Emotion Prompting for Speech Emotion Recognition

Xingfa Zhou1, Min Li2, Lan Yang1, Rui Sun3, Xin Wang4, Huayi Zhan1,∗

1Sichuan Changhong Electronics Holding Group Co. Ltd., China
2Xinjiang University of Finance & Economics, China
3Leshan Normal University, China
4Southwest 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 re-formulates 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
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 Wav2Vec2 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 textual 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 $\mathcal{X}$ denotes instance space and $\mathcal{Y} = \{y_1, y_2, ..., y_C\}$ denotes label space with $C$ possible emotion categories. EmotionPrompt learns a function $f_k(\cdot): \mathcal{X} \rightarrow \mathcal{Y}$ 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 Wav2vec2 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

![Figure 1: Network modules and the overall architecture.](image-url)
entailment dataset $\hat{D} = \{ (\hat{x}_i, \hat{y}_i) \}_{i=1}^N$.

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

$$\hat{y}_i = \begin{cases} 
1, & \text{if } k == i \\
0, & \text{otherwise} 
\end{cases}$$ (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 Wav2Vec2’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 $g_0$ is applied 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 $M_{pool}$. After obtaining the single feature vector $z$, we apply another fully-connected layer $g_{d'}$ that maps $z$ to emotion logits $y_e \in \mathbb{R}^{C}$.

$$y_a = g_0(h)$$ (3)

$$y_e = g_{d'}(z)$$ (4)

$$z = M_{pool}(h)$$ (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:

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

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

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

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

$$L = L_e + \alpha L_a$$ (8)

where $\alpha$ is weight to combine $L_a$ and $L_e$ into together.

At inference time, we drop the ASR module $g_0$, and just keep the $g_{d'}$ module to predict emotions. Specifically, the inference process is as Algorithm 1.

### 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:

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

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

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

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample frequency</td>
<td>16k Hz</td>
</tr>
<tr>
<td>training epochs</td>
<td>100</td>
</tr>
<tr>
<td>optimizer</td>
<td>AdamW</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.1</td>
</tr>
<tr>
<td>learning rate</td>
<td>$5 \times 10^{-5}$</td>
</tr>
<tr>
<td>batch size</td>
<td>2</td>
</tr>
<tr>
<td>gradient accumulation steps</td>
<td>4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>emotion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>it is a angry mood</td>
</tr>
<tr>
<td>happy</td>
<td>it is a happy mood</td>
</tr>
<tr>
<td>neutral</td>
<td>it is a neutral mood</td>
</tr>
<tr>
<td>sad</td>
<td>it is a sad mood</td>
</tr>
</tbody>
</table>

#### Table 3: Baseline methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wu et al. [24]</td>
<td>capsule network</td>
</tr>
<tr>
<td>Sajjad et al. [25]</td>
<td>ResNet-101 + bi-LSTM</td>
</tr>
<tr>
<td>Lu et al. [26]</td>
<td>ResNet-101 + bi-LSTM</td>
</tr>
<tr>
<td>Cai et al. [3]</td>
<td>Wave2vec2+ASR</td>
</tr>
<tr>
<td>Liu et al. [11]</td>
<td>cascaded attention network</td>
</tr>
<tr>
<td>Baruah et al. [13]</td>
<td>Variational RNN</td>
</tr>
<tr>
<td>Kim et al. [12]</td>
<td>CNN+BiLSTM+attention</td>
</tr>
</tbody>
</table>
In this paper, we study speech emotion recognition. We propose a novel approach that prompts the pre-trained speech model Wave2Vec2 for SER. Specifically, speech emotion recognition is reformulated into an entailment 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. 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 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 study, we propose a 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


