ECAI 2024
U. Endriss et al. (Eds.)

© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA241034

CA-SER: Cross-Attention Feature Fusion for Speech Emotion Recognition

Bashar M. Deeb[a,*], **Andrey Savchenko**[b,c,**] **and Ilya Makarov**[d,e]

[a]MIPT, Moscow, Russia
[b]Laboratory of Algorithms and Technologies for Network Analysis, HSE University, Nizhny Novgorod, Russia
[c]Sber AI Lab, Moscow, Russia
[d]Artificial Intelligence Research Institute (AIRI), Moscow, Russia
[e]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
ORCID (Andrey Savchenko): https://orcid.org/0000-0001-6196-0564, ORCID (Ilya Makarov): https://orcid.org/0000-0002-3308-8825

Abstract. In this paper, we introduce a novel tool for speech emotion recognition, CA-SER, that borrows self-supervised learning to extract semantic speech representations from a pre-trained wav2vec 2.0 model and combine them with spectral audio features to improve speech emotion recognition. Our approach involves a self-attention encoder on MFCC features to capture meaningful patterns in audio sequences. These MFCC features are combined with high-level representations using a multi-head cross-attention mechanism. Evaluation of speech emotion recognition on the IEMOCAP dataset shows that our system achieves a weighted accuracy of 74.6%, outperforming most existing techniques.

1 Introduction

Emotions are complex psychological states encompassing various subjective experiences, physiological changes, and behavioral responses [21]. They are crucial in human interactions, influencing our thoughts, actions, and well-being. Understanding emotions is fundamental to human communication and has garnered significant interest in various scientific disciplines [33, 37]. Emotions can be recognized from many modalities, e.g., audio, body, psychological signals, and faces [10, 12, 27, 29, 30]. One of the most practically important tasks is Speech Emotion Recognition (SER) [7, 32, 35], in which it is required to predict emotions conveyed in human speech (sadness, anger, fear, happiness..) [2]. Its importance stems from numerous applications, including Human-Machine interaction systems, virtual assistants, mental health surveillance, etc. [18, 19, 28, 31].

Nowadays, many successful techniques for speech processing are based on Self-Supervised Learning (SSL) [25]. Unfortunately, one major challenge of using SSL models is that they are trained on clean audio data, spoken passively without intonations [9]. In contrast, emotional datasets contain audio signals with various intonation, pitch, and intensity.

To mitigate this problem, in this paper, we introduce a novel pipeline that includes a feature fusion layout of spectral features and SSL representations using a cross-attention mechanism. We enhance the spectral feature representations by allowing them to attend to the representation of SSL models. The source code is publicly available to reproduce our experiments[1]. The demonstration video for our framework is available at [2].

2 Related Works

In the SER task, predicting the emotion label for each speech signal from the test set $V = \{v_1, v_2, ..., v_M\}$ of M new utterances is required. We assume the speech training dataset $U = \{u_1, u_2, ..., u_N\}$ is available. It contains N labeled utterances. Each utterance $u_i \in \mathbb{R}^{t_i}$ is associated with an emotion label e_i.

Conventional approach [15, 20, 40] extracts emotional labels from Spectral features such as spectrograms and Mel-Frequency Cepstral Coefficients (MFCC) to obtain good performance. These features are considered domain-agnostic because they contain information about audio itself, regardless of the underlying task. In [15], they used a Convolutional Neural Network (CNN) architecture to process audio spectrograms, followed by a pooling method to identify the emotion of each utterance. In [40], they devised an attention-based bidirectional long short-term memory for SER. The model's input is speech spectrograms, and the model shows good accuracy. An attention-based bidirectional long short-term memory (BLSTM) neural network was combined in [40] with a connectionist temporal classification (CTC) objective function. Finally, transformers have been widely used nowadays. For example, a transformer encoder with an added focus score was implemented in [13].

It is well-known that remarkable performance in Automatic Speech Recognition (ASR) is achieved by SSL [3, 11, 16, 22] techniques such as wav2vec. They are pre-trained on extensive audio data to capture contextualized representations from raw audio inputs. A transfer learning method for SER was proposed in [25], where features extracted from pre-trained wav2vec 2.0 models are modeled using simple neural networks. SSL was used for SER [38] to represent speech utterance for classification, but it has not provided high accuracy compared to existing state-of-the-art models. In our paper, we try to fill the gap between SER and SSL.

* Corresponding Author. Email:bashar.deeb7@yandex.ru
** Corresponding Author. Email: avsavchenko@hse.ru.

[1] https://github.com/BasharBetta7/SER
[2] https://youtu.be/vOREtCpDBwE

3 Proposed Model

Fig. 1 illustrates the proposed Cross-Attention Speech Emotion Recognition (CA-SER) architecture. At first, the raw audio of each utterance is segmented into sequences of fixed length. These sequences are fed into the wav2vec 2.0 model, followed by a preprocessing module, which extracts contextualized representation of each sequence [4]. Meanwhile, MFCC features are extracted and fed into the feature encoder module to enhance the semantic feature representations. Then, we obtain fused representations using cross-attention fusion module [34]. The final classifier performs global average pooling across sequences, followed by a linear layer to generate a probability distribution over emotional labels.

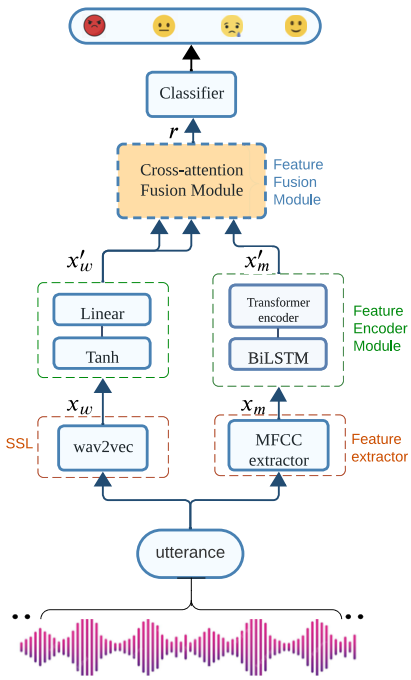

Figure 1. Proposed CA-SER architecture.

SSL representation. Wav2vec 2.0 [4] is a framework for self-supervised learning of audio representations. The model expects raw audio as an input, which is transformed into audio embedding using the CNN feature extraction layer. Audio embeddings are then passed through a contextualized encoder consisting of several self-attention transformer encoders. The base model uses 12 transformer encoders with 8 attention heads each. Each encoder outputs a representation of the original audio. Previous studies show that the first transformer encoders capture low-level information about the audio [23], while the last encoder captures high-level semantic relations in the audio. Since we are interested in capturing the emotional context of the utterance u_i, we suspect that we need to use representations from the middle encoders. We followed study [23] in capturing information from the ninth encoder as our speech representations $x_w \in \mathbb{R}^{t_w \times d_w}$, which tends to give a good balance between low-level and high-level relations within the utterance. Here, $d_w = 768$ is the feature dimension of each output sequence obtained from the pre-trained wav2vec 2.0 base model. We apply $tanh$ activation to re-scale the values of representations, followed by $Linear$ layer to transform the utterance representation into $x'_w \in \mathbb{R}^{t_w \times d}$.

Feature Extractor. In addition to wav2vec representations, a well-known feature representation for audio signals, MFCC, is used. It involves converting the audio signal into the frequency domain using the Fourier Transform, applying a filter-bank based on the Mel scale, taking the logarithm of the filter-bank energies, and applying the Discrete Cosine Transform (DCT) to obtain the final coefficients [1]. We extract 40 coefficients for each 10-millisecond sequence of the utterance, which results in feature representation $x_m \in \mathbb{R}^{t_m \times d_m}$ where $d_m = 40$.

Feature Encoder Module (FEM). The objective of FEM is to enhance the descriptiveness of utterance's feature representations by allowing them to attend to each other. The module consists of two sub-modules: Bidirectional Long Short-Term Memory ($BiLSTM$) with a hidden size 256, followed by the self-attention encoder to capture short and long semantic relations among different sequences. The output of the feature encoder is represented as $x'_m \in \mathbb{R}^{t_m \times d}$.

Feature Fusion Module (FFM). We propose implementing a multi-head cross-attention mechanism from a transformer decoder [34, 36] to fuse MFCC feature representations x'_m with wav2vec representations x'_w. Cross-attention is a dynamic weighted average between two types of sequences:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \quad (1)$$

It consists of queries $Q \in \mathbb{R}^{t_q \times d_k}$ taken from the first type, while keys $K \in \mathbb{R}^{t_k \times d_k}$ and values $V \in \mathbb{R}^{t_v \times d_v}$ are extracted from the other type. We implement multi-head cross-attention [36] with n total heads, allowing the model to attend to information from different representation subspaces, thus enhancing the contextual representation of the output. We then apply a feed-forward network with two fully connected layers, separated by a Gaussian Error Linear Unit ($GeLU$) activation. As a result, we obtain a fused representation of the utterance $r \in \mathbb{R}^{t_m \times d}$.

Classifier. Fused representations r are converted into a probability distribution over emotion labels. Our model has 108 M parameters. We train it to minimize categorical cross-entropy loss between predictions and targets over all dataset samples.

4 Demo System

We develop a special demo system to implement the CA-SER model. Fig. 2 shows the pipeline of our tool. It consists of the pre-processing unit, followed by the model. Our software can classify recorded speech utterances into one of four emotional labels: angry, happy, sad, and neutral. The audio can be specified by its file path or recorded directly using an internal microphone. The system can also accept multiple audio files at once. The CA-SER model performs preprocessing of an input signal, including resampling scaling and extracting MFCC features. After that, an emotional label for audio is obtained by calling the method (*predict_emotion*). In addition, the utterance representation can be extracted and used for Downstream audio tasks. We have trained the model on the IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset [5]. Until now, we have released our best pre-trained checkpoint of the model, which is called (*caser*). In addition to the tool, our released repository includes instructions on training, evaluating, and reproducing our results. We supply a Jupyter notebook (*package_test*) for testing SER.

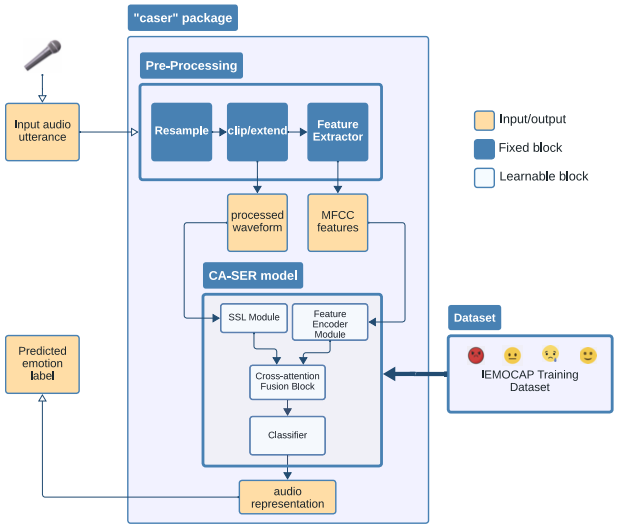

Figure 2. CA-SER demo system pipeline

5 Experimental Results

This section describes the setup and datasets we used to evaluate the model's performance.

Dataset. We train and evaluate our model on one of the most widely used emotion recognition datasets, IEMOCAP. It contains about 12 hours of audiovisual data, recorded from ten actors (5 females and 5 males) with audio, video, transcriptions, and motion-capture information [5]. The final dataset contains 5531 utterances, classified into one of five labels (happy, angry, sad, neutral, and excited). For consistent comparison with previous works, we merge "happy" and "excited" into the category of "happy", and we consider the 5531 acoustic utterances from 4 emotion labels: (1636 happy, 1084 sad, 1103 angry, and 1708 neutral). To evaluate the performance of the model, we implemented cross-validation in two configurations that are used in existing papers that work with the IEMOCAP dataset, namely, 5-fold leave-one-session-out cross-validation (CV-5) and 10-fold leave-one-actor-out cross-validation (CV-10). In each configuration, we use the average value of the metric across all folds as our final result.

Metrics. We evaluate the model using two metrics: Unweighted accuracy (UA) takes the average of the class-wise accuracy without weighting by class size. Weighted accuracy (WA) accounts for class imbalances by assigning weights to each class based on their relative sizes. In addition, we compare our best-performing model with the IEMOCAP benchmark on Weighted F1 score.

Training details. All audio files are sampled with a sampling rate of 16kHz. Audio files with over 8 seconds are clipped to fit into available memory. Also, short audio files (less than 3 seconds) are extended by concatenating the audio until it reaches 3 seconds. 40 MFCC features are calculated using a hop length of 10 milliseconds, which means that each 10-millisecond window is treated as one sequence of the utterance. We train the model with a batch size of 2 and update the gradients after every four batches. This approach compensated for the small batch size because of the GPU memory bottleneck. Each fold was trained for 20 epochs with early stopping on the weighted accuracy. We used the AdamW optimizer with learning rate $4e{-}5$.

Performance evaluation. We compared our model with several SER techniques including previous attempts [25, 38] that used SSL for this task. The experimental results are presented in Tables 1,2. Our technique reached a weighted accuracy of 72.34% in 5-fold cross-validation, which is better than most attention-based models. The model from [6] achieved the highest accuracy by using a much larger HuBERT model and adding the Speaker Normalization task. The HuBERT Large model has 317 million parameters [8], which is three times larger than our model. Our unweighted accuracy is the highest among the recorded baselines. In Table 2, the proposed model achieved the highest evaluation metrics among audio-only models, with 2.59% improvement on WA and 0.67% on UA.

Table 1. Weighted Accuracy [WA] and Unweighted Accuracy [UA] metrics with 5-fold-cross-validation.

Model	WA	UA	Modality
TDNN-LSTM-Attention [25]	66.3	60.3	A
CTC+Attention [40]	67.0	69.0	A
wav2vec 2.0-PT [25]	67.2	-	A
HuBERT Base [38]	68.9	-	A
CNN TF Att.pooling [15]	71.75	68.06	A
CNN-DARTS [26]	72.55	69.36	A
MPT-HCL [41]	72.83	-	A+T+V
SDT [17]	73.95	-	A+T+V
HuBERT Large + SN [6]	**74.2**	-	A
CA-SER [OURS]	72.34	**71.53**	A

Table 2. Weighted Accuracy [WA] and Unweighted Accuracy [UA] metrics of audio-only models, 10-fold-cross-validation.

Model	WA	UA
audio-BRE [39]	64.60	65.20
Audio-CNN-xvector [24]	66.60	68.40
MHSA-FACA [13]	72.01	72.83
CA-SER [OURS]	**74.60**	**73.50**

6 Conclusion and Future Work

In this study, we introduced a novel SSL-based SER model (Fig. 1) and its demo system (Fig. 2). Our approach involved the implementation of a feature encoder, followed by a multi-head cross-attention fusion module. The proposed model demonstrated competitiveness and achieved high recognition accuracy when compared to other audio-only models (Tables 1,2).

Our system can be applied in various domains. For instance, within Human-Machine Interaction Systems, it can serve as a sub-module, enabling virtual assistants to generate responses based on the emotional state of customers. This capability enhances the overall user experience by allowing the system to respond appropriately and empathetically, leading to more effective and engaging interactions. Our approach can assess the mental health of patients and aid them in improving their emotional well-being. Furthermore, researchers can leverage our model's speech representation as an intermediate step for other downstream tasks. By providing a reliable and robust speech representation, our model opens up avenues for further research and advancements in related fields.

In the future, we plan to add a real-time SER to the software, which will track the dynamical change of emotional state in ongoing conversation via microphone and export them into output files for evaluation. Moreover, it is necessary to study how generalizable our model is to other languages [14], considering that the IEMOCAP dataset was recorded in English.

References

[1] Z. K. Abdul and A. K. Al-Talabani. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158, 2022.

[2] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati. Speech emotion recognition through hybrid features and convolutional neural network. *Applied Sciences*, 13, 4 2023.

[3] B. T. Atmaja and A. Sasou. Evaluating self-supervised speech representations for speech emotion recognition. *IEEE Access*, 10:124396–124407, 2022.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. 6 2020. URL http://arxiv.org/abs/2006.11477.

[5] 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, 2007.

[6] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory. Speaker normalization for self-supervised speech emotion recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7342–7346. IEEE, 2022.

[7] A. Hashem, M. Arif, and M. Alghamdi. Speech emotion recognition approaches: A systematic review. *Speech Communication*, page 102974, 2023.

[8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. 6 2021. URL http://arxiv.org/abs/2106.07447.

[9] J. Hyeon, Y. H. Oh, Y. J. Lee, and H. J. Choi. Improving speech emotion recognition by fusing self-supervised learning and spectral features via mixture of experts. *Data and Knowledge Engineering*, 150, 3 2024.

[10] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. *IEEE Transactions on Instrumentation and Measurement*, 2023.

[11] A. Karpov and I. Makarov. Exploring efficiency of vision transformers for self-supervised monocular depth estimation. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 711–719. IEEE, 2022.

[12] Y. I. Khokhlova and A. Savchenko. About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems. *Optical Memory and Neural Networks*, 23(1):34–42, 2014.

[13] J. Kim, Y. An, and J. Kim. Improving speech emotion recognition through focus and calibration attention mechanisms. In *Proceedings of INTERSPEECH*, volume 2022-September, pages 136–140. International Speech Communication Association, 2022.

[14] V. Kondratenko, N. Karpov, A. Sokolov, N. Savushkin, O. Kutuzov, and F. Minkin. Hybrid dataset for speech emotion recognition in russian language. In *Proc. INTERSPEECH 2023*, pages 4548–4552, 2023.

[15] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai. An attention pooling based representation learning method for speech emotion recognition. In *Proceedings of INTERSPEECH*, volume 2018-September, pages 3087–3091. International Speech Communication Association, 2018.

[16] A. Luginov and I. Makarov. Swiftdepth: An efficient hybrid cnn-transformer model for self-supervised monocular depth estimation on mobile devices. In *Proceedings of International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 642–647. IEEE, 2023.

[17] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 2023.

[18] I. Makarov and D. Zuenko. Style-transfer autoencoder for efficient deep voice conversion. In *Proceedings of the 21st International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 000121–000126. IEEE, 2021.

[19] I. Makarov, N. Veldyaykin, M. Chertkov, and A. Pokoev. American and russian sign language dactyl recognition. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 204–210, 2019.

[20] Q. Mao, M. Dong, Z. Huang, and Y. Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16:2203–2213, 12 2014.

[21] H. L. Meiselman. *Emotion measurement ((Second Edition)*. Elsevier, 2021.

[22] E. Morais and H. Aronowitz. Speech emotion recognition using self-supervised features edmilson morais , ron hoory , weizhong zhu , itai gat , matheus damasceno and hagai aronowitz. pages 2–6, 2022.

[23] N. Naderi and B. Nasersharif. Cross corpus speech emotion recognition using transfer learning and attention-based fusion of wav2vec2 and prosody features. *Knowledge-Based Systems*, 277, 10 2023.

[24] Z. Peng, Y. Lu, S. Pan, and Y. Liu. Efficient speech emotion recognition using multi-scale cnn and attention. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021-June:3020–3024, 2021.

[25] L. Pepino, P. Riera, and L. Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. 4 2021. URL http://arxiv.org/abs/2104.03502.

[26] T. Rajapakshe, R. Rana, S. Khalifa, B. Sisman, and B. Schuller. Enhancing speech emotion recognition through differentiable architecture search. *arXiv preprint arXiv:2305.14402*, 2023.

[27] A. Savchenko and L. Savchenko. Audio-visual continuous recognition of emotional state in a multi-user system based on personalized representation of facial expressions and voice. *Pattern Recognition and Image Analysis*, 32(3):665–671, 2022.

[28] A. V. Savchenko. Phonetic words decoding software in the problem of russian speech recognition. *Automation and Remote Control*, 74:1225–1232, 2013.

[29] A. V. Savchenko. MT-EmotiEffNet for multi-task human affective behavior analysis and learning from synthetic data. In *Proceedings of European Conference on Computer Vision (ECCV) Workshops*, pages 45–59. Springer, 2022.

[30] A. V. Savchenko. EmotiEffNets for facial processing in video-based valence-arousal prediction, expression classification and action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5716–5724, 2023.

[31] A. V. Savchenko and L. V. Savchenko. Towards the creation of reliable voice control system based on a fuzzy approach. *Pattern Recognition Letters*, 65:145–151, 2015.

[32] L. Savchenko and A. V Savchenko. Speaker-aware training of speech emotion classifier with speaker recognition. In *Proceedings of the 23rd International Conference on Speech and Computer (SPECOM)*, pages 614–625. Springer, 2021.

[33] K. R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44:695–729, 12 2005.

[34] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara. Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition. *Proceedings of INTERSPEECH*, pages 3755–3759, 2020.

[35] 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 *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204. IEEE, 2016.

[36] 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 (NeurIPS)*, 30, 2017.

[37] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022.

[38] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

[39] S. Yoon, S. Byun, S. Dey, and K. Jung. Speech emotion recognition using multi-hop attention mechanism. In *Proceedings of International conference on acoustics, speech and signal processing (ICASSP)*, pages 2822–2826. IEEE, 2019.

[40] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller. Attention-enhanced connectionist temporal classification for discrete speech emotion recognition. In *Proceedings of INTERSPEECH*, volume 2019-September, pages 206–210. International Speech Communication Association, 2019.

[41] S. Zou, X. Huang, and X. Shen. Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5994–6003, 2023.