Interspeech 2024
1-5 September 2024, Kos, Greece

An Effective Local Prototypical Mapping Network for Speech Emotion Recognition

Yuxuan Xi[1], *Yan Song*[1], *Lirong Dai*[1], *Haoyu Song*[2], *Ian McLoughlin*[1,3]

[1]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China.
[2]The Australian National University. [3]ICT Cluster, Singapore Institute of Technology, Singapore.
xyxah96@mail.ustc.edu.cn, {songy, ivm, lrdai}@ustc.edu.cn, u7439298@anu.edu.au

Abstract

Speech emotion recognition (SER) systems are generally optimized through utterance-level supervision, but emotion is complex and often varies within an utterance. This paper propose a local prototypical mapping network (LPMN) to model frame-level emotional variance and better exploit within-frame dynamics to improve performance. Specifically, a codebook of prototypes is first constructed to characterize complex frame-level features output from a pre-trained backbone network. An utterance-level embedding is obtained by selecting the most emotion-related mappings via a similarity measure between features and prototypes, motivated by multiple instance learning algorithms. Prototypes can be jointly optimized with quantization loss and CE loss. A prototype selection scheme is further proposed to select emotion-aware prototypes to reduce bias caused by irrelevant factors. Evaluations on IEMOCAP and MER2023 benchmarks demonstrate the effectiveness of LPMN.

Index Terms: speech emotion recognition, convolutional neural network, multiple-instance learning

1. Introduction

Emotion is an important internal state experienced by humans, and is often reflected through speech content or its spoken characteristics. Speech emotion recognition (SER) analyses utterances to automatically determine which emotion(s) they convey. This has importance in emerging applications such as voice call handling in automated call centers, mental health diagnoses, remote education, detection of mistruth from voice, and so on.

Motivated by the recent success of deep learning in related domains, SER methods based on Deep Neural Networks (DNN), Convolutional Neural Networks (CNN) [1, 2], and Recurrent Neural Networks (RNN) [3] have been proposed. They generally follow traditional machine learning pipelines: extract frame-level features via a deep neural network architecture followed by a pooling layer to obtain utterance-level embeddings that are classified to obtain human emotion labels.

Existing methods [4, 5, 6, 7, 8] mainly optimize network parameters under utterance-level supervision because datasets generally provide utterance labels. Datasets are also extremely limited in size. This, coupled with the complexity of human emotion makes the task challenging. Furthermore, SER can be considered as a detection task in essence, where the occurrence or preponderance of certain emotion types within an utterance may decide its overall classification result. For example, if most of a speech utterance is "neutral" while only a small segment shows an obvious "happy" emotion, it may be categorized overall as "happy". Thus, aggregation methods like average- and max- pooling may not optimally capture the emotional "significance" of small segments.

To address the significance issue, it is easy to see why attention mechanisms have been applied in previous studies, over time and/or frequency dimensions [4, 5, 6]. These usually apply self-attention to calculate the weights of frame-level features before pooling. Alternatively, multiple instances learning (MIL) based methods were recently proposed [7, 8] to capture the most emotion relevant segments. They view utterances as 'bags' containing the segments as 'instances'. Following the MIL algorithms, the emotion class of an utterance is determined by aggregating the decisions across the segments. However, these methods mainly rely on emotion class over a whole utterance, which may be sub-optimal without taking advantage of local linguistic and para-linguistic information.

In order to overcome those disadvantages, we propose a novel local prototypical mapping network (LPMN) to model the local emotional variance. This is illustrated in Fig.1. In this proposed SER system, a well pre-trained HuBERT [9] model is used to extract frame-level features of emotional speech. A codebook with K prototypes is initially constructed to characterize the distribution of frame-level features, spanning the linguistic and para-linguistic characteristics such as content, speaker, and speaking style. A vector quantization-variational autoencoder (VQ-VAE) [10] like objective is adopted as quantization loss for online prototype learning. Motivated by the MIL algorithm, each frame-level feature is then mapped to the prototypes by measuring the instance-to-prototypes similarities to characterize local emotion variance. Finally, an utterance embedding is obtained by selecting the most emotion-related local prototypical mappings, in time and frequency, for determining emotion class. Furthermore, cross entropy (CE) can be incorporated with quantization loss to learn discriminative prototypes. Note that SER tasks have a distribution mismatch caused by different speakers and speaking styles. A prototype selection scheme based on importance measure with respect to the Fisher information metric (FIM) estimate [11] of learned prototypes is additionally proposed to address this issue. The contributions of our proposed methods can be summarized as follows.

(1) Unlike existing MIL-based methods [7, 8], LPMN exploits prototypes to model the local feature distribution, which enables large-scale unlabeled corpus use in addition to smaller emotion datasets in a "pretraining + finetuning" paradigm.

(2) Following the MIL based aggregator, local prototypical mapping takes the prototypes as target classes, instead of utterance-level emotion types, effectively capturing complex local emotion variance to improve the performance of utterance-level classification.

(3) A prototype selection scheme is developed for the SER task, that addresses the emotion bias issue caused by irrelevant factors in an efficient way.

Experimental results on IEMOCAP and MER2023 bench-

10.21437/Interspeech.2024-1374

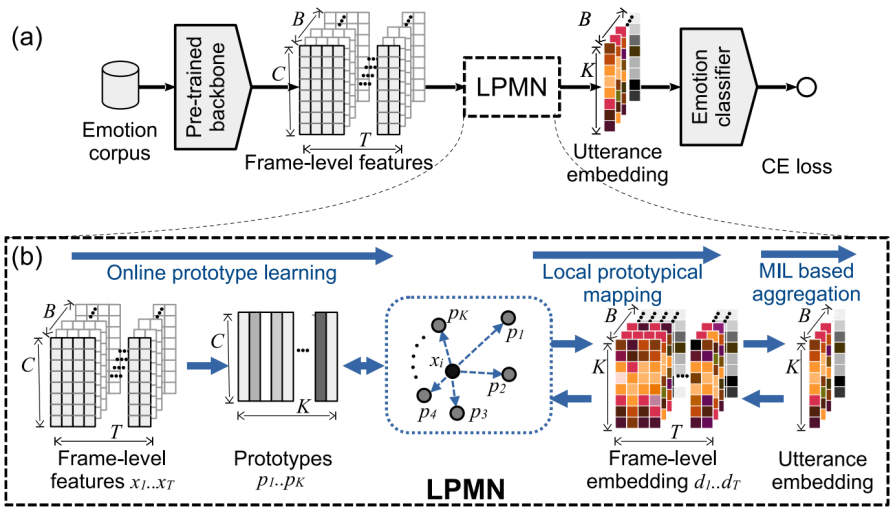

Figure 1: *Illustration of the proposed SER system utilizing a local prototypical mapping network (LPMN): (a) the overall SER pipeline, (b) Detail of LPMN's 3 stages: (i) online prototype learning, (ii) local prototypical mapping, and (iii) MIL-based aggregation.*

marks demonstrate the superiority of the proposed LPMN based SER method. For IEMOCAP [12], a performance of 75.71%, 77.42% and 75.82% in terms of F1-score, weighted accuracy (WA) and unweighted accuracy (UA) respectively can be obtained. For MER2023 [13], a performance of 70.51% and 0.9953 in terms of F1-score and valance mean-square error (MSE) respectively can be achieved.

2. Overview of the proposed framework

In this section, we will briefly describe the pipeline of the proposed SER system, as shown in Fig. 1 (a). To address the limited amount of labeled data, we exploit a pretrained network,*i.e.*, HuBERT [9], for extraction of frame-level features. This follows a general "pretraining + finetuning" paradigm, where a pre-trained backbone can extract rich linguistic and para-linguistic information useful for downstream tasks in various domains [14, 15, 16, 17, 18].

As aforementioned, emotion is complex, and SER is essentially a detection task. It is difficult for existing methods to achieve consistently satisfactory performance under utterance-level supervision, which reduces potentially rich local information to a single label from a predefined class set. To capture complex emotional variance within an utterance, a novel LPMN is designed and inserted before classifier to model rich para-linguistic and linguistic information in local feature space, as shown in Fig. 1(b). The three stages in LPMN, namely (i) online prototype learning, (ii) local prototypical mapping, and (iii) MIL-based aggregation. In addition, we proposed a prototype selection method in the post-processing phase. These stages perform as follows:

Online prototype learning. A codebook of K prototypes $\{p_j, j = 1...K\}$ is constructed, to act as the universal model of frame-level feature distribution. Since the prototypes can be learned in an unsupervised way, we can utilize large-scale unlabeled speech datasets. A VQ-VAE [10] like objective is adopted as quantization loss to jointly optimize the prototypes and local features, while utterance-level CE loss can be added as described below. Section 4 presents ablation experiments on different finetuning schemes and Ks, the number of prototypes.

Local prototypical mapping and **MIL-based aggregation.** Following the MIL algorithm, we treat the utterance as a 'bag' containing a set of instances (*i.e.*, extracted frame-level features) $\{x_i, i = 1...T\}$. The main difference from existing MIL based SER methods [7, 8] is that the *prototypes*, instead of *emotion types*, are used as target concepts. Prototypical mapping is calculated as the probability of instance x_i being recognized as belonging to the $j-th$ prototype p_j. Similar to [7, 8], MIL-aggregation selects the most emotion-relevant mappings with a max-pooling along time axis, followed by the emotion classifier. In implementation, for the SER task, the number of prototypes K can be fixed (e.g. $K = 1024$ or 512). Prototypes can be optimized with both quantization loss in frame-level space and cross-entropy (CE) loss under utterance-level supervision.

Prototype selection. Since prototypes are designed to model the local feature distribution, most of them may be less relevant to the downstream SER task. Furthermore, for specific sections, there exists an emotion bias issue caused by irrelevant factors such as different voices, speaking styles, and themes. To address this, the importance measure for prototypes according to FIM [11] is estimated via exponential moving average (EMA) during the previous training procedure. The most relevant top-K' prototypes are considered to be emotion-aware for specific section and speaker. The emotion classifier is fine-tuned after prototype selection.

3. Methods

3.1. Online prototype learning

Let $\mathbf{X} = \{x_i \in \mathbb{R}^C, i = 1...N\}$ denote a set of C-dimensional frame-level features output from the backbone network architecture, a codebook of K prototypes $\mathbf{P} = \{p_j \in \mathbb{R}^C, j = 1...K\}$ is constructed. To jointly optimize the local features $\mathbf{X}$ and prototypes $\mathbf{P}$, a quantization loss following VQ-VAE [10] is used as follows.

$$L_{OPL} = ||sg(X) - P||_2 + ||X - sg(P)||_2 \quad (1)$$

where $sg(.)$ denotes the stop-gradient operator. The eqn.(1) loss aims to minimize quantization error between local features and prototypes, and can be performed iteratively in an online fashion.

3.2. Local prototypical mapping

Given an utterance containing a sequence of T frame-level features, denoted by $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^C, i = 1...T\}$. The codebook of K prototypes $\mathbf{P}$ are used as the target concepts, following the diversity density (DD) [19] algorithm. Local prototypical mapping is defined as,

$$\begin{aligned} cosine(x_i, p_j) &= \frac{x_i \cdot p_j}{||x_i||\,||p_j||} \\ s(x_i, p_j) &= \frac{\exp(cosine(x_i, p_j)/\tau)}{\sum_{l=1}^{K} \exp(cosine(x_i, p_j)/\tau)} \\ d_i &= [s(x_i, p_1), s(x_i, p_2), ..., s(x_i, p_K)] \end{aligned} \tag{2}$$

where $cosine(x_i, p_j)$ is the cosine similarity measure between frame-level feature x_i and jth prototype p_j. The softmax of $cosine(x_i, p_j)$ is $s(x_i, p_j)$, which can be considered as the probability of x_i belonging to the target concept j, τ is a temperature hyper-parameter, set by default to $\tau = 1.0$. The aim of more effectively describing complex frame-level emotion variance in feature space compares to existing MIL based methods with utterance-level supervision [7, 8].

We further perform MIL based aggregation for SER, with the assumption that a bag is positive if at least one of its instances is positive; otherwise, the bag is negative, since that SER task is essentially a detection task, Specifically, given the mappings $\{\mathbf{D} = d_i, i = 1...T\}$, MIL-based aggregation can be implemented via max-pooling along the T dimension, *i.e.*, $\mathbf{Y} = maxpool([d_1, d_2, ..., d_T])$. $\mathbf{Y} \in \mathbf{R}^K$ is then used for SER emotion classification, resulting in a classification loss L_{CE}. As previously noted, L_{CE} can be combined with the L_{OPL} in eqn. 1 to learn the codebook of prototypes $\mathbf{P}$.

3.3. Prototype selection

For the SER task, the learned prototypes can be further selected according to their relevance measures. Given input features $\mathbf{X} = \{x_i, i = 1...n\}$, let θ denotes the network parameters. A family of probability distribution functions, each pertaining to a different realization of the network, can be denoted as $D = p(y \mid x; \theta)$, where y denotes output of the network. The total number of parameters in the network is set to n. A $n \times n$ symmetric positive semi-definite FIM F can be evaluated at on θ, where each entry $F_{i,j}$ can be written as

$$F_{i,j}(\theta) = E[(\frac{\partial \log p(y \mid x; \theta)}{\partial \theta_i})(\frac{\partial \log p(y \mid x; \theta)}{\partial \theta_j})^\top] \tag{3}$$

where F provides an estimate of how much information a random variable carries about a parameter of the distribution. In the context of DNN, F is a natural importance metric of network parameters θ. However, estimating a $n \times n$ matrix is prohibitive for networks with large number of parameters. In [11], the authors proposed to reconstruct just the diagonal entries of F by iteratively perturbing each parameter in the network and generating a new data set from the network output for each perturbation. We focus on importance measuring of the learned prototypes $\mathbf{P}$. That is, the diagonal elements of F can be calculated as the importance measure for each prototype p_j, which is as follows,

$$F(p_j) = \frac{1}{B * T} \sum_{i=1}^{B*T} (\frac{\partial \log p(y_i \mid x_i; p_j)}{\partial p_j})^2 \tag{4}$$

where B, T denote the batch size and number of local features in each utterance, and y_i is the emotion label. In each iteration, we use the emotion labels to synchronously calculate the $F(p_j)$ of each prototype and CE loss. The estimate of F is iteratively accumulated via EMA during training procedure.

In implementation, we perform prototype selection during the post-processing phase. With the estimated importance measure F of each prototype according to the specific SER task, the top-K' prototypes can be selected for local prototypical mapping and MIL-based aggregation. A new emotion classifier with random initialized parameters is utilized via a linear probing scheme [20]. In the future, we can further apply the FIM for network pruning and try fine-tuning schemes.

4. Experiments

4.1. Datasets and system description

We conduct experiments on Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [12] and Multi-modal Emotion Recognition Challenge 2023 (MER2023) [13] benchmarks to evaluate the effectiveness of the LPMN based SER system.

IEMOCAP dataset contains approximately 12 hours of audio-visual data recorded by 10 skilled actors. The database is divided into 5 sections, each containing one male and one female actor. According to the recording scenarios, it can be further subdivided into either improvised or scripted speech sections. Each utterance is annotated by multiple annotators with 8 emotion labels. For fair comparison with previous works, 4 emotion types (namely *neutral*, *happy*, *angry* and *sad*) are chosen in experiments.

Performance evaluation of IEMOCAP. A 5-fold cross-validation is performed using a leave-one-out strategy, where utterances from 4 sections are used for training, and the remain section is used for validation and test. The metrics WA, UA, and F1-score are used for performance evaluation.

MER2023 dataset is an extended version of CHEAVD [21]. In MER2023, 3373 pieces of video clip are provided, which is nearly 4 hours duration in total. Each video clip is labeled by at least three annotators using six categories (namely *neutral*, *anger*, *happiness*, *sadness*, *worry*, and *surprise* . In addition, each video clip is annotated with a 1-dimensional value *Valence* to evaluate the positive and negative estimate of emotion. In experiments, the speech data extracted from the video clips is used as an input for SER.

Performance evaluation of MER2023. Following the baseline settings in [13], we randomly divided the labeled train and validation datasets into five parts, and conducted a 5-fold cross-validation for performance evaluation. The F1-score and MSE metrics are used to evaluate the performance of discrete emotion classification and Valance respectively.

System description. As shown in Fig. 1, a pretrained HuBERT-large model [9] is exploited to extract the frame-level features. The HuBERT network is mainly divided into two parts: waveform encoder and BERT encoder. The waveform encoder is composed of seven 512-channel layers with strides [5,2,2,2,2,2,2] and kernel widths [10,3,3,3,3,2,2]. It generates a feature sequence at a 20ms frame rate for audio sampled at 16kHz. The BERT encoder consists of 24 transformer blocks. It generates a 1024-dimensional frame-level feature sequence.

We obtain IEMOCAP performance using the baseline system (with an average pooling layer) and evaluate the performance of fine-tuning different HuBERT networks with various numbers of pretrained blocks, for SER. It has been shown that the output of HuBERT with 19 blocks can have better performance than other arrangements. In following experiments, we

Table 1: *IEMOCAP baseline with different fine-tuning schemes for HuBERT-large and average pooling.*

Finetuning schemes	WA	UA	F1-score
Fixed	73.43%	67.71%	68.73%
19th block	**74.18**%	67.32%	68.95%
18th & 19th blocks	74.05%	**67.78**%	**69.10**%

Table 2: *Performance comparison between baseline with average pooling (AP), max pooling (MP), our proposed LPMN, and counterparts on IEMOCAP, in terms of WA(%), UA(%), F1-score (FS)(%) and MSE.*

Method	IEMOCAP			MER2023	
	WA	**UA**	**FS**	**FS**	**MSE**
Co-att [23]	71.64	72.70	-	-	-
W2v2-PT [22]	67.20	-	-	-	-
Spk-norm [24]	74.20	-	-	-	-
GLRF [25]	72.81	73.39	72.92	-	-
Baseline-AP	74.05	67.78	69.10	65.63	1.0976
Baseline-MP	73.15	68.01	68.99	61.16	1.1659
our LPMN	**77.42**	**75.82**	**75.71**	**70.51**	**0.9953**

take it as the pretrained backbone.

Since it is infeasible to fine-tune the HuBERT network due to the limited dataset for our task (the entire model size is about 317M), we further conduct ablation experiments on different fine-tuning schemes, where the transformer blocks to be fine-tuned are much smaller (*i.e.*, about 1%) compared to the remaining components. Table 1 presents finetuning results on the last two transformers (*i.e.*,18th,19th) blocks, showing a slight performance improvement. This fine-tuning scheme is thus adopted for the remaining experiments.

All SER systems are implemented in a PyTorch deep learning framework[1]. The Adam optimiser is used with a mini-batch size of 128. For the IEMOCAP dataset, the system is trained over 50 epochs with an initial learning rate of 0.0001. For the MER2023 dataset, all parameters are the same as IEMOCAP, except 100 training epochs are used.

4.2. Main Results

Experiments on IEMOCAP and MER2023. As shown in Table 2, for IEMOCAP, the best performance is obtained, achieving 77.42%, 75.82% and 75.71% for WA, UA and F1-score respectively. This remarkably outperforms the baseline system by about 4%, 6%, 6% absolute, respectively. For MER2023, an SER performance of 70.51% and 0.9953 is achieved in F1-score and valance MSE respectively. This outperforms the baseline system without LPMN. These results show the effectiveness of LPMN's ability to introduce prototypes that help to capture complex emotion variance.

Comparison to state-of-the-art systems. Since there are many SER methods and evaluation metrics in the literature, only those adopting the same standard settings are listed in Table 2. In [22], the pre-trained wav2vec2.0 backbone (with similar transformer based network) was exploited for SER, but the performance is 67.2% in WA, much worse than our proposed system. As can be see, the frame-level emotion variance modeling we effectively exploit can allow us to achieve significant performance improvement over existing methods.

[1]https://pytorch.org/

Table 3: *Performance before prototype selection with various number K of prototypes in LPMN, in terms of WA(%), UA(%), F1-score (FS)(%) and MSE.*

Method	IEMOCAP			MER2023	
	WA	**UA**	**FS**	**FS**	**MSE**
LPMN-128	75.07	72.49	74.73	68.83	1.0346
LPMN-256	75.16	75.04	75.10	68.11	**0.9928**
LPMN-512	76.56	**75.46**	74.97	68.84	1.0068
LPMN-1024	**76.85**	74.50	74.88	**69.68**	0.9961
LPMN-2048	76.16	74.82	**75.32**	68.34	0.9956

4.3. Ablation experiments

Table 3 lists SER results *before* prototype selection, with various number for prototypes K, ranging from 128 to 2048, denoted as LPMN-K. It is shown that when LPMN-1024 achieves almost the best performance on both IEMOCAP and MER2023 datasets in the most important metrics *WA* and *FS*. In Table 4, performance is shown *after* prototype selection, as detailed in subsection 3.3. The experiment is conducted to select the prototypes from the LPMN-1024 system, denoted as LPMN-ps-K', where K' prototypes are selected according to the estimated importance measure to eliminate the bias caused by irrelevant factors. It is shown by dropping most of the learned prototypes, the performance can still be improved. This may be attributed to the emotion complexity, and limited size of IEMOCAP. The prototype selection is performed as a post-processing step, which may not be optimal in practice. It would be interesting to incorporate this criterion into the online prototype learning in future.

Table 4: *Performance of SER with different selected prototypes K', in terms of WA(%), UA(%), F1-score (FS)(%) and MSE.*

Method	IEMOCAP			MER2023	
	WA	**UA**	**FS**	**FS**	**MSE**
LPMN-1024	76.85	74.50	74.88	69.68	0.9961
LPMN-ps-512	76.12	74.55	74.94	69.67	1.0111
LPMN-ps-200	76.97	74.05	74.07	69.89	0.9984
LPMN-ps-100	77.39	75.51	75.42	70.14	**0.9953**
LPMN-ps-60	**77.42**	**75.82**	**75.71**	**70.51**	0.9968
LPMN-ps-30	69.28	68.53	67.71	56.56	1.2413

5. Conclusion

In this paper, we proposed a Local Prototypical Mapping Network (LPMN) to model meaningful within-utterance emotional variance. Specifically, a codebook of prototypes were online learned to model frame-level features output from a previously well-trained backbone network. With the help of these prototypes, we can select the most emotion-related mappings from the similarity measure between features and prototypes, where this was motivated by multiple instance learning (MIL) algorithm. In addition, the importance measure for each prototype on a SER task, was estimated and accumulated during the training procedure. This enables the prototype selection for eliminating the emotion bias, as could be caused by irrelevant factors such as speech content, voice and occasion. Extensive experiments on IEMOCAP and MER2023 clearly demonstrate the effectiveness of this proposed LPMN approach for SER.

6. References

[1] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. Interspeech 2017*, 2017, pp. 1089–1093.

[2] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Proc. Interspeech 2018*, 2018.

[3] J. Lee and I. Tashev, "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition," in *Proc. Interspeech 2015*, 2015, pp. 1537–1540.

[4] Y. Yu and Y.-J. Kim, "Attention-lstm-attention model for speech emotion recognition and analysis of iemocap database," *Electronics*, vol. 9, no. 5, p. 713, 2020.

[5] C.-S. Ahn, C. Kasun, S. Sivadas, and J. Rajapakse, "Recurrent multi-head attention fusion network for combining audio and text for speech emotion recognition," in *Proc. Interspeech 2022*, 2022, pp. 744–748.

[6] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022*, 2022, pp. 7367–7371.

[7] C. Fu, C. Liu, C. T. Ishi, and H. Ishiguro, "Maec: Multi-instance learning with an adversarial auto-encoder-based classifier for speech emotion recognition," in *ICASSP 2021*, 2021, pp. 6299–6303.

[8] S. Mao, P. Ching, and T. Lee, "Deep Learning of Segment-Level Feature Representation with Multiple Instance Learning for Utterance-Level Speech Emotion Recognition," in *Proc. Interspeech 2019*, 2019, pp. 1686–1690.

[9] 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," pp. 3451–3460, 2021.

[10] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, 2017, p. 6309–6318.

[11] M. Tu, V. Berisha, M. Woolf, J.-s. Seo, and Y. Cao, "Ranking the parameters of deep neural networks using the fisher information," in *ICASSP 2016*, 2016, pp. 2647–2651.

[12] Busso, Carlos, Bulut, Murtaza, Lee, C. Chun, Kazemzadeh, Abe, Mower, Emily, Kim, Samuel, Chang, J. N., Lee, Sungbok, Narayanan, and S. S., "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[13] Z. Lian, H. Sun, L. Sun, K. Chen, M. Xu, K. Wang, K. Xu, Y. He, Y. Li, J. Zhao, Y. Liu, B. Liu, J. Yi, M. Wang, E. Cambria, G. Zhao, B. W. Schuller, and J. Tao, "Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23, 2023, p. 9610–9614.

[14] K. He, R. Girshick, and P. Dollar, "Rethinking imagenet pre-training," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4917–4926.

[15] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, "Seeing out of the box: End-to-end pre-training for vision-language representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 12 976–12 985.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, and X. Xiao, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE journal of selected topics in signal processing*, 2022.

[17] Y.-X. Xi, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu, "Frontend attributes disentanglement for speech emotion recognition," in *ICASSP 2022*. IEEE, 2022, pp. 7712–7716.

[18] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7169–7173.

[19] O. Maron and T. LozanoPérez, "A framework for multiple-instance learning," in *MIT press*, 1998, pp. 570–576.

[20] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9620–9629.

[21] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "Cheavd: a chinese natural emotional audio–visual database," *Journal of Ambient Intelligence and Humanized Computing*, no. 4, pp. 1–12, 2016.

[22] P. Leonardo, R. Pablo, and F. Luciana, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[23] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7367–7371.

[24] 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)*, 2022, pp. 7342–7346.

[25] D. Chaoyue, L. Jiakui, Z. Daoming, L. Baoxiang, Z. Tianhao, and Z. Qunyan, "Stable Speech Emotion Recognition with Head-k-Pooling Loss," in *Proc. Interspeech 2023*, 2023, pp. 661–665.