Hybrid Contrastive Learning Decoupling Speech Emotion Recognition

Li, Chenyu; Gu, Yu; Zhang, He; Liu, Linsong; Lin, Haixiang; Wang, Shuang

DOI
10.1109/ICASSP49660.2025.10889881

Publication date
2025

Document Version
Final published version

Published in
ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings

Hybrid Contrastive Learning Decoupling Speech Emotion Recognition

Chenyu Li[1], Yu Gu[1,*], He Zhang[2], Linsong Liu[1], Haixiang Lin[3], Shuang Wang[1]
[1]School of Artificial Intelligence, Xidian University, Xi'an, China
[2]School of Journalism and Communication, Northwest University, Xi'an, China
[3]Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands

ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889881

***Abstract*—Speech signals contain rich information, such as textual content, emotion, and speaker identity. To extract these features more efficiently, researchers are investigating joint training across multiple tasks, like Speech Emotion Recognition (SER) and Speaker Verification (SV), aiming to improve performance by decoupling task-specific knowledge. Traditional multitask decoupling methods in SER typically use orthogonalization to increase the distance between parameter vectors in the feature space. In this paper, we introduce a novel Hybrid instance-level Contrastive Decoupling Loss. This method leverages supervised labels to effectively decouple SER and SV. Unlike previous approaches, it is not restricted to dual-stream models with identical architectures and can be easily integrated with leading models for each sub-task. Experimental results show that our proposed Hybrid Contrastive Learning Decoupling (HCLD) method significantly outperforms traditional orthogonal decoupling approaches.**

***Index Terms*—speech emotion recognition, feature decoupling, speaker verification.**

I. Introduction

Speech Emotion Recognition (SER) is crucial for interpreting emotional cues in human speech, and it's applied in areas like human-computer interaction, speech synthesis, and intent detection. It's vital for advancing intelligent robots and AI. Daily speech not only conveys meaning through language but also communicates speaker emotions, which can be inferred from unique speech characteristics.

The speech in daily conversations contains rich information, including not only the meaning represented by the language itself, but also the speaker's emotions, and can reflect the speaker's information through unique characteristics. Recent research has led to the design of various tasks, including SER, Automatic Speech Recognition (ASR) [1], [2], Speaker Verification (SV) [3], Gender Identification [4], and Keyword Spotting [3], to exploit the diverse information within speech, thereby enhancing the performance of each task. However, task interference poses a challenge to multitasking advancements. To overcome this, some studies have explored decoupling methods [3], [5] to isolate relevant features in high-dimensional spaces, thereby enabling models to concentrate on specific objectives. While these methods, often involving tailored architectures and parameter-level orthogonalization, have promoted multitask learning in SER, they are limited by their reliance on specific architectures and complex orthogonalization techniques, hindering generalizability across different tasks and models.

To enhance the adaptability and broad applicability of feature decoupling techniques, this paper presents an Instance-level Hybrid Contrastive Learning Decoupling framework, referred to as HCLD. Specifically, we employ the HuBERT [6] pre-training model to extract audio features and then refine them further. In the decoupling representation learning phase, we create positive and negative sample pairs by combining different speakers and emotions. The Instance-level Contrastive Learning Loss (ICLD) is then used to decouple the Speech Emotion Recognition (SER) and Speaker Verification (SV) tasks. To ensure emotional distinctiveness in the decoupled feature space, we introduce two additional losses: the Label-based Contrastive Learning Loss (LCL) and the Supervised Contrastive Learning Loss (SupCon). These losses leverage emotion labels obtained from the RoBERTa model to serve as a benchmark for emotion classification, thereby boosting the SER performance. Our proposed HCLD method demonstrates a significant improvement over single-task SER on the IEMOCAP dataset.

Specific Contributions of This Paper:

- We propose a novel instance-level Contrastive Learning Decoupling Loss (ICLD), which differs from existing methods based on model parameters or feature orthogonalization. It can be flexibly applied to different-architectures SER and SV models.
- To prevent the ICLD loss from muddling the feature space, we've added hybrid contrastive learning losses, including: Label-based Contrastive Learning Loss (LCL) and the Supervised Contrastive Learning loss (SupCon Loss). We introduce emotional semantic labels as a benchmark, futher to reduce the distance between features belonging to the same category, while expanding the distance between features from different categories.
- We tested our proposed HCLD on the challenging IEMOCAP dataset [7]. Compared to single-task SER, our method achieved a maximum performance boost of 5.77%.

This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0110400 and Grant 2021ZD0110404; in part by the National Natural Science Foundation of China under Grant 62271377 and Grant 62201407; in part by the Key Research and Development Program of Shannxi Program under Grant 2021ZDLGY01-06, Grant 2022ZDLGY01-12, Grant 2023YBGY244, Grant 2023QCYLL28, Grant 2024GX-ZDCYL-02-08, and Grant 2024GX-ZDCYL-02-17.

* Corresponding author. e-mail: guyu@xidian.edu.cn

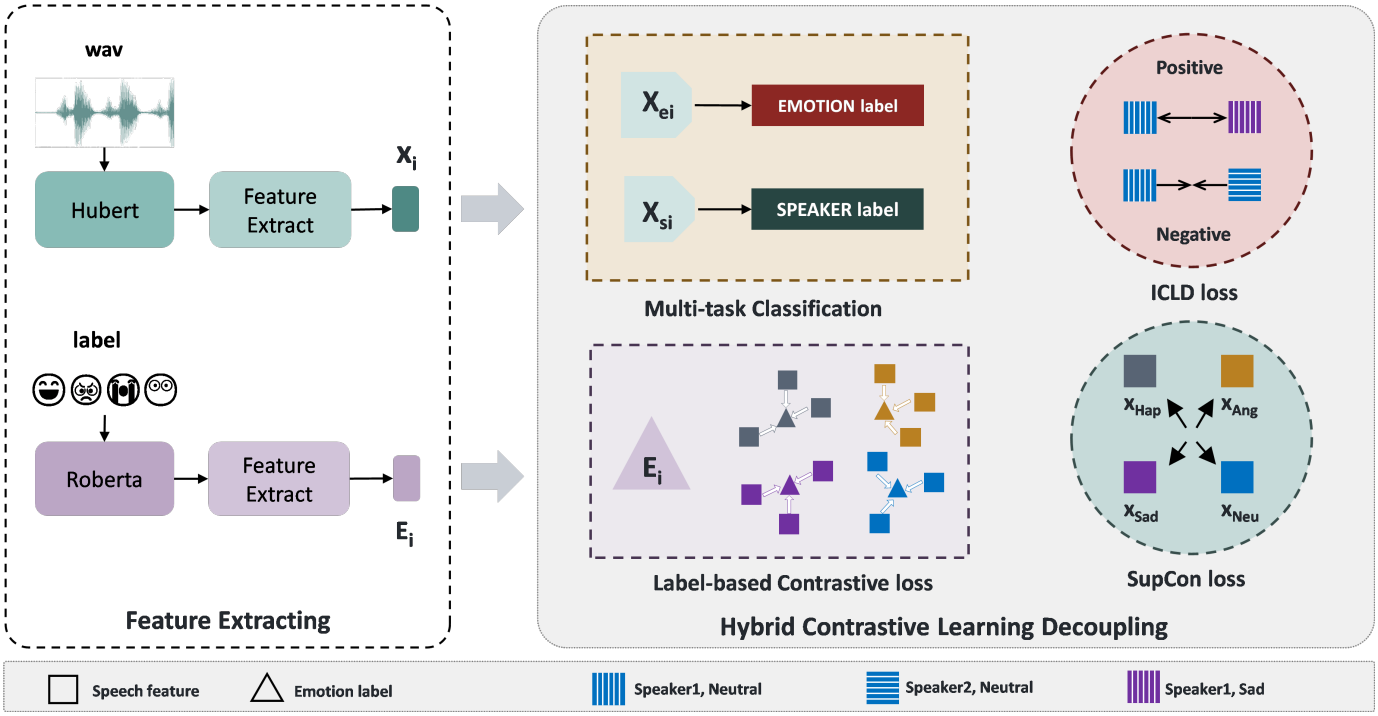

Fig. 1. Illustration of our proposed Method. On the left is the multi-task feature extraction part, which uses the Hubert pre-training model for feature extraction and employs models based on CNN/LSTM/Transformer to extract speech emotion features X_e and speaker feature X_s. The proposed HCLD loss further decouples the extracted X_e and X_s in the feature space.

II. Related Work

A. Speech Emotion Recognition

The task of SER involves identifying a speaker's emotional state from speech signals. As deep learning advances, the models and features employed for SER are continually evolving. On one hand, Transformer-based models [8]–[10] have surpassed CNN and RNN-based models [11]–[15] due to their ability to capture both global and local information. On the other hand, pre-trained models like Wav2Vec [16] and HuBERT [17]are increasingly replacing traditional features like MFCC [18] and Spectrograms [19], thus enhancing SER accuracy and performance.

B. Multi Task Learning and Decoupling in SER

In recent years, researchers have explored integrating SER with other tasks, such as Cai et al. [2] jointly training SER and ASR, and Pan et al. [4] combining SER with gender recognition. These approaches have bolstered model generalization by leveraging correlations between feature representations and tasks. However, multitasking progress is impeded by task interference. To tackle this, some have employed feature decoupling strategies, like Wang et al. [3] using a dual-stream model for KWS and SV tasks, or Liu et al. [5] suppressing irrelevant information with feature map flows.

Current SER multi-task decoupling strategies often optimize at the parameter level, limiting their application across various tasks. This paper proposes an instance-level multi-task decoupling framework that can be easily integrated into different tasks and networks, enhancing performance.

C. Contrastive Learning in SER

Contrastive learning has gained traction in SER tasks. Li et al. [20] were among the first to use SupCon loss for emotion recognition in conversational contexts. Tu et al. [21] introduced unsupervised contrastive learning to understand the role of context and common sense in emotional judgment. Pan et al. [4] aligned audio and text features across modalities. Ye et al. [22] designed a module for comparative emotion decoupling. Building on these, this paper delves into contrastive learning's application in instance-level multi-task decoupling.

III. Proposed Methodology

As shown in Figure 1, we propose a Hybrid Contrastive Learning Decoupling (HCLD) method for Speech Emotion Recognition (SER) and Speaker Verification (SV). The model consists of three parts: feature extraction, multi-task classification, and hybrid contrastive learning decoupling loss (HCLD). Below we will present the task configuration and each of these three parts separately.

A. Task Configuration

Suppose we have a dataset D consisting of N utterances $u_1, u_2, ..., u_N$, each with corresponding emotion labels $e_1, e_2, ..., e_N$ and speaker labels $s_1, s_2, ..., s_N$. We define the SER task as assigning an emotion label e_i to each utterance u_i, and the SV task as assigning a speaker label s_i to

each utterance u_i. The decoupling goal is to leverage the differential features focused on by the two tasks to improve the performance of the main task (the SER task).

B. Feature Extraction

We use the pre-trained HuBert-base model [6] as the raw waveform encoder. Specifically, we employ the checkpoint pre-trained on the 960-hour LibriSpeech dataset [23] released by torchaudio as the feature extraction model. Subsequently, we design three feature extractors based on popular network architectures (CNN, RNN, transformers) to extract emotion features X_e and speaker features X_s from x_i.

$$X_i = Encoder(x_i)$$

where $x_i \in (batchsize, length, dim)$ represents the audio features extracted by the pre-trained model hubert.

C. Multi-task Classification

Having extracted the emotion features X_e and speaker features X_S, we predict emotion and speaker categories through Linear network and softmax layer, with cross-entropy loss. We use L_{task} to realize basic multi-task learning:

$$\widetilde{y_e} = softmax(w * X_e + b), L_{emo} = -\sum_{i=1}^{N} y_{ei} \log(\widetilde{y_e})$$

$$\widetilde{y_s} = softmax(w * X_s + b), L_{spe} = -\sum_{i=1}^{N} y_{si} \log(\widetilde{y_s})$$

$$L_{task} = L_{emo} + L_{spe}$$

where y_{ei} the true emotion label, y_{si} is the true speaker label, $\widetilde{y}$ is the predicted probability distribution from the $softmax$ layer, w and b are the learned model parameters, and N is the total number of samples used in training.

D. Hybrid Contrastive Learning Decoupling

1) Instance-level Contrastive Learning Decoupling: To achieve decoupling between SER and SV tasks, we construct positive samples as samples from the same person expressing different emotions, and negative samples as samples from different people expressing the same emotion. We then calculate the loss of positive samples and negative samples to obtain the overall contrastive decoupling loss L_{CLD}. This makes samples of the same speaker and different emotion closer in feature space, and samples of different speaker and same emotion further apart.

$$L_{\text{positive}} = \frac{1}{|x^+|} \sum_{(i,j)\in x^+} (1 - \cos(X_e[i], X_e[j]))^2$$

where $|x^+|$ represents the number of positive sample pairs, and $|x^-|$represents the number of negative sample pairs. For each pair of positive samples x^+, we want the emotion features to $\hat{X}_e$ be closer. For each pair of positive samples x^+, we calculate the cosine similarity between their emotion features, hoping that the higher the similarity, the better.

$$L_{\text{negative}} = \frac{1}{|x^-|} \sum_{(i,j)\in x^-} \max(0, Margin - \cos(X_s[i], X_s[j]))^2$$

For each pair of negative samples x^-, we want the speaker features $\hat{f}_{spe}$ to be further apart. For each pair of negative samples x^-, we calculate the cosine similarity between their speaker features, hoping that this similarity is less than a preset threshold (Margin).

The total decoupling loss is:

$$L_{\text{CLD}} = L_{\text{positive}} + L_{\text{negative}}$$

2) Label-based Contrastive Loss: We use the roberta-base model to extract the features E_i of the text labels e_i, as a benchmark in the feature decoupling stage. Inspired by the anchor loss [24], based on the SupCon loss, we construct positive and negative sample pairs using the features E_i of the labels and the audio features X_e, to compute the contrastive learning loss based on labels:

$$L_{Label} = -\sum_{i\in I} \frac{1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(\hat{E}[i] \cdot \hat{X}e[p]/\tau)}{\sum_{a\in A(i)} \exp(\hat{E}[i] \cdot \hat{X}e[a]/\tau)}$$

where $\hat{X}_e[i]$ represents the emotional feature of the i th sample.

3) Supervised Contrastive Loss: To avoid the CLD loss scrambling the feature space, we also introduce a supervised contrastive learning loss, the SupCon loss. This aims to minimize the distance between features of the same category and maximize the distance between features of different categories.

$$L_{SCL} = -\sum_{i\in I} \frac{1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(\hat{X}s[i] \cdot \hat{X}s[p]/\tau)}{\sum_{a\in A(i)} \exp(\hat{X}s[i] \cdot \hat{X}s[a]/\tau)}$$

where $\hat{X}_s[i]$ represents the feature representation of the i sample (which could be the speech feature vector extracted by the deep learning model). $P(i)$ represents the set of all positive samples of the same category as sample i. $A(i)$ represents the set of all samples available for comparison, including positive and negative samples, but excluding sample i itself.

Particularly, contrastive learning methods typically require a larger $batchsize$; otherwise, it can result in a significant decrease in training performance. We replicate and downsample the features X_e and X_s , ensuring that each training sample has at least one positive sample.

Therefore, the total loss can be expressed as:

$$L = L_{task} + L_{\text{CLD}} + L_{SCL} + L_{Label}$$

IV. Experiments and Results

A. Dataset

IEMOCAP [7] is a widely-used benchmark SER dataset with a total length of approximately 12 hours. It contains five sessions, each featuring one male and one female participant-for a total of ten speakers.

TABLE I
MODEL PERFORMANCE COMPARISONS ON THE IEMOCAP DATASET

Model	WA(%)	UA(%)
CNN-LSTM [25]	65.4	66.9
LGFA [26]	73.29	62.63
ATFNN [27]	73.81	64.48
TFF [28]	74.43	63.90
TAP [29]	-	74.2
TAPT [30]	-	74.3
HCLD-CNN	75.52	76.08
HCLD-RNN	75.70	75.9
HCLD-transformers	73.80	74.59

TABLE II
ABLATION STUDY

Type	Model	Emotion		Speaker	
		WA(%)	UA(%)	WA(%)	UA(%)
EMO-only	CNN	70.84	71.61	-	-
	LSTM	70.13	70.95	-	-
	Transformer	68.03	68.73	-	-
orth	CNN	73.35	74.04	74.04	74.83
	LSTM	74.44	75.24	78.67	78.48
	Transformer	70.67	71.50	81.66	81.52
wo.SupCon	CNN	74.35	74.8	81.57	81.47
	LSTM	75.16	75.71	80.94	80.77
	Transformer	71.18	72.02	81.84	81.69
wo.LCL	CNN	75.03	74.27	74.98	75.01
	CNN	**75.88**	**76.51**	84.46	84.27
	LSTM	75.43	75.78	77.6	77.69
	Transformer	72.99	73.62	81.84	81.72
HCLD	CNN	75.52	76.08	75.70	75.90
	LSTM	**75.70**	**75.90**	78.95	79.00
	Transformer	**73.80**	**74.59**	83.02	82.72

We utilize 5531 audio samples from IEMOCAP, which encase four types of emotions: hap (1636), ang (1084), sad (1103), and neu (1708), with all ten speaker labels being used. The dataset is randomly divided into a training set (80%) and a test set (20%), and a five-fold cross-validation method is employed for training. Each sample is trimmed to 7.5 seconds with a sampling rate set at 16kHz.

B. Experimental Setup

After using the pre-trained HuBert-Base [6] for feature extraction, the features are normalized and mapped to the dimensions of $[batchsize, length, channel]$. Specifically, we set the $batchsize$ to 32, and the $length$ and dim are set to 374 and 768, respectively.

We employ the Adam optimizer with a learning rate of $5e{-}4$ for model training. The model is trained for 100 epochs with a batch size of 32, of which 5 epochs are dedicated to linear warm-up. Weighted accuracy (WA) and unweighted accuracy (UA) were used as the evaluation metrics.

C. Results and Analysis

As shown in Table 1, following the introduction of multi-task auxiliary SER training, the accuracy of emotion recognition significantly improved. Within the CNN-based model, the incorporation of multi-tasking and our proposed HCLD loss resulted in an increase of WA from 70.84% to 75.52% (+4.68). In the LSTM-based feature extraction model, the HCLD loss boosted WA from 70.13% to 75.70% (+5.57). Within the transformer model, WA rose from 68.03% to 72.99% (+3.96), and UA from 68.73% to 73.80% (+5.07). The significant performance improvements brought about by the introduction of HCLD indicate the effectiveness of the proposed HCLD method, coupling multi-task joint training with feature decoupling. Compared to the orth method, our HCLD method also led to performance improvements, proving the effectiveness of our proposed method.

Furthermore, we conducted ablation experiments to validate the effectiveness of the Supcon loss and label-based contrast loss. After removing the SupCon loss L_{SCL}, the WA decreased by 1.53%, 0.27%, and 1.81% respectively, thus proving the actual existence of L_{SCL} constraint on labels in the feature space. After removing the label-based loss L_{Label}, the WA in models with LSTM and Transformer as feature extractors dropped by 0.27% and 0.81% respectively. However, in the CNN-based feature extraction model, the performance actually increased by 0.36%. This might be due to performance bottlenecks in the data. Additionally, we found that when using transformers, the performance of the SER task was poorer, but the SV task performed better than the CNN or LSTM architectures. We suspect this might be due to imperfect network architecture or loss function design, and we will continue to investigate this in our future research.

D. Comparison with State-of-the-Arts

To validate the effectiveness of our proposed method, we compared our HCLD multitask feature decoupling method with six kinds of unimodal SER methods. Compared to the LGFA ATFNN TFF model, our WA scores have been increased by 2.41%, 1.89% and 1.27% respectively. Notably, there is a significant increase in UA scores, with increments of 13.45%, 11.6% and 12.18% respectively. Compared to the TAPT, our method improved the UA by 2.3%. This indicates that exploring multitask joint training and multitask decoupling, particularly the decoupling of emotional features and other features in speech, can significantly enhance the performance of SER. Especially in the simultaneous improvement of WA and UA, our method performs better. Multitask decoupling learning has tremendous research potential.

V. CONCLUSION AND LIMITATIONS

In this paper, we presents an instance-level Hybrid Contrastive Learning Decoupling method, which effectively enhances the performance of SER tasks by introducing a Speaker Verification. Different from traditional methods based on orthogonalization, this approach can operate independently of specific network architecture designs and effectively disentangle emotion features from speaker features. The effectiveness of the HCLD method has been validated on the IEMOCAP dataset.

In the future, we aim to explore more possibilities of this contrastive decoupling method and attempt to introduce more subtasks.

REFERENCES

[1] S. Ghosh, U. Tyagi, S. Ramaneswaran, H. Srivastava, and D. Manocha, "Mmer: Multimodal multi-task learning for speech emotion recognition," *arXiv preprint arXiv:2203.16794*, 2022.

[2] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning." in *Interspeech*, vol. 2021. Brno, 2021, pp. 4508–4512.

[3] L. Wang, R. Gu, W. Zhuang, P. Gao, Y. Wang, and Y. Zou, "Learning decoupling features through orthogonality regularization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7562–7566.

[4] Y. Pan, Y. Hu, Y. Yang, J. Yao, W. Fei, L. Ma, and H. Lu, "Gemo-clap: Gender-attribute-enhanced contrastive language-audio pretraining for speech emotion recognition," *arXiv preprint arXiv:2306.07848*, 2023.

[5] T. Liu, R. K. Das, M. Madhavi, S. Shen, and H. Li, "Speaker-utterance dual attention for speaker and utterance verification," *arXiv preprint arXiv:2008.08901*, 2020.

[6] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[7] 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.

[8] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.

[9] X. Wang, M. Wang, W. Qi, W. Su, X. Wang, and H. Zhou, "A novel end-to-end speech emotion recognition network with stacked transformer layers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6289–6293.

[10] Y. Wang, Y. Gu, Y. Yin, Y. Han, H. Zhang, S. Wang, C. Li, and D. Quan, "Multimodal transformer augmented fusion for speech emotion recognition," *Frontiers in neurorobotics*, vol. 17, p. 1181598, 2023.

[11] M. R. Amer, B. Siddiquie, C. Richey, and A. Divakaran, "Emotion detection in speech using deep networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 3724–3728.

[12] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.

[13] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[14] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech emotion recognition using speech feature and word embedding," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 519–523.

[15] 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.

[16] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[17] M. A. Pastor, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, "Cross-corpus speech emotion recognition with hubert self-supervised representation," in *IberSPEECH 2022*. ISCA, 2022, pp. 76–80.

[18] M. Jain, S. Narayan, P. Balaji, A. Bhowmick, R. K. Muthu *et al.*, "Speech emotion recognition using support vector machine," *arXiv preprint arXiv:2002.07590*, 2020.

[19] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[20] S. Li, H. Yan, and X. Qiu, "Contrast and generation make bart a good dialogue emotion recognizer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 10, 2022, pp. 11 002–11 010.

[21] G. Tu, B. Liang, R. Mao, M. Yang, and R. Xu, "Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 14 054–14 067.

[22] J. Ye, Y. Wei, X.-C. Wen, C. Ma, Z. Huang, K. Liu, and H. Shan, "Emodna: Emotion decoupling and alignment learning for cross-corpus speech emotion recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5956–5965.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] F. Yu, J. Guo, Z. Wu, and X. Dai, "Emotion-anchored contrastive learning framework for emotion recognition in conversation," *arXiv preprint arXiv:2403.20289*, 2024.

[25] Y. Xia, L.-W. Chen, A. Rudnicky, R. M. Stern *et al.*, "Temporal context in speech emotion recognition." in *Interspeech*, vol. 2021, 2021, pp. 3370–3374.

[26] C. Lu, H. Lian, W. Zheng, Y. Zong, Y. Zhao, and S. Li, "Learning local to global feature aggregation for speech emotion recognition," *arXiv preprint arXiv:2306.01491*, 2023.

[27] C. Lu, W. Zheng, H. Lian, Y. Zong, C. Tang, S. Li, and Y. Zhao, "Speech emotion recognition via an attentive time–frequency neural network," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 6, pp. 3159–3168, 2022.

[28] Y. Wang, C. Lu, Y. Zong, H. Lian, Y. Zhao, and S. Li, "Time-frequency transformer: A novel time frequency joint learning method for speech emotion recognition," in *International Conference on Neural Information Processing*. Springer, 2023, pp. 415–427.

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

[30] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.