*mathematics*

MDPI

*Article*

# Multimodal Prompt Learning in Emotion Recognition Using Context and Audio Information

**Eunseo Jeong †, Gyunyeop Kim † and Sangwoo Kang ***

School of Computing, Gachon University, Seongnam-si 13120, Republic of Korea; msjung153@gachon.ac.kr (E.J.); gyop0817@gachon.ac.kr (G.K.)
* Correspondence: swkang@gachon.ac.kr
† These authors contributed equally to this work.

**Abstract:** Prompt learning has improved the performance of language models by reducing the gap in language model training methods of pre-training and downstream tasks. However, extending prompt learning in language models pre-trained with unimodal data to multimodal sources is difficult as it requires additional deep-learning layers that cannot be attached. In the natural-language emotion-recognition task, improved emotional classification can be expected when using audio and text to train a model rather than only natural-language text. Audio information, such as voice pitch, tone, and intonation, can give more information that is unavailable in text to predict emotions more effectively. Thus, using both audio and text can enable better emotion prediction in speech emotion-recognition models compared to semantic information alone. In this paper, in contrast to existing studies that use multimodal data with an additional layer, we propose a method for improving the performance of speech emotion recognition using multimodal prompt learning with text-based pre-trained models. The proposed method is using text and audio information in prompt learning by employing a language model pre-trained on natural-language text. In addition, we propose a method to improve the emotion-recognition performance of the current utterance using the emotion and contextual information of the previous utterances for prompt learning in speech emotion-recognition tasks. The performance of the proposed method was evaluated using the English multimodal dataset MELD and the Korean multimodal dataset KEMDy20. Experiments using both the proposed methods obtained an accuracy of 87.49%, $F_1$ score of 44.16, and weighted $F_1$ score of 86.28.

**Keywords:** multimodal; prompt learning; speech emotion recognition; audio processing; natural language processing

**MSC:** 68T50



**Citation:** Jeong, E.; Kim, G.; Kang, S. Multimodal Prompt Learning in Emotion Recognition Using Context and Audio Information. *Mathematics* **2023**, *11*, 2908. https://doi.org/10.3390/math11132908

Academic Editors: Junaid Baber, Ali Shariq Imran, Sher Doudpota and Maheen Bakhtyar

Received: 29 May 2023
Revised: 16 June 2023
Accepted: 26 June 2023
Published: 28 June 2023



## 1. Introduction

Several studies on multimodal machine learning have proposed deep learning to train models using multimodal data, such as audio, image, text, and video. In previous studies, tasks using unimodal data, such as image classification and dialogue systems, were mainly performed. However, the accumulation of large datasets and the development of deep learning have led to research on various multimodal tasks, such as video classification tasks that use audio and image data and speech emotion recognition in conversation data using audio and text information. Multimodal deep-learning models have been adopted instead of existing unimodal models to solve multimodal tasks. Multimodal deep-learning models that collect information from multiple modalities, such as visual, auditory, and sensor data, can provide enhanced predictive performance because they are trained with more diverse information than unimodal models are trained with. For example, in natural language emotion-recognition tasks, models are likely to perform better when using audio data together with conversation text data than when using natural-language text for model training. Emotion-recognition models have often been trained using natural-language

semantic information. However, audio data such as pitch, tone, and intonation can help to predict emotions more effectively. Speech emotion recognition in conversation data research [1] also demonstrated performance improvement in multimodal approaches compared to prediction using natural-language text as unimodal data.

Since the advent of the GPT-2 [2] model in the field of natural-language processing, the size of the language model has gradually increased, making it difficult to fine-tune due to problems such as the computational resources and time costs being increased by a vast number of parameters and massive datasets. Since then, in contrast to the general fine-tuning method that updates all the parameters of a model, methodologies that perform downstream tasks without updating the model parameters have been actively proposed [3,4]. In addition, a methodology for prompt learning that trains in a process similar to the pre-training method and uses natural-language templates and generation-based language models has been proposed. Prompt learning is a methodology for performing downstream tasks with only pre-training and not fine-tuning by setting questions or requests for downstream tasks as a template for natural-language text. Prompt learning uses a pre-trained language model trained with a large amount of data. In general, however, many pre-trained language models train models using unimodal data, which makes the use of data modalities not used in pre-training in downstream tasks challenging.

The proposed method solves speech emotion recognition in conversation data, which is an audio-text multimodal task. The proposed method was trained with prompt learning to solve speech emotion-recognition tasks using data on conversations. In this study, prompt learning was performed with a language model that is pre-trained on natural-language text. However, as natural-language models are pre-trained with unimodal data, including additional modalities that are not used in the pre-training as an input for a language model is difficult. Several studies [5,6] have attached additional classification layers suitable for each modality to language models to train them with different types of data. However, in prompt learning, adding an additional classification layer to a model is challenging because the pre-training and fine-tuning processes are performed similarly. In other words, prompt learning using a unimodal pre-trained language model is problematic, and processing multimodal tasks such as speech emotion recognition is difficult.

In this study, we propose a method that can solve the above-mentioned difficulty using both text and audio data in prompt learning of a text-based pre-trained language model for speech emotion recognition from conversation data. The proposed method adds audio information to prompt learning using a text-based pre-trained language model with a self-attention mechanism and uses text and audio information together to obtain additional information. In addition, continuous emotions in previous utterances can affect the current emotion in speech emotion-recognition tasks. Thus, we propose a method to deduce emotion and contextual information from previous utterances to be used as important information for training a model to perform emotion classification for input utterances by adjusting the loss function of a language model to improve its performance.

Our contributions can be summarized as follows:

- The proposed method is trained on multimodal data in prompt learning using a text-based pre-trained language model. By adding audio information for self-attention in a text-based pre-trained language model, speech emotion recognition in conversation data can be performed with multimodal data. Prompt learning using multimodal data is expected to improve the predictive performance of models by utilizing information from more diverse modalities.
- The proposed method deduces emotions and contextual information of the previous utterances as important information for predicting emotions of the current utterance in speech emotion recognition from conversation data. Owing to the context of conversation data, the emotions and contextual information of previous utterances can assist in predicting the emotions of the current utterance.

The remainder of this paper is organized as follows: Section 2 provides an overview of multimodal deep learning and prompt learning. Section 3 describes our proposed multimodal method. Section 4 reports the experiments, settings, and results of the proposed method. Finally, Section 5 concludes the paper.

## 2. Related Works

### *2.1. Multimodal Deep Learning*

Recently, several multimodal deep-learning models have been designed to be trained concurrently with data of multiple modalities, such as vision, auditory, and sensor data. In particular, many studies [7–11] have proposed training speech-recognition models using various forms of data such as audio and text. Trained multimodal deep-learning models can train from diverse information and achieve a high prediction accuracy. For example, the sentence "Yeah, that's a lot of fun" provides only semantic information, whereas an audio recording of a woman laughing provides auditory information. However, using both text and audio data to train a model enables improved diversity of information, such as semantic and auditory information in this case. In addition, previous studies on the use of tri-modality data to perform downstream emotion-recognition tasks include DialogueTRM [12], EmoCaps [13], UniMSE [14], and M2FNet [15]. Tri-modality allows the model to train on more information than bi-modality, which can be expected to improve its performance. However, multimodal deep-learning models involve the problem that each modality has a different embedding vector space. Therefore, different embedding vectors must be aligned or fused for each modality. Many studies [16–19] on how to fuse the multimodal data have also been proposed. The model's predictive performance can affect how to fuse the different modalities.

Recently, multimodal language models combining text and audio data have been proposed with the development of automatic speech recognition (ASR). The multimodal deep-learning model, which uses audio to train based on the transformer model, trains by dividing the audio into units, similar to the image-training method. The audio data are divided into units and pass through a transformer model, which is trained jointly with the text data. Multimodal models that simultaneously use text and audio include VATT [20], SPCL-CL-ERC [21], MM-DFN [22], and SMIN [23]. Some studies [12–15,22,23] train the model with additional classification layers to classify the emotion for speech emotion recognition in conversation data using audio and text information.

### *2.2. Prompt Learning*

Since the advent of the GPT-2 [2] model, many studies have proposed solving downstream tasks without fine-tuning. A training methodology for fine-tuning a process similar to pre-training has also been proposed to achieve a high prediction performance with small datasets. In existing fine-tuning methods, language models are trained using additional layers to solve downstream tasks. For example, an additional classification layer that determines the probability of each class can be attached to a pre-trained model to solve the downstream classification task. A model incorporating an additional layer was trained using a dataset to solve the downstream task. However, prompt learning has been proposed as a fine-tuning method similar to models based on language generation, such as masked language modeling. This method attempts to use the information obtained in the pre-training step for fine-tuning. As a result, a better performance was achieved with a small amount of data. In addition, many methods [3] related to prompt learning have been proposed because they exhibit high performance even with rich data and can handle a wide variety of tasks.

Prompt learning is a methodology that uses a natural-language template as input to a model, which then obtains results through natural-language generation. The natural-language template describes the downstream task performed by the model. For example, the natural-language prompt template is configured as "the emotion in the next sentence is <MASK>" in the emotion-recognition task. The result is gained by generating natural-

language tokens to enter <MASK>, such as "happiness". Natural-language templates for prompt learning can be organized by humans or automatically generated by training models. There are LAMA [24] and GPT-3 [4], in which humans directly generate templates suitable for their data. The AutoPrompt [25], LM-BEF [26], and P-tuning [27] models have been proposed to generate proper natural-language templates or template embeddings for the given input data. These methodologies can achieve a high performance with a small amount of data by reducing the gap between pre-training and downstream tasks. This is possible because the pre-training and fine-tuning of the language model are performed similarly in prompt learning.

## 3. Methodology

The proposed emotion-recognition training process based on multimodal prompt learning is described in Section 3.1. In Section 3.2, we explain the application of contextual information to the training process such that continuous emotions in previous utterances can affect current emotions. The overall architecture of the proposed method is illustrated in Figure 1.

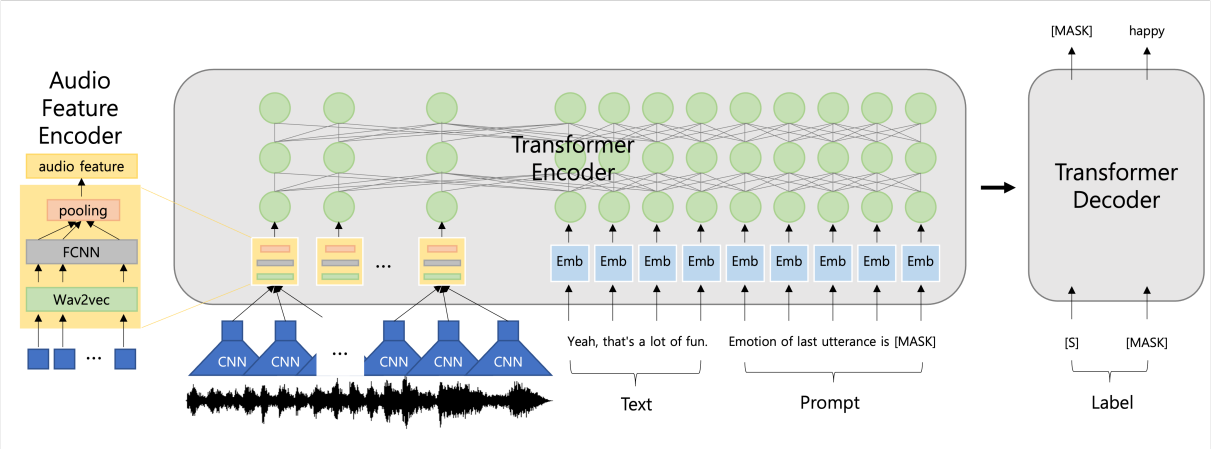


**Figure 1.** Overall architecture of proposed method.

### *3.1. Prompt with Audio Feature*

Among multimodal emotion-recognition tasks, speech emotion recognition predicts emotions using natural-language speech text and audio information. Since audio includes voice information such as pitch, tone, and intonation, a model can be trained with more information than with text only. Therefore, using text and audio together to predict emotions is expected to produce a higher performance in emotion prediction. In addition, downstream task training with prompt learning has recently demonstrated high performance. However, prompt learning in a language model pre-trained on natural-language text cannot use multiple modalities because the fine-tuning process is similar to pre-training. Prompt learning in language models and additional deep-learning layers cannot be used because the input and output of the model originate from natural-language text. Therefore, prompt learning in a language model pre-trained on natural-language text is problematic because it is difficult to use modalities other than those trained in the downstream task.

The proposed method jointly uses natural-language text and audio in the prompt-learning process of a text-based pre-trained model for speech emotion-recognition tasks. In other words, the proposed method performs self-attention using both natural-language text and audio in prompt learning using models pre-trained on text. The overall architecture of the proposed method is illustrated in Figure 1. First, audio feature vectors are extracted using the *Wav2Vec* 2.0 [28] model, a feed-forward neural network (FFCN), and the pooling layer. The natural-language speech text and prompt template are configured with the given conversation data. Subsequently, audio feature vectors, natural-language speech text, and natural-language prompts are used as inputs to models based on the transformer

encoder–decoder model [29]. The natural-language text and audio used as inputs to the model can simultaneously train text and audio features by performing *STSelfAttention* inside the transformer model.

The proposed method uses an audio feature encoder to extract audio features and inputs the features into a prompt-based language model. First, the audio features are extracted by inputting the audio data $A$ into *Wav2Vec* 2.0 and FFNN. The pooling layer is then used to extract the audio features by compressing the number of existing $n$ audio embeddings to $l$. The compression of the number of audio embeddings to $l$ is intended to change the form that can be entered into the transformer model through the audio feature encoder.

$$S' = Wav2Vec(A) \tag{1}$$

$$S'' = S'W_s + b_s \tag{2}$$

$$s_x = \frac{1}{\lfloor n/l \rfloor} \sum_{i=0+x\lfloor n/l \rfloor}^{\lfloor n/l \rfloor + x\lfloor n/l \rfloor} e_i \tag{3}$$

$$S = \{s_0, s_1, \ldots, s_{l-2}, s_{l-1}\} \tag{4}$$

Equation (1) extracts the audio embedding by inputting the given audio data $A$ into the *Wav2Vec* 2.0 model. Subsequently, the input audio embeddings are extracted from *Wav2Vec* 2.0 to FFNN, such as Equation (2), to extract $n$ audio embedding vectors $S'' = \{e_0, e_1, \ldots, e_{n-2}, e_{n-1}\}$. $e_i$ is the $i$-th audio embedding that passes the *Wav2Vec* 2.0 model and FFNN. $n$ denotes the number of extracted audio embedding vectors. $W_s$ and $b_s$ are learnable parameters. Equation (3) compresses the existing $n$ audio embedding vectors to $l$. $s_x$ is the $x$-th audio embedding. The hyper-parameter $l$ is the number of audio-embedding vectors to be compressed. $S$ in Equation (4) represents the $l$ audio feature passing through the *Wav2Vec* 2.0 model, FFNN, and pooling layer. Pooling is performed using mean pooling.

Prompt learning is a methodology in which a model specifies the tasks to be performed and returns the result in text written in natural language. In this paper, speech emotion recognition was solved using prompt learning. In the proposed method, "Emotion of the last utterance is <MASK>" is used as a prompt for the speech emotion recognition task. The transformer decoder then generates an emotion token (neutral, joyful, etc.) corresponding to the prompt <MASK>. The natural-language speech text $T_d$ and natural-language prompt $T_p$ are used as inputs for prompt learning. Subsequently, the natural-language speech text and natural-language prompts are converted into the word embedding for each token for use as input to the transformer model with *Embedding*. This process is expressed as follows:

$$D = Embedding(T_d) \tag{5}$$

$$P = Embedding(T_p) \tag{6}$$

$T_d$ in Equation (5) represents the natural-language speech text data in the speech emotion-recognition task. $D$ is a word embedding for each token of $T_d$. $T_p$ in Equation (6) is a natural-language prompt such as "Emotion of current utterance is <MASK>". $P$ denotes the word embedding for each token in the prompt. For example, "Yeah, that's a lot of fun" is a speech text $T_d$ and the number of speech text tokens is $a$. "Emotion of last utterance is <MASK>" is a prompt $T_p$ and the number of prompt tokens is $b$. The speech text and prompt are tokenized by a tokenizer. Then, the word embedding $D = \{T_d^1, T_d^2, \ldots, T_d^{a-1}, T_d^a\}$ of the utterance text and the word embedding $P = \{T_p^1, T_p^2, \ldots, T_p^{b-1}, T_d^b\}$ of the prompt are extracted.

In this paper, prompt learning was performed using transformer-based encoder–decoder models, such as T5 [3] and BART [30]. In the self-attention step of the text-based pre-trained transformer encoder, the query, key, and value are all trained with unimodal data, such as natural-language text embedding. Therefore, it is difficult to use multimodal data as the input in prompt learning of text-based pre-trained transformer-based encoder–decoder models. The proposed method is *STSelfAttention*, which uses natural-language text and audio together in the prompt learning of text-based pre-trained models. The proposed *STSelfAttention* method performs self-attention using audio, natural-language text, and natural-language prompt embeddings concurrently to generate the query, key, and value. Subsequently, self-attention is performed with the query, key, and value containing each modality and data feature. In this paper, various modalities of information were used in prompt learning using text-based pre-trained models by replacing the existing transformer encoder's self-attention with *STSelfAttention*. This process is expressed as follows:

$$Q = [S; D; P]W_q \tag{7}$$

$$K = [S; D; P]W_k \tag{8}$$

$$V = [S; D; P]W_v \tag{9}$$

$$STSelfAttention(Q, K, V) = softmax(\frac{Q^T K}{\sqrt{d_k}})V \tag{10}$$

Equations (7)–(9) generate the query, key, and value in *STSelfAttention*. $W_q$, $W_k$, and $W_v$ are learnable parameters. Equation (10) is an expression of *STSelfAttention*. In this paper, self-attention was replaced with *STSelfAttention* in the existing transformer encoder.

### *3.2. Prompt with Previous Context*

In speech emotion-recognition tasks, the emotions of the previous utterance can affect the emotion of the current utterance. Therefore, in this study, we propose a method to deduce contextual information on emotions and previous utterances as important information for predicting the emotions expressed by a given utterance. To use the contextual information of emotions and previous utterances, the previous and current utterances are used concurrently as inputs in the transformer model. It was also configured to simultaneously predict the emotions for current and previous utterances while training the model. In addition, "Emotion of current utterance is <MASK>" and "Emotion of last utterance is <MASK>" are used as prompts to predict emotions of the previous and current utterances, respectively. In this paper, $[T_d^{x-k}; T_p^{x-k}; T_d^{x-(k-1)}; T_p^{x-(k-1)}; \ldots; T_d^x; T_p^x]$ is the input to the transformer encoder to predict the $x$-th utterance. In this case, $T_d^x$ is the natural-language speech text for the $x$-th utterance, and $T_p^x$ is a natural-language prompt for predicting the emotion of the $x$-th utterance. Hyper-parameter $k$ is the number of previous utterances to be used. Thereafter, the emotional words for each <MASK> are predicted in reverse order in the transformer decoder. As a result, $[y^x; y^{x-1}; \ldots; y^{x-k}]$ is generated as the transformer decoder output. $y^x$ is an emotion token (e.g., neutral, joyful) for the $x$-th utterance.

In addition, we added emotions from previous utterances to the training loss function to induce training on emotion-change information over time. To differentiate between the learning weight and the contextual information of the previous and current utterance, predictive loss functions for each of the current and previous emotions were constructed. The predictive loss function of the current emotion is an LM loss, which predicts the emotion word of the current utterance by receiving the speech and prompt for the previous and current utterances and the audio of the current utterance. The predictive loss function of the previous emotions is an LM loss for generating $[y^{x-1}; \ldots; y^{x-k}]$ auto-regressively. In the process of constructing the training loss, we differentiate between the predictive

loss functions of the current and previous emotions, causing the model to focus more on predicting the current emotion. This process is expressed as follows:

$$L_c = \prod_{i=1}^{m} p(y_i^x | y_{<i}^x, D^{x-k \leq \leq x}, P^{x-k \leq \leq x}, A^x) \tag{11}$$

$$L_p = \prod_{j=1}^{k} d * \prod_{i=1}^{m} p(y_i^{x-j} | y_{<i}^{x-j}, y^{x-j<\leq x}, D^{x-k \leq \leq x}, P^{x-k \leq \leq x}, A^x) \tag{12}$$

$$L = L_c + L_p \tag{13}$$

Equation (11) is a loss function for the current emotion prediction, and Equation (12) is a loss function for the previous emotion prediction. $D^x$, $P^x$, and $A^x$ are natural-language text, natural-language prompts, and audio data for the $x$-th utterance, respectively. $y^x$ is an emotion word for the $x$-th utterance. The hyper-parameter $d$ is a constant that regulates the weight of the loss function relative to the previous emotions. Equation (13) represents the final loss function for training. In the reference step, we generate only the emotion word $y^x$ for the current utterance.

## 4. Experiment

In this section, we present the results of experiments on the proposed method. Experiments were conducted to compare and analyze the performance of the proposed method. Section 4.1 describes the information of datasets and experimental settings. Section 4.2 analyzes the results of experiments using the proposed methods in this paper. Section 4.2.1 presents the results of the experiments on how to use text and audio data together in the prompt learning of text-based pre-trained models. Section 4.2.2 presents the results of the experiments on predicting the emotion of the current utterance using the emotion and contextual information from the previous utterances. All experiments were conducted using MELD, which is an English dataset, and KEMDy20, which is a Korean dataset. Section 4.2.3 presents the experimental results obtained using the proposed method. The proposed method uses audio data and previous utterances together for emotion recognition of the current utterance.

### *4.1. Datasets and Experiment Settings*

In this experiment, the Multimodal Emotion Lines Dataset (MELD) [31] and the Korean Emotion Multimodal Dataset in 2020 (KEMDy20) [32] were used as datasets for performance evaluation. The information of datasets are depicted in Table 1.

**Table 1.** Information of datasets.

| | | # Dialogue | # Utterances | # Label | Speech-to-Text | Data Source | Language |
|---|---|---|---|---|---|---|---|
| MELD | train | 1039 | 9989 | 7 | subtitle | *Friends* TV series | English |
| | dev | 114 | 1109 | | | | |
| KEMDy20 | train | 30 | 10304 | 7 | human annotation | Free conversation on the subject of two people | Korean |
| | dev | 10 | 3158 | | | | |

MELD is an English multimodal dataset for emotion analysis with speech text, audio, and video collected from the *Friends* TV series. In this study, the experiment was conducted using only the text and audio data. Of the 1153 dialogues, 1039 were used for training and 114 for evaluation. The training and evaluation dialogue data consisted of 9989 and 1109 utterances, respectively. The utterance texts are crawled through the subtitle files of all the episodes, which contain the beginning and the end timestamps of the utterances. Each utterance was classified into one of seven emotions: anger, disgust, fear, joy, neutral, sadness, and surprise. The data are unbalanced, with the utterances for neutral and joy accounting for approximately 65% of the total data.

KEMDy20 is a Korean multimodal dataset for analyzing the association between text and audio of speech, biosignals, and emotions from free conversation on the subject of two people. The experiment was conducted using only speech text and audio. The dataset was randomly separated and tested because the training and test sets were not separated in advance. Of the total forty conversation sessions, thirty were used for training and ten were used for evaluation. The training and evaluation conversation data consisted of 10,304 and 3158 utterances, respectively. The utterance texts are crawled through human annotation. Each utterance was classified into one of seven emotions: happy, fear, surprise, angry, neutral, sad, and disgust. The labels were translated into Korean emotional words as "기쁨, 두려움, 놀람, 분노, 중립, 슬픔, and 논쟁". The data are unbalanced, with the utterances for neutral and happy accounting for approximately 95% of the total data.

The baseline model of this experiment used a pre-trained T5-base [3], pko-T5 [33]. The pre-trained audio models were wav2vec2-base-960 h and wav2vec2-large-xlsr-korean. In the experiment using the English dataset MELD, the prompt for the previous utterances was set to "Emotion of current utterance is <MASK>", and the prompt for the current utterance was set to "Emotion of last utterance is <MASK>". Additionally, in the experiment using the Korean dataset KEMDy20, "현재 발화의 감정은 <MASK>" was used as a prompt for the previous utterances, and "마지막 발화자의 감정은 <MASK>" was used for the current utterance. Experiments using previous context used two previous utterances by setting $k$ to 2. The emotion answer token for MELD was tested by setting it the same as the label in the dataset. However, KEMDy20 translated the answer to emotions into Korean emotional words and used them as "기쁨, 두려움, 놀람, 분노, 중립, 슬픔, 논쟁". The metrics used for the performance evaluation were accuracy, $F_1$ score, and weighted $F_1$ score, which is typically used for the performance evaluation of unbalanced data. The model was trained with a batch size of eight examples, a learning rate of $1 \times 10^{-5}$ , and epochs of 30. Then, we used the highest performance of the epochs.

### *4.2. Experimental Results*

In this section, we present the results of the experimental evaluation of the proposed method. AF is a method that uses text and audio together in prompt learning of the text-based pre-training model described in Section 3.1. PC is a method that additionally uses previous utterances and contextual information to predict the current-utterance emotion, described in Section 3.2. AF + PC uses both text and audio, current and previous utterances, and previous contextual information for current emotion prediction, which is the proposed method in this paper.

#### 4.2.1. Experimental Results of Audio Features

First, an experiment was conducted on prompts with audio features, corresponding to Section 3.1. Table 2 presents the performance-evaluation results for the baseline and T5 prompt + AF using MELD. AF is a method that uses text and audio together in prompt learning of the text-based pre-training model. T5 prompt + AF is the result of the proposed method in Section 3.1 added to the baseline model. The audio information was compressed by setting the number of existing $n$ audio embeddings to $l$. The hyper-parameter $l$ was tested by setting it to 1, 10, and 30. Prompt learning using additional audio information showed similar or higher performance than prompt learning using only text. When the number of audio embeddings was set to 30 with prompt learning using additional audio information, the highest performance score was an accuracy of 64.56, $F_1$ score of 49.37, and weighted $F_1$ score of 62.93. In addition, the results of the additional experiment using KEMDy20 are shown in Table 3. The results of the experiments with T5 prompt + AF, which uses text and audio together, showed a higher accuracy than the baseline. However, it showed a lower $F_1$ score than the baseline in some cases. When the number of audio embeddings was set to 30, the results of the experiments showed the highest performance score with an accuracy of 87.42, $F_1$ score of 40.94, and a weighted $F_1$ score of 85.94. In the

two experiments, audio embedding performed the best at 30, and it performed better than the baseline in all performance metrics.

**Table 2.** Experimental results table of audio features with prompt learning with MELD.

| Model | L | Accuracy (%) | $F_1$ | Weighted $F_1$ |
|---|---|---|---|---|
| T5 prompt (Baseline) | | 63.93 | 47.73 | 62.03 |
| T5 prompt + AF | 1 | 63.30 | 47.77 | 61.65 |
| | 10 | 62.30 | 48.71 | 62.05 |
| | 30 | 64.56 | 49.37 | 62.93 |

**Table 3.** Experimental results of audio features with prompt learning with KEMDy20.

| Model | L | Accuracy (%) | $F_1$ | Weighted $F_1$ |
|---|---|---|---|---|
| T5 prompt (Baseline) | | 87.11 | 38.43 | 85.71 |
| T5 prompt + AF | 1 | 87.36 | 37.54 | 85.89 |
| | 10 | 87.40 | 37.40 | 85.93 |
| | 30 | 87.42 | 40.94 | 85.94 |

4.2.2. Experimental Results of Previous Context

In this section, experiments were conducted on prompt learning with the previous context, corresponding to Section 3.2. Table 4 presents the performance-evaluation results for the baseline and T5 prompt + PC using MELD. In Table 4, T5 prompt + PC additionally uses previous utterances and contextual information to predict the emotion of the current utterance. The hyper-parameter $k$ was set to 2 in all the experiments, which means that two previous utterances were used. In addition, experiments on various $d$ values adjusted the proportion of the loss function to the previous emotions. $d$ was tested by setting it to 0.15, 0.35, and 0.5. As a result of this experiment, the results of T5 prompt + PC, a method using emotions and contextual information of previous utterances for predicting the emotion of the current utterance, showed a higher performance than the baseline in all performance metrics. When $d$ was designated as 0.5, the results of the experiments showed an accuracy of 65.10, $F_1$ score of 52.54, and weighted $F_1$ score of 63.96, with the highest $F_1$ score and weighted $F_1$ score. In addition, from Table 5, the results for KEMDy20 show a similar performance to the baseline in the experimental results of T5 prompt + PC, which uses previous emotions and utterances for current-emotion prediction. The accuracy and weighted $F_1$ score showed a lower performance than the baseline, while the $F_1$ score showed a higher performance than the baseline. Similar to the experimental results for MELD, when the ratio of the loss function to the previous-emotion prediction was set to 0.5, the results of the experiments showed an accuracy of 86.51, $F_1$ score of 39.13, and weighted $F_1$ score of 85.66, with the highest $F_1$ score.

**Table 4.** Experimental results of previous context with prompt learning with MELD.

| Model | D | Accuracy (%) | $F_1$ | Weighted $F_1$ |
|---|---|---|---|---|
| T5 prompt (Baseline) | | 63.93 | 47.73 | 62.03 |
| T5 prompt + PC | 0.15 | 65.28 | 52.03 | 63.47 |
| | 0.35 | 64.47 | 50.56 | 63.38 |
| | 0.5 | 65.10 | 52.54 | 63.96 |

**Table 5.** Experimental results of previous context with prompt learning with KEMDy20.

| Model | D | Accuracy (%) | $F_1$ | Weighted $F_1$ |
|---|---|---|---|---|
| T5 prompt (Baseline) | | 87.11 | 38.43 | 85.71 |
| T5 prompt + PC | 0.15 | 86.42 | 37.79 | 85.50 |
| | 0.35 | 86.85 | 38.70 | 85.63 |
| | 0.5 | 86.51 | 39.13 | 85.66 |

#### 4.2.3. Experimental Results of Proposed Method

We conducted experiments by applying both Sections 3.1 and 3.2 to evaluate the performance of the proposed method. The proposed method is T5 prompt + AF + PC, as shown in Table 6. In Table 6, the proposed method showed a higher performance than previous studies [12–15,22,23]. A methodology for prompt learning trains the model in a similar process as pre-training. Since downstream tasks in prompt learning are performed using pre-trained information without additional classification layers for emotion recognition, it performs better than previous studies. T5 prompt + AF + PC uses both text and audio information of two previous utterances and the current utterance to predict the current emotion in the speech emotion-recognition task. As hyper-parameters for each proposed method, 30 for $l$ and 0.5 for $d$ were used, which showed the best performance in Sections 4.2.1 and 4.2.2 using each proposed method. T5 prompt + AF + PC performed better than T5 prompt + AF, which uses additional audio information, and T5 prompt + PC, which uses previous utterances and emotions. The performance scores of T5 prompt + AF + PC were an accuracy of 66.72, $F_1$ score of 52.66, and weighted $F_1$ score of 65.42. The experiments showed a higher performance score for T5 prompt + AF + PC using the two proposed methods together than T5 prompt + AF and T5 prompt + PC using each method independently. The proposed method showed a similar or lower performance than the tri-modal model using audio, text, and vision in existing studies. However, it showed a higher performance than the bi-modal model using only audio and text. The results of the proposed method using KEMDy20 are shown in Table 7. The results showed the highest performance for T5 prompt + AF + PC. T5 prompt + AF + PC performed better than when each method was used independently, with an accuracy of 87.49, $F_1$ score of 44.16, and weighted $F_1$ score of 86.28. In both experiments, when each method was used alone, the performance improvement was small or decreased in some cases. However, when the two methods were used together, they showed a much higher performance than the baseline model.

**Table 6.** Experimental results of proposed method with MELD. * is a bi-modal model using audio and text; ** is a tri-modal model using audio, text, and vision.

| Model | Accuracy (%) | $F_1$ | Weighted $F_1$ |
|---|---|---|---|
| MM-DFN [22] * | 62.49 | - | 59.46 |
| SMIN [23] * | - | - | 64.50 |
| DialogueTRM [12] ** | 64.60 | - | 63.20 |
| EmoCaps [13] ** | - | - | 64.00 |
| UniMSE [14] ** | 65.09 | - | 65.51 |
| M2FNet [15] ** | 67.85 | - | 66.71 |
| T5 prompt (Baseline) | 63.93 | 47.73 | 62.03 |
| T5 prompt + AF | 64.56 | 49.37 | 62.93 |
| T5 prompt + PC | 65.10 | 52.54 | 63.96 |
| T5 prompt + AF + PC (Proposed model) | 66.72 | 52.66 | 65.42 |

**Table 7.** Experimental results of proposed method with KEMDy20.

| Model | Accuracy (%) | $F_1$ | Weighted $F_1$ |
|---|---|---|---|
| T5 prompt (Baseline) | 87.11 | 38.43 | 85.71 |
| T5 prompt + AF | 87.42 | 40.94 | 85.94 |
| T5 prompt + PC | 86.51 | 39.13 | 85.66 |
| T5 prompt + AF + PC (Proposed model) | 87.49 | 44.16 | 86.28 |

## 5. Conclusions

In this paper, we proposed two methods to improve the performance of emotion analysis in speech emotion-recognition tasks. The first method is combining audio and text data for prompt learning. Prompt learning using a text-based pre-trained model is problematic. Because the model is pre-trained using only text, using modalities other than those involved in the pre-training process to solve the downstream tasks is difficult. The proposed method solves this problem and performs a self-attention mechanism inside the transformer model using text and audio data as inputs to the transformer-based encoder-decoder model in prompt learning. It can train with audio and text data together for prompt learning with text-based pre-trained models by performing self-attention using both text and audio. The second uses emotions and contextual information from previous utterances to predict the emotions of a given utterance. Continuous emotional changes linked to previous utterances can also affect the classification of current emotions in speech emotion-recognition tasks. In this study, we have proposed a method to adjust the loss function of a pre-trained language model to deduce the emotion and context of previous utterances, which can be used as important information for training a model to classify the emotions expressed in input utterances. Model training and evaluation were conducted using the MELD English multimodal dataset and the KEMDy20 Korean multimodal dataset. As a result, T5 prompt + AF + PC performed better than the baseline model with an accuracy of 66.72, $F_1$ score of 52.66, and weighted $F_1$ score of 65.42 on MEDL. It also showed a higher performance than the bi-modal model using the existing audio and text. In addition, prompt + AF + PC showed the highest performance, with an accuracy of 87.49, $F_1$ score of 44.16, and weighted $F_1$ score of 86.28 on KEMDy20. The experimental results of this paper confirm that both proposed methods showed the highest performance in the speech emotion-recognition task. However, the experiments in the paper were performed with audio-text bi-modality. It is expected that the speech emotion-recognition task using more diverse modalities such as vision, heart rate, etc. will enable a higher performance of emotion recognition. We intend to perform speech emotion recognition using three or more modalities.

**Author Contributions:** Methodology, E.J. and G.K.; Software, G.K.; Data curation, E.J.; Writing—original draft, E.J.; Writing—review & editing, G.K.; Supervision, S.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C1005316). This work was also supported by the Gachon University research fund of 2021. (GCU-202109980001).

**Data Availability Statement:** The Multimodal Emotion Lines Dataset (MELD) is publicly available in [31]; https://affective-meld.github.io/, accessed on 4 October 2018. The Korean Emotion Multimodal Dataset in 2020 (KEMDy20) is publicly available in [32]; https://nanum.etri.re.kr/share/kjnoh/KEMDy20?lang=ko_KR, accessed on 21 December 2021.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Chu, I.H.; Chen, Z.; Yu, X.; Han, M.; Xiao, J.; Chang, P. Self-supervised Cross-modal Pretraining for Speech Emotion Recognition and Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 5105–5114.
2. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
3. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
4. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.
5. Ma, X.; Xu, P.; Wang, Z.; Nallapati, R.; Xiang, B. Domain Adaptation with BERT-based Domain Classification and Data Selection. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), Hong Kong, China, 3 November 2019; pp. 76–83. [CrossRef]
6. Xu, L.; Bing, L.; Lu, W.; Huang, F. Aspect Sentiment Classification with Aspect-Specific Opinion Spans. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 3561–3567. [CrossRef]
7. Chumachenko, K.; Iosifidis, A.; Gabbouj, M. Self-attention fusion for audiovisual emotion recognition with incomplete data. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 2822–2828. [CrossRef]
8. Liu, Y.; Li, J.; Wang, X.; Zeng, Z. EmotionIC: Emotional Inertia and Contagion-driven Dependency Modelling for Emotion Recognition in Conversation. *arXiv* **2023**, arXiv:2303.11117.
9. Lee, J.; Lee, W. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 5669–5679. [CrossRef]
10. Joshi, A.; Bhat, A.; Jain, A.; Singh, A.; Modi, A. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 4148–4164. [CrossRef]
11. Dutta, S.; Ganapathy, S. HCAM—Hierarchical Cross Attention Model for Multi-modal Emotion Recognition. *arXiv* **2023**, arXiv:2304.06910.
12. Mao, Y.; Liu, G.; Wang, X.; Gao, W.; Li, X. DialogueTRM: Exploring Multi-Modal Emotional Dynamics in a Conversation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November 2021; pp. 2694–2704. [CrossRef]
13. Li, Z.; Tang, F.; Zhao, M.; Zhu, Y. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 1610–1618. [CrossRef]
14. Hu, G.; Lin, T.E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 7837–7851.
15. Chudasama, V.; Kar, P.; Gudmalwar, A.; Shah, N.; Wasnik, P.; Onoe, N. M2FNet: Multi-Modal Fusion Network for Emotion Recognition in Conversation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, New Orleans, LA, USA, 19–20 June 2022; pp. 4652–4661.
16. Liu, D.; Wang, Z.; Wang, L.; Chen, L. Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning. *Front. Neurorobotics* **2021**, *15*, 697634. [CrossRef] [PubMed]
17. Wu, W.; Zhang, C.; Woodland, P.C. Emotion Recognition by Fusing Time Synchronous and Time Asynchronous Representations. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6269–6273. [CrossRef]
18. Atmaja, B.T.; Akagi, M. Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM. *Speech Commun.* **2021**, *126*, 9–21. [CrossRef]
19. Xie, B.; Sidulova, M.; Park, C.H. Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion. *Sensors* **2021**, *21*, 4913. [CrossRef] [PubMed]
20. Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.H.; Chang, S.F.; Cui, Y.; Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv* **2021**, arXiv:2104.11178.
21. Song, X.; Huang, L.; Xue, H.; Hu, S. Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 5197–5206.
22. Hu, D.; Hou, X.; Wei, L.; Jiang, L.; Mo, Y. MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 7037–7041. [CrossRef]

23. Lian, Z.; Liu, B.; Tao, J. SMIN: Semi-supervised Multi-modal Interaction Network for Conversational Emotion Recognition. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1. [CrossRef]
24. Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. Language Models as Knowledge Bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 2463–2473. [CrossRef]
25. Shin, T.; Razeghi, Y.; Logan, R.L., IV; Wallace, E.; Singh, S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020.
26. Gao, T.; Fisch, A.; Chen, D. Making Pre-trained Language Models Better Few-shot Learners. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; pp. 3816–3830. [CrossRef]
27. Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; Tang, J. GPT Understands, Too. *arXiv* **2021**, arXiv:2103.10385.
28. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 12449–12460.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
30. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7871–7880. [CrossRef]
31. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 527–536. [CrossRef]
32. Noh, K.J.; Jeong, H. Korean Multimodal Emotion Dataset 2020 (KEMDy20). 2021. Available online: https://nanum.etri.re.kr/share/kjnoh/KEMDy20?lang=ko_KR (accessed on 28 May 2023).
33. Park, D. pko-t5: PAUST Korean T5 for Text-to-Text Unified framework. 2022. Available online: https://github.com/paust-team/pko-t5 (accessed on 28 May 2023).