감정 인지를 위한 음성 및 텍스트 데이터 퓨전: 다중 모달 딥 러닝 접근법

에드워드 카야디 [1], 송미화 [1]
[1] 세명대학교 스마트 IT 학부
e-mail : edw.chydi@gmail.com, mhsong@semyung.ac.kr

Speech and Textual Data Fusion for Emotion Detection: A Multimodal Deep Learning Approach

Edward Dwijayanto Cahyadi[1], Mi-Hwa Song[1]
[1]School of Smart IT , Semyung University

Summary

Speech emotion recognition(SER) is one of the interesting topics in the machine learning field. By developing multi-modal speech emotion recognition system, we can get numerous benefits. This paper explain about fusing BERT as the text recognizer and CNN as the speech recognizer to built a multi-modal SER system.

1. Introduction

Emotions are a fundamental aspect of human communication, profoundly influencing our interactions, decisions, and relationships. Recognizing and understanding these emotions is not only an important human ability but also a critical requirement in the development of artificial intelligence-based technology, such as virtual assistants and mental health support systems. Recently, deep learning algorithms have successfully addressed problems in various fields, such as image classification, speech recognition, text-to-speech generation, and other machine learning-related areas. Also, the use case of transformer-based system has been very popular in the last couple of years mostly in the area of text and image generation. These fundamental has motivated us to build a multi-modal emotion recognition machine learning system.

2. Related Research

S.Y et al[1]. Explains how speech and text processing can be done. It includes several steps. The first step is to build an Audio Recurrent Encoder. The second step is to build the Text Recurrent Encoder (TRE), And for the fusion model, the writer builds the Multimodal Dual Recurrent Encoder (MDRE) and Multimodal Dual Recurrent Encoder with Attention (MDREA). Convolutional Neural Networks (CNN) are a type of systematic neural network that successively consists of many layers. The CNN model typically consists of a SoftMax unit, several convolution layers, pooling layers, and fully linked layers. This sequential network creates an abstract model of the input using a feature extraction method. BERT is an open-source machine learning framework for

natural language processing (NLP). The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with question-and-answer datasets. BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based on their connection (In NLP, this process is called attention.)

3. Research Method

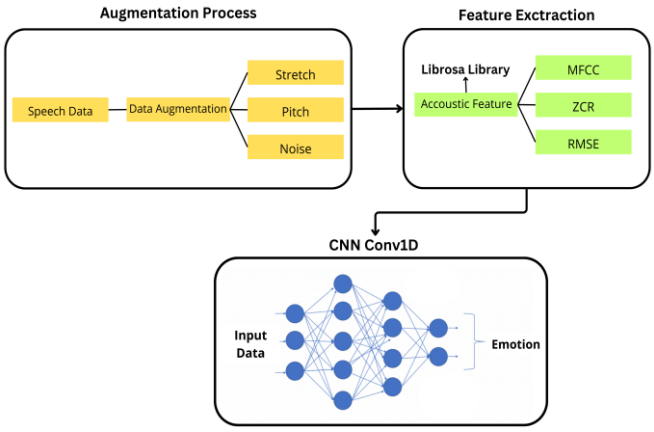

Figure 1. SER Proposed System Structure

This research uses a total of 12,242 Voices data images divided into 5 categories of emotions [happy, sad, angry, fearful, and surprised]. We used four types of different datasets, the first one is Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D), the second is Ryerson Audio-Visual Database (RAVDESS), the third one is Toronto emotional speech set (TESS), the last one is Surrey Audio-Visual Expressed Emotion (SAVEE), all data were obtained from Kaggle. To create the machine learning model, we use Python as the programming language, Jupyter Notebook, and tensor flow as the machine learning model library. Before we do the feature extraction, we have done an augmentation process first. The goal of this process is to give more variation of data to the model, the augmentation includes Noise, Pitch, and stretching audio augmentation. The features that we extract from the sound are ZCR (Zero Crossing Rate), RMSE (Root Mean Square Energy), and MFCC (Mel-Frequency Cepstral Coefficients). Using some function in the librosa library we extract the features and then we input it into the CNN model.

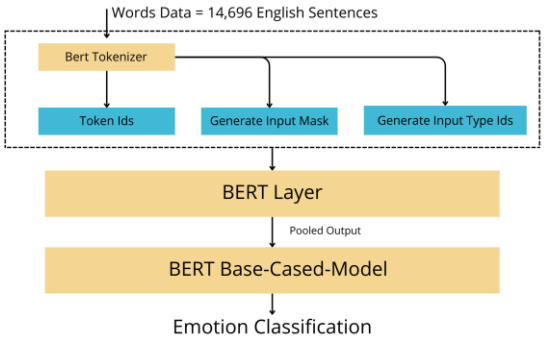

Figure 2. TER Proposed System Structure

For the text emotion recognition (TER) model we use the BERT pre-trained model to do the classification because this type of model already has the ability to convert text into some word embeddings and also it has the ability to recognize not just the words but the whole context of a sentence. We trained the model using 14,696 sentence data that has already been labeled with 5 types of emotions.

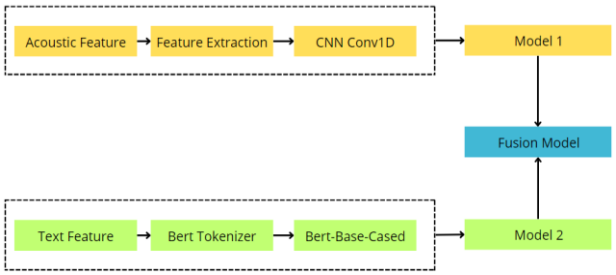

Figure 3. Fusion Model Proposed System Structure

Lastly, we build the late fusion model by creating another ANN (Artificial Neural Network) model that receives input from the SER and TER, it has been trained with features from the output of those 2 model also we add some features like the length of the soundtrack. The model consists of 3 dense layers and has the type of sequential model.

Metrics / Model	Precisions	Recall	Accuracy	F1
SER	84.16%	65%	65%	67.7%
BERT	91.4%	80%	80%	83%
Fusion Model	92%	90%	90%	88%

Figure 4. Model Result Comparison

As we see in the table above, the SER model has the least effective performance compared to other models, followed by the TER model, and lastly, the fusion model has the most effective performance. This happened because in some conditions SER model has some difficulty translating the signals coming from the data, of course This will also affect the features and the result of the model, but on the other hand, TER can still get the result right because it translate the sound into words first and classified it. But in some cases, SER can be more effective than TER, The TER model has a dependency on the speech-to-text recognizer system if this error then the input also will be disturbed, but the signal from sound will remain the same and SER model can outperform TER in this case. And that is why the fusion model can achieve the best performance of all other models. Fusion model also has the ability to enhanced accuracy and robustness, combining words with audio data can provide a more understanding of the speaker's emotional state.

4. Conclusion

In this paper, we propose a multimodal speech emotion recognizer using 2 different inputs, text, and sound signal. By combining CNN and BERT model we create a late fusion model. Extensive experiments show that our proposed model outperforms another model with an accuracy of 90% while the SER model has the least effective performance with 65% accuracy followed by the BERT model with 80% accuracy. By combining both textual and audio data, the SER system can more accurately handle scenarios where acoustic cues alone may not suffice, such as sarcasm or situations where audio and text convey different emotions, resulting in a more context-aware and effective system.

Reference

[1] Seunghyun Yoon, Seokhyun Byun, Kyomin Jung. "Multimodal Speech Emotion Recognition Using Audio and Text", pp. 1-7, 2018.

[2] Eu Jin Lok, "Surrey Audio-Visual Expressed Emotion (SAVEE)",Kaggle,https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee

[3] Eu Jin Lok, " Crowd Sourced Emotional Multimodal Actors ,Kaggle,https://www.kaggle.com/datasets/ejlok1/cremad

[4] Steven R.Livingstone, " RAVDESS Emotional speech

audio”Kaggle,https://www.kaggleg/ravdessemotionalspeecaudio
[5] Eu Jin Lok, “ Toronto emotional speech set (TESS),Kaggle,https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess