KoMPT: A Multimodal Emotion Recognition Model Integrating KoELECTRA, MFCC, and Pitch with Multimodal Transformer

MoungHo Yi , KeunChang Kwak , JuHyun Shin *

Posted Date: 16 September 2024

doi: 10.20944/preprints202409.1178.v1

Keywords: KoELECTRA; MFCC; Pitch; Cross Modal Attention; Multimodal Transformer

Article

KoMPT: A Multimodal Emotion Recognition Model Integrating KoELECTRA, MFCC, and Pitch with Multimodal Transformer

MoungHo Yi [1], KeunChang Kwak [1] and JuHyun Shin [2,*]

[1] Department of Electronic Engineering, Chosun University, Gwangju 61452, Korea; audgh3710@gmail.com(M.Y.); kwak@chosun.ac.kr(K.K.)
[2] Department of New Industry Convergence, Chosun University, Gwangju 61452, Korea
* Correspondence: jhshinkr@chosun.ac.kr

Abstract: With the development of human-computer interaction, the importance of emotion recognition is increasing. Emotion recognition technology provides practical benefits in various industries such as improving user experience, education, and organizational productivity. In education, emotion recognition can enable real-time understanding of students' emotional states and provide tailored feedback, and in the workplace, by monitoring employees' emotional states, it can improve work performance and job satisfaction. Therefore, multimodal-based research that combines text, speech, and video data for emotion recognition is being conducted in various industries. In this study, we propose an emotion recognition method that combines text and speech data, reflecting the characteristics of the Korean language. For text, KoELECTRA is used to perform embedding, and for speech, MFCC and pitch analysis are used to extract features. Finally, a multimodal transformer model is proposed that combines these two data types to perform emotion recognition. The multimodal transformer model processes text and speech data separately, and then learns the interaction between the two modalities through a Cross-Modal Attention mechanism. Through this, the complementary information from text and speech is effectively combined, improving emotion recognition performance. Experimental results show that the proposed model outperforms single-modality models, achieving a high accuracy of 73.13% and an F1-Score of 0.7344 in emotion classification. This study contributes to the advancement of emotion recognition technology by combining various language and modality data, and higher performance can be expected through the integration of additional modalities in the future

Keywords: KoELECTRA; MFCC; pitch; cross modal attention; multimodal transformer

1. Introduction

Human-Computer Interaction (HCI) is becoming increasingly important with the development of digital devices and software. HCI is moving away from simple command input and output and moving toward providing a better user experience by having machines understand human emotions and intentions and respond appropriately. Emotion Recognition (ER) plays a key role in this interaction, and technology that understands human emotions and responds appropriately is becoming an important element in improving the user experience [1]. Emotion recognition technology improves the quality of HCI and plays a role in increasing usability, accessibility, and efficiency in various industries, including education and organizational productivity [2]. In the field of education, emotion recognition technology can monitor students' emotional states and positively impact learning performance. Based on this, it can provide customized learning approaches and adaptive feedback. In an e-learning environment, utilizing multimodal emotion recognition technology can help identify students' emotions in real time and adjust learning difficulty or provide appropriate feedback [3]. This technology can contribute to increasing learning engagement and

 doi:10.20944/preprints202409.1178.v

reducing learning dropouts due to stress and frustration. In the area of organizational productivity, emotion recognition technology also contributes to stress management, fatigue control, and improving job satisfaction by monitoring the emotional state of employees in the workplace. Since the emotional state of employees is closely related to work performance, it is important to understand it in real time and take appropriate measures. Especially in a remote work environment, where face-to-face interaction is limited, emotion recognition systems become even more necessary. Through this, productivity can be maintained and improved by monitoring the emotional state of employees and providing necessary support even in a non-face-to-face environment. As such, the need for emotion recognition technology will continue to increase in various fields in the future, and it will play an important role in improving work efficiency and performance in each industry.

Korean has unique linguistic characteristics for emotion recognition. First, in terms of text, Korean is an agglutinative language that performs various grammatical functions through word inflection. This characteristic makes it essential to understand the entire context of a sentence because the word order in a sentence is flexible, and subjects and objects are often omitted. For example, the meaning of a sentence can change depending on the position or inflection of the same word, and emotional expression can also change subtly. Therefore, in Korean text emotion analysis, it is necessary to deeply analyze the structure and context of the sentence beyond the vocabulary level. In terms of phonetics, Korean intonation and pronunciation patterns play an important role in emotional expression. Even the same word can convey emotions differently depending on intonation, and the pronunciation system of Korean (long, short, aspirated, liquid sounds, etc.) provides important clues for understanding emotional states. For example, emotions in Korean change depending on the pitch of the speech, pronunciation stress, and speaking speed, and emotional states can be predicted through this. Scherer's [4] study explains that the pitch and speed of speech can indicate emotional state, and this can also be an important factor in Korean speech emotion recognition.

Existing emotion recognition studies have mainly focused on one modality, either text or speech, but showed limitations in capturing important emotion information. For example, Bharti's [5] study combined various text data sets (ISERA, WASSA, Emotion-stimulus) and used a combination of CNN (Convolutional Neural Network) and Bi-GRU (Bidirectional Gated Recurrent Unit) models to recognize emotions. This study classified emotions based on text data, but did not consider speech or other modalities, and thus could not sufficiently reflect the complexity of emotions. Similar limitations also appear in speech emotion recognition. Kim's [6] study extracted speech features such as Mel-Spectrogram and MFCC (Mel-Frequency Cepstral Coefficients) using the Emo-DB and RAVDESS data sets, and then combined BiLSTM-Transformer and 2D CNN to recognize emotions. However, this study also focused only on speech data and did not consider interactions with other modalities such as text. Thus, single-modality approaches have limitations in sufficiently reflecting the complex characteristics of emotions. In addition, many emotion recognition studies have been conducted based on data in German or English, and emotion recognition studies on Korean are relatively lacking. To solve this problem, a multimodal approach that combines various modalities such as text, speech, and video has recently attracted attention. A multimodal approach is a methodology that combines the unique emotional information of each modality to increase the accuracy of emotion recognition. For example, in the study by H. Park [7], deep learning models were trained for text and speech separately, and then a weighted average ensemble was used to improve emotion recognition performance. Y. Kim [8] also trained deep learning models for text and speech separately, and then used an ensemble method by averaging, showing higher performance than with a single modality. However, multimodal studies tend to rely mainly on simple combination methods such as average ensemble or weighted average ensemble. This may limit emotion recognition performance because it does not sufficiently reflect the interaction between modalities. To complement this, this study proposes a deep learning model that incorporates preprocessing reflecting the characteristics of the Korean language, adds a transformer encoder to each modality, and enhances the interaction between text and speech through cross-modal attention.

 doi:10.20944/preprints202409.1178.v

This study aims to design an emotion recognition model that reflects the unique linguistic and **speech** characteristics of Korean. Korean is an agglutinative language with many grammatical variations, and context-dependent analysis is essential. To this end, text data **are** processed through the KoELECTRA model, and speech data **are analyzed** using MFCC and Pitch to extract key features of speech signals. KoELECTRA is a BERT-based Korean-specific model that learns Korean honorifics and informal speech, as well as complex lexical changes. Through this, text data are converted into high-dimensional vectors, and speech data are vectorized by extracting MFCC representing timbre and Pitch reflecting intonation information. A multimodal transformer model is used to combine text and speech data. This model consists of a transformer encoder that processes each modality individually and a cross-modal attention mechanism that combines text and speech data. Cross-modal attention learns the interaction between the two modalities by using text embeddings as queries and speech embeddings as keys and values. This effectively combines complementary information between text and speech, and improves the accuracy of emotion recognition. Therefore, this study proposes a multimodal transformer emotion recognition model that considers the interaction between Korean text and speech to reflect the characteristics of the Korean language and overcome the limitations of existing studies. This study will contribute to overcoming the limitations of existing emotion recognition studies by reflecting the linguistic and **speech** characteristics of the Korean language and improving the performance of Korean-based multimodal emotion recognition. In addition, it will contribute to the development of emotion recognition technology that can be applied to various languages and modality data in the future.

The structure of this paper is as follows. In Section 2, we define the contents of existing studies related to text, speech, and multimodal emotion recognition, and in Section 3, we explain the proposed multimodal transformer model. In Section 4, we evaluate the proposed model through experiments and verify the performance of the model by comparing it with existing emotion recognition studies using objective performance indicators. Finally, in Section 5, we present the conclusions of the stud

2. Related Work

This chapter covers the latest research trends in emotion recognition technology using text, speech, and multimodal data. First, the main principles and performance of emotion recognition technology that analyzes text and speech separately using a single-modality approach are explained. Then, multimodal emotion recognition that combines text and speech data is discussed, and the limitations of existing research and potential improvements are presented.

2.1. Text Emotion Recognition

Text Emotion Recognition (TER) is a technology that analyzes text data to identify emotional states. It works by extracting emotional patterns from text and learning them. Early text emotion recognition mainly utilized keyword-based analysis methods, but these methods had the limitation of not sufficiently reflecting the context. In particular, such a simple keyword method is not appropriate for languages such as Korean, where sentence structures are flexible and subjects and objects are often omitted. To solve this problem, deep learning-based natural language processing techniques have been widely used recently. Commonly used text embedding techniques include classical methods such as Word2Vec, GloVe, and FastText, and context-based embedding techniques such as BERT. In Sabbeh's [9] study, BERT embedding showed better performance when compared to various word embedding techniques. BERT achieves higher accuracy by efficiently processing contextual information, whereas classical embedding techniques have limitations in handling complex contexts. In Mutinda's [10] study, the proposed LeBERT model performed sentiment analysis by combining BERT-based embedding and lexicon-based features, and achieved excellent performance in experiments using large-scale review datasets such as Yelp and IMDB. In Bharti's [5] study, high performance was achieved through analysis by combining various text sentiment recognition datasets using a combination of CNN and Bi-GRU, and in Li's [11] study, a BERT-based deep learning model also recorded high accuracy. These studies show that text sentiment recognition

 doi:10.20944/preprints202409.1178.v

technology is evolving from simple keyword-based to deep learning-based complex context analysis, and various text embedding techniques play an important role in improving performance.

2.2. Speech Emotion Recognition

Speech Emotion Recognition (SER) is a technology that analyzes speech data to identify the speaker's emotions. In speech emotion recognition, speech signals are analyzed, emotional patterns are extracted from their characteristics, and a model is trained to predict emotional states based on this data. Representative techniques for extracting speech characteristics include MFCC and Mel-Spectrogram. MFCC analyzes the frequency components of speech signals to extract important information such as the timbre of the speech, and the Mel-Spectrogram is an important tool for analyzing emotional states by visualizing the temporal changes in frequency components. In Reggiswarashari's study [12], speech data were processed using frame-based segmentation and overlap segmentation techniques, and the data converted to MFCC were used for emotion recognition through CNN, and various datasets (RAVDESS, SAVEE, TESS) were combined to achieve an accuracy of 83.69%. In Kim's [6] study, various features of speech signals were extracted through Mel-Spectrogram, MFCC, and Spectral Contrast, and BiLSTM-Transformer and 2D CNN were combined for learning, and as a result, accuracies of 95.65% and 80.19% were recorded on the Emo-DB and RAVDESS datasets, respectively. In Hazra's [13] study, after extracting emotional features of speech data using MFCC, five deep learning models—CNN, LSTM, ANN, MLP, and CNN-LSTM—were used to conduct experiments on the TESS, SAVEE, and RAVDESS datasets, and among them, the CNN-LSTM model achieved the highest accuracy of 84.35%. In this way, recent speech recognition studies are improving the performance of emotion recognition by combining various deep learning models based on feature extraction techniques such as MFCC and Mel-Spectrogram.

2.3. Multimodal Emotion Recognition Using Text and Speech Data

Multimodal emotion recognition is a technology that recognizes emotions by combining various modality data such as text, speech, and images. According to the study of H. Park [7], when text and speech data were processed separately and then the average ensemble method was applied, the performance actually decreased, but when the weighted average ensemble was used, the F1-Score improved by 3%. In the study of Y. Kim [8], each text and speech model was trained independently, and then the outputs of the two models were combined using the average ensemble method, which showed a performance of 78.51%, higher than that achieved when using a single modality. Currently, most multimodal emotion recognition studies mainly use ensemble techniques that combine the results of each modality. This approach improves performance but has the limitation of not sufficiently reflecting the interaction between modalities. In addition, if the quality or amount of data between each modality is imbalanced, it can have a negative impact on overall model performance [14,15]. To address this, this study proposes a method to perform emotion recognition by adding a transformer encoder to each modality and combining features between modalities through Cross-Modal Attention. This method enables higher accuracy in emotion recognition by facilitating interaction between text and speech data.

3. Materials and Methods

3.1. Overall Process

In this paper, we propose a multimodal transformer emotion recognition model that considers the interaction between Korean text and speech. Text data are embedded using the KoELECTRA model that reflects the characteristics of Korean, and speech data are vectorized using MFCC and Pitch. Each text and speech dataset is processed by a separate transformer encoder, and the model recognizes emotions by combining features through Cross-Modal Attention. Figure 1 shows the overall research structure of the proposed method.

 doi:10.20944/preprints202409.1178.v

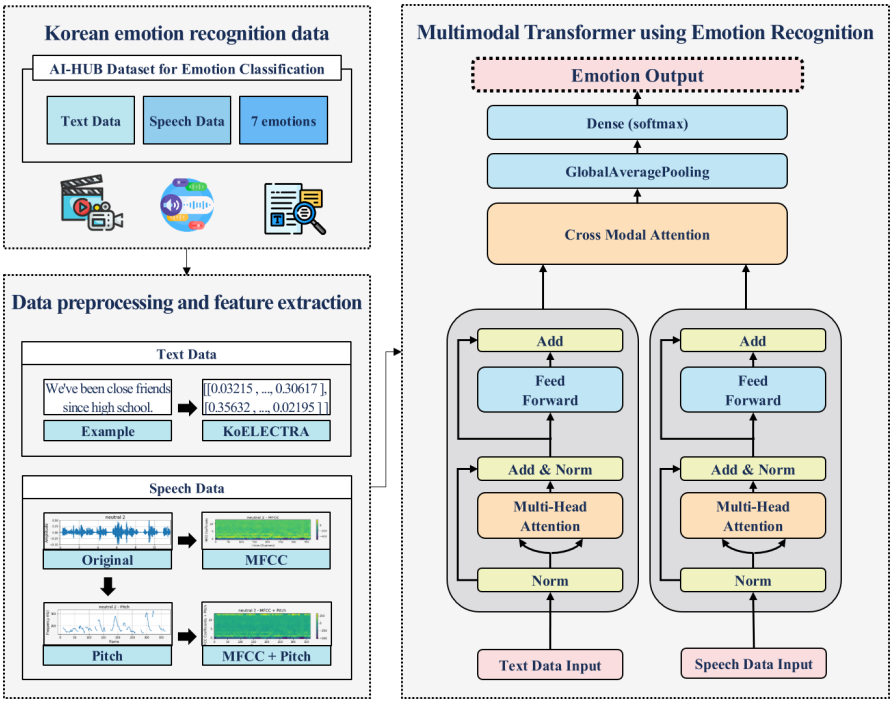

Figure 1. Research structure.

3.2. Preprocessing and Feature Extraction

This section covers two modalities (text and speech) and explains the process of extracting important features from each modality and converting them into a form that the model can learn. Text data are embedded using the KoELECTRA model to reflect the linguistic characteristics of Korean, and speech data features are extracted by combining MFCC and Pitch. The two datasets are adjusted to the same size to be used in training with the multimodal transformer model.

3.2.1. KoELECTRA, Which Takes Korean Characteristics into Account

This section describes the text embedding process using the KoELECTRA model to effectively process Korean data. KoELECTRA is designed to reflect the contextual features of Korean more accurately, and is based on the structure of BERT and ELECTRA models. First, the text is analyzed morphologically to determine the role of each word in the sentence. Then, the KoELECTRA model is used to convert this morphologically analyzed text into a high-dimensional vector. This vector contains information including the semantic relationships and context between words, and is in a form that the model can process. The converted high-dimensional embedding vector generates vectors of different sizes depending on the length of each sentence. For example, the vector size of the sentence “Oh, you clean it for me!” is (11, 768), which means that a sentence consisting of 11 tokens is embedded into a 768-dimensional vector. The vector size of the sentence “We first met at the academy and started dating because we liked each other” is (18, 768), which means that the sentence consisting of 18 tokens is embedded as a 768-dimensional vector. The size of this embedding vector varies depending on the length of the sentence, and in order to use it as input for a deep learning model, the length of all sentences must be the same. To match the input dimensions, the [PAD] token is inserted into the parts where the number of tokens is insufficient to match the length. The [PAD] token is masked so that it is not learned during the model’s calculation process. This process sets the weight for the [PAD] token to 0 in the attention mechanism, helping the model focus only on the actual information of the sentence. In other words, the [PAD] token is only used to match the input size, and does not affect the important information that the model needs to learn. As a result, the deep learning model can receive inputs of consistent size without change, which increases computational efficiency and maintains the actual information of the sentence. In this study, in order to make the high-dimensional vectors for all sentences the same size, the number of tokens is

 doi:10.20944/preprints202409.1178.v

calculated for each sentence, and then the size of the vector is adjusted based on the sentence with the largest number of tokens. First, the number of tokens is calculated using the following formula. A given sentence S consists of words $W_1, W_2, \ldots, W_n$. Each word W_I is split into multiple subword tokens $t_{i1}, t_{i2}, \ldots, t_n$ by the WordPiece tokenizer, and this can be expressed as Equation (1). T represents the total number of tokens generated from the entire sentence, n is the number of words in the sentence, and k_I represents the number of subword tokens into which each word w_i is divided.

$$T = \sum_{i=1}^{n} k_i \quad (1)$$

Also, when inputting data to the model, the final number of tokens is calculated as in Equation (2) by adding the [CLS] token to the beginning of the sentence and the [SEP] token to the end of the sentence. This is the number of tokens calculated using the tokenizer for the actual data used. After tokenizing the entire dataset, the longest sentence was found to have 73 tokens.

$$T_{\text{final}} = T + 2 \quad (2)$$

In order to match the number of tokens in all sentences to the longest sentence, the dimensions of the vectors were unified by adding [PAD] tokens to the shorter sentences.

In Table 1, we present examples of original Korean sentences and their corresponding embedding vectors generated using the KoELECTRA model. Since KoELECTRA is designed specifically for the Korean language, the WordPiece tokenizer splits the Korean words into subwords, as indicated by the '##' symbols. These subwords are then embedded into 768-dimensional vectors. To ensure uniform input size for the deep learning model, shorter sentences are padded with the [PAD] token to match the longest sentence, resulting in a final vector size of (73, 768) for all sentences. For example, in the sentence "어, □□ 니가 대신 해 줘!" ("Oh, you clean it for me!"), the word "니가" is split into two subword tokens, "니" and "##가". Similarly, "처음 학원에서 만났다 가 서로 □□□□ □□□ □□□□" ("We first met at the academy and started dating because we liked each other") contains 18 tokens, including several subwords like "##에서" and "##게". By preserving the Korean subword structure in the embedding process, KoELECTRA captures the unique linguistic characteristics of the Korean language, such as its agglutinative nature, where suffixes like "##게" and "##□□" carry significant grammatical meaning. This allows the model to process the intricate relationships between words and subwords, ultimately leading to more accurate understanding of Korean text in emotion recognition tasks.

Table 1. Final Results of KoELECTRA Embedding.

Original Text (in Korean)	Embedding Vector	Vector Size
어, 청소 니가 대신 해 줘!	['[CLS]', '어', ',', '청소', '니', '##가', '대신', '해', '줘', '!', '[SEP]', '[PAD]', ..., , '[PAD]']	(73, 768)
둘 다 하기 싫어서 화내	['[CLS]', '둘', '다', '하', '##기', '싫', '##어', '##서', '화', '##□', '.', '[SEP]', '[PAD]', ..., , '[PAD]']	(73, 768)
□□ □□□□ □□□□ □□ □□□□ □□□ □□□	['[CLS]', '처음', '학원', '##에', '##서', '만났', '##다가', '서로', '좋', '##아', '##해서', '사귀', '##게', '되', '##었', '##지', '.', '[SEP]', '[PAD]', ..., , '[PAD]']	(73, 768)
□□□□□ □□ □□ 싸우게 돼	['[CLS]', '룸', '##메이트', '##와', '너무', '자주', '싸우', '##게', '돼', '.', '[SEP]', '[PAD]', ..., , '[PAD]']	(73, 768)
□□□□□ □□ 말 안했는데 □□□□ □□□	['[CLS]', '부모', '##님', '##한', '##테', '아직', '말', '안', '##했', '##는데', '말', '##하', '##기', '##가', '두려워', '.', '[SEP]', '[PAD]', ..., , '[PAD]']	(73, 768)

3.2.2. Combining MFCC and Pitch While Considering the Characteristics of the Korean Language

 doi:10.20944/preprints202409.1178.v

Speech data is an important element in emotion recognition, enabling the frequency characteristics and intonation information of the speech to be reflected simultaneously. To this end, MFCC and Pitch are combined to extract the features of the speech signal, and then the two features are combined and used. MFCC analyzes the frequency components of the speech signal and generates a feature vector that reflects the timbre of the speech. The speech signal is first enhanced through Pre-emphasis, and then the signal is divided into short frames, with a Hamming Window applied to each frame. A Mel-scale filter bank is applied to the signal converted to the frequency domain using FFT, and then the final MFCC coefficients are obtained through DCT. Pitch plays an important role in intonation and emotional expression, and is extracted by analyzing the periodicity of the speech signal. Pitch extraction calculates the period T_0 of the speech signal using the Autocorrelation method, and based on the obtained T_0, the Pitch frequency F_0 is derived to reflect emotion-specific characteristics. Pitch represents the pitch (fundamental frequency) of the speech and is used as an additional dimension that provides important information in emotion recognition. Figure 2 visualizes the results of analyzing the data using MFCC and Pitch, respectively.

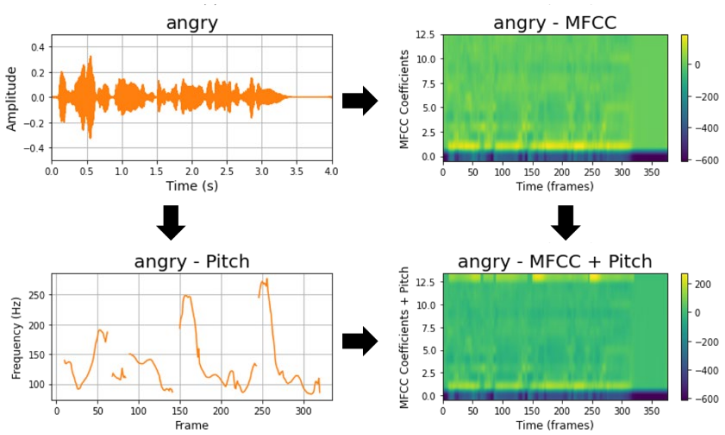

Figure 2. MFCC and Pitch Analysis.

In this paper, we adjust the dimensions of MFCC and Pitch to match the text and speech data to the same dimension. Text data were converted to a vector of size (73, 768) based on the maximum word length of 73 using the KoELECTRA model. Accordingly, the dimension of MFCC was adjusted to match the speech data to the same dimension as the text data. MFCC generated 72 Mel frequency components by setting the number of Mel filters to 72, and each frequency component was adjusted to 768 dimensions to match a vector of size (73, 768), which is the same as the text data. Pitch was also extracted as a one-dimensional feature and combined to match (1, 768) to maintain consistency with the text data. Finally, MFCC and Pitch were combined to generate a vector of size (73, 768) that includes both frequency components and intonation information of the speech signal. Through this process, the speech data has the same vector size as the text data, so text and speech data can be consistently combined in a multimodal emotion recognition model. In this process, the important characteristics of the speech signal are preserved, and information combination between text and speech is possible, which can improve the efficiency of multimodal learning. The final dimension of the speech data after combining MFCC and Pitch can be expressed as Equation (3).

$$\begin{aligned} &\mathrm{X}_{\mathrm{MFCC}} \in R^{73\times768}, X_{Pitch} \in R^{1\times768} \\ &\mathrm{X}_{\mathrm{final}} = [X_{MFCC}, X_{Pitch}] \Rightarrow X_{final} \in R^{73\times768} \end{aligned} \tag{3}$$

In this study, both text data and speech data were used as inputs to the model by aligning them to vectors of size (73, 768). This enabled us to effectively combine text and speech data in a multimodal emotion recognition model by maintaining consistent dimension configurations between the two modalities. This dimension unification prevented information loss due to dimension mismatch and maximized emotion recognition performance when the model learned the interaction between the two modalities.

 doi:10.20944/preprints202409.1178.v

3.3. Multimodal Transformer for Emotion Recognition

This section describes a multimodal transformer model that recognizes emotions by combining text and speech data. The model is designed to take text data and speech data as input and improve the emotion recognition performance by considering the interaction between the two modalities. In particular, it effectively learns complementary information between text and speech data by combining the Multi-Head Attention and Cross-Modal Attention mechanisms, and performs the final emotion classification through Global Average Pooling. The proposed multimodal transformer model has two transformer encoders that independently process text and speech data. Multi-Head Attention in each encoder is a core mechanism of the transformer model and is used to simultaneously learn various patterns in the input data. Each head learns multiple perspectives in the text or speech data to obtain richer representations. Multi-head attention calculates the interaction using query (Q), key (K), and value (V) vectors as in Equation (4), and the result is passed to the next layer through a residual connection and layer normalization. This process is repeated multiple times for each modality to effectively extract features from the input data.

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

Cross Modal Attention learns the interaction between text and speech modalities, and effectively combines the information from each modality. Cross Modal Attention uses the embedding vector extracted from the text modality as a query (Q), and the vector extracted from the speech modality as a key (K) and value (V) to learn the interaction between the two modalities. This strengthens the relationship between text and speech data, and enables more sophisticated emotion recognition. The calculation process of Cross Modal Attention is as shown in Equation (5), and it learns which part to pay attention to when important information in the text is given based on speech data. This strengthens the interaction between speech and text, and complements the characteristics of each modality to derive better emotion recognition performance.

$$\text{Attention}_{\text{cross}}(\text{Q}_{\text{text}}, \text{K}_{\text{speech}}, \text{V}_{\text{speech}}) = softmax\left(\frac{Q_{text}K_{speech}^T}{\sqrt{d_k}}\right)V_{speech} \quad (5)$$

The proposed model consists of two transformer encoders and one Cross-Modal Attention module. Text data are vectorized using KoELECTRA, and speech data are converted into MFCC and pitch features. The converted data are processed by the transformer encoder of each modality, and the interaction is learned through Cross-Modal Attention. The learned features are summarized through Global Average Pooling, and classified into seven emotion classes through the final output layer. The transformer encoder of each modality consists of a multi-layer Multi-Head Attention and Feed-Forward Network (FFN). The Multi-Head Attention stage processes text and speech data in parallel to help interpret the data from multiple perspectives, and automatically learns what information is important in each modality. Next, the FFN learns higher-dimensional representations based on the output values obtained from Multi-Head Attention to enhance the detailed features of each modality. Finally, Layer Normalization and Residual Connection are applied to reduce the difference between input and output values, and to increase the stability of the learning process. Layer Normalization prevents the output values of each layer from becoming excessively large or small, and Residual Connection minimizes information loss during learning by allowing the input values to be directly transmitted to the next layer. Through this structure, the complex features of text and speech data are effectively learned, and the final emotion recognition performance is improved by strengthening the interaction between the two modalities. After passing through the Cross-Modal Attention module, the combined text and speech feature maps summarize the most important information through Global Average Pooling and express it as a single vector. The feature vector summarized through Global Average Pooling finally passes through the Dense layer and Softmax activation function to calculate the probability distribution for the emotion class. This model classifies seven emotion classes, and the Softmax function calculates the probability value for each emotion class, and then determines the class with the highest probability as the final emotion

 doi:10.20944/preprints202409.1178.v

prediction. This model learns emotion recognition from text and speech data, and for this purpose, it uses the categorical cross-entropy loss function to minimize the difference between the model's prediction results and the actual labels. In addition, the Adam optimizer is used to update the model's weights during the learning process, and the learning rate is adaptively adjusted. Text and speech data are learned together, and the interaction between these two modalities plays an important role in improving the model's performance. The following is a pseudocode that describes the entire proposed multimodal transformer model.

Algorithm 1. Pseudocode of Multimodal Transformer Model for Emotion Recognition

Input: Text Features X_{text}, Speech Features X_{speech}
Output: Emotion Classification Y
// Initialize speech and text transformers
Text Encoder : $Z_{text} = Transformer_{text}(X_{text})$
Speech Encoder : $Z_{speech} = Transformer_{speech}(X_{speech})$

// Feature Extraction and Input
$F_{text} = KoELECTRA(X_{text})$
$F_{speech} = MFCC(X_{speech}) + Pitch(X_{speech})$
$Z_{text} = Transformer_{text}(F_{text})$
$Z_{speech} = Transformer_{speech}(F_{speech})$

// Cross Modal Attention
$AttentionScore = \text{softmax}(\frac{Q_{text} \circ K_{speech}^{T}}{\sqrt{d_k}}) \circ V_{speech}$
$CombinedFeatures = AttentionScore \circ V_{speech}$

// Global Average Pooling and Final Output
$Z_{pooled} = \frac{1}{N}\sum_{i=1}^{N} Z\_i$
$y_{pred} = softmax(W_{dense} Z_{pooled} + b_{dense})$
$Y = \text{argmax}(y_{pred})$

4. Experiments and Assessment

4.1. Data Set

The dataset used is the Korean emotion recognition dataset provided by AI-HUB. The dataset was built for the purpose of developing a multi-emotion recognition model, and conversations spoken in various situations were collected through an application. The collected data were labeled with seven emotions (angry, disgust, fear, happiness, neutral, sadness, surprise), and each utterance was assigned an emotion by five experts. In this study, in order to increase the accuracy of emotion recognition, the emotion most frequently selected among the emotions assigned by the five experts was chosen as the final emotion. If two or more emotions were selected the same number of times, both emotions were assigned. However, to optimize the consistency and performance of the emotion recognition model, data that were assigned two or more emotions were removed, and only data that were clearly assigned a single emotion were used. The data used included 14,606 entries from the 4th year, 10,011 entries from the 5th year, and 19,374 entries from the 5th year (2nd). After removing the data with multiple emotions, a total of 36,888 entries were used in the experiment. Looking at the emotion-specific data in Table 2, we can see that the data are severely imbalanced. To address this, we conducted the experiment by adjusting the number of all emotion data to the median value of 3,412. In this way, we minimized the imbalance between emotion classes, allowing the emotion recognition model to learn in a balanced way for each emotion

Table 2. Number of data by emotion.

Emotion	Collected Data	Final Data Used
Angry	7,126	3,412

 doi:10.20944/preprints202409.1178.v

Disgust	2,265	2,265
Fear	2,534	2,534
Happiness	3,412	3,412
Neutral	5,611	3,412
Sadness	15,074	3,412
Surprise	866	866
Total	36,888	19,313

4.2. Experimental Results

In this section, we propose a multimodal emotion recognition model using 19,313 text and speech data provided by AI-HUB. The data are divided into 17,381 training data and 1,932 test data, and the training and test data were split in a ratio of 9:1. The model was trained with the training data, and the performance was evaluated using the test data. In order to evaluate the performance of the proposed model, we compare the performance of the emotion recognition model using only text-only and speech-only models, the ensemble model that recognizes emotions by averaging the results of the two models, and finally the proposed multimodal transformer. First, the performance of each model was evaluated based on F1-Score and Accuracy, and the results are shown in Table 3. The text-only model and the speech-only model showed relatively low performance, and the ensemble model improved performance by combining the two models, but the proposed multimodal transformer model recorded the highest accuracy.

Table 3. Performance evaluation for each model.

Division	F1-Score	Accuracy
Text Data : KoELECTRA	0.6994	0.6951
Speech Data : MFCC + Pitch	0.4005	0.3012
Balanced Ensemble	0.7098	0.7075
Multimodal Transformer	**0.7344**	**0.7313**

Next, to compare the changes in Accuracy and Loss based on the Epoch, a graph visualizing the learning process of each model was added to Figure 3. This graph shows the performance changes during the learning process of the model, and in particular, it can be seen that the proposed multimodal transformer model converges quickly and maintains higher accuracy. In addition, the Loss graph visually confirms that the model is stably optimized during the learning process.

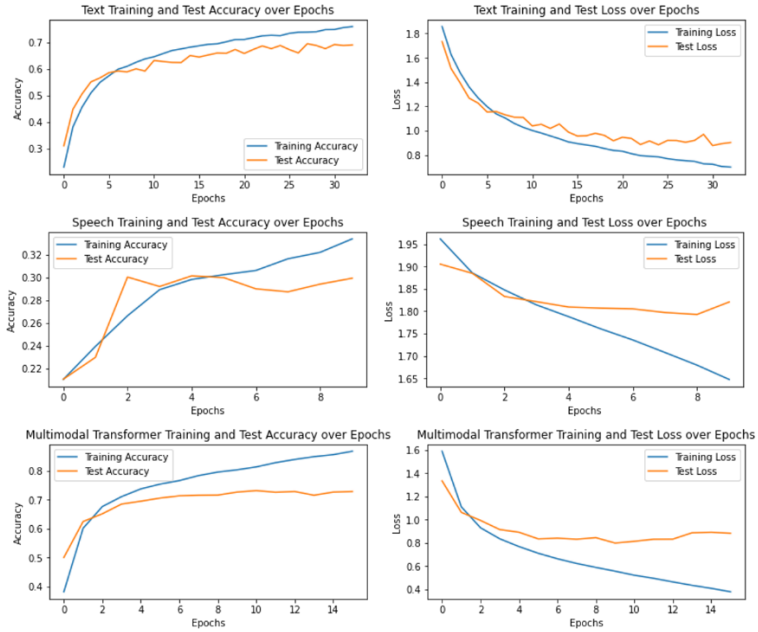

Figure 3. Accuracy and loss change graph based on epoch by model.

 doi:10.20944/preprints202409.1178.v

To emphasize that the proposed multimodal transformer model achieved the best performance, the confusion matrix that analyzes the performance of the model in detail for each emotion class is visualized in the form of a heatmap in Figure 4. This heatmap clearly shows how the model performed for each emotion, and in particular, it can be confirmed that the proposed model accurately predicts various emotion classes.

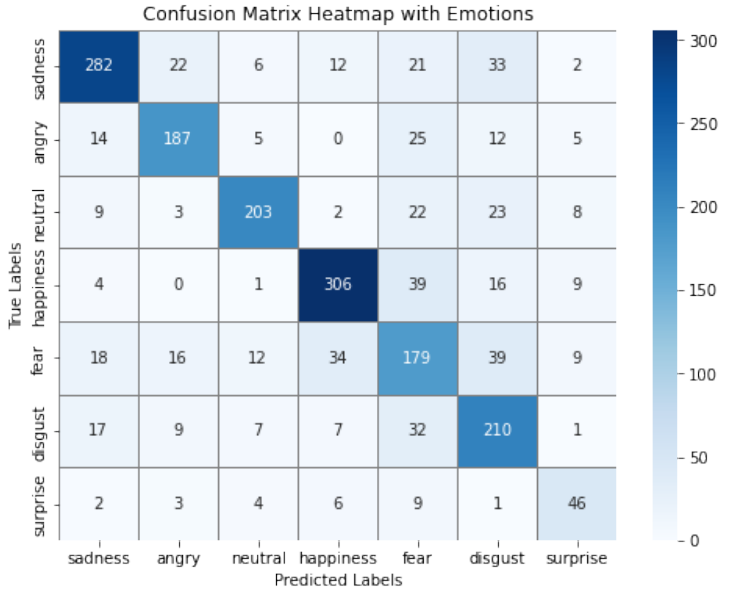

Figure 4. Confusion Matrix heatmap for each emotion class.

Finally, Figure 5 provides a bar graph that comprehensively visualizes the Accuracy, Precision, Recall, and F1-Score of each model, clearly showing that the multimodal transformer model achieved the highest performance. The proposed model shows excellent results in all performance metrics, which proves that the Cross-Modal Attention technique, which effectively learns the interaction between text and speech, played a significant role in improving the performance. These results suggest that the multimodal approach can significantly improve emotion recognition performance compared to single-modality models.

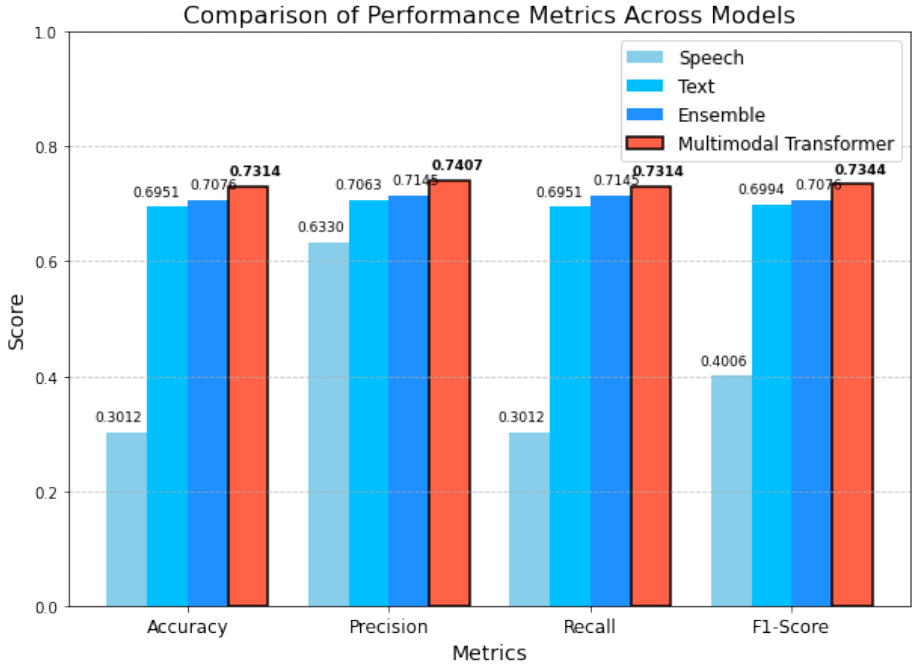

Figure 5. Comprehensive results of Accuracy, Precision, Recall, and F1-Score by model.

 doi:10.20944/preprints202409.1178.v

The experimental results show that text data has relatively low emotion recognition accuracy, and speech data shows no learning at all. On the other hand, the proposed multimodal transformer model significantly improves the performance by combining the strengths of both modalities. Compared with the Balanced Ensemble model, the multimodal transformer model also showed better performance, suggesting that the approach that considers the interaction between the two modalities greatly contributes to improving emotion recognition performance.

5. Conclusion

In this paper, we propose KoMPT: A Multimodal Emotion Recognition Model Integrating KoELECTRA, MFCC, and Pitch with a Multimodal Transformer model, which improves the performance of emotion recognition. The proposed model combines text embedding based on KoELECTRA and speech features using MFCC and Pitch, and effectively learns the interaction between text and speech using the Cross-Modal Attention mechanism. This allows us to achieve higher accuracy and efficiency in emotion recognition. The experimental results show that the multimodal approach outperforms the single modality model. In particular, the Multimodal Transformer model recorded higher accuracy and F1-Score than when text and speech data were used alone, and achieved an accuracy of 73.13% in emotion classification. This shows that the performance of emotion recognition can be significantly improved by combining complementary information from text and speech. In addition, the preprocessing process and model design that reflect the linguistic and phonetic characteristics of Korean enhanced the performance of emotion recognition. KoELECTRA provided an embedding that reflected the context of Korean well, and MFCC and Pitch successfully extracted the frequency components and intonation information of the speech, thereby improving the performance of speech emotion recognition. This study made an important contribution to research on multimodal emotion recognition based on Korean, and will be able to contribute to the development of emotion recognition technology that combines various language and modality data in the future. In future studies, it is expected that emotion recognition performance can be further improved by combining new modalities such as video data or applying more advanced multimodal combination techniques.

References

1. Xie, Z.; Guan, L. Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. In Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15–19 July 2013; IEEE, 2013; pp. 1–6.
2. Kalateh, S.; Estrada-Jimenez, L.A.; Pulikottil, T.; Hojjati, S.N.; Barata, J. The human role in Human-centric Industry. In Proceedings of the 48th Annual Conference of the IEEE Industrial Electronics Society (IECON 2022), Brussels, Belgium, 17–20 October 2022; IEEE, 2022.
3. Bahreini, K.; Nadolski, R.; Westera, W. Towards multimodal emotion recognition in e-learning environments. Interact. Learn. Environ. 2016, 24, 590–605.
4. Scherer, K.R.; Johnstone, T.; Klasmeyer, G. Vocal expression of emotion. In Handbook of Affective Sciences; Oxford University Press: Oxford, UK, 2003; pp. 433–456.
5. Bharti, S.K.; Varadhaganapathy, S.; Gupta, R.K.; Shukla, P.K.; Bouye, M.; Hingaa, S.K.; Mahmoud, A. Text-Based Emotion Recognition Using Deep Learning Approach. Comput. Intell. Neurosci. 2022, 2022, 2645381.
6. Kim, S.; Lee, S.-P. A BiLSTM–Transformer and 2D CNN Architecture for Emotion Recognition from Speech. Electronics 2023, 12, 4034.
7. Park, H. Enhancement of Multimodal Emotion Recognition Classification Model through Weighted Average Ensemble of KoBART and CNN Models. Korean Institute of Information Scientists and Engineers, 2023, pp. 2157-2159.
8. Kim, Y.-J.; Roh, K.; Chae, D. Feature-based Emotion Recognition Model Using Multimodal Data. Korean Institute of Information Scientists and Engineers 2023, 2157–2159.
9. Kim, Y.; Roh, K.; Chae, D. Feature-based Emotion Recognition Model Using Multimodal Data. Korean Institute of Information Scientists and Engineers 2023, 6, 2169–2171.
10. Mutinda, J.; Mwangi, W.; Okeyo, G. Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network. Appl. Sci. 2023, 13, 1445.
11. Li, H.; Ma, Y.; Ma, Z.; Zhu, H. Weibo Text Sentiment Analysis Based on BERT and Deep Learning. Appl. Sci. 2021, 11, 10774.

 doi:10.20944/preprints202409.1178.v

12. Reggiswarashari, F.; Sihwi, S.W. Speech emotion recognition using 2D-convolutional neural network. Int. J. Electr. Comput. Eng. 2022, 12, 6594–6601.
13. Hazra, S.K.; Shubham, M.; Kaushal, C.; Prabhakar, N. Emotion recognition of human speech using deep learning method and MFCC features. Radioelectron. Comput. Syst. 2022, 4, 161-172.
14. Poria, S.; Majumder, N.; Hazarika, D.; Cambria, E.; Gelbukh, A.; Hussain, A. Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines. IEEE Intell. Syst. 2018, 33, 17–25.
15. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multimodal Sentiment Analysis. arXiv 2017, arXiv:1707.07250.