion classes, and z_i^j is the output logits of the classifier for the i-th class in the j-th segment. For an utterance labeled as emotion class i, we first select the K segments based on the posterior probabilities by:

$$\{\ell^j\}_{j=1}^{K} = rank(k)\left(\text{sigmoid}(z_i^1), \dots, \text{sigmoid}(z_i^N)\right) \quad (1)$$

where $rank(k)$ is an operator that returns the indices of the top $k\%$ elements ($k \in [10, 100], K = \lceil k\% \cdot N \rceil$), the ℓ^j indexes the selected segment. Then the head-k-pooling loss for a single utterance can be defined as:

$$\mathcal{L} = \frac{1}{K}\frac{1}{C}\sum_{j=1}^{K}\sum_{i=1}^{C}[y_i^{\ell j}\log(1+e^{-z_i^{\ell j}})+(1-y_i^{\ell j})\log(1+e^{z_i^{\ell j}})] \quad (2)$$

Negative Sampling. We perform negative sampling in unselected segments. Two situations need to be carefully considered: 1) if the emotional state of an utterance is labeled as `neutral`, it can be used as a negative sample for any other category; and 2) if the current label of an utterance is non-neutral, it can appear in the negative sample pool excluding current class and neutral class. For example, segments labeled as `happy` in an utterance are unlikely to be `fear`, `sad` or `angry`, but some of them may be `neutral`.

3. Experiments

3.1. Datasets and Evaluation Protocols

We evaluate our model on the Interactive Emotional Dyadic Motion Capture database (**IEMOCAP**) [21] and our collected Chinese Mandarin dataset **ST-EMO**. Specifically, **IEMOCAP** contains 12 hours of English audiovisual data divided into five sections, each with scripted and improvised scenes recorded by a male and a female professional actor. Scripted parts are performed for predetermined emotions, while improvised parts are closer to natural speech. Following previous works [2, 9, 22], we first merge the two classes of excited and happy as they are close in the valence and arousal (VA) domain [23], then select four types of emotions (*i.e.*, angry, happy, sad, and neutral) for experiments, and evaluate on the full (scripted+improvised) dataset. **ST-EMO** comprises 153 hours of audio data, which are divided into five emotion classes (*i.e.*, angry, happy, sad, neutral, and fear), and were recorded by 50 male actors and 50 female actresses in the text-independent, text-dependent, and improvised scenario, respectively. Speaker age ranges from 20 to 41. ST-EMO was collected in several small rooms with no background noise, using Android smartphones (48kHz, 16-bit) fixed at a distance of 25cm directly in front of the actors/actresses. The duration of all utterances in ST-EMO ranges between 1.0 and 17.2 seconds, with an average duration of 5.6 seconds, and over 91.2% of them range between 3 and 10 seconds. Text-independent and text-dependent scripts are relevant to everyday life, which resembles the SMP2020-EWECT [24] and CPED [25] datasets. Statistically, the text-independent section consists of 100 scripts, each of which was performed by all actors/actresses with five distinct emotions, *e.g.*, saying the script "原来是有人送你回来" with five distinct emotions. The text-dependent section includes 500 emotion-related scripts, with each emotion corresponding to 100 scripts, and all actors/actresses are required to perform each script with the specific emotion, *e.g.*, saying the script "这的确是个好消息，赶紧告诉我的好朋友去" with `happy` emotion. The improvised section contains 100 in-car scenarios, with each emotion corresponding to 20 scenarios. For each scenario, actors and actresses are asked to speak freely with the specified emotion. For instance, let the actor imagine a driving scene like "当你开车上坡时，汽车在坡上溜车了" and act freely with the emotion of fear. "好可怕，差点碰到后面的车了" may be a possible response utterance.

Table 1: *Statistics of the newly collected ST-EMO dataset.*

	Angry	**Happy**	**Neutral**	**Sad**	**Fear**	**All**
Independent	9,132	5,818	9,644	8,760	7,048	40,402
Dependent	9,711	7,323	9,587	9,134	7,402	43,157
Improvisation	1,978	1,781	1,980	1,641	1,800	9,180
Full Dataset	20,821	14,922	21,211	19,535	16,250	92,739

Table 2: *Performance comparison between GLRF and its counterparts in terms of WA (%), UA (%), and Micro-F1 (%).*

	IEMOCAP			**ST-EMO**		
	WA	**UA**	**Micro-F1**	**WA**	**UA**	**Micro-F1**
Deep-CNN	64.59	65.35	64.43	68.57	68.96	69.46
L-SERNet	68.21	68.10	68.15	71.23	70.01	71.52
GLAM	69.74	71.03	69.70	71.79	70.72	71.88
DRN-MHSA	66.67	67.56	66.73	69.92	68.64	70.48
AA-CNN	70.38	71.71	70.13	72.53	71.21	72.69
GLRF	**72.81**	**73.39**	**72.92**	**75.33**	**73.88**	**75.29**

After collecting the raw data, we denoised the data. For each piece of data, we consulted four experts to judge its emotion category. The data is considered valid only when the categories given by at least two experts are consistent with the original label of the data. By this means, we got a total of 92,739 pieces of data. Table 1 lists the detailed statistics of the dataset.

Three commonly-used metrics are adopted for assessment: weighted accuracy (WA), unweighted accuracy (UA), and micro-F1 score. The difference between WA and UA is that UA considers label imbalance in computing accuracy, while UA does not. In the testing phase, the final prediction is obtained by averaging the predictions over all segments in an utterance. In addition, to gain a better understanding of the proposed loss, we develop a new metric, called *pure accuracy*, to measure the smoothness of model predictions. Under the criterion of pure accuracy, a model's prediction is deemed correct only if it satisfies one of the following conditions: *i*) for a test utterance labeled as `non-neutral`, the model must hit the current label at least once across all segments while allowing only the `neutral` class to be present in the prediction results except the current class; *ii*) for a test utterance labeled as `neutral`, the model's prediction for all segments must be `neutral`. Denote by N_f the number of all utterances, N_c the number of utterances identified as correct, and we have *Pure-Acc* $= (N_c/N_f) \cdot 100\%$.

3.2. Baselines and Implementation Details

We take several strong SER baselines into account for model comparison, including Deep-CNN [14], L-SERNet [9], GLAM [23], DRN-MHSA [4], and AA-CNN [19]. It is noteworthy that GLRF exhibits a significant departure from the above methods. While certain methods like AA-CNN and GLAM also use the attention mechanism or gMLP [26] to capture long-distance relationships, their attention is built on top of multiple CNN layers. In contrast, GLRF runs the attention and

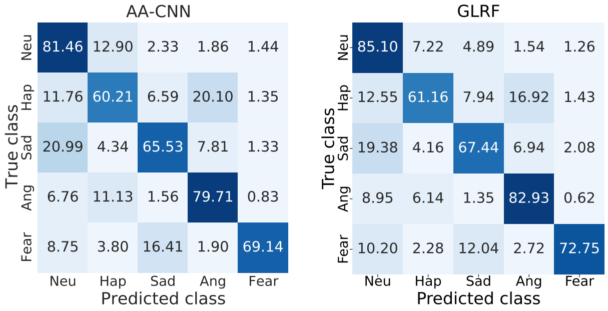

Figure 3: *Confusion matrix on the ST-EMO dataset.*

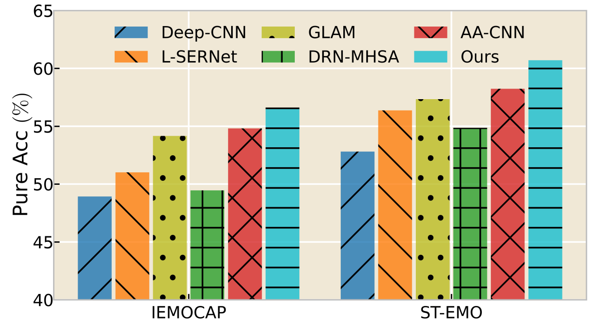

Figure 4: *Comparison of GLRF and other baselines on the IEMOCAP and ST-EMO datasets with respect to pure accuracy.*

CNN in parallel across various layers, thereby capturing global and local contexts separately and allowing us to fuse emotional information at different scales in a more fine-grained manner.

In our experiments, we apply the 5-fold cross-validation to ensure reliability on IEMOCAP. We randomly select 80% of the data for training and the remaining 20% for testing, following [14, 19, 27]. We split the ST-EMO dataset into training and test sets in a ratio of 8:2 with no speaker overlap. All compared methods are run on the same data split while using their suggested hyperparameter configurations. Furthermore, we finetune all models to achieve their optimal performance. During training, each utterance is split into 2s segments with 1.8s overlap, and during testing, each utterance is divided into 2s segments with 1.6s overlap. The MFCCs features of each segment are extracted as input. Finally, the scores of all segments within the same utterance are averaged to yield the prediction result.

3.3. Main Results

Table 2 illustrates that GLRF outperforms the state-of-the-art methods by a substantial margin in all the listed metrics. Notably, GLRF exhibits a 2.43% absolute improvement in WA on IEMOCAP compared to the current leading method, AA-CNN, and a 2.8% absolute improvement in WA on ST-EMO. These results fully demonstrate the superiority and effectiveness of our model. For an in-depth analysis, we show the confusion matrices derived from our model and AA-CNN in Fig. 3, to clarify the model's predictions for specific emotion categories. It is evident that the model's precision has exhibited the most significant enhancement in the three emotional categories of `neutral`, `angry`, and `fear` compared to AA-CNN, with a remarkable surge of 3.64%, 3.22%, and 3.61%, respectively.

Pure accuracy is an SER evaluation protocol that takes the predictions for each segment into account and is particularly sensitive to sharp predictions (*i.e.*, those with significant differences in the valence and arousal domain). Higher pure accuracy indicates smoother and more stable model predictions. As illustrated in Fig. 4, our model outperforms other methods in pure accuracy, implying that GLRF is more suitable for streaming deployment. In streaming SER scenarios, the model is required to provide real-time predictions for each segment. Models with higher pure accuracy tend to avoid the jump of predictions between different classes, leading to a better user experience.

Table 3: *Ablations of the* $\boldsymbol{k}$ *on the IEMOCAP dataset.*

$\boldsymbol{k}$	**WA**	**UA**	**Micro-F1**	**Pure-Acc**
10	68.11	69.39	67.91	51.14
30	69.83	71.24	69.62	53.96
50	72.09	73.11	72.15	55.87
70	**72.81**	**73.39**	**72.92**	**56.62**
90	71.77	72.40	71.84	55.42

Table 4: *Ablation results of the GLRF's core components on the IEMOCAP dataset.*

Ablation Settings	**WA**	**UA**	**Micro-F1**
GLRF	**72.81**	**73.39**	**72.92**
wo/ Fusion Block	71.73	72.80	71.75
wo/ Global-Local	71.27	72.65	71.24
wo/ Head-k-Pooling	70.86	71.78	71.03

3.4. Ablation Study

This section provides ablation studies to shed light on the effect of each component of GLRF, as well as the proposed loss.

Impact of *k* in Loss Function. Table 3 presents the results of GLRF trained on a range of k values. Note that k in the head-k-pooling loss function denotes the percentage of segments that contribute to the loss calculation (a.k.a. *decision makers*). As k approaches 100, the head-k-pooling loss function degenerates into the conventional cross-entropy loss. The table shows that setting $k = 70$ yields the best performance, so we use it by default in our experiments. Increasing k beyond this value leads to a decrease in pure accuracy.

Effectiveness of Key Components. To elucidate the influence of individual components in the GLRF, we conduct a comprehensive component analysis by iteratively replacing each component with one from the full GLRF and evaluating the resultant performance, as depicted in Table 4. Specifically, setting (1) entails the substitution of summation for the fusion manner in GLRA; setting (2) involves the replacement of global and local branches with two convolutions (conv3×3 and conv5×5) in GLRA; setting (3) corresponds to the substitution of head-k-pooling loss with cross-entropy loss. Overall, the model performance exhibits a considerable decline as each component is replaced, verifying the efficacy of the proposed components.

4. Conclusion

In this work, we present GLRF, a novel architecture for speech emotion recognition. It combines the strengths of convolution and self-attention, performing local context modeling and long-distance relationship modeling in parallel while capturing both globality and locality. Furthermore, we tailor a head-k-pooling loss to facilitate training on SER tasks, enabling the model to teach itself by picking the most confident top k% segments. Besides, we introduce ST-EMO, the largest Chinese Mandarin speech emotion dataset for SER research. Experimental results on two speech emotion benchmarks demonstrate the superiority and stability of our model.

5. References

[1] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, 2014, pp. 223–227.

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

[3] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto, and S. Makino, "Perfor-mance improvement of speech emotion recognition by neutral speech detection using autoencoder and intermediate representation," in *Proc. INTERSPEECH*, 2022, pp. 4700–4704.

[4] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. ICASSP*, 2019, pp. 6675–6679.

[5] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[6] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *Proc. ICASSP*, 2019, pp. 7405–7409.

[7] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.

[8] B. T. Atmaja, A. Sasou, and M. Akagi, "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion," *Speech Communication*, vol. 140, pp. 11–28, 2022.

[9] A. Aftab, A. Morsali, S. Ghaemmaghami, and B. Champagne, "LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition," in *Proc. ICASSP*, 2022, pp. 6912–6916.

[10] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, 2016, pp. 5200–5204.

[11] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *Proc. ICASSP*, 2017, pp. 5115–5119.

[12] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[13] K. Chauhan, K. K. Sharma, and T. Varma, "Speech emotion recognition using convolution neural networks," in *International Conference on Artificial Intelligence and Smart Systems*, 2021, pp. 1176–1181.

[14] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[15] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with long-short range attention," *arXiv preprint arXiv:2004.11886*, 2020.

[16] R. Dai, S. Das, K. Kahatapitiya, M. S. Ryoo, and F. Brémond, "MS-TCT: Multi-scale temporal convtransformer for a lightweight model based on separable convolution for speech emotion recognition detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 041–20 051.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, p. 5998–6008.

[18] M. Xu, F. Zhang, and S. U. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *Annual Computing and Communication Workshop and Conference*, 2020, pp. 1058–1064.

[19] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *Proc. ICASSP*, 2021, pp. 6319–6323.

[20] P. E. Bestelmeyer, S. A. Kotz, and P. Belin, "Effects of emotional valence and arousal on the voice perception network," *Social Cognitive and Affective Neuroscience*, vol. 12, pp. 1351–1358, 2017.

[21] 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, pp. 335–359, 2008.

[22] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Proc. ICASSP*, 2017, pp. 2741–2745.

[23] W. Zhu and X. Li, "Speech emotion recognition with global-aware fusion on multi-scale feature representation," in *Proc. ICASSP*, 2022, pp. 6437–6441.

[24] "The evaluation of weibo emotion classification technology, SMP2020-EWECT," *The Ninth China National Conference on Social Media Processing*, 2020. [Online]. Available: https://smp2020ewect.github.io/

[25] Y. Chen, W. Fan, X. Xing, J. Pang, M. Huang, W. Han, Q. Tie, and X. Xu, "CPED: A large-scale chinese personalized and emotional dialogue dataset for conversational AI," *arXiv preprint arXiv:2205.14727*, 2022.

[26] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to MLPs," in *Advances in Neural Information Processing Systems*, 2021, pp. 9204–9215.

[27] Y. Liu, H. Sun, W. Guan, Y. Xia, and Z. Zhao, "Discriminative feature representation based on cascaded attention network with adversarial joint loss for speech emotion recognition," in *Proc. INTERSPEECH*, 2022, pp. 4750–4754.