음성을 통한 감정 해석: 감정 인식을 위한 딥 뉴럴 네트워크 예비 연구

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

Unraveling Emotions in Speech: Deep Neural Networks for Emotion Recognition

Edward Dwijayanto Cahyadi, 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 SER, we can get numerous benefits. By using a convolutional neural network and Long Short Term Memory (LSTM) method as a part of Artificial intelligence, the SER system can be built.

1. Introduction

Our Human emotions are complex and diverse, playing a crucial role in our daily lives and influencing our behavior, decision-making, and interactions. Speech Emotion Recognition is a task of computational paralinguistics and speech processing that seeks to identify and classify the emotions expressed in spoken language. The objective is to infer a speaker's emotional state such as joy, rage, grief, or frustration from their speech patterns, including prosody, pitch, and rhythm. Emotion recognition using DNNs is a cutting-edge research area that holds great promise for various applications, from human-computer interaction and affective computing to mental health diagnosis. According to Taiba Majid Wani (2021), the research team explains how speech processing can be done. It includes several steps, preprocessing, farming, normalization, noise reduction. Long Short-Term Memory (LSTM) is made up of three gates (forget, input, and output) and one cell state. The input gates determine what fresh information to remember, the forget gate determines which information from previous inputs to forget, and the output gate determines which portion of the cell state to output. Convolutional Neural Networks (CNN) are a type of systematic neural network that consists of many layers. The CNN model typically consists of a SoftMax unit, several convolution layers, pooling layers, and fully linked layers.

2. Research Method

This research uses a total of 12,162 Voices data images divided into two categories of gender and emotions [neutral, happy, sad, angry, fearful, disgusted, and surprised]. We used four types of different datasets, and the first one is Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) (consist of 7,442 clips of voice), the second is Ryerson Audio-Visual Database (RAVDESS) consist of 1,440 clips of voice, the third one is Toronto emotional speech set (TESS) consist of 2,820 clips of voice, the last one is Surrey Audio-Visual Expressed Emotion (SAVEE) consist of 480 clips of voice.

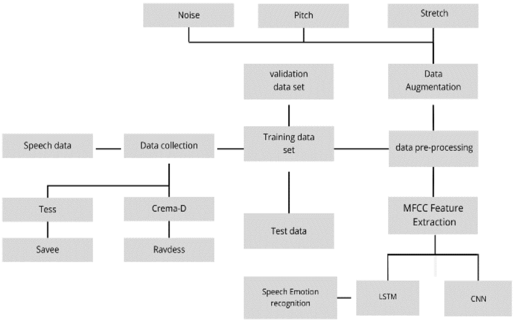

Figure 1. Proposed System Structure

To create the machine learning model, we use Python as the programming language and Google Collaboratory as the compiler. According to the previous study, the CNN and LSTM algorithm performs well in classifying voice data based on emotions and gender. Based on that, we tried to make a CNN and LSTM machine-learning model. Before building the model, we have done data augmentation process.

First, as we are working with four different datasets, we need to create a data frame storing all emotions of the data in the data frame with their paths, and we will use this data frame to extract features for our model training. The next process is to do data augmentation, and data augmentation is the process by which we create new synthetic data samples by adding small perturbations to our initial training set. To generate syntactic data for audio, we apply noise injection, shifting time, and changing the pitch. The objective of this process is to make our model invariant to those perturbations and enhance its ability to generalize, and we use the Librosa function to do three data augmentation. Extraction of features is a very important part of analyzing and finding relations between different things.

As we already know that the data provided by audio cannot be understood by the models directly, so we need to convert them into an understandable format in which feature extraction is used. In this experiment, we extract Mel Frequency Cepstral Coefficients to be our feature. Then, we split the data into three parts. 70% is for training data sets,15% is for the validation data, and the last 15% is for the testing data.

3. Result

For the LSTM algorithm model, we set the epoch to 100, and the percentage of accuracy in the test data that we received is 62.407%, 0.6806 of F1 score, 0.5533 of Recall, and 0.6806 of Precision. On the other hand, we also build a model using the CNN algorithm, and the result of the accuracy is 96.526%, 0.9653 of F1 Score, 0.9643 of Recall, and 0.9664 of Precision.

Hyperparameter	LSTM	CNN
Batch size	65	64
Activation	Relu	Softmax, Relu
Optimizer	Adam	Rmsprop
Epoch	100	50
Dropout	No	No
Loss	Mean squared error	Categorical crossentropy

Figure 2. LSTM and CNN Hyperparameter

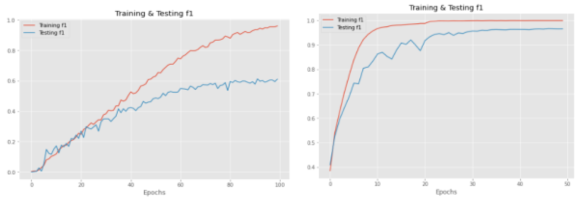

Figure 3. LSTM and CNN F1 Score

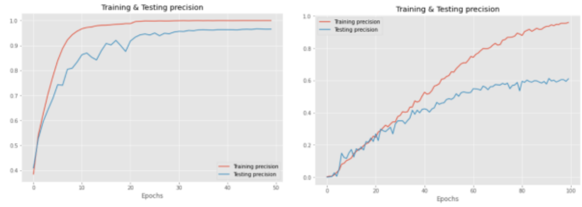

Figure 4. LSTM and CNN Precision

Method	Accuracy	Precision	Recall	F1 Score
CNN	96.526%	0.9664	0.9643	0.9653
LSTM	61.018%	0.6806	0.5533	0.6806

Figure 5.CNN and LSTM Model Comparison

4. Limitation

Human emotion is something that is really complex and sometimes can be very subjective. Although deep learning can make good classification regarding the labels, it still can't completely understand the meaning of human emotion and how to precisely tell the specific one. A lot of datasets and good computation power are needed to build a good deep learning model for a speech emotion recognition system.

5. Conclusion

The intricacy and variety of human emotions are pivotal in our everyday existence, shaping our actions, choices, and social interactions. Speech Emotion Recognition, a field within computational paralinguistics and speech processing, aims to detect and categorize the emotions conveyed through spoken language. Speech emotion recognition can be built through traditional methods of machine learning or modern deep learning. LSTM and CNN are one of the deep learning methods that can solve the speech emotion recognition problem. Through an experiment, both of the algorithms work well on this topic, where LSTM gains 61.018% of accuracy, and on the other hand, CNN gets a better result at 96.526% of accuracy.

Reference

[1] Taiba Majid Wani.Teddy Surya Gunawan.Syed Asif Ahmad Qadri.Mira Kartiwi “A Comprehensive Review of Speech Emotion Recognition Systems”, pp. 2-3, 2021.

[2] Eva Lieskovská. Maroš Jakubec. Roman Jarina. Michal Chmulík “A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism”, pp. 4-7, 2021.

[3] Javier de Lope. Manuel Graña, “An ongoing review of speech emotion recognition”, pp. 3-7, 2023.

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

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

[6] Steven R.Livingstone, “ RAVDESS Emotional speech audio”Kaggle,https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

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