Interspeech 2024
1-5 September 2024, Kos, Greece

Iterative Prototype Refinement for Ambiguous Speech Emotion Recognition

Haoqin Sun[1], *Shiwan Zhao*[1], *Xiangyu Kong*[2], *Xuechen Wang*[1], *Hui Wang*[1], *Jiaming Zhou*[1], *Yong Qin*[1,*]

[1]Nankai University, Tianjin, China
[2]University of Leicester, Leicester, United Kingdom
qinyong@nankai.edu.cn

Abstract

Recognizing emotions from speech is a daunting task due to the subtlety and ambiguity of expressions. Traditional speech emotion recognition (SER) systems, which typically rely on a singular, precise emotion label, struggle with this complexity. Therefore, modeling the inherent ambiguity of emotions is an urgent problem. In this paper, we propose an iterative prototype refinement framework (IPR) for ambiguous SER. IPR comprises two interlinked components: contrastive learning and class prototypes. The former provides an efficient way to obtain high-quality representations of ambiguous samples. The latter are dynamically updated based on ambiguous labels—the similarity of the ambiguous data to all prototypes. These refined embeddings yield precise pseudo labels, thus reinforcing representation quality. Experimental evaluations conducted on the IEMOCAP dataset validate the superior performance of IPR over state-of-the-art methods, thus proving the effectiveness of our proposed method.

Index Terms: speech emotion recognition, iterative prototype refinement, contrastive learning

1. Introduction

The advancement of affective computing [1, 2] has spurred the development of a wide array of emotion recognition corpora [3, 4, 5]. Despite this progress, the intricate and costly process of data annotation, combined with the inherent ambiguity of emotional expressions, poses significant challenges. Traditional speech emotion recognition (SER) systems commonly rely on the assumption that emotions can be clearly categorized, attributing a precise or singular label to each vocal expression. For instance, Sun et al. [6] have engaged in representation learning with IEMOCAP [7], while Liu et al. [8] have explored metric learning on EmoDB [9] and CAISA [10]. Nevertheless, these approaches may fall short of capturing the full spectrum of emotional ambiguity. This observation is supported by psychological and statistical research [11, 12], as well as machine learning studies [13], which collectively suggest that the boundaries between different emotional categories are not always distinct. The complexity and nuance of emotional expressions, coupled with the diversity in corpus composition, demand more sophisticated SER methodologies capable of more effectively navigating the intricate emotional landscape. Such advancements would transcend the constraints of singular emotion labeling, thereby enhancing the precision and broad applicability of emotion recognition across a varied range of emotional states.

* Corresponding author.

Recently, researchers have attempted to address the challenges in ambiguous SER by modeling emotion ambiguity. For example, Lotfian et al. [14] propose a multitask learning framework to learn the primary and secondary emotions of an utterance, ignoring information about other minor emotions. In contrast, Ando et al. [15] propose soft-target label learning – estimating the proportion of all emotions, which increases the complexity of soft label learning. Moreover, Fei et al. [16] propose a multi-label emotion classification method to represent the ambiguity of emotions. However, these methods are primarily applicable to corpora retaining expert voting information during annotation. In the absence of annotated records, the generalizability of these methods is challenged.

To address these challenges, Zhou et al. [17] propose a Multi-Classifier Interaction Learning (MCIL) framework. This framework emulates the expert annotation process through multiple classifiers. On the one hand, it represents the annotation process using the classification results of multiple classifiers, facilitating soft label learning. On the other hand, it employs the majority voting mechanism to generate a precise label, enabling the study of traditional SER techniques. It is important to note that the efficacy of MCIL is heavily contingent on the performance of individual classifiers. Poor performance of these classifiers may result in unsatisfactory data labeling quality and emotion recognition performance. Conversely, if a particular classifier exhibits exceptional performance, the effectiveness of interactive learning may be diminished. Furthermore, simply using classifiers to annotate ambiguous samples and retraining the model may compromise the discriminability of the model in this type of data.

Inspired by the work [18], in this paper, we propose an iterative prototype refinement framework (IPR) for ambiguous SER. IPR comprises three key phases: class prototype learning, class prototype updating, and contrastive learning. In the class prototype learning phase, precise samples facilitate the initial acquisition of class prototype embeddings via a warm-up mechanism. Subsequently, during the class prototype updating phase, unlabeled ambiguous samples are introduced to participate in training process. We calculate the distance of the samples from each class prototype to construct ambiguous soft labels and update the class prototypes by the proportion of classes in the soft labels. Finally, contrastive learning is employed to augment the discriminative capability of the model, updating class prototypes when ambiguous samples and their enhanced counterparts points to one prototype. The main contributions are summarized as follows:

(1) We propose a novel iterative prototype refinement framework for ambiguous SER. Contrastive learning aids in the learning of ideal representations, improving the quality of class prototypes. Subsequently, enhanced class prototypes guide the

10.21437/Interspeech.2024-2525

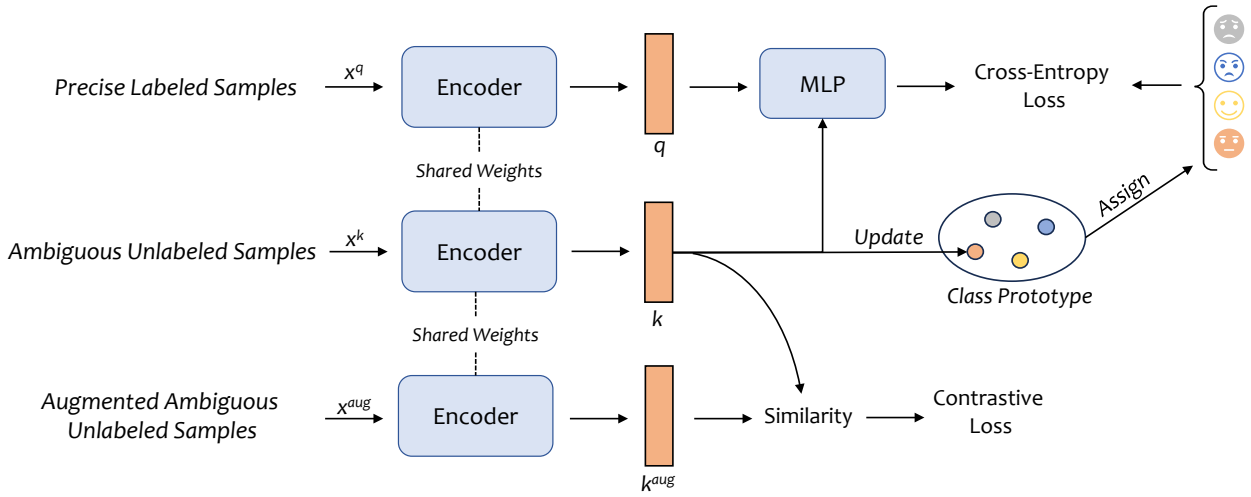

Figure 1: *The overall architecture of IPR, which consists of class prototype learning phase, class prototype updating phase and contrastive learning phase.*

model in generating high-quality representations.

(2) IPR is trained on a small amount of high-quality data, enabling the construction of ambiguous soft labels and single labels for a large amount of data, thereby possessing better generalization capability.

(3) Experimental results on IEMOCAP demonstrate that IPR outperforms the current state-of-the-art methods, achieving an accuracy of 70.75%, with an absolute improvement of 2.00%.

2. Proposed Methods

Illustrated in Fig. 1, our proposed framework, IPR, comprises three key phases: class prototype learning, class prototype updating, and contrastive learning.

2.1. Class Prototype Learning

First, we initially train the encoder using the precise labeled samples x_i^q. Meanwhile, for each class $c \in \{1, 2, ..., C\}$, the prototype vectors p_c are initialized based on the precise samples as representative embeddings.

$$q_i = \text{Encoder}\left(x_i^q; \theta\right), \tag{1}$$

$$p_c = \frac{1}{n} \sum_{i \in c} q_i, \tag{2}$$

where Encoder represents the pre-trained model Wav2vec2.0 [19], which is commonly used as a feature extractor [20, 21]. θ represents the trainable parameters. n represents the number of samples belonging to class c.

Since, when the class prototype is not adequately learned, it has a negative impact in guiding the model to learn unlabeled data. Therefore, we set up the warm-up mechanism so that the encoder learns the better representation of the class prototypes on the precise labeled data.

2.2. Class Prototype Updating

Once we acquire a prototype representation trained on precise labeled samples, the training process of the model will incorporate ambiguous unlabeled samples x_i^k. Then, we compute the similarity between k_i and each class prototype vector p_c to construct ambiguous soft label $\{s_1, s_2, ..., s_C\}$, where s_c is $k_i^\top p_c$. The class prototype vector with the largest similarity to k_i is the pseudo label z_c of that representation.

$$z_c = \arg\max k_i^\top p_c, \tag{3}$$

where $\boldsymbol{p}_c$ represents the c-th class. max represents maximization operation.

In the subsequent step, following the assignment of pseudo labels to ambiguous unlabeled samples, we dynamically adjust all class prototype vectors based on the ambiguous labels to incorporate the emotional information introduced by these samples. Traditional methods for calculating prototype embeddings require iterative updates, which is time-consuming and computationally intensive. To avoid this problem, we use moving average as the update method for prototype embedding.

$$\begin{aligned} p_c &= \text{Normalize}\left(\gamma p_c + (1-\gamma) k_i s_c\right), \\ &\text{if } c = \max \text{MLP}\left(k_i\right) = \max \text{MLP}\left(k_i^{aug}\right), \end{aligned} \tag{4}$$

where Normalize represents the normalization operation. γ represents a tunable hyperparameter. MLP stands for the classifier.

2.3. Contrastive learning

To enhance the quality of learned emotion representations in ambiguous samples, we integrate contrastive learning into IPR. For each ambiguous sample x_i^k, we initiate the process by generating an augmented version, denoted as x_i^{aug}, through a mixing method. This method encompasses various techniques such as adding noise, changing volume, adding reverberation, and changing pitch. Subsequently, we employ class prototype embeddings to assign pseudo labels for both x_i^k and x_i^{aug}. In a batch of $2N$ samples, comprising x_i^k and x_i^{aug}, we differentiate between positive examples, which share the same pseudo label, and negative examples, which possess different pseudo labels. The loss is computed as follows:

$$\mathcal{L}_{con} = -\log \frac{\exp\left(\left(x_i^k \cdot x_i^{aug}\right)/\tau\right)}{\sum_{j \in \{k, aug\}} \exp\left(\left(x_i^k \cdot x_j^{aug}\right)/\tau\right)}, \tag{5}$$

where τ is a temperature parameter.

Table 1: *The sizes of training and testing datasets.*

Dataset	D1	D2	D3
IEMOCAP	1710	3421	400

2.4. Training Objective

To train our proposed IPR, three distinct loss functions are employed: cross-entropy loss for precise labeled samples, cross-entropy loss for ambiguous samples, and contrastive learning loss. The overall training loss function is formulated as:

$$\mathcal{L} = \alpha \mathcal{L}_{cls}\left(x^q, y\right) + \beta \mathcal{L}_{cls}\left(x^k, \hat{y}\right) + \mu \mathcal{L}_{con}, \qquad (6)$$

where α, β, and μ represent the tunable trade-off factors. y and $\hat{y}$ represent ground-truth and pseudo labels, respectively.

3. Experiments

3.1. Dataset

IEMOCAP [7] stands out as a widely used dataset in SER. Developed by the Signal Analysis and Interpretation Laboratory at the University of Southern California, this dataset comprises emotionally charged dialogues enacted by 10 actors representing diverse demographics in terms of gender and age. These dialogues are orchestrated using constraint-based scripts crafted to mimic authentic interaction scenarios. we focus on four emotions: anger, happiness, sadness, and neutral, where excitement class is merged into the happiness class.

3.2. Experimental Settings

As shown in Table 1, following the methodology proposed by [17], we partition the IEMOCAP dataset into three distinct subsets tailored for addressing ambiguous SER scenarios. Specifically, D1 encompasses precise labeled samples exhibiting minimal ambiguity in emotional classification. D2 comprises unlabeled samples with moderate ambiguity in emotion. D3, on the other hand, includes labeled samples characterized by the highest degree of ambiguity in emotion, serving as the evaluation set. The determination of emotion ambiguity is predicated on the level of consensus among experts during the annotation process. For each utterance in IEMOCAP, multiple experts contribute their assessments, and the consistency of their judgments dictates the attribution of the sample. A high level of agreement among annotators results in D1, whereas increased variability in assessments places a sample within D2. Samples demonstrating substantial disagreement among annotators are designated to D3.

In our experimental setup, we establish three baseline systems: a baseline system, a baseline$^+$ system, and a supervised baseline system. The baseline system is trained exclusively on D1 and subsequently evaluated on D3. In contrast, the baseline$^+$ system utilizes both D1 and unlabeled D2 during training, which employs model-generated soft labels instead of the prototype-assigned pseudo labels employed in our proposed method. Finally, the supervised baseline system is trained on a combination of labeled D1 and D2 data and tested on D3. This system consists of an encoder and a MLP without the incorporation of class prototype and contrastive learning.

Our proposed framework, IPR, is trained on A100 utilizing PyTorch. We adopt a batch size of 8 and set the maximum training epoch to 50, incorporating a warm-up mechanism for the initial 10 epochs. We employ the AdamW optimizer with an initial learning rate of 10^{-5}, setting γ to 0.99. The trade-off parameters α and μ are fixed at 1.0 and 0.2, respectively. Initially, β is set to 0.0 and updated according to Eq. 8, gradually increasing based on the current and maximum epoch numbers. The increment of β accelerates with each epoch until reaching 0.5.

$$\beta = \min\left(\frac{\text{weight}_m \times \left(r^{\frac{\text{epoch}_c}{\text{epoch}_m/2}} - 1\right)}{r-1}, \text{weight}_m\right), \qquad (7)$$

where the weight parameter weight_m is set to 0.5, denoting its maximum value. Here, epoch_m represents the maximum epoch, while epoch_c denotes the current epoch. The exponential growth factor r governs the rate of increase. Consistent with prior research [17], we adopt accuracy (Acc) as the primary metric for assessing the efficacy of our proposed method. The final Acc is the average results across five seeds.

Table 2: *Performance comparison of our proposed methods with SOTA approaches on IEMOCAP.*

Method	**Description**	**Acc(%)**
Cummins et al.[22]	AlexNet	59.20
Li et al. [23]	CNN	58.00
Ando et al. [24]	Multi-label	57.80
Liu et al. [25]	CapsNet	58.55
Zhou et al. [17]	Majority voting	65.00
Zhou et al. [17]	MCIL	67.00
Ghifary et al. [26]	DaNN	65.00
Yu et al. [27]	DAAN	65.75
Ganin et al. [28]	DANN	68.50
Cui et al. [29]	BNM	68.75
IPR	**Prototype learning**	**70.75**

3.3. Comparison with State-of-the-Art Methods

To demonstrate the effectiveness of IPR, we compare IPR with many state-of-the-art (SOTA) methods. Table 2 lists a brief description and accuracy results of these methods, where the first half is SER method, retrained following MCIL training algorithm, and the second half is the classical semi-supervised methods.

We could find that our method achieves the optimal performance on Acc, reaching 70.75%, which is significantly better than SOTA methods (with an absolute improvement of 2.00%). This performance is mainly attributed to the collaborative effect of class prototype learning and contrastive learning, which reinforce each other, thus allowing the model to capture weak dominant emotions from ambiguous utterances. In comparison with the SER approaches, IPR demonstrates effectiveness and reliability in generating pseudo labels. With the guidance of pseudo labels, IPR achieves a 3.75% improvement on Acc. Furthermore, in the comparison with the semi-supervised methods, we find that IPR is improved by 2.00% on Acc. These semi-supervised methods only utilize a large amount of unlabeled data for the purpose of improving recognition performance. It does not seem to tackle the problem of annotation on large amounts of data by learning on a small number of precise samples. The experimental results provide strong evidence that our proposed IPR could provide ideal performance for ambiguous SER and permit more reliable data annotation.

Table 3: *Experimental results of different baseline systems on IEMOCAP dataset. The baseline$^+$ represents training using model-generated pseudo labels. $\star$ indicates that p-value < 0.05 (compared with baseline). ★ indicates that p-value < 0.05 (compared with baseline and baseline$^+$).*

Method	Train	Acc(%)
baseline	D1	64.25
baseline$^+$	D1 and unlabeled D2	66.85*
IPR	**D1 and unlabeled D2**	**70.75**★
supervised baseline	both labeled D1 and D2	72.30

3.4. Baseline System Analysis

We report the experimental results of IPR with different baseline systems as shown in Table 3. First, we observe that by integrating these unlabeled samples, the performance of IPR improves significantly, by 6.50% over baseline on Acc. This highlights the importance of using unlabeled ambiguous data to optimize model. Second, IPR outperforms baseline$^+$. This proves that IPR, an iterative prototype refinement framework, outperforms model-generated soft labels in the generation of pseudo labels. Notably, the baseline$^+$ also outperforms the majority voting method of MCIL in SOTA. Thus, when faced with a corpus without voting information, our method utilizes a small amount of data for training and could achieve a large number of data annotation, both single and ambiguous soft labels. Finally, by comparing IPR and supervised baseline, we could notice that on Acc, the proposed method reduces 1.55%, while MCIL reduces 5.30%. As a result, IPR is closer to the performance of supervised system, while it produces more precise and reliable pseudo labels when annotating unlabeled samples.

3.5. Class Prototype Analysis

Figure 2 shows the trend of similarity between class prototype embeddings during the training process. In the initial 10 epochs of training, we observe a transient increase in the similarity between the neutral prototype and other class prototype vectors. This phenomenon arises from the localized occurrence of emotional expression within speech segments. Additionally, given the similarity in activation levels between happiness and neutrality, they exhibit the highest prototype similarity, aligning with prior finding [30]. Intuitively, anger and sadness represent contrasting emotions, thus exhibiting a declining trend in the similarity between their prototypes. In subsequent training stages, prototype-assigned pseudo labels are highly inaccurate, resulting in unlabeled data negatively impacting the model and thus reducing the quality of the learned class prototypes. Furthermore, all prototype embeddings are soft updated according to the proportion of each emotion in ambiguous soft labels, which leads to an increase in the similarity of the emotion prototypes. Ultimately, we observe that the class prototypes stabilize at a relatively steady state with a similarity greater than 0, affirming the inherent ambiguity in emotion expression.

3.6. Pseudo Label Analysis

Figure 3 shows agreement rate of model-generated pseudo labels and prototype-assigned pseudo labels with ground-truth labels. We observe that the upward trend in the agreement rate between prototype-assigned and ground-truth labels is significantly greater than the agreement rate between model-generated

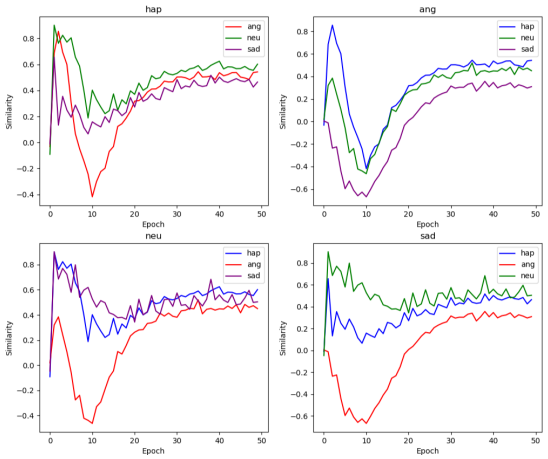

Figure 2: *Similarity between class prototype embeddings during training.*

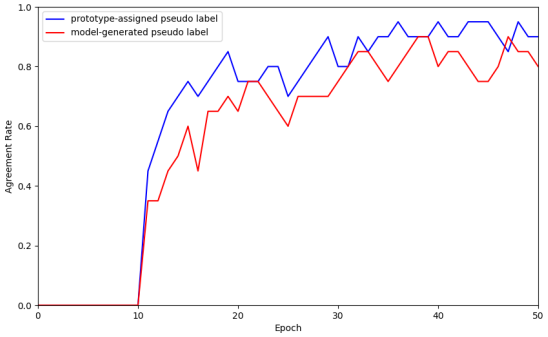

Figure 3: *Blue (red) represents the agreement rate between model-generated (prototype-assigned) pseudo labels and ground-truth labels.*

soft labels and ground-truth labels. Additionally, the prototype-assigned pseudo labels exhibit a higher agreement rate with the ground-truth labels. On the one hand, IPR could achieve ambiguous soft label annotation on a large amount of unlabeled data by training on a small amount of precise labeled data. On the other hand, it could assign more precise single labels while maintaining the desired recognition performance.

4. Conclusions

In this paper, we propose a novel iterative prototype refinement framework leveraging class prototypes and contrastive learning for ambiguous SER. IPR capitalizes on the clustering effect induced by contrastive learning to generate optimal representations. These representations facilitate the refinement and updating of class prototype embeddings, thereby enhancing the precision of assigned pseudo labels. Consequently, improved pseudo labels contribute to the augmentation of representation quality, establishing a positive feedback loop. Through a comprehensive array of comparative experiments, ablation studies, and visualization analyses conducted on the IEMOCAP benchmark dataset, we substantiate the effectiveness of IPR.

5. Acknowledgments

This work was supported in part by NSF China (Grant No. 62271270).

6. References

[1] T. Hu, A. Xu, Z. Liu, Q. You, Y. Guo, V. Sinha, J. Luo, and R. Akkiraju, "Touch your heart: A tone-aware chatbot for customer care on social media," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–12.

[2] H. Arsikere, E. Shriberg, and U. Ozertem, "Computationally-efficient endpointing features for natural spoken interaction with personal-assistant systems," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3241–3245.

[3] S. Haq, P. J. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition." in *AVSP*, vol. 2009, 2009, pp. 53–58.

[4] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[5] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[6] H. Sun, S. Zhao, X. Wang, W. Zeng, Y. Chen, and Y. Qin, "Fine-grained disentangled representation learning for multimodal emotion recognition," *arXiv preprint arXiv:2312.13567*, 2023.

[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. Liu, H. Sun, W. Guan, Y. Xia, Y. Li, M. Unoki, and Z. Zhao, "A discriminative feature representation method based on cascaded attention network with adversarial strategy for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1063–1074, 2023.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech in ninth european conference on speech communication and technology," 2005.

[10] J. Zhang and H. Jia, "Design of speech corpus for mandarin text to speech," in *The blizzard challenge 2008 workshop*, 2008.

[11] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 2022.

[12] J. Tao, Y. Li, and S. Pan, "A multiple perception model on emotional speech," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–6.

[13] X. Wang, S. Zhao, and Y. Qin, "Supervised Contrastive Learning with Nearest Neighbor Search for Speech Emotion Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 1913–1917.

[14] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multi-task learning," *Interspeech 2018*, 2018.

[15] A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, Y. Ijima, and Y. Aono, "Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4964–4968.

[16] H. Fei, D. Ji, Y. Zhang, and Y. Ren, "Topic-enhanced capsule network for multi-label emotion classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1839–1848, 2020.

[17] Y. Zhou, X. Liang, Y. Gu, Y. Yin, and L. Yao, "Multi-classifier interactive learning for ambiguous speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 695–705, 2022.

[18] J. Zhou, S. Zhao, N. Jiang, G. Zhao, and Y. Qin, "Madi: Inter-domain matching and intra-domain discrimination for cross-domain speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[19] 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.

[20] S. Dang, T. Matsumoto, Y. Takeuchi, H. Kudo, T. Tsuboi, Y. Tanaka, and M. Katsuno, "Using self-learning representations for objective assessment of patient voice in dysphonia," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 359–363.

[21] Y. Gao, H. Shi, C. Chu, and T. Kawahara, "Enhancing two-stage finetuning for speech emotion recognition using adapters," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 316–11 320.

[22] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 478–484.

[23] P. Li, Y. Song, I. V. McLoughlin, W. Guo, and L.-R. Dai, "An attention pooling based representation learning method for speech emotion recognition," 2018.

[24] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Speech emotion recognition based on multi-label emotion existence model." in *INTERSPEECH*, 2019, pp. 2818–2822.

[25] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7174–7178.

[26] M. Ghifary, W. B. Kleijn, and M. Zhang, "Domain adaptive neural networks for object recognition," in *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings 13*. Springer, 2014, pp. 898–904.

[27] C. Yu, J. Wang, Y. Chen, and M. Huang, "Transfer learning with dynamic adversarial adaptation network," in *2019 IEEE international conference on data mining (ICDM)*. IEEE, 2019, pp. 778–786.

[28] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[29] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3941–3950.

[30] Y. Liu, H. Sun, W. Guan, Y. Xia, and Z. Zhao, "Discriminative Feature Representation Based on Cascaded Attention Network with Adversarial Joint Loss for Speech Emotion Recognition," in *Proc. Interspeech 2022*, 2022, pp. 4750–4754.