Emotion Recognition in Human Speech

Tarun Ravada

1 Introduction

Speech is a dense form of communication that can convey information effectively. It mainly consists of two types of information, linguistic and para-linguistic. Linguistic information refers to the verbal content of the speech, where as para-linguistic information refers to the implicit information embedded in speech such as tone, speaking style, emotion, and intent of the speaker [1]. The expression conveyed in speech can change with changes in the para-linguistic information. As such it is crucial to understand both the linguistic and para-linguistic information of speech in order to comprehend the message being conveyed.

Speech Emotion Recognition is a machine learning problem tasked with extracting the para-linguistic information in speech, specifically the emotion. Such information carries a lot of value in voice based systems such as virtual assistants and language translation systems. Although several advances have been made in speech recognition systems, emotion recognition remains a challenging task due to the complexity of emotional expressions [2]. However, with improvements in computational power and available data it is now possible to recognize the emotion in speech with a certain degree of accuracy.

This study explores a Convolution Neural Network approach to train a model that can classify different emotions in speech, and compare the effects of feature selection and data size on the accuracy of the model. The final model is able to classify 7 different emotions - calm, happy, sad, angry, fearful, disgust, and surprised with an overall accuracy of 74%.

2 Methods

2.1 Data

The data for this study was obtained from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [1].

The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes neutral, calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. The data is available in 3 different modality formats, audio-only, audio-video, and video-only.

[1] "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" by Livingstone & Russo is licensed under CC BY-NA-SC 4.0. `https://zenodo.org/record/1188976`

Emotion	No. of files
Neutral	96
Calm	192
Happy	192
Sad	192
Angry	192
Fearful	192
Surprise	192
Disgust	192
Total	1440

Table 1: Class distribution of different emotions in RAVDESS speech

For the purpose of this study the speech audio-only portion of RAVDESS was used. This consists of 1440 files, 60 trials per actor x 24 actors. The class distribution of the audio-only data can be see in in Table 1. The calm and neutral emotions are both baseline emotions [3]. The data is uniformly distributed across all classes except the calm class. Since both the neutral and calm class are similar the neutral class was not used in this study, to maintain class balance. Figure 1 shows the waveform of a sample audio file from the data-set.

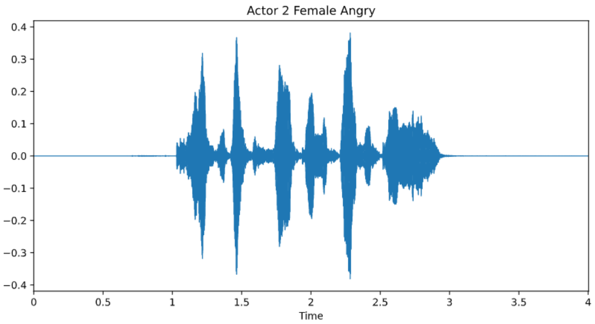


Figure 1: Audio waveform - Female-Angry-Actor 2

2.2 Feature Selection and Extraction

In order to perform learning and classification on the speech data, relevant features must be extracted from the data. These features will serve as the input to the learning and classification model. This study focuses on the use of two types of features that can be used to represent audio, namely Mel-Frequency Cepstral Coefficients and Log Mel spectograms.

The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC.

Mel-spectogram is a time vs frequency representation of the audio signal on a Mel-scale. The logarithmic form of mel-spectogram is used since humans perceive sound in logarithmic scale.

Librosa [2] is a python package for music and audio analysis. Librosa's feature extraction methods were used to extract MFCCs and Log Mel-spectograms from all the audio files. Vectors consisting of 40 features were generated from the MFCCs and Log Mel-spectograms, which were used as input to the learning algorithm. Figure 2 shows the generated mfccs for the audio waveform shown in Figure 1.

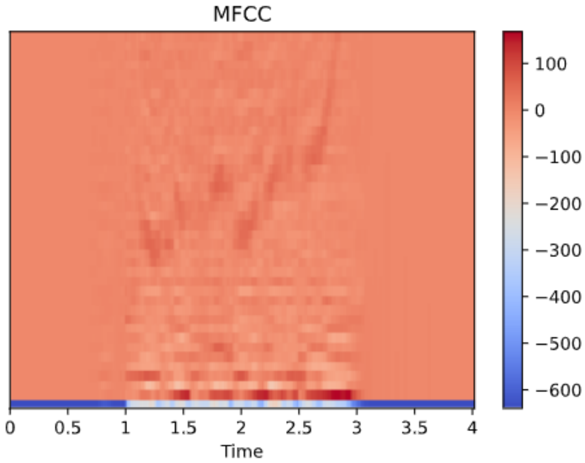


Figure 2: 40 MFCCs generated for audio file Female-Angry-Actor 2

2.3 Model Selection

In order to understand the training complexity of the data and to develop a baseline, 3 different off the shelf classifiers - Gaussian Naive Bayes classifier, Decision Tree Classifier, and Multi Layer Perceptron classifier were trained on the input features and their accuracies were assessed, as shown in Table 2.

Classifier	MFCC	Mel-spectrogram
Naive Bayes	0.37	0.25
Decision Tree	0.43	0.31
MLP	0.61	0.47

Table 2: Baseline Accuracies

The Multilayer Perceptron architecture showed the best performance. This is expected, since multilayer back-propagation networks are able to learn complex, high-dimensional, non-linear

[2] Librosa version 0.8.0 `http://doi.org/10.5281/zenodo.3955228`

mappings from data. This makes them excellent candidates for image and speech recognition [4].

Based on the findings from the baseline architectures a 1D Convolution Neural Network architecture was chosen for the classification task. Different layer sizes, kernel sizes, model depth, and regularization techniques were explored which resulted in the final model shown in Figure 3.

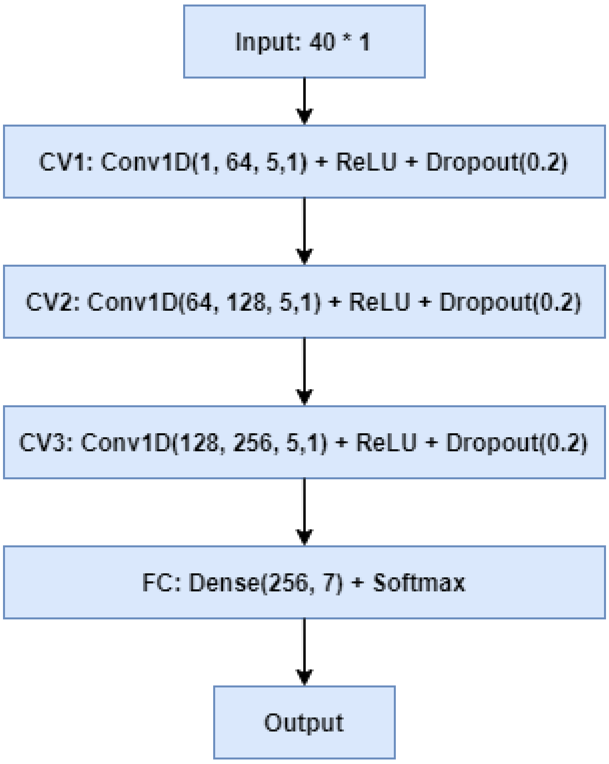

Input: 40 * 1
CV1: Conv1D(1, 64, 5,1) + ReLU + Dropout(0.2)
CV2: Conv1D(64, 128, 5,1) + ReLU + Dropout(0.2)
CV3: Conv1D(128, 256, 5,1) + ReLU + Dropout(0.2)
FC: Dense(256, 7) + Softmax
Output

Figure 3: 1D CNN Model Architecture

2.4 Data Augmentation and Training

The data for the 7 chosen emotions contribute to 1344 audio files usable for training and testing. The 1344 original files were split into 336 test files and 1008 train files. The data was shuffled and a stratified split was performed in order to maintain uniform class distribution in test and train data.

Neural Network models work better with more input data, and less input data can lead to over-fitting on the training data. Since this study is focused on using the RAVDESS dataset, rather than introduce new input data, existing data was augmented to increase the input data size.

Data augmentation techniques considered were pitch modulation, time shift, and waveform stretch [3]. The presence of pitch modulated data did not affect the performance of the CNN model, and therefore this augmentation technique was excluded to save training time.

The final train data consisted of the original 1008 audio files, 1008 audio files time shifted by 1 second, 1008 audio files stretched by a factor of 0.5, and 1008 audio files stretched by a factor of 1.5 giving a total train dataset size of 4032. The feature extraction techniques discussed in Section 2.2 were applied to the train and test files to generate the MFCC and Log Mel-spectrogram train and test input vectors.

The training data was further split into training and validation data to perform parameter tuning on the model. The model was trained using a batch size of 16 over a maximum of 100 epochs. Categorical cross entropy loss was used as the loss function and Adam optimization was used. Early stopping techniques were employed and training was terminated early if the validation loss was stagnant over 5 epochs.

3 Results and Discussion

A total of 4 different models were trained for comparison. Two models were trained using the MFCCs as input features, one on the original dataset and one on the augmented dataset. Similarly, two models were trained using the Log Mel-spectrogram features as input.

	Original Dataset	Augmented Dataset
Train accuracy	0.96	0.99
Test accuracy	0.68	0.74

Table 3: MFCCs Model performance on original data vs augmented data

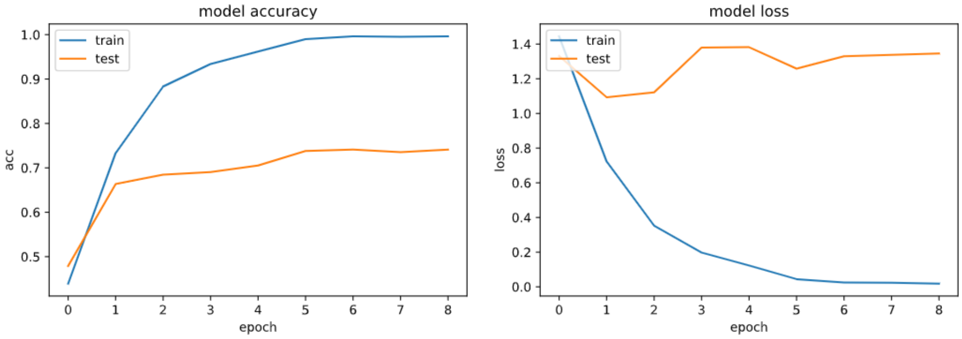
Table 3 shows the accuracy of the MFCCs model on the original dataset and the augmented dataset. Adding additional inputs improved the performance of the model, however the improvement in performance was not in proportion with the increase in dataset size.

	MFCC	Log Mel-spectogram
Test Accuracy	0.74	0.64

Table 4: Comparison of MFCC vs Log Mel-spectogram as input features

The architecture saw better performance on both the original and augmented dataset when using MFCCs as input features. Perhaps a different model architecture would be suited to using Log Mel-spectogram as inputs, since it is continuous in nature, unlike the MFCCs which are discrete coefficients.

[3] The techniques used are discussed in this Medium article `https://medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6`



(a) Train/Validation accuracy

(b) Train/Validation Loss

Figure 4: Best Model - MFCC on augmented data

Of the 4 models, the model using the MFCCs as input, trained on the augmented dataset provided the best overall test accuracy. Figure 4 shows the train/valiation accuracy and loss for this model.

Table 5 shows the performance of the model on individual class predictions. The model was good at identifying the emotions calm, and angry with at least 80% accuracy. Strong emotions such as anger and disgust have defining features in speech, which are particularly easy for humans to recognize. However, the results of this classification show that the model does not classify emotions similar to humans.

Emotion	Precision	Recall	F1-Score
Calm	0.78	0.90	0.83
Happy	0.65	0.75	0.70
Sad	0.81	0.62	0.71
Angry	0.74	0.88	0.80
Fearful	0.75	0.75	0.75
Disgust	0.76	0.54	0.63
Surprised	0.72	0.75	0.73
accuracy			0.74
macro avg	0.75	0.74	0.74
weighted avg	0.75	0.74	0.74

Table 5: Model accuracy on individual class predictions

4 Conclusion

This study examined the performance of a multilayerd CNN model on the task of Speech Emotion Recognition. Two different input features were considered, Mel Frequency Cepstrum coefficients and Log Mel-spectograms. The model performed significantly better when using MFCCs as the input feature. This could be attributed to the fact that MFCCs are discrete coefficients and are easily representable, whereas Log Mel-spectograms are not discrete features and might require a different model architecture for good performance. Data augmentation techniques were explored and employed to increase the size of the training set. This increased the model performance from 68% to 74%. Despite the addition of augmented data, the model continues to overfit on training data. Overfitting can be regularized by adding new data and retraining the model. However, as the scope of this study is to utilize the RAVDESS data set, the option to add additional data was not explored.

References

[1] Y. Yamashita, "A review of paralinguistic information processing for natural speech communication," *Acoustical Science and Technology*, vol. 34, no. 2, pp. 73–79, 2013.

[2] S. Suganya and E. Y. A. Charles, "Speech emotion recognition using deep learning on audio recordings," in *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, vol. 250, 2019, pp. 1–6. DOI: `10.1109/ICTer48817.2019.9023737`.

[3] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, pp. 1–35, May 2018. DOI: `10.1371/journal.pone.0196391`. [Online]. Available: `https://doi.org/10.1371/journal.pone.0196391`.

[4] Y. Lecun and Y. Bengio, "Convolutional networks for images, speech, and time-series," Jan. 1995.