INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Emotion Prompting for Speech Emotion Recognition

Xingfa Zhou[1], *Min Li*[2], *Lan Yang*[1], *Rui Sun*[3], *Xin Wang*[4], *Huayi Zhan*[1,*]

[1]Sichuan Changhong Electronics Holding Group Co. Ltd., China
[2]Xinjiang University of Finance & Economics, China
[3]Leshan Normal University, China
[4]Southwest Petroleum University, China

{xingfa.zhou, lan.yang, huayi.zhan}@changhong.com, limin_xju@xjufe.edu.cn,
ruisun@whu.edu.cn, xinwang@swpu.edu.cn

Abstract

Speech Emotion Recognition (SER) classifies speech into emotion categories such as: Happy, Angry. Most prior works for SER focused on how to mine compelling features to improve performance. However, these methods ignore the influence of emotional label information on SER. Recent studies have attempted to prompt pre-trained language models and yield good performance for NLP tasks. Nevertheless, few works have attempted to prompt pre-trained speech models (PSM) on speech tasks. In light of these, we propose a simple but effective prompt-based method that prompts PSM for SER. Firstly, we reframe SER as an entailment task. Next, we generate speech prompts and combine them with the raw audio to form the input for PSM. Finally, we build a multi-task learning framework to extract more compelling features by simultaneously performing automatic speech recognition (ASR) and SER. Experiments on the IEMOCAP benchmark show that our method outperforms state-of-the-art baselines on the SER task.

Index Terms: speech emotion recognition, prompt, entailment task, multi-task learning

1. Introduction

With the development of artificial intelligence, emotion recognition is attracting more and more researchers' interest. Emotion recognition is an integral part of human-computer interaction, which aims to identify meaningful emotional information from the face, speech, or text data. Among these data, speech is a significant carrier of information, encompassing both semantic and emotional aspects. In recent years, Speech Emotion Recognition (SER) has gained widespread adoption in various domains, including intelligent customer service, distance education, and intelligent healthcare. However, due to the severe confusion between some emotions, it is still challenging to recognize the emotions expressed in speech.

The early SER systems usually consist of two primary cascading components: feature extraction and classification. These systems utilize explicit features, such as spectral, prosodic, and speech quality, to recognize emotion [1, 2]. However, these methods require strong domain knowledge and a profound speech understanding. Furthermore, such cascading methods are prone to error propagation. In light of these, more and more SER systems have begun to try end-to-end methods with the development of deep learning. In particular, due to larger model capacity and efficient training algorithms, end-to-end deep neural models can automatically extract efficient features and often outperform those traditional systems based on well-designed features. Therefore, end-to-end deep neural models have become the preferred methods for SER [3, 4].

Most prior SER works concentrate on how to mine compelling features to improve performance, such as taking a pre-trained speech model as a feature extractor. Although these works achieve good performance, they ignore the impact of emotion prompts on SER. Recent studies have attempted prompt-based fine-tuning learning to prompt pre-trained models and yield good performance. Nevertheless, such a paradigm is little studied in the speech community. In light of this, to further improve the performance of SER, we propose a simple yet effective prompt-based model EmotionPrompt to prompt PSM with the emotional information for SER. EmotionPrompt reformulates SER into an entailment task. Specifically, we construct a speech emotion prompt and concatenate it with the raw input waveform to create the final input for PSM. Moreover, to enhance the ability to extract speech features and improve performance, EmotionPrompt simultaneously performs ASR and SER.

Experiments on the popular SER benchmark dataset IEMOCAP demonstrate that EmotionPrompt achieves state-of-the-art results. The main contributions of our work are as follows:

- We propose a speech emotion recognition method, EmotionPrompt, a prompt-based model that prompts PSM with emotional label information. It is an early attempt to prompt PSM for speech tasks.
- An end-to-end multi-task learning deep neural model is built to extract more effective features.
- Intensive experimental studies are conducted on the popular SER benchmark dataset to show the effectiveness of our proposed approach.

The rest of the paper is organized as follows: in section 2, we first present recent related work on SER, prompt learning, and the pre-trained models. Next, section 3 describes the proposed model and the training and inference processes. Empirical results and analysis are introduced in section 4. Finally, conclusions are drawn in section 5.

2. Related work

2.1. Speech Emotion Recognition

Early SER works take steps such as preprocessing, feature extraction and classification to detect emotion. Feature extraction is a vital process, which aims to produce effective feature representations for different emotions. These works are mainly based on temporal features of spectral features [1, 5, 6]. Besides, some other new features [7, 8] like low-level descriptive features, high-level prosody features of energy and pitch contours, linguistic features, and utterance level features such as

*Corresponding author

10.21437/Interspeech.2023-1385

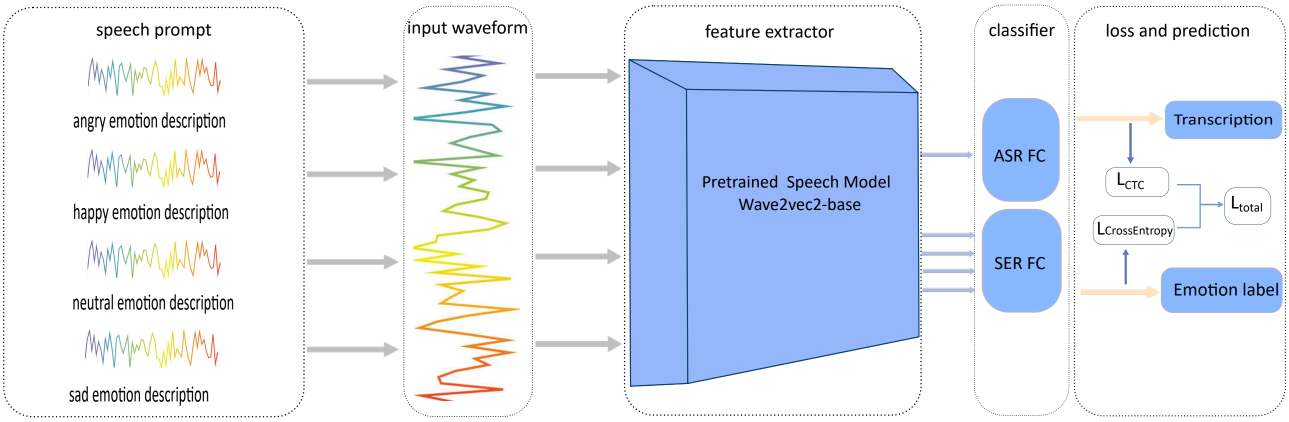

Figure 1: *Network modules and the overall architecture.*

speaking rate are also extracted to build SER systems. Various machine learning algorithms are used to classify emotions from these features, such as support vector classification algorithm (SVM), hidden Markov model (HMM), and Ada Boosted decision tree. Some researchers also apply ensemble techniques to improve performance [9, 10].

Traditional machine learning methods require strong feature extractors and are prone to perform worst. With the development of deep neural networks and efficient learning algorithms, researchers increasingly use various deep learning architectures to extract features implicitly. [11] proposes a cascaded attention network (CAN) to extract effective emotional features. Furthermore, an adversarial joint loss strategy is introduced to distinguish the emotional embeddings with high similarity by the generated hard triplets in an adversarial fashion. In [12], the authors use a Focus-Attention (FA)mechanism and a novel Calibration-Attention (CA) mechanism in combination with the multi-head self-attention for SER. In [13], an attention-based model is proposed. The work uses MFCC as the input speech representation and a variational RNN as the key ML component for SER. [3] builds an end-to-end model for SER. the study leverages the pre-trained wav2vec-2.0 for speech feature extraction, and fine-tune on SER data through two tasks: SER and speech recognition (ASR).

2.2. Prompt Learning

Prompt-based learning (PBL) is a new paradigm of fine-tuning approach inspired by GPT-3 [15]. PBL formalizes downstream tasks as language mask prediction tasks by leveraging prompts. An appropriate prompt contributes to the improved adaptation of pre-trained language models (PLM) to specific tasks. Recent studies have prompted PLMs for various NLP tasks such as text classification [16], generation [17], and sentiment analysis [18], and yield good performance. There have been extensive efforts in prompt mining to create more effective. Manually designing prompts that consist of discrete tokens is used by [19]. Since manual prompt mining is both time-consuming and hard to find the best prompts, [16, 20] propose to generate prompts automatically to avoid heavy prompt engineering.

2.3. Pre-trained Model

The pre-trained models have shown to be very effective feature extractors [21]. These models are trained in an unsupervised manner in the pretraining phase. Once pretraining is done, the model could be finetuned for specific downstream tasks using a relatively small amount of labeled training data. This finetuning paradigm based on a pre-trained model has achieved excellent performance in most NLP tasks. Due to this outstanding performance, this paradigm is moving from the natural language processing domain to the speech domain. Wav2Vec2 [22] is such a pre-trained speech model that learns speech representations by pretraining on large quantities of audio data. It tries to recover the randomly masked portion of the encoded audio feature. In this paper, we take Wav2Vec2 to extract speech features for SER.

3. Proposed Method

We propose an end-to-end multi-task learning speech prompting framework to adapt Wave2Vec2 to SER. Figure 1 illustrates the architectural framework of our proposed EmotionPrompt model. Two critical steps in the EmotionPrompt are the construction of the emotion label prompt and the task conversion of SER, which reformulates SER into an entailment task. A text-to-speech model first encodes the textural emotion label prompt into speech. And then, we prepend the input of raw waveform with the speech prompt as the input of Wave2Vec2. Next, EmotionPrompt performs entailment prediction and speech-to-text recognition simultaneously. Finlay, the entailment prediction results are transformed into emotion labels. In the following, we describe the framework in detail.

3.1. Entailment Prediction and Speech-to-text Recognition

Suppose X denotes instance space and $\mathcal{Y} = \{y_1, y_2, ..., y_C\}$ denotes label space with C possible emotion categories. EmotionPrompt learns a function $f_\theta(\cdot) : X \rightarrow Y$ from training data $D = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is the raw speech waveform, y_i is the corresponding emotion category, and N is the size of dataset.

Constructing speech prompts for SER poses a significant challenge. In this work, we construct C textual emotion label descriptions $\{p_k\}_{k=1}^{C}$ as the prompt to prompt Wave2vec2 the emotional label information. Similar to the existing work [18, 19], we hand-craft the descriptions. For instance, we can choose *it is a sad mood* to describe the sadness category. Details can be seen in Table 2. We then transfer the textual descriptions into speech by google text-to-speech model. Based on the speech emotion label descriptions, we reshape D as an entailment dataset $\hat{D} = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{N}$.

$$\hat{x}_i = Concat(p_k, x_i) \tag{1}$$

$$\hat{y}_i = \begin{cases} 1, & if \quad k == i \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where p_k is sampled from $\{p_k\}_{k=1}^{C}$, and the $Concat$ operation denotes prepending x_i with p_k.

We denote the pre-trained Wav2Vec2 model as M. Similar to the work [3], we extract Wave2Vec2's last hidden layer's output $h = M(\hat{x}_i) \in \mathbb{R}^{L\times d}$ as the input feature of Emotion-Prompt, where L is the total sampling length, and d is the hidden dimension of M. For the ASR task, a fully-connected layer g_θ is appied to map h to character logits $y_a \in \mathbb{R}^{L\times V}$, where V is the size of characters vocabulary. For the SER, we first convert the sequence vector h into a single vector z by a meaning pool operation $Mpool$. After obtaining the single feature vector z, we apply another fully-connected layer $g_{\theta'}$ that maps z to emotion logits $y_e \in \mathbb{R}^{C}$.

$$y_a = g_\theta(h) \tag{3}$$

$$y_e = g_{\theta'}(z) \tag{4}$$

$$z = Mpool(h) \tag{5}$$

3.2. Training and Inference

In the training stage, we train the SER and ASR simultaneously. That is, SER and ASR share the same feature extractor M. At the end of both SER and ASR, we apply a softmax operator on both y_a and y_e to convert them to probability vectors. For the ASR encoding, the Connectionist Temporal Classification (CTC) loss is used to train the model against the encoding of the given gold transcription and y_a. Details could be found in [23]. We obtain the ASR loss as:

$$L_a = CTC(\hat{y}_a, t) \tag{6}$$

where t is the given gold transcription, and $\hat{y}_a = softmax(y_a)$. Similarly, we calculate the cross-entropy between the predicted emotion logits and the true emotion labels to optimize the SER model's parameters.

$$L_e = CrossEntropy(\hat{y}_e, l), \tag{7}$$

where l is the true emotion label, and $\hat{y}_e = softmax(y_e)$. Finally, we define the final training loss as follows:

$$L = L_e + \alpha L_a \tag{8}$$

where α is weight to combine L_a and L_e into together.

At inference time, we drop the ASR module g_θ, and just keep the $g_{\theta'}$ module to predict emotions. Specifically, the inference process is as Algorithm 1.

Algorithm 1 The emotion recognition inference process

```
initialization empty list L
for each p_k ∈ {p_k}_{k=1}^C do
    x̂_i ← Concat(p_k, x_i)
    h ← M(x̂_i)
    z ← Mpool(h)
    y_e^i ← g_θ'(z)
    L.append(y_e^i[1])
end for
y = argmax(L)
```

4. Experiments

4.1. Dataset and Experimental Setup

Dataset. We evaluate the proposed method on the IEMOCAP benchmark dataset [27]. The dataset contains about 12 hours of speech from 10 performers. IEMOCAP contains 10,039 utterances. Each utterance is labeled with one of the following emotions: neutral, sad, happy, angry, surprised, excited, fearful, frustrated, disgusted, and others. Following much of the works on SER [11, 12, 13, 3], Our experiment considers four emotions: angry, happy, sad, and neutral, where the excitement class is merged into the happy class. Finally, 5,531 utterances are selected to evaluate our method.

Metrics. To compare with previous approaches under the same conditions, we perform 5-fold and 10-fold cross-validation as experimental results. We compute the final weighted accuracy(WA) as follows:

$$acc = \frac{1}{N_{total}} \sum_{k=1}^{10} N_k \tag{9}$$

where N_{total} is the number of correct emotion category predictions in the k-th fold.

Table 1: *the hyper-parameters setting for experiments.*

Parameter Name	Value
sample frequency	16k Hz
training epochs	100
optimizer	AdamW
α	0.1
learning rate	$5e^{-5}$
batch size	2
gradient accumulation steps	4

Table 2: *The emotion label descriptions used for prompt*

emotion	Description
angry	it is a angry mood
happy	it is a happy mood
neutral	it is a neutral mood
sad	it is a sad mood

Table 3: *Baseline methods*

Method	Description
Wu et al. [24]	capsule network
Sajjad et al. [25]	ResNet-101 + bi-LSTM
Lu et al. [26]	ResNet-101 + bi-LSTM
Cai et al. [3]	Wave2vec2+ASR
Liu et al. [11]	cascaded attention network
Baruah et al. [13]	Variational RNN
Kim et al. [12]	CNN+BiLSTM+attention

Method	cross-validation	acc%
Wu et al. [24]	10-fold	72.73
Sajjad et al. [25]	5-fold	72.25
Lu et al. [26]	10-fold	72.6
Cai et al. [3]	10-fold	78.15
Liu et al. [11]	5-fold	76.17
Baruah et al. [13]	5-fold	65.5
Kim et al. [12]	10-fold	72.83
Ours	5-fold	**79.52**
Ours	10-fold	**79.99**

Table 4: *Classification accuracy performance of different baseline approaches for speech emotion recognition on the IEMOCAP dataset. The bold front denotes the best performance.*

Training. Wav2Vec2-base is used as the pre-trained model to extract features for all experiments. We take the implementation and pre-trained weights from the Huggingface Transformers library [28]. Table 1 lists all the hyper-parameter settings we use in the experiments. All models are trained on one Tesla P100 GPU.

Baseline methods. The state-of-the-art models compared with our model are shown in Table 3. Note that all the results of the baselines method are directly copied from their papers.

Prompts. Recent studies have investigated the development of automatic generation methods [16] to design more effective prompts. As our work represents the first study that utilizes speech prompting for SER, we adopt a simple yet effective approach by manually defining all the descriptions. The label descriptions are provided in Table 2. After obtaining the textual emotion label descriptions, we convert them into speech using Google's text-to-speech API.

4.2. Results and Analysis

Overall Performance. We compare our method with recent SER works under a similar experimental setup. As shown in Table 4, Our method surpasses all the baseline methods on the IEMOCAP dataset under both 5-fold cross-validation and 10-fold cross-validation settings. These results prove that our proposed method can produce a promising performance for SER. [3] proposes a multi-task learning framework to simultaneously perform ASR and SER tasks with an end-to-end deep neural model based on Wav2Vec2. This strategy is similar to our work. Compared to other baselines, [3] significantly improves the state-of-the-art performance. This result proves that Wav2Vec2 and ASR can produce a promising performance for SER. Compared to [3], our proposed method shows a definite improvement of 1.37% and 1.84%, respectively. It indicates the effectiveness of speech emotion label prompts.

Ablation Study. To assess the contribution of each framework component, we conduct an ablation study. As shown in Table 5, We sequentially remove two key components: ASR and prompt. We set α to 0 to remove the ASR component. Table 5 shows that accuracy drops by 2.82% and 2.69% on 5-fold cross-validation and 10-fold cross-validation settings, respectively. The results verify that the ASR component helps to improve emotion recognition performance. The reason is that the ASR component enhances the feature extraction ability of Wave2Vec2 when our model performs ASR and SER simultaneously in the training stage. Next, we remove the prompt component. Note that the SER model then degenerates into a multi-classification task rather than an entitlement task. When the prompt is removed, we can see that the performance deteriorates, which verifies the significant impact of the emotion label prompt on our model.

Method	5-fold	10-fold
Ours	79.52	79.99
-ASR	76.7	77.3
-ASR&prompt	72.1	73.5

Table 5: *Ablation study over two main components of proposed framework.*

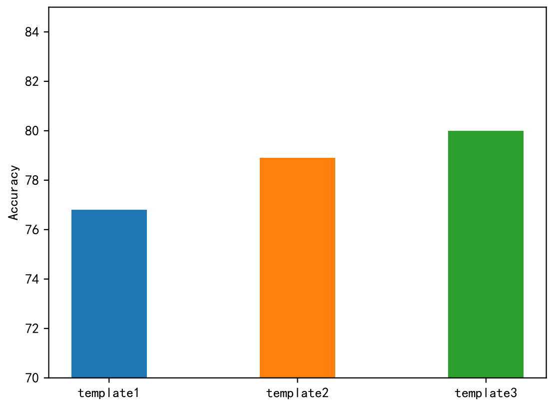

Figure 2: *Influence of prompt's logic under the 10-fold cross-validation setting.*

Influence of Prompt. [29] points out that prompt template significantly influences performance. We conduct an experiment to validate the influence. We manually construct three prompt templates where the semantic logic correlation is template-3>template-2>template-1. As shown in Figure 2, we can see that the accuracy has increased significantly with the prompt templates more logically. Since manual prompt mining is a time-consuming and challenging process to identify the best prompts, we will explore how to generate prompts automatically to avoid heavy prompt engineering in future work.

5. Conclusions

In this paper, we study speech emotion recognition. We propose a simple but effective prompt-based method that prompts the pre-trained speech model Wave2Vec2 for SER. Specifically, speech emotion recognition is reformulated into an entailment task. Next, we generate speech prompts and combine them with the raw audio to form the input for PSM. Finally, an end-to-end multi-task learning framework is built to simultaneously perform ASR and SER to extract compelling features. Comprehensive experiments are conducted on the IEMOCAP benchmark dataset. Experimental results demonstrate that the proposed method significantly outperforms strong baselines. The ablation study establishes the effectiveness of the prompt and multi-task learning framework.

6. References

[1] S. Yun and C. D. Yoo, "Speech emotion recognition via a max-margin framework incorporating a loss function based on the watson and tellegen's emotion model," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, 2009, pp. 4169–4172.

[2] Q. Mao and Y. Zhan, "A novel hierarchical speech emotion recognition method based on improved DDAGSVM," *Comput. Sci. Inf. Syst.*, vol. 7, no. 1, pp. 211–222, 2010.

[3] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, 2021, pp. 4508–4512.

[4] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, "Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1448–1460, 2022.

[5] J. Lin, C. Wu, and W. Wei, "Emotion recognition of conversational affective speech using temporal course modeling," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 1336–1340.

[6] H. Lee, T. Hu, H. Jing, Y. Chang, Y. Tsao, Y. Kao, and T. Pao, "Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 215–219.

[7] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Commun.*, vol. 52, no. 7-8, pp. 613–625, 2010.

[8] J. Deng and B. W. Schuller, "Confidence measures in speech emotion recognition based on semi-supervised learning," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, 2012, pp. 2226–2229.

[9] C. Lee, E. Mower, C. Busso, S. Lee, and S. S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 2009, pp. 320–323.

[10] C. Wu and W. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affect. Comput.*, vol. 2, no. 1, pp. 10–21, 2011.

[11] 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 *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, 2022, pp. 4750–4754.

[12] J. Kim, Y. An, and J. Kim, "Improving speech emotion recognition through focus and calibration attention mechanisms," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. ISCA, 2022, pp. 136–140.

[13] M. Baruah and B. Banerjee, "Speech emotion recognition via generation using an attention-based variational recurrent neural network," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. ISCA, 2022, pp. 4710–4714.

[14] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence LSTM architecture," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, 2020, pp. 6474–6478.

[15] T. Brown, B. Mann, N. Ryder, M. Subbiah *et al.*, "Language models are few-shot learners," in *Proceedings of NIPS*, 2020.

[16] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proceedings of ACL/IJCNLP*, 2021, pp. 3816–3830.

[17] F. Carlsson, J. Öhman, F. Liu, S. Verlinden, J. Nivre, and M. Sahlgren, "Fine-grained controllable text generation using non-residual prompting," in *Proceedings of ACL*. Association for Computational Linguistics, 2022, pp. 6837–6857.

[18] H. Wu and X. Shi, "Adversarial soft prompt tuning for cross-domain sentiment analysis," in *Proceedings of ACL*. Association for Computational Linguistics, 2022, pp. 2438–2447.

[19] K. Qi, H. Wan, J. Du, and H. Chen, "Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates," in *Proceedings of ACL*, 2022, pp. 1910–1923.

[20] J. Li, T. Tang, J. Nie, J. Wen, and X. Zhao, "Learning to transfer prompts for text generation," in *Proceedings of NAACL*. Association for Computational Linguistics, 2022, pp. 3506–3518.

[21] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[23] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, ser. ACM International Conference Proceeding Series, vol. 148. ACM, 2006, pp. 369–376.

[24] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu, and H. Meng, "Speech emotion recognition using capsule networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2019, pp. 6695–6699.

[25] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep bilstm," *IEEE Access*, vol. 8, pp. 79 861–79 875, 2020.

[26] Z. Lu, L. Cao, Y. Zhang, C. Chiu, and J. Fan, "Speech sentiment analysis via pre-trained features from end-to-end ASR models," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 7149–7153.

[27] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008. [Online]. Available: https://doi.org/10.1007/s10579-008-9076-6

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Oct. 2020, pp. 38–45.

[29] D. Tam, R. R. Menon, M. Bansal, S. Srivastava, and C. Raffel, "Improving and simplifying pattern exploiting training," in *Proceedings of EMNLP*. Association for Computational Linguistics, 2021, pp. 4980–4991.