A Multi-modal Approach to Speech Emotion Recognition

Jiayu Zhou, Paras Sibal, Rohit Bernard, Sanjana Mallikarjuna, Swapnil Mallick
University of Southern California
{zhoujiay,psibal,rpbernar,moodbagi,smallick}@usc.edu

1 Motivation

Speech Emotion Recognition (SER) attempts to identify emotions from speech irrespective of the semantic content. The primary objective of SER was to improve the man-machine interface. This objective motivates the use of SER. The enhancement in SERs made them widely used for monitoring psychophysiological state, improving customer service, and forensics.

Recognition of emotions in speech presents several challenges. The emotions are very subjective, and sometimes difficult for humans to identify appropriate emotions expressed in normal communication. And, the ability to automate such a process is still an ongoing research subject. This project aims to design a multi-model on labeled data instead of deep learning models used by existing research.

1.1 Novel Contributions

Firstly, in addition to deriving features directly from speech, we will convert speech-to-text and then use text semantics as part of features. Secondly, we compare the two models to check for a better result. Lastly, we aim to merge the two results to design a multi-modal.

2 Design

2.1 Materials

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) has been envisaged as our source of data. The RAVDESS (Livingstone and Russo, 2018) is a multimodal dataset of emotional speech and songs. The human speech corpus encompasses basic emotions such as calmness, happiness, anger, sadness, fear, surprise and disgust whereas songs consist of emotions like calmness, anger, happiness, sadness and fear. Every expression in the dataset has two levels of emotional intensity along with a neutral expression.

2.2 Methods

In the development of a SER, preprocessing is the crucial phase to tune the audio data. The raw audio data is sampled and padded with zeros to shorter audio clips for uniform standardized dimensions. The mel spectrograms, generated using tuned data clips, are converted to mel scales and normalized by the mean and variance so that the perpetual frequencies difference becomes smaller. Further, we may compress them into a short-time spectrum to extract only the essential coefficients.

With generated mel spectrograms, we plan to design two basic models. The first model takes spectrograms as input and then classifies emotions using CNN layers or biLSTM. The second model extracts text from the spectrogram and then classifies emotions from the text semantics. In this model, audio is transformed to text using the HMM or DL methods. Further, we will use biLSTM layers to get character features and classify emotions using vectorized texts from word embedding layers and pre-trained word vector space. We will experiment with different methods to merge these two models by concatenating the character features or using an extra model to fuse the results with multi-sensor data fusion techniques.

2.3 Baselines and Evaluation Protocols

Baselines We aim to compare the performance of our model against other research which use techniques such as LSTM's (Sahu, 2019), RNN's (Yoon et al., 2018), and Deep CNN's (Issa et al., 2020). It is important to note that different research consider different emotions for classification, and to make a fair comparison, we must consider the same emotions as each research, respectively.

Evaluation Protocols We aim to use supervised learning techniques using cross-validation, since the size of the dataset isn't very large. Our model will generate a multi-class classification, the results of which will be reported using a confusion matrix, along with the F1 score.

3 Timeframe and Division of Work

3.1 Timeframe

We will analyze the data and create a pipeline to preprocess it in the first week. And then, we will build a pipeline to train the model and evaluate the result from the model. A simple model will be designed to test the quality of the pipeline. We may use another two weeks to build more models and try different methods. If both models can get a good result, we will design a fusion model that will work as a multi-model for our project. We may use another data source to compare the generalization performance of our models. Meanwhile, we will start working on the final report.

3.2 Difficulties and Reactions

Since other researchers have achieved satisfactory results in this domain, the project itself is feasible but we need to work towards improving the model performance. Reading research papers and handling model design difficulties will be our priorities. We can adjust our models first to see if there are some improvements. If it doesn't work, we may change the data source.

References

Dias Issa, Muhammed Demirci, and Adnan Yazici. 2020. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59(101894).

Steven R. Livingstone and Frank A. Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, 13.

Gaurav Sahu. 2019. Multimodal speech emotion recognition and ambiguity resolution. *CoRR*, abs/1904.06022.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. *CoRR*, abs/1810.04635.