Speech Emotion Recognition:

A Comparative Study

Tushar Choudhary
2K18/EE/225

Abstract

The pursuit of capturing and classifying the emotional state of the person from speech utterances is one of the centerpieces of research in the systems of human-machine interaction (HMI). The study bases itself upon comparing the hybrid architectures of CNN from a custom-designed CNN model for the task of speech emotion recognition. This project presents three different architectures with a lens of comparison. The three architectures used are a CNN architecture, a hybrid system of CNN and KNN architecture, and a hybrid cascaded system of CNN and LSTM.

I. Introduction

Emotion recognition from speech utterances is one of the primary components of human-computer interaction today. It has been a wide area of research for more than two decades now. systems were developed and tested. Many systems were developed and tested but their performance couldn't be improved over a limit. This was because the traditional features for speech weren't capable of detecting complex details like emotional states [1]. With the advent of disruptive deep learning techniques for speech emotion recognition and feature extraction, many performance barriers were overshadowed. Some of those techniques have stuck around while the new ones keep shedding more light on how to improve the performance of speech emotion recognition models.

II. THEORY

Various deep learning techniques have been proposed over the years for recognizing the emotion of the speaker from speech utterances. The most popular ones are the deep neural network, convolution neural network, and recurrent neural network (more specifically LSTM) architectures [2]. Let us take a look at some of the features of CNN and LSTM since we'll be using them in our project.

A. Convolution Neural Network

A Convolution Neural Network, also known as ConvNet, is one of the most popular variants of the feed-forward deep learning architecture which has the ability to assign learnable weights and biases to various aspects in the image and differentiate between them. It uses the mathematical convolution operation on the inputs as opposed to the affine transformation performed by deep neural networks. CNNs provide a higher level of feature extraction with a lesser need for data pre-processing.

The convolution layer can be mathematically written as [4]

$$(h_k)_{ij} = (W_k * q)_{ij} + b_k,$$

Where, $(h_k)_{ij}$ denotes the (i,j) element of the kth output feature map, q represents the input feature maps, W_k and b_k denote the kth filter and bias, respectively.

A ConvNet enables the extraction of Spatial and Temporal dependencies in an image with the help of convolutional filters. The local connections in the architecture are built upon filters that are convolved with the input and have the same weight and bias, that are each of size a-b-1. The Convolutional layer computes the dot products between the provided inputs and the weights. So, the parameters for weight W_i and biasing n_i for generation of maps z_i for i features with sizes a−b−1 can be given as [5]

$$z_i = g\,(W_i * r + n_i)$$

B. Long Short Term Memory

RNN suffers from the problem of vanishing gradients, which means that it cannot learn long-term dependencies. In order to rectify the vanishing gradient problem, endless efforts were employed and LSTM is one such solution.[6] They have a more complex cell structure than a normal recurrent neuron, which helps them to better control how they learn and forget from various input sources.

The cell state (cell memory), the horizontal line

running through the top of the diagram through which information flows, and the internal mechanism called gates that can control the flow of information. These are the two most important aspects of LSTMs. The LSTM cell's output is the hidden state, which is used for prediction. It includes previous input information (from cell state/memory) as well as current input information (decided according to which context is important).

III. Project Workflow

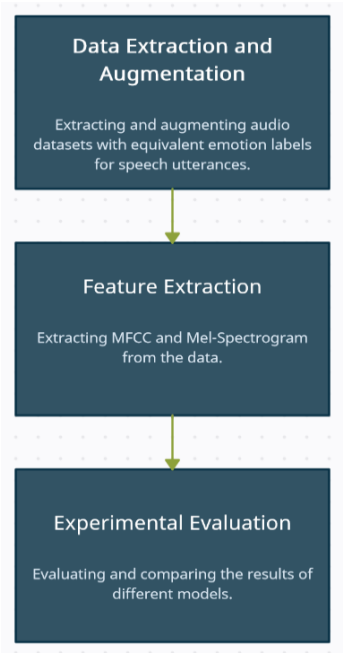

Fig. Project Workflow

IV. Data Description

We have used the augmentation of two of the most renowned speech emotion datasets: RAVDESS and TESS.

A. RAVDESS

This dataset includes 1440 speech files and 1012 Song files from the recordings of 24 professional actors, vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each file was rated 10 times on emotional validity, intensity, and genuineness.

B. TESS:

A set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total. Two actresses were recruited from the Toronto area.

V. Feature Extraction

Continuous features, qualitative features, and spectral features are some of the most common acoustic features used to recognize speech emotion. To recognize speech emotion, many aspects have been explored. Although several researchers weighed the benefits and drawbacks of each characteristic, no one has been able to determine which group is the best until now. [7]

Once we have the datasets and we're done with preprocessing, we will extract sound features from the data points using the librosa library.

A. Mel Spectrogram

A spectrogram is a representation of a signal (e.g. an audio signal) that shows the evolution of the frequency spectrum in time. The mel scale is a non-linear transformation of frequency scale based on the perception of pitches. The mel scale is calculated so that two pairs of frequencies separated by a delta in the mel scale are perceived by humans as being equidistant. Mel Spectrogram gives us a visualization of frequency spectrum on the mel scale.

B. Mel Frequency Cepstral Coefficients

MFCCs are coefficients that represent audio based on perception with their frequency bands logarithmically positioned and mimics the human vocal response. The MFCC feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT.

VI. Experimental Evaluation

In the project, we have compared the performance of three techniques on our models. These three techniques are:

1. A Custom-Designed CNN Model
2. A CNN+KNN Model
3. A CNN+LSTM Hybrid Model

A. CNN Model

In this model, I have designed a CNN architecture from scratch. It has three 1D convolutional layers with 64, 128, and 256 filters

respectively. The optimizer used in the model is RMSprop with a learning rate of 0.00005.

conv1d_6 (Conv1D)	(None, 168, 64)	
activation_8 (Activation)	(None, 168, 64)	
dropout_6 (Dropout)	(None, 168, 64)	
max_pooling1d_4 (MaxPooling1	(None, 42, 64)	
conv1d_7 (Conv1D)	(None, 42, 128)	
activation_9 (Activation)	(None, 42, 128)	
dropout_7 (Dropout)	(None, 42, 128)	
max_pooling1d_5 (MaxPooling1	(None, 10, 128)	
conv1d_8 (Conv1D)	(None, 10, 256)	
activation_10 (Activation)	(None, 10, 256)	
dropout_8 (Dropout)	(None, 10, 256)	
flatten_2 (Flatten)	(None, 2560)	
dense_4 (Dense)	(None, 64)	
dense_5 (Dense)	(None, 8)	
activation_11 (Activation)	(None, 8)	

Fig. CNN Model Architecture

The CNN model trained fairly well and fast. The results obtained while training the CNN model are given below.

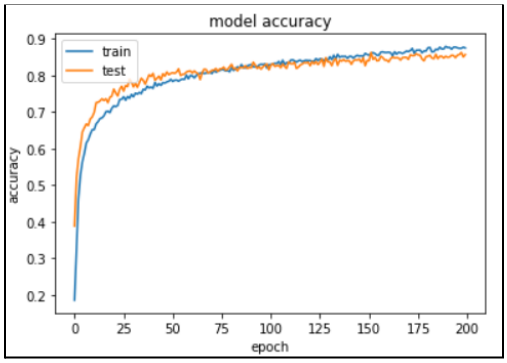

Fig. Model Accuracy

	precision	recall	f1-score	support
0	0.82	0.91	0.86	114
1	0.81	0.79	0.80	85
2	0.93	0.84	0.88	164
3	0.81	0.86	0.84	153
4	0.88	0.88	0.88	150
5	0.92	0.80	0.85	164
6	0.85	0.86	0.85	108
7	0.81	0.92	0.86	113
accuracy			0.86	1051
macro avg	0.85	0.86	0.85	1051
weighted avg	0.86	0.86	0.86	1051

Fig. Classification Report

B. CNN+KNN Model

The inspiration of this model arose from the system proposed in [8]. In this model, we extract the features from the second last dense layer of the previously designed CNN architecture. Then we use those features to train a KNN model with the number of neighbours equal to the number of emotion labels i.e. 8 number of neighbors

Here is the classification report of the CNN+KNN hybrid system.

	precision	recall	f1-score	support
0	0.82	0.89	0.85	114
1	0.70	0.84	0.76	85
2	0.84	0.79	0.82	164
3	0.81	0.80	0.81	153
4	0.89	0.89	0.89	150
5	0.84	0.80	0.82	164
6	0.87	0.84	0.85	108
7	0.86	0.85	0.85	113
accuracy			0.83	1051
macro avg	0.83	0.84	0.83	1051
weighted avg	0.84	0.83	0.83	1051

Fig. Classification Report

C. CNN+LSTM Model

A convolutional neural network being a feed-forward network, filters spatial data whereas the recurrent neural network (LSTM) feeds data back into itself.[6]

In this model, I have designed a cascaded architecture of CNN and LSTM respectively from scratch. It has two 1D convolutional layers with 64 and 128 filters respectively followed with an LSTM layer with 64 units. The optimizer used in the model is SGD with a learning rate of 0.0001 with a decay of 1e-6.

conv1d_14 (Conv1D)	(None, 168, 64)	
activation_20 (Activation)	(None, 168, 64)	
dropout_16 (Dropout)	(None, 168, 64)	
max_pooling1d_14 (MaxPooling	(None, 42, 64)	
conv1d_15 (Conv1D)	(None, 42, 128)	
activation_21 (Activation)	(None, 42, 128)	
dropout_17 (Dropout)	(None, 42, 128)	
max_pooling1d_15 (MaxPooling	(None, 10, 128)	
lstm_7 (LSTM)	(None, 64)	
dense_8 (Dense)	(None, 8)	
activation_22 (Activation)	(None, 8)	

Fig. CNN+LSTM Model Architecture

The CNN+LSTM model trained faster than the CNN model but was less stable and accurate comparatively. The results obtained while training the CNN+LSTM model are given below.

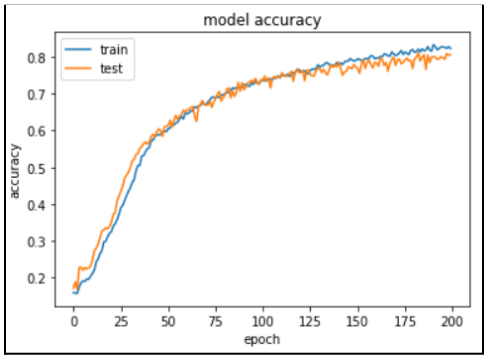

Fig. Model Accuracy

	precision	recall	f1-score	support
0	0.97	0.82	0.89	114
1	0.62	0.92	0.74	85
2	0.80	0.80	0.80	164
3	0.85	0.72	0.78	153
4	0.82	0.87	0.84	150
5	0.78	0.73	0.75	164
6	0.89	0.79	0.83	108
7	0.77	0.88	0.82	113
accuracy			0.81	1051
macro avg	0.81	0.82	0.81	1051
weighted avg	0.82	0.81	0.81	1051

Fig. Classification Report

VII. Conclusion

Three methods involving deep learning techniques for speech emotion recognition were compared on the basis of their performance to detect each of the eight emotions considered in the project. Here are the conclusions of the project:

The model with the highest overall performance was the CNN model with an accuracy of 85%, followed with the CNN+KNN model with an accuracy of 83%. The least accurate model was the one with the CNN+LSTM architecture.

Each model performed better than the others for different sets of emotions. The bifurcations of the results might be helpful for developing Speech Emotion Recognition Models with different applications.

The CNN architecture gave out the best results for the emotion labels calm, happy, and fear.

The CNN+KNN model gave out the best results for the emotion labels angry and surprised.

The CNN+LSTM architecture gave out the best results for the emotion labels neutral, sad, and disgust.

References

[1] 1] S. Zhang, S. Zhang, T. Huang, and W. Gao, “Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching,” *IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576–1590, 201*

[2] Pandey S, Shekhawat H, Prasanna S, “Deep learning techniques for speech emotion recognition: A review,” in the proceedings of *2019 29th International Conference Radioelektronika, RADIOELEKTRONIKA 2019 - Microwave and Radio Electronics Week, MAREW 2019, (2019), 177-183*

[3] W. Zheng, J. Yu, and Y. Zou, “An experimental study of speech emotion recognition based on deep convolutional neural networks,” in Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on. IEEE, 2015, pp. 827–831.

[4] H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for speech emotion recognition,” Neural Networks, vol. 92, pp. 60–68, 2017.

[5] Khalil R, Jones E, Babar M, *et al.*, “Speech Emotion Recognition Using Deep Learning Techniques: A Review,” IEEE Access, (2019), 7, doi: 10.1109

[6] Qazi H, Kaushik B, “A Hybrid Technique using CNN+LSTM for Speech Emotion Recognition,” in International Journal of Engineering and Advanced Technology, (2020), 9(5)

[7] Zhao J, Mao X, Chen L, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks” in Biomedical Signal Processing and Control, (2019), 47

[8] Akila, A., Umamaheswari, J., “An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN” in the proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019