



Label Semantic-Driven Contrastive Learning for Speech Emotion Recognition

Jiayi Hu¹, Leyuan Qu¹, Haoxun Li¹, Taihao Li^{1,*}

¹Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, China

hujiayi23@mails.ucas.ac.cn, leyuan.qu@ucas.ac.cn, lihaoxun23@mails.ucas.ac.cn,
lith@ucas.ac.cn

Abstract

Speech Emotion Recognition (SER) is crucial for human-computer interaction applications. However, SER remains a challenging task due to limited datasets and ambiguous emotion boundaries. While Self-Supervised Learning (SSL) models have demonstrated considerable success in speech processing tasks, existing approaches still struggle to distinguish subtle emotional variations. In this paper, we propose a novel **Label Semantic-driven Contrastive Learning** framework (LaSCL) that integrates emotion label semantic embeddings into speech representation learning. Our method uses label embeddings as semantic anchors to explicitly model relationships between emotions and employ a label divergence loss to better establish clearer emotion boundaries. Experiments on the widely used IEMOCAP benchmark indicate that LaSCL achieves state-of-the-art performance compared with previous methods.

Index Terms: Speech emotion recognition, supervised contrastive learning, label embedding

1. Introduction

Speech is one of the most direct and convenient means of communication. Speech Emotion Recognition (SER) aims to establish relationships between speech signals and emotional states, enabling machines to recognize and interpret human emotions. As an emerging field within artificial intelligence, SER has attracted considerable research interest for its applications ranging from human-computer interaction, online education, healthcare and information security. However, SER remains a challenging task due to the limited availability of labeled datasets compared to other speech-related tasks, which constrains the ability to train robust and generalizable models.

Recent advances in Self-Supervised Learning (SSL) have demonstrated promising results in speech processing tasks. SSL models such as Wav2vec2.0 [1], HuBERT [2] and WavLM [3] have achieved remarkable performance in Automatic Speech Recognition (ASR) by learning robust speech representations from large-scale unlabeled data. As SSL and pre-training techniques continue to evolve, researchers attempt to apply pre-trained models to SER tasks. Boigne et al. [4] proposed a transfer learning approach that combines acoustic features from the Wav2vec model and linguistic features from the BERT model using only a small amount of training data. Pepino et al. [5] explored the effectiveness of Wav2vec2.0 for SER by proposing a weighted combination of multi-layer features and evaluating different pre-training strategies. Chen et al. [6] developed a fine-tuning strategy that modifies task-specific pre-training to learn better emotion representations from Wav2vec2.0.

While these studies have demonstrated the effectiveness of pre-trained SSL models for SER tasks, most of them still struggle with class imbalance issues, which can significantly impact the model's performance on minority classes. Moreover, the inherent complexity and subtlety of emotional expressions often lead to ambiguous boundaries between emotions, particularly when emotional intensity is low or semantic content is not clearly expressed. For example, emotions such as *happy* and *excited* are frequently misclassified as *neutral* [7].

In recent years, Contrastive Learning (CL) [8] has achieved great success in self-supervised representation learning in various domains, including computer vision [9, 10], speech [11, 12], and natural language processing (NLP) [13, 14, 15]. Based on this, Supervised Contrastive Learning (SCL) [16, 17, 18] has been proposed to extend CL by utilizing label information. By leveraging supervised information, SCL brings samples from the same class (positive pairs) closer, while simultaneously pushing apart samples from different classes (negative pairs) within a batch. To address the problem of SER, researchers have employed SCL to learn robust and generalized feature representations. Alaparthi et al. [19] proposed SCL on top of the Wav2vec2.0 transformer with custom augmentations during the fine-tuning stage. Wang et al. [20] combined cross-entropy loss with SCL loss during fine-tuning, and employed the k-nearest neighbors algorithm in the inference stage to refine the model's predictions.

Despite the promising advantages of these SCL methods, they still have some limitations. Labels are typically provided in the form of text. However, they are often treated as simple categorical indices, neglecting the rich semantic information and inherent relationships embedded in these labels. Moreover, how to effectively leverage semantic relationships between emotion labels becomes particularly crucial when dealing with limited training data.

In this work, we propose a novel Label Semantic-driven Contrastive Learning framework (LaSCL) for SER. Inspired by recent advances in label embedding for NLP tasks [21], our approach integrates emotion label embeddings into the contrastive learning process while dynamically aligning acoustic and linguistic knowledge. Specifically, our framework introduces the following innovations: (1) We use a text encoder to transform emotion labels into a semantic space and explicitly model relational structures between emotions. These semantic embeddings can serve as anchors or positive / negative samples in the SCL objective and mitigate class imbalance issues. It promotes intra-class compactness by forming coherent clusters for identical labels and enhances inter-class separability by maximizing the distances between representations of different emotions. (2) We adopt a mixed data augmentation strategy and align speech representations extracted from pre-trained SSL models with la-

*Corresponding author.

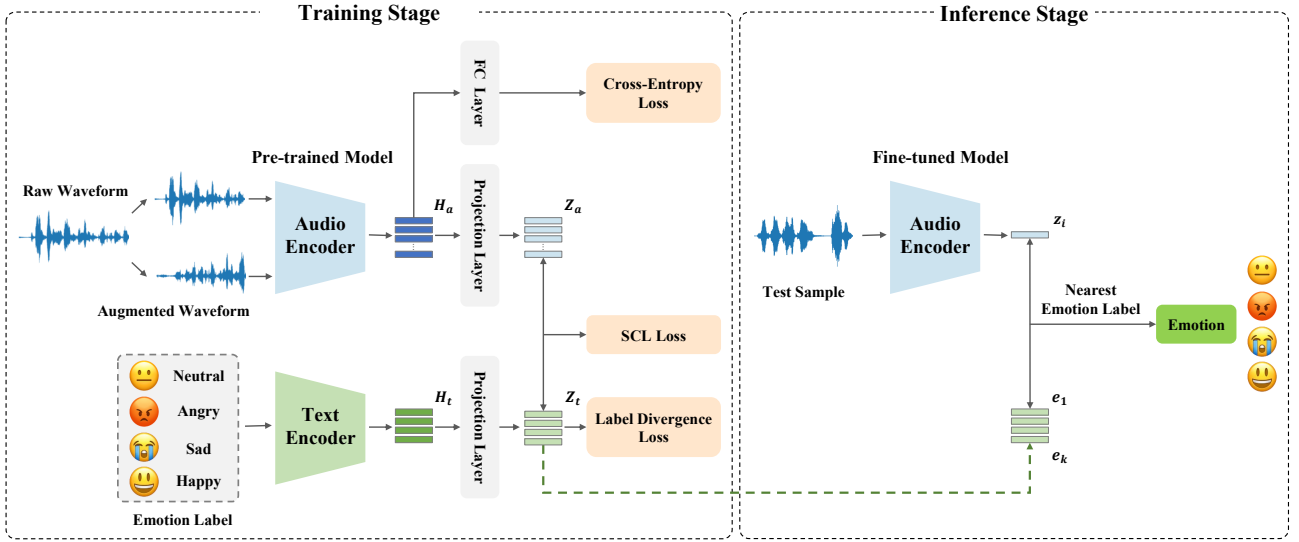


Figure 1: The architecture of our proposed method LaSCL.

bel embeddings in a unified semantic space. This alignment bridges the gap between acoustic patterns and emotion semantics, enabling the model to capture emotional relationships more effectively. (3) We implement a label divergence loss that maximizes the separation between label embeddings in the representation space. This creates more discriminative boundaries between different emotions.

To summarize, the main contributions of this work are as follows:

- We propose a novel label semantic-driven contrastive learning framework that effectively incorporates emotion label semantic embeddings to establish clearer emotion boundaries, which is the first framework to explicitly integrate label semantics into contrastive learning for SER.
- We combine SCL loss, cross-entropy loss and label divergence loss during the fine-tuning stage to improve performance by learning better feature representations.
- Experiments conducted on the widely used IEMOCAP dataset demonstrate that our proposed method surpasses state-of-the-art (SOTA) performance.

2. Method

In this section, we introduce the overall architecture of our proposed method LaSCL. As illustrated in Figure 1, LaSCL consists of two stages: a training stage, which integrates multiple learning objectives, and an inference stage, which leverages the learned representations for emotion classification.

2.1. Model Architecture

Audio Encoder To address the challenge of data scarcity in SER and extract effective speech representations from raw waveforms, we adopt pre-trained SSL models (e.g. Wav2vec2.0 [1], HuBERT [2] and WavLM [3]) as our audio encoder. Additionally, we apply data augmentation techniques such as adding noise, pitch shifting and reverberation to enhance the diversity of the training data. For a batch of N samples, the augmented training batch \mathbf{X} consists of $2N$ instances, including both original and augmented waveforms. The augmented waveforms inherit the same labels as their original counterparts. Then, we feed the processed input into the pre-trained audio encoder and obtain the last hidden states of the outputs as follows:

$$\mathbf{H}_a = \text{AudioEncoder}(\mathbf{X}) \quad (1)$$

where $\mathbf{H}_a \in \mathbb{R}^{2N \times d_a}$ denotes the speech representations of dimensionality d_a . To fully leverage the representation power of SSL models, the audio encoder is fine-tuned during training.

Text Encoder To effectively encode semantic information from labels, we employ RoBERTa [22] as our text encoder, which has demonstrated superior performance in various NLP tasks. Given a set of K -class emotion labels $\mathcal{L} = \{l_1, l_2, \dots, l_K\}$, the text encoder extracts semantic features and generates the corresponding label embeddings as follows:

$$\mathbf{H}_t = \text{TextEncoder}(\mathcal{L}) \quad (2)$$

where $\mathbf{H}_t \in \mathbb{R}^{K \times d_t}$ denotes the text representations of dimensionality d_t . To ensure we get stable label representations, the text encoder is frozen during the training process.

To align the speech and label embeddings in a shared semantic space, we utilize separate projection layers for the two modalities. Specifically, the audio/text projection layer consists of two learnable multi-layer perceptron (MLP) layers with GELU activation function in between. The audio and text embeddings are mapped into a joint multimodal space of dimension d as follows:

$$\mathbf{Z}_a = \text{MLP}_a(\mathbf{H}_a); \quad \mathbf{Z}_t = \text{MLP}_t(\mathbf{H}_t) \quad (3)$$

where $\mathbf{Z}_a \in \mathbb{R}^{2N \times d}$ and $\mathbf{Z}_t \in \mathbb{R}^{K \times d}$ denote the final projected representations of the audio and text embeddings respectively. This projection mechanism helps bridge the modality gap between acoustic and semantic features.

2.2. Training Stage

Supervised Contrastive Learning With audio and text embeddings ($\mathbf{Z}_a, \mathbf{Z}_t$) mapped into the same semantic space, their distance can be measured using cosine similarity. Given a batch of N samples, let $\mathbf{Z} = \{z_1, z_2, \dots, z_{2N+K}\} = \mathbf{Z}_a \cup \mathbf{Z}_t$ denote the unified set of projected features from both acoustic embeddings and label embeddings. For each anchor, positive pairs are constructed from samples sharing the same emotion label (including both speech samples and their corresponding label embeddings), while all other samples are treated as negative pairs. The SCL loss can be formulated as follows:

$$\mathcal{L}_{\text{SCL}} = \sum_{i \in \mathcal{I}} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(z_i, z_p)/\tau)}{\sum_{j \in Z(i)} \exp(\text{sim}(z_i, z_j)/\tau)} \quad (4)$$

where $i \in \mathcal{I} = \{1, \dots, 2N + K\}$ denotes the index of an anchor instance, $Z(i)$ denotes all indices except i , and $P(i)$ represents the set of all positive pairs with the same emotion label as anchor i . τ is a temperature parameter that scales and controls the concentration level of the distribution.

Label Divergence Loss To further enhance the discriminative power of label embeddings and establish clearer decision boundaries between emotion categories, we introduce a label divergence loss. This auxiliary objective maximizes the separation between all pairs of distinct emotion label embeddings, ensuring a more uniform distribution while increasing inter-class distances in the semantic space. The label divergence loss is defined as follows:

$$p_{\text{Label}}^{(i,j)} = \frac{\exp(\mathbf{e}_i \cdot \mathbf{e}_j)}{\sum_{\mathbf{e}_k \in E_i} \exp(\mathbf{e}_i \cdot \mathbf{e}_k) + 1} \quad (5)$$

$$\mathcal{L}_{\text{Label}} = - \sum_{\mathbf{e}_i \in E} \sum_{\mathbf{e}_j \in E_i} \log \left(1 - p_{\text{Label}}^{(i,j)} \right) \quad (6)$$

where \mathbf{e}_i represents the embedding of the i -th emotion label, $E_i = \{\mathbf{e}_j \mid j \neq i\}$ is the set of all label embeddings except \mathbf{e}_i . \cdot denotes the dot product. As shown in Eq. 5, when calculating similarity scores between dissimilar emotion labels, we add a constant term of 1 to the denominator to enhance numerical stability during training.

Cross-Entropy Loss To provide direct supervision for emotion classification and enhance the model’s discriminative ability, we incorporate the traditional cross-entropy (CE) loss during the contrastive learning process. A fully connected (FC) layer is applied to the speech representations for classification, minimizing the difference between the predicted probabilities and ground-truth labels. The CE loss is defined as follows:

$$\hat{\mathcal{Y}} = \text{softmax}(\text{FC}(\mathbf{H}_a)) \quad (7)$$

$$\mathcal{L}_{\text{CE}} = - \frac{1}{2N} \sum_{i=1}^{2N} \sum_{j=1}^K y_{ij} \log \hat{y}_{ij} \quad (8)$$

where $\hat{\mathcal{Y}} \in \mathbb{R}^{2N \times K}$ denotes the predicted probability distribution over K emotion classes, \hat{y}_{ij} represents the predicted probability for class j of sample i , and y_{ij} is the ground-truth label.

The overall training objective of the LaSCL combines the SCL loss, label divergence loss and CE loss as follows:

$$\mathcal{L} = \lambda_1 (\mathcal{L}_{\text{SCL}} + \lambda_2 \mathcal{L}_{\text{Label}}) + (1 - \lambda_1) \mathcal{L}_{\text{CE}} \quad (9)$$

where λ_1 and λ_2 are trade-off hyperparameters that control the balance between each loss term. This joint optimization enables the model to learn discriminative emotion representations while maintaining the underlying semantic relationships.

2.3. Inference Stage

After fine-tuning, we adopt a similarity-based strategy for emotion recognition. Given a test utterance, we extract its speech representation using the fine-tuned audio encoder and projection layer. Instead of using traditional softmax-based classification, we utilize the learned semantic relationships by computing similarities between the speech representation and all emotion label embeddings. The final prediction is obtained by matching each speech representation to the nearest label embedding in the semantic space as follows:

$$\text{pred} = \arg \max_j \text{sim}(z_i, e_j) \quad (10)$$

where z_i denotes the representation of the i -th speech sample, e_j represents the emotion label embedding for class j .

3. Experiments

3.1. Dataset

We evaluate our proposed method on IEMOCAP (Interactive Emotional Dyadic Motion Capture) [23] dataset. IEMOCAP is the most widely used benchmark for SER. The dataset contains approximately 12 hours of audio-visual recordings from 10 professional actors, organized into dyadic sessions. We adopt the common practice of merging *happy* and *excited* into one emotion class *happy*, resulting in 5531 utterances with four primary emotion categories: *happy* (1636), *sad* (1084), *angry* (1103) and *neutral* (1708). For pre-processing, we set the max length of all utterances to 10 seconds due to computational limitations.

3.2. Experimental Setup

In this work, we use pre-trained models from HuggingFace such as Wav2vec2.0-Base/Large, HuBERT-Base/Large and WavLM-Base/Large as our audio encoder. Inspired by Wang et al. [24], we adopt a partial fine-tuning strategy to leverage the learned low-level acoustic features effectively. Specifically, we froze the CNN-based feature encoder and only fine-tuned the parameters of the Transformer-based contextualized encoder. All models are trained on 2×NVIDIA A100 GPUs with a batch size of 32. The number of training epochs is set to 50. We use the AdamW optimizer with a CosineAnnealingWarmup learning rate scheduler, and set the learning rate to 1×10^{-4} for fine-tuning the pre-trained model and 1×10^{-3} for other components. As for the hyperparameters in Eq. 4 and Eq. 9, we explore the optimal values for τ in $\{0.05, 0.07, 0.1, 0.15, 0.2\}$, λ_1 in $\{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ and λ_2 in $\{0, 0.01, 0.1, 1.0\}$. We evaluate the performance of our method using Weighted Accuracy (WA) and Unweighted Accuracy (UA) as metrics.

3.3. Results and Analysis

3.3.1. Performance

Following previous works [25, 6, 26, 27, 28, 29], we evaluate the performance of our proposed method LaSCL on the IEMOCAP dataset using 5-fold cross-validation (CV) in the leave-one-session-out scheme. The results are shown in Table 1. It can be observed that our proposed method outperforms recent SOTA approaches utilizing pre-trained SSL models, with improvements of up to 2.04% and 2.0% on WA and UA over the best previous method respectively.

Table 1: Comparison with SOTA SER approaches using pre-trained SSL models on the IEMOCAP dataset in 5-fold CV.

Approach	Year	WA(%)	UA(%)
Zou et al. [25]	2022	69.80	71.05
Chen et al. [6]	2023	-	74.30
Fang et al. [26]	2023	74.95	74.03
Gao et al. [27]	2023	74.94	76.10
Chen et al. [28]	2024	73.70	74.30
Hu et al. [29]	2024	75.75	76.42
Ours	2025	77.79	78.42

To further evaluate the performance of LaSCL on each emotional class, we present the confusion matrix and t-SNE visualization [30] for different fine-tuning methods, as shown in Fig. 2 and Fig. 3. When fine-tuning the WavLM model with only CE loss, the model struggles to distinguish between emotional states and tends to misclassify emotions, especially between *neutral* and other categories. In contrast, fine-tuning with our proposed method demonstrates that LaSCL effectively

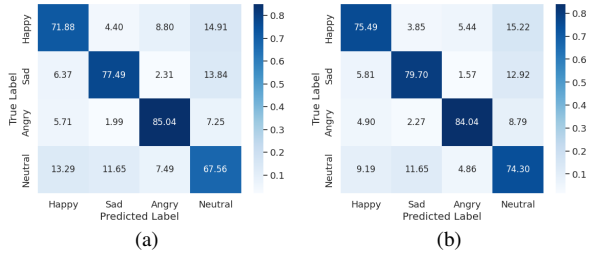


Figure 2: Comparison of confusion matrices (%) between different fine-tuning methods. (a): Results after fine-tuning WavLM with CE loss. (b): Results after fine-tuning with LaSCL.

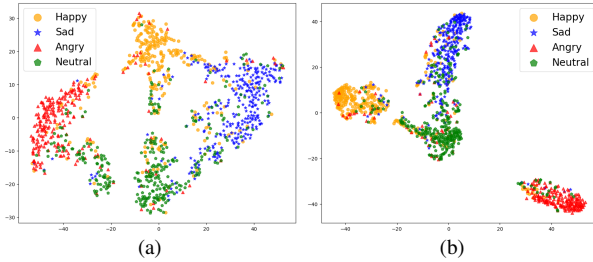


Figure 3: The t-SNE visualization of speech representations. (a): Outputs after fine-tuning WavLM with CE loss. (b): Outputs after fine-tuning with LaSCL.

establishes more distinct decision boundaries between different emotions in the representation space, thereby improving the model’s ability to differentiate subtle emotional variations.

3.3.2. Comparison of Different Data Augmentations

In this section, we conduct extensive experiments comparing eight different augmentation methods in a supervised contrastive setting to investigate the impact of data augmentation strategies on SER tasks. The results are summarized in Table 2. It could be found that the mixed augmentation strategy that combines multiple augmentation methods surpasses all individual methods and achieves the best overall performance. It demonstrates the effectiveness of diverse augmentation methods in improving model performance, particularly within a contrastive learning framework. The increased variety in training samples helps the model learn more robust and generalizable features by generating more informative positive / negative pairs during contrastive learning.

Table 2: Performance with different data augmentation strategies in the training stage.

Augmentation Methods	WA (%)	UA (%)
Polarity Inversion	72.70	73.68
Noise	76.54	77.28
Gain	74.26	75.25
High Low Pass	75.40	75.96
Delay	75.11	76.14
Pitch Shift	74.98	76.23
Reverberation	73.85	74.81
Mixed Augmentations	77.79	78.42

Additionally, the consistent performance across WA and UA metrics (with differences below 1%) suggests that our framework can naturally alleviate both data scarcity and class imbalance issues while improving the model’s classification ability. By treating label embeddings as anchors or positive / negative samples, the framework ensures that each emotion class is equally represented in the semantic space during the whole training, regardless of its frequency in the dataset.

Table 3: Performance of our proposed method with different audio encoders on IEMOCAP.

Audio Encoder	Text Encoder	WA (%)	UA (%)
Wav2vec2.0 Base		68.83	69.64
Wav2vec2.0 Large		72.26	73.28
HuBERT Base	RoBERTa Large	68.26	69.57
HuBERT Large		73.73	74.77
WavLM Base		73.03	74.58
WavLM Large		77.79	78.42

3.3.3. Comparison of Different Audio Encoders

To comprehensively evaluate the effectiveness of different pre-trained SSL models, we conduct experiments using different audio encoders within our proposed framework. As shown in Table 3, when using the pre-trained RoBERTa Large model as the text encoder, WavLM-based approaches generally obtain superior recognition results under all paradigms. Specifically, WavLM-Large achieves the best performance, surpassing Wav2vec2.0 and HuBERT by substantial margins. These results demonstrate that the choice of pre-trained audio encoder has a significant impact on the model’s emotion recognition capabilities. Models with advanced architecture and larger parameters show better performance in speech representation learning.

3.3.4. Ablation Study

To validate the effectiveness of each component in our proposed framework, we conduct comprehensive ablation studies by removing different components individually. The results shown in Table 4 demonstrate that each component contributes positively to the overall performance. First, we remove data augmentation and observe a significant drop in the performance of nearly 3.3%, indicating the critical role of our mixed augmentations strategy in learning robust emotional representations, particularly in the context of limited training data. We then remove the text encoder for label embedding and observe a moderate decrease. This result demonstrates the effectiveness of incorporating emotional semantic information into the SCL framework. Furthermore, we remove the L_{Label} which encourages uniform distribution of label embeddings in the representation space. It can be seen that L_{Label} helps establish clearer decision boundaries between emotion categories. The lack of L_{CE} also demonstrates that the classification objective is essential for the contrastive learning framework, particularly in optimizing the fine-tuning process of pre-trained models for SER tasks.

Table 4: Ablation studies of different key components in LaSCL.

Model	WA (%)	UA (%)
LaSCL	77.79	78.42
w/o Data Augmentation	74.56	75.06
w/o Label Embedding	76.23	76.78
w/o Label Divergence	76.74	77.49
w/o Classification Objective	75.36	75.99

4. Conclusions

In this paper, we presented LaSCL, a novel label semantic-driven contrastive learning framework for SER, which integrates emotion label embeddings as anchors and explicitly integrates label semantics to guide speech representation learning. Experimental results on IEMOCAP demonstrate the effectiveness of LaSCL. In the future, we plan to incorporate prosodic information into our method and extend this framework to multi-modal emotion recognition scenarios.

5. Acknowledgements

This work was supported in part by the Scientific Research Starting Foundation of Hangzhou Institute for Advanced Study (2024HIASC2001), in part by Zhejiang Provincial Natural Science Foundation of China (No. LQN25F020001), and in part by the Key R&D Program of Zhejiang (2025C01104).

6. References

- [1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [2] 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.
- [3] 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.
- [4] J. Boigne, B. Liyanage, and T. Östrem, “Recognizing more emotions with less data using self-supervised transfer learning,” *ArXiv*, vol. abs/2011.05585, 2020.
- [5] L. Pepino, P. Riera, and L. Ferrer, “Emotion recognition from speech using wav2vec 2.0 embeddings,” in *Interspeech 2021*, 2021, pp. 3400–3404.
- [6] L.-W. Chen and A. I. Rudnický, “Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition,” *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2021.
- [7] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, “Emotion recognition from variable-length speech segments using deep learning on spectrograms,” in *Interspeech 2018*, 2018, pp. 3683–3687.
- [8] P. H. Le-Khac, G. Healy, and A. F. Smeaton, “Contrastive representation learning: A framework and review,” *Ieee Access*, vol. 8, pp. 193 907–193 934, 2020.
- [9] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2019.
- [10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [11] A. Saeed, D. Grangier, and N. Zeghidour, “Contrastive learning of general-purpose audio representations,” *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3875–3879, 2020.
- [12] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, “Contrastive unsupervised learning for speech emotion recognition,” *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6329–6333, 2021.
- [13] H. Fang, S. Wang, M. Zhou, J. Ding, and P. Xie, “Cert: Contrastive self-supervised learning for language understanding,” *arXiv preprint arXiv:2005.12766*, 2020.
- [14] T. Gao, X. Yao, and D. Chen, “Simcse: Simple contrastive learning of sentence embeddings,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 2021, pp. 6894–6910.
- [15] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, “Consert: A contrastive framework for self-supervised sentence representation transfer,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Association for Computational Linguistics, 2021, pp. 5065–5075.
- [16] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.
- [17] B. Guneļ, J. Du, A. Conneau, and V. Stoyanov, “Supervised contrastive learning for pre-trained language model fine-tuning,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.
- [18] V. Suresh and D. Ong, “Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 4381–4394.
- [19] V. S. Alaparthi, T. R. Pasam, D. A. Inagandla, J. Prakash, and P. K. Singh, “Scser: Supervised contrastive learning for speech emotion recognition using transformers,” *2022 15th International Conference on Human System Interaction (HSI)*, pp. 1–7, 2022.
- [20] X. Wang, S. Zhao, and Y. Qin, “Supervised contrastive learning with nearest neighbor search for speech emotion recognition,” in *Interspeech 2023*, 2023, pp. 1913–1917.
- [21] Z. Zhang, Y. Zhao, M. Chen, and X. He, “Label anchored contrastive learning for language understanding,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*. Association for Computational Linguistics, 2022, pp. 1437–1449.
- [22] Y. Liu, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.
- [23] 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.
- [24] Y. Wang, A. Boumadane, and A. Heba, “A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding,” *arXiv preprint arXiv:2111.02735*, 2021.
- [25] H. Zou, Y. Si, C. Chen, D. Rajan, and C. E. Siong, “Speech emotion recognition with co-attention based multi-level acoustic information,” *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7367–7371, 2022.
- [26] Y. Fang, X. Xing, X. Xu, and W. Zhang, “Exploring downstream transfer of self-supervised features for speech emotion recognition,” in *Proc. INTERSPEECH*, vol. 2023, 2023, pp. 3627–3631.
- [27] Y. Gao, C. Chu, and T. Kawahara, “Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with asr and gender pretraining,” in *Proc. Interspeech*, 2023, pp. 3637–3641.
- [28] W. Chen, X. Xing, P. Chen, and X. Xu, “Vesper: A compact and effective pretrained model for speech emotion recognition,” *IEEE Transactions on Affective Computing*, vol. 15, pp. 1711–1724, 2023.
- [29] Y. Hu, H. Yang, H. Huang, and L. He, “Cross-modal features interaction-and-aggregation network with self-consistency training for speech emotion recognition,” in *Interspeech 2024*, 2024, pp. 2335–2339.
- [30] L. van der Maaten and G. E. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.