Interspeech 2024
1-5 September 2024, Kos, Greece

DropFormer: A Dynamic Noise-Dropping Transformer for Speech Emotion Recognition

Jialong Mai[1], *Xiaofen Xing*[1], *Weidong Chen*[1], *Xiangmin Xu*[2]

[1]School of Electronic and Information Engineering, South China University of Technology, China
[2]School of Future Technology, South China University of Technology, China

202320111090@mail.scut.edu.cn, xfxing@scut.edu.cn, eewdchen@mail.scut.edu.cn, xmxu@scut.edu.cn

Abstract

Speech Emotion Recognition (SER) is an important component for human-computer interaction. Recently, various optimized Transformer variants have been applied to SER. However, most of studies use all the information in the sample and tend to overlook local details, making it difficult to perceive emotional information that is present locally in speech. While there are studies exploring how to utilize local information, their approaches are not flexible enough or are overly complex. To address the issues, we propose DropFormer, a new architecture that focuses only on the emotional segments by dynamically dropping non-emotional information. DropFormer consists of two main components: (1) Drop Attention, proficient in capturing local emotion and highlighting emotion-related segments, (2) Token Dropout Module, capable of dropping tokens lacking emotional information. Experimental results show that our DropFormer achieves state-of-the-art performance on the IEMOCAP and MELD benchmarks.

Index Terms: speech emotion recognition, transformer, attention machanism

1. Introduction

Human emotion plays a critical role in communication. Speech Emotion Recognition (SER), an essential tool to inform intelligent systems about the feelings of users [1, 2], is widely used in numerous applications, such as intelligent robots, automated call centers, and distance education [3, 4].

Since human emotions are complex and nuanced, researchers have worked hard to improve system performance in various ways. [5] suggested traditional emotion labels oversimplify the problem and proposed to use word embeddings obtained from a Language Model (LM) as labels for SER. [6] argued that stacking convolutions overlooks global information and improved it by combining attention mechanism. [7] proposed that multi-view Speech Emotion Recognition is complex and suggested learning emotion-related information from two feature views using a concise method. Different from the above directions aiming to enhance SER, Transformer [8], as an efficient architecture, attracts researchers in the field of SER.

Attention-based Transformer [8, 9, 10, 11] has become the dominant backbone in natural language processing (NLP) and computer vision (CV). However, its applications to the SER task remain limited because human emotions are inherently complex and ambiguous. The attention mechanism is the core of the Transformer, which empowers Transformer to capture long-range dependency via the global receptive field. However, the full attention mechanism computes pairwise token affinity across all spatial locations, which overlooks that not all the information in audio is related to emotion. As shown in Figure 1, considering a sample "I'm so happy for you. [LAUGHTER]" in IEMOCAP [12], the tone fluctuates at the end of the utterance because of the laughter. In addition, the segment corresponding to "happy" is spoken with stress. Therefore, these segments in the sample convey the majority of emotional information. However, the segments with flat tone, such as "I'm" and "you", contain little emotional information and are useless to the SER system. Consequently, the full attention mechanism employed in the vanilla Transformer is suboptimal for the SER tasks.

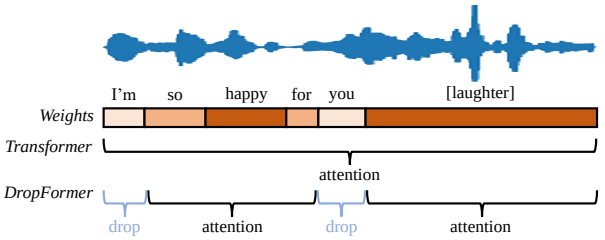

Figure 1: *The utterance "I'm so happy for you. [LAUGHTER]" from IEMOCAP. Speech segments with strong emotions (indicated by dark colours) should be given more attention weight, while non-emotional speech segments (indicated by light colours) should be given less weight.*

To alleviate the above problem, a promising solution is to change the receptive field manually. Li *et al.* [13] applied static attention windows across multiple time scales. However, these fixed windows lack the flexibility to capture emotional segments with various durations. There are also works trying to dynamically constrain the attention scopes [14, 15, 16]. However, [14] relied on an additional decision network to determine the size and the position of the window. [15] processed local and global information asynchronously, which complicated the overall system. [16] retained weights on all queries, making the model sensitive to the non-emotional information.

In this work, we propose a dynamic Transformer, named DropFormer, for SER. DropFormer is mainly composed of Drop Attention and Token Dropout Module. Specifically, DropFormer employs Drop Attention to process local and global information synchronously without introducing additional networks. Token Dropout Module dynamically drops noise tokens, which also reduces the computational costs. The contributions can be summarized as follows:

- We introduce the Drop Attention, which is proficient in capturing local emotion and dynamically highlighting emotion-related segments.
- We introduce the Token Dropout Module, dropping tokens lacking emotional information in speech to reduce the impact

10.21437/Interspeech.2024-651

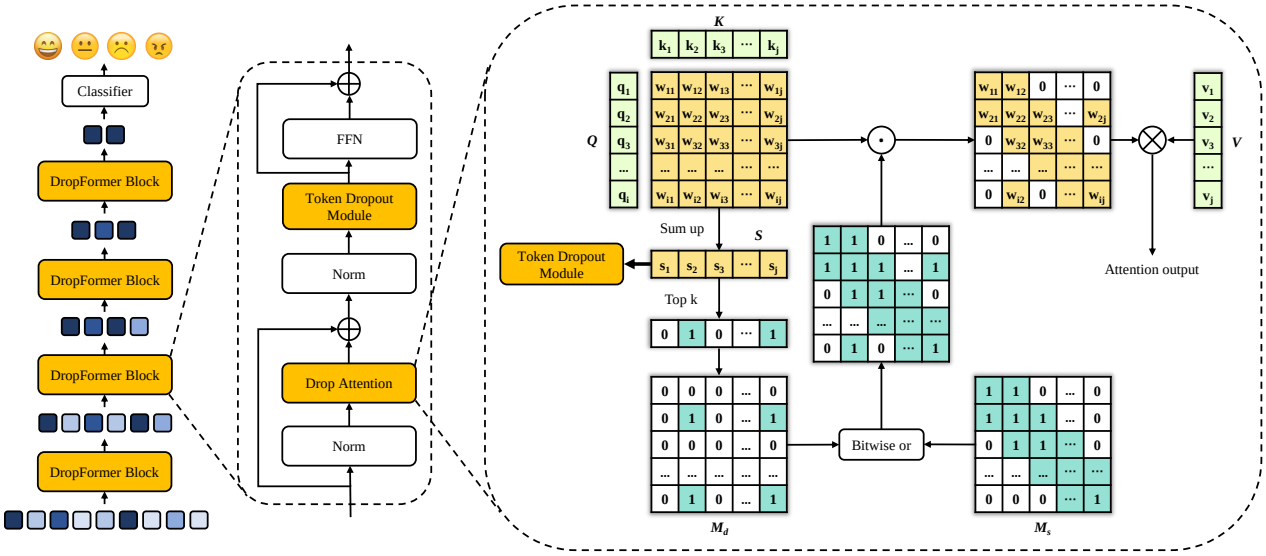

Figure 2: *The left side shows that DropFormer consists of four layers of DropFormer Blocks, with each layer discards noise tokens proportionally, the deeper the color, the more important the token is. On the middle side, the details of the DropFormer Block are presented. We introduce Drop Attention to focus on emotion and Token Dropout Module to discard noise. The right side shows the details of Drop Attention, S represent a discriminator for the importance of each token, $\odot$ and $\otimes$ represent position-wise multiplication and matrix multiplication, M_d and M_s represent dynamic mask and static mask, respectively.*

of noise and improve the model efficiency.

- Experimental results show that our DropFormer achieves state-of-the-art performance on the IEMOCAP and MELD [17] benchmarks.

2. Methodology

The proposed DropFormer, which is composed of multiple stacked DropFormer Blocks, is illustrated in Figure 2. Each DropFormer Block consists of a Drop Attention module and a Token Dropout Module. In which, the Drop Attention module dynamically focuses on the emotional information in the input speech signal, and the Token Dropout Module is used to remove the non-emotional tokens progressively. More details will be introduced in the following subsections.

2.1. Drop Attention

Drop Attention (DA), illustrated on the right side of Figure 2, is at the core of the proposed DropFormer. Different from the previous attention mechanisms [13, 14, 16], DA is able to capture local details while dynamically focusing on the emotional information in the global context.

The inputs of DropFormer are first linearly projected into Q, K, and V, which represent Query, Key, and Value vectors, respectively. Subsequently, the weight matrix W is computed as the normalized dot product between the Q and K:

$$W = softmax(\frac{QK^T}{\sqrt{dQ}}) \tag{1}$$

where d_Q is the dimension of the Query vector. Inspired by [16], DA utilizes the attention weights in W to calculate the importance of each token. We note that the elements in the same column of the weight matrix W are utilized to weight the corresponding value vector. This implies that the elements in the l-th column of the weight matrix W indicate the importance of the l-th token. By summing up the weights column by column, the resulting summation S serves as a discriminator for the importance of each token. The indices of the *top-k* largest value in S are recorded as n_1, n_2,..., n_k, indicating the positions that contain the most emotional information in the speech. The dynamic mask (M_d) is defined as:

$$m_{ab} = \begin{cases} 1 & \text{if a} \in \text{N and b} \in \text{N} \\ 0 & \text{else} \end{cases}, m_{ab} \in M_d \tag{2}$$

where m_{ab} is the element in row a and column b of the M_d, and N represents the set of n_1, n_2,..., n_k.

The proposed M_d resets the attention output of the unimportant frames to zero while maintaining the attention output of the important frames. Therefore, the M_d can determine the emotional positions in the global perspective and enable dynamic modeling of speech emotion.

To better capture the local information in speech, we introduce a static mask (M_s). Specifically, emphasizing the diagonal of the weight matrix W, which represents the local weights of each query vector, helps the model capture local information. We also set parts of M_s to 0 to make it sparser and avoid redundant due to the similarity between adjacent speech frames. The M_s is illustrated in Figure 3. Empirically, the similarity between adjacent speech frames of pre-trained features is high, but after going through the encoder, the similarity will decrease. What's more, the proposed Token Dropout Module (described in the following subsection) is able to concentrate on emotional frames by discarding 'noise' frames, we expanded the diagonal on the last two layers.

Finally, we perform a bitwise *or* operation ($\vee$) on the corresponding elements of the M_d and M_s to obtain the final mask M.

$$M = M_d \vee M_s \tag{3}$$

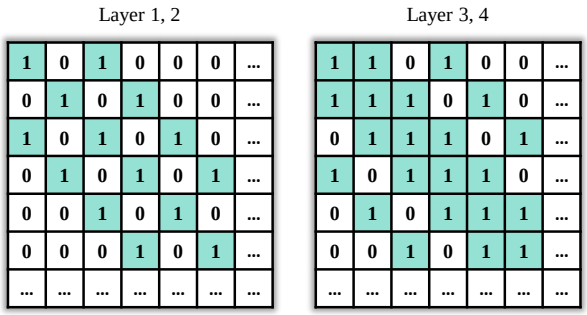

Figure 3: M_s *on different DropFormer layers. Compared to the first two layers, we expand the diagonal on the last two layers to accommodate the variation of the pre-trained features as they pass through the encoder.*

The output of DA is calculated as follow:

$$attn = (W \cdot M) \times V \tag{4}$$

2.2. Token Dropout Module

DA is utilized to help the model focus on emotional information in speech. To further mitigate the impact of noise on the system and enhance model efficiency, we propose the Token Dropout Module (TDM), which filters out non-emotional tokens. The structure of the TDM is depicted in Figure 4.

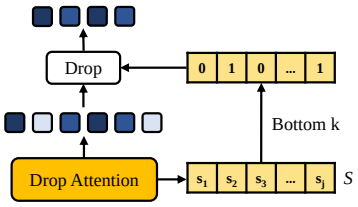

Figure 4: *TDM determines the discarded tokens based on* S *from the DA. The deeper the color, the more important the token is. For simplicity, the residual connection and layer normalization are not plotted.*

TDM reuses the vector S from the DA, which represents the importance of each speech frame. Feature x from DA removes noise through TDM, which is defined as:

$$Indices_k = Bottom\ k(S) \tag{5}$$

$$TDM(x) = Drop(x, Mask(Indices_k)) \tag{6}$$

where Bottom k represents the operation of getting the indices of the first k minimum values of the vector S from DA, Mask denotes setting the index positions to 1 and the remaining positions to 0, Drop means dropping the token marked 1 in the Mask from x to get the output of TDM.

Moreover, the TDM in DropFormer helps reduce the computational cost by reducing the length of the token sequence.

3. Experiments

3.1. Datasets

IEMOCAP dataset is used following the same way as in previous studies [13, 15, 16, 14]. We merge excitement into happiness category and select 5,531 utterances from happy, angry, sad and neutral classes. The experiments are conducted using the leave-one-session-out cross-validation strategy.

MELD comprises 13,708 utterances across 7 emotion classes. It is officially divided into training, validation, and testing sets. We use the validation set for hyperparameters tuning, and present the scores on the testing set.

3.2. Experiment Setup

Pre-trained self-supervised WavLM [18] is adopted to extract the acoustic features. SGD optimizer with a learning rate of $5e^{-4}$ is used for IEMOCAP, and Adam optimizer with a learning rate of $1e^{-3}$ is used for MELD to optimize the model, training for 100 and 120 epochs, respectively. The learning rate is adjusted using the cosine annealing schedule. The batch size is 32. The number of DropFormer blocks is 4.

In DA, the Keep Rate (KR) for emotional tokens in each layer is set to 10%. The larger KR we set, the less attention weights are reset to zero. In TDM, the Drop Rate (DR) for 'noise' tokens in each layer is set at 10%, resulting in DropFormer dropping 35.39% of tokens in total.

In IEMOCAP, we use Weighted Accuracy (WA) and Unweighted Accuracy (UA) as metrics, which helps us analyze model performance taking class distribution into account. In MELD, we use Weighted F1 (WF1) as metric, contributing to the comprehensive evaluation of models.

3.3. Experimental Results and Analysis

3.3.1. Ablation Study

This section presents ablation studies to illuminate the effect of each component of DropFormer. As demonstrated in Table 1, we conduct a comprehensive component analysis by iteratively replacing each component with one from the full DropFormer and evaluating the resultant performance. Specifically, in setting (1), the TDM in DropFormer has been removed; in setting (2), the DropFormer is replaced with a Transformer that only uses the proposed M_s;in setting (3), the DropFormer is replaced with a Transformer that only utilizes the proposed M_d; and in setting (4), DA is removed from DropFormer. Overall, the model performance exhibits a considerable decline as each component is replaced, verifying the efficacy of the proposed components.

We also compare the DropFormer with the vanilla Transformer. The results reported in Table 1 show that in IEMOCAP, DropFormer substantially surpasses the vanilla Transformer with a 2.4% improvement in WA and a 2.53% improvement in UA.

3.3.2. Comparison with Some Known Systems

We compare the DropFormer with the lastest known systems on the IEMOCAP and MELD datasets. From Table 2, it can be seen that our method gives the best WA and UA on IEMOCAP, outperforming not only the latest Transformer variants applied to the SER[19, 13, 14, 15], but also over other recent research[5, 7, 6, 20] dedicated to improving the SER. Furthermore, our method achieves the highest WF1 on MELD.

Table 1: *Ablation results of the DropFormer's core components on the IEMOCAP.*

Model	WA	UA
DropFormer (Ours)	**75.29**	**76.60**
DropFormer (Ours) w/o TDM	74.95	76.20
DropFormer (Ours) w/o TDM & M_d	75.13	75.85
DropFormer (Ours) w/o TDM & M_s	74.72	75.36
DropFormer (Ours) w/o DA	74.06	74.52
Vanilla Transformer	72.89	74.07

Table 2: *Comparison with known state-of-the-art systems on IEMOCAP and MELD.*

IEMOCAP			
Method	Year	WA	UA
Word embeddings [5]	2023	68.47	69.68
SpeechFormer++ [19]	2023	70.50	71.50
MSTR [13]	2023	70.60	71.60
DST [14]	2023	71.80	73.60
DCW [7]	2023	72.08	72.17
DWFormer [15]	2023	72.30	73.90
GLRF [6]	2023	72.81	73.39
SMW_CAT [20]	2023	73.80	74.25
DropFormer (Ours)	2024	**75.29**	**76.60**
MELD			
Method	Year	WF1	
MSTR [13]	2023	46.15	
SpeechFormer++ [19]	2023	47.00	
DWFormer [15]	2023	48.50	
DST [14]	2023	48.80	
DropFormer (Ours)	2024	**49.25**	

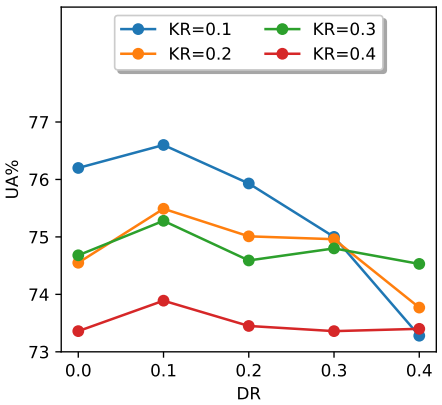

Figure 5: *The relationship between KR and DR. We use UA as a metric to account for category imbalance.*

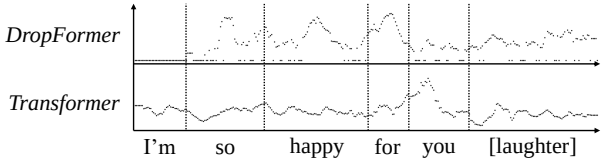

Figure 6: *The attention weights of the utterance "I'm so happy for you. [LAUGHTER]" from IEMOCAP. Compared to the vanilla Transformer, the proposed DropFormer focus on emotional segments and discard noise.*

3.3.3. Model Hyperparameter Analysis

The ablation study validates the effectiveness of DA and TDM. Therefore, it is necessary to analyze the relationship between KR (the hyperparameter determining DA) and DR (the hyperparameter determining TDM). As illustrated in Figure 5, on IEMOCAP, we examine the effect of different KR and DR values on DropFormer. We show part of the results in the figure (both KR and DR are less than 0.5) because they reflect the regularity while giving a concise picture. As shown by the highest point of the blue line, DropFormer achieves the best performance when KR = 0.1 and DR = 0.1, meaning a total of 34.39% of tokens are discarded. This shows that focusing on emotional information and discarding noise tokens is an effective way to improve emotion recognition. In addition, as KR increases, system performance decreases. This indicates that the DA captures the noise beyond the emotional information, which is detrimental to performance. When DR reaches 0.1, the performance of DropFormer decreases as DR increases. However, as KR increases, the trend of performance decrease slows down. This implies that the TDM, which discards information exceeding the proportion of noise, is harmful to performance. The more information captured dynamically by DA, the greater the tolerance for discarding information.

3.3.4. Visualization Analysis

To visually measure DropFormer's capabilities in dynamically highlighting emotion-related segments (effectiveness of DA) and dropping tokens lacking emotional information (effectiveness of TDM). We intuitively compare the attention weights in each attention mechanism by visualization, as depicted in Figure 6. The weight of full attention is distributed on all tokens of the speech, and worse, mainly on the fragment 'you', which is inevitably affected by noise and hard to grasp the emotional segments. However, DropFormer concentrates on emotional segments (centered on 'happy' and [laughter]) and sets part of the weight of the noise segment (centered on 'I'm' and 'you') to 0, successfully learning emotions dynamically and discarding noise. The above results intuitively validate the efficacy of DA to focus on emotional information and the TDM to discard noise.

4. Conclusions

In this paper, we propose DropFormer, a new architecture designed to focus only on emotion-related segments. Its core is the Drop Attention mechanism, proficient at capturing local emotion and dynamically highlighting emotion-related segments. It also contains the Token Dropout Module, capable of dropping tokens lacking emotional information. Experimental results on the IEMOCAP and MELD corpora demonstrate the effectiveness of the proposed DropFormer. Ablation studies confirm the effectiveness of both the Drop Attention mechanism and the Token Dropout Module. In the future, we plan to investigate how to capture more emotional details in speech to further improve the system performance.

5. Acknowledgements

The work is supported in part by Natural Science Foundation of Guangdong Province 2022A1515011588; in part by Nansha key project 2022ZD011; in part by the Science and Technology Project of Guangzhou 202103010002; in part by the Guangdong Provincial Key Laboratory of Human Digital Twin 2022B1212010004. Xiaofen Xing is the corresponding author.

6. References

[1] J. J. Gross, H. Uusberg, and A. Uusberg, "Mental illness and well-being: an affect regulation perspective," *World Psychiatry*, vol. 18, no. 2, pp. 130–139, 2019.

[2] J. J. Gross and R. F. Muñoz, "Emotion regulation and mental health." *Clinical psychology: Science and practice*, vol. 2, no. 2, p. 151, 1995.

[3] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6675–6679.

[4] 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.

[5] E. Stanley, E. DeMattos, A. Klementiev, P. Ozimek, G. Clarke, M. Berger, and D. Palaz, "Emotion label encoding using word embeddings for speech emotion recognition."

[6] C. Ding, J. Li, D. Zong, B. Li, T. Zhang, and Q. Zhou, "b. stable speech emotion recognition with head-k-pooling loss." INTERSPEECH, 2023.

[7] K. Liu, D. Wang, D. Wu, and J. Feng, "Speech emotion recognition via two-stream pooling attention with discriminative channel weighting," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] 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.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[11] 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.

[12] 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.

[13] Z. Li, X. Xing, Y. Fang, W. Zhang, and H. Fan, "Multi-scale temporal transformer for speech emotion recognition," in *Proc. Interspeech*, 2023.

[14] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "Dst: Deformable speech transformer for emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] S. Chen, X. Xing, W. Zhang, W. Chen, and X. Xu, "Dwformer: Dynamic window transformer for speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6897–6901.

[17] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[18] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[19] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "Speechformer++: A hierarchical efficient framework for paralinguistic speech processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 775–788, 2023.

[20] Y. He, N. Minematsu, and D. Saito, "Multiple acoustic features speech emotion recognition using cross-attention transformer," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.