Speech Emotion Recognition from RAVDESS dataset using prasodic and spectral features

Harsh V. Singh, Het Dave, Diya Fursule

{singh.187,dave.2,fursule.1}@iitj.ac.in

Indian Institute of Technology, Jodhpur

Dept. of Computer Science and Engineering - CSL2050 Major proejct

Contents

1 Introduction and background **1**

2 Preprocessing and Feature Extraction **2**

2.1 Pre-emphasis 2

2.2 Short-term spectral features 3

2.3 Prasodic features 4

3 Modeling **4**

3.1 Exploratory Model training 5

3.2 1D CNN on prasodic inputs 7

3.3 2D CNN on Short-term spectral feature inputs 8

4 Conclusion **9**

1 Introduction and background

Speech emotion recognition task is one of the most important problems in the field of paralinguistics. The goal of speech emotion recognition is to predict the emotional content of speech and to classify speech according to one of several labels (i.e., happy, sad, neutral, and angry etc.). As opposed to speech recognition tasks, SER dilutes the focus on time-series behaviour and stresses more on prasodic features.

The dataset we have used is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).

The portion we are concerned with contains 1440 files: 60 trials per actor x 24 actors = 1440. 24 professional actors (12 female, 12 male), vocalize two lexically-matched statements in a neutral North American accent.

2 Preprocessing and Feature Extraction

Raw audio is loaded as a series of samples with varying amplitudes. Typical dimensions of input audio are : 22.05kHz x 3.5sec duration = 77,175 samples. For any model to perform reassonably well, dimensionality reduction and feature extraction are prerequisite to model training.

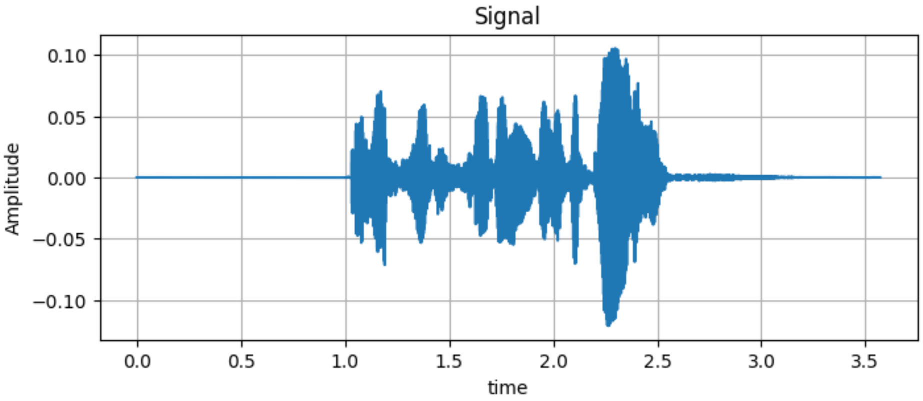


Figure 1: Sample audio waveform

2.1 Pre-emphasis

A pre-emphasis filter is useful in many ways: (1) balance the frequency spectrum since high frequencies usually have smaller magnitudes compared to lower frequencies. (2) may also improve the Signal-to-Noise Ratio (SNR).
The most commonly used filter is shown below :

$$H(z) = 1 - 0.95z^{-1} \tag{1}$$

The emphasised signal is then used for feature extraction.

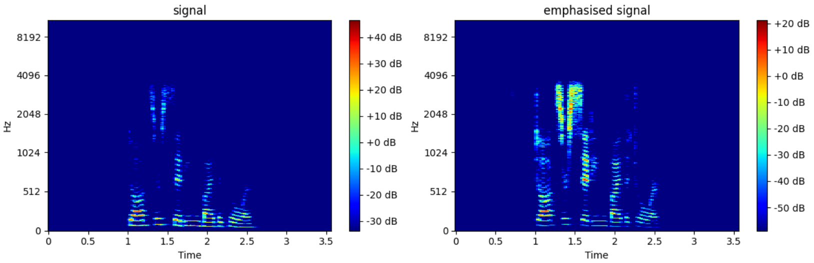


Figure 2: Effect of pre-emphasis on the spectrogram of a sample audio.

2.2 Short-term spectral features

CNN -based models have been trained on information derived from raw audio signals using spectrograms or audio features such as Mel-frequency cepstral coefficients (MFCCs).

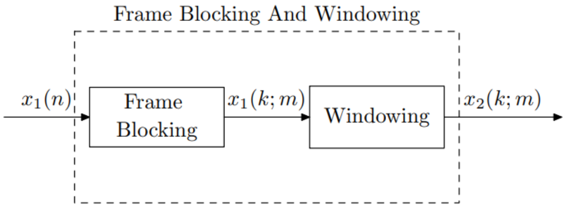


Figure 3: Audio segmentation for short term analysis

Both librosa and pyAudioAnalysis offer inbuilts functionaities for extracting these. MFCCs and Mel-spectrograms utilise Short term Fourier transform, since the standard transform assumes that the signals extend to infinity and computes frequency weights averaged over the entire duration.
Treatment of spectrograms as images utilise CNN for training, as demonstrated in the sections below.

Human perception of the frequency contents of sound does not follow a linear scale. Hence a non-linear transformation to the analysed frequencies is applied -

$$F_{mel} = 2595 \cdot log_{10}(1 + \frac{F_{Hz}}{700}) \tag{2}$$

MFCCs are obtained by applying a series of triangular filter banks in the mel scale onto the STFT followed by Discrete cosine transform. Implementation details are not included in this report but can be accessed through [4].
We will proceed by using the feature selection module of librosa that directly performs these computations for us. Chroma features are also used to reduce time correlation and bring out pitch variations.

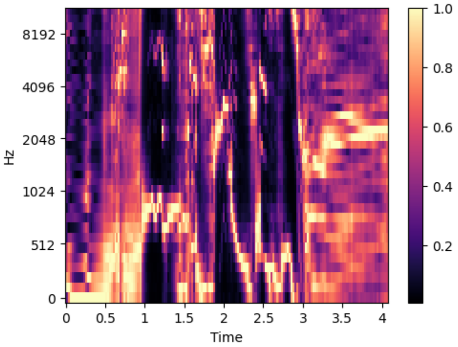
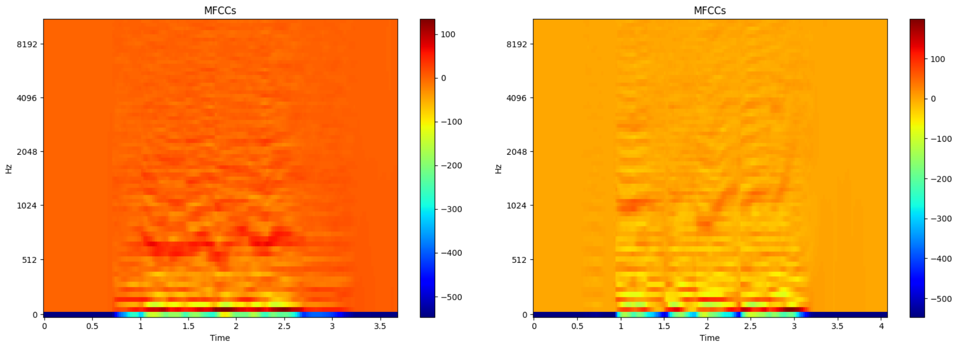


Figure 4: Chromagram for angry emotion sample



(a) angry

(b) fear

Figure 5: MFCC representations for 2 of the provided samples

2.3 Prasodic features

Speech produced in a state of fear, anger, or joy becomes loud and fast, with a higher and wider range in pitch, whereas emotions such as sadness or tiredness generate slow and low-pitched speech.
This intuition suggests us to stress more on overall behaviour of the audio sample rather than time variabilities. We have thus averaged 20 MFCCs, 12 chroma features and overall frequency intensities from Mel spectrogram.

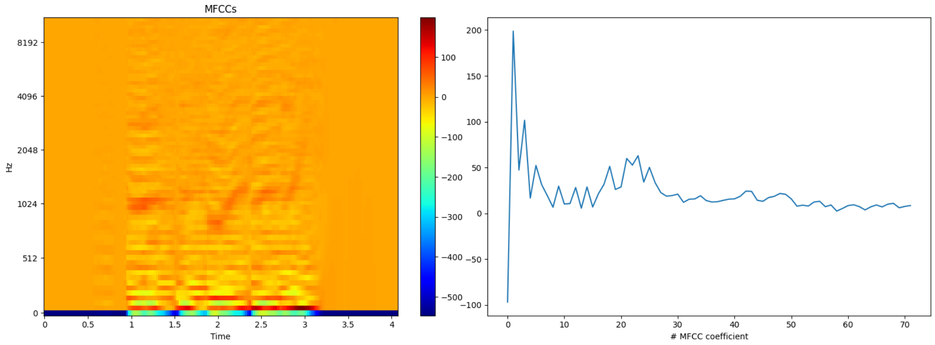


Figure 6: Time averaged MFCC coefficients

To conclude, prasodic feature extraction results in 180 features being analysed, whereas short term features viz. Mel spectrograms of dimension 128 x 130 are set apart for training CNN and visualising the differences between the two.

3 Modeling

The training procedure extracts features from the provided speech samples, apply a 64:16:20 train-val-test split and compute the performance of the chosen model. Comparisions and inferences have been analalysed in the end.

3.1 Exploratory Model training

We first train the following models using 1-dimensional prasodic features to check for class separability in the data.

MLP - Multi Layer Perceptron

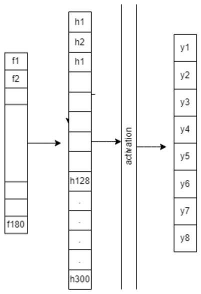
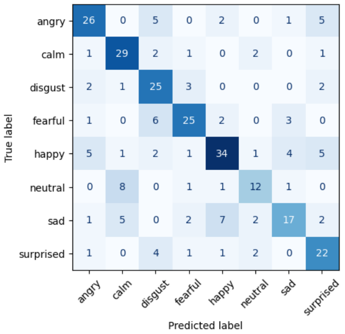


Figure 7: MLP architecture

Hyperparameters - input nodes : 180, hidden nodes = 300, output nodes = 8

Activation	train	test
ReLU	65.10 %	47.22 %
Logistic	99.91 %	64.58 %
tanh	100 %	62.85 %
identity	41.41 %	33.33 %

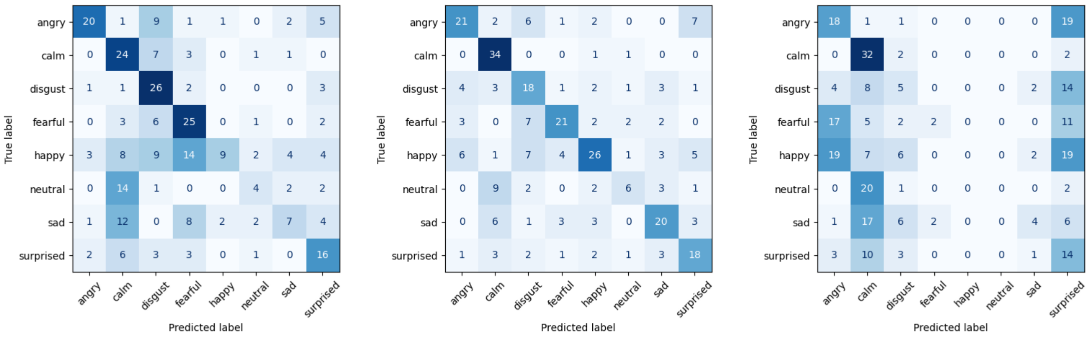


(a) MLP performance with different activations (b) CM for the best activation

Figure 8: Performance of MLP

RFC - Random Forest Classifier
Base model derived from decision tree classifier, intuitively makes classification based on comparision of features such as pitch and amplitude.
training Accuracy: 100.0 %
testing Accuracy: 56.94 %

SVM - Support Vector Classifier
training Accuracy: 31.34 %
testing Accuracy: 26.04 %



(a) MLP (relu)

(b) RFC

(c) SVC

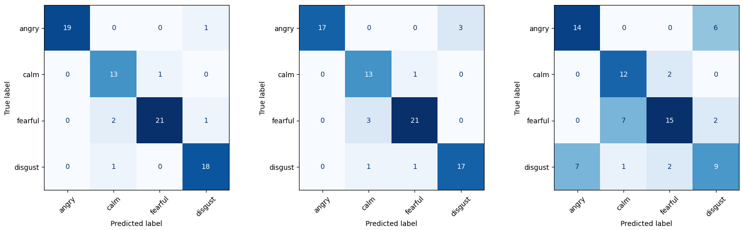
Figure 9: Confusion Matrices for intermediate performers

MLP is found to outperform the other two by a substantial margin due to its ability of capturing non-linearities. Tanh and logistic (sigmoid) activations are better learners than identity or ReLU. Among the 8 classes, the emotions 'anger,'calm','disgust' and 'fearful' are highly distinguishable and separable, as seen from the matrices above.
We refine our dataset to include only these emotions for training and analyse improvements to the resulting design.
RAVDESS audio file names may also be filtered according to the intensity of emotion expressed by the 4th quantifier in the file name : 01 for weak and 02 for strong. Again taking the stronger emotions before proceeding, we expect significantly better results.

Performance on filtered dataset

If we select only the emotions with strong intensity and those among the more distinguishable ones, we find that the models learn much better and with greater generalisability.

Model	**train**	**test**
MLP(logistic)	100.00 %	92.21 %
RFC	100.00 %	88.31 %
SVC	68.98 %	64.94 %



(a) MLP(logistic)

(b) RFC

(c) SVC

Figure 10: Performance on well-separable emotions

3.2 1D CNN on prasodic inputs

We have used 1 convolutional Layer followed by max pooling, which is flattened into the 1st fully connected layer of size 128x1. The second FC layer is of size nx1 corresponding to the n emotions (of the refined dataset).

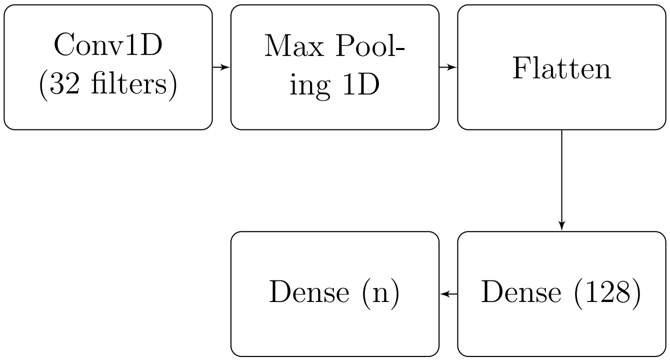
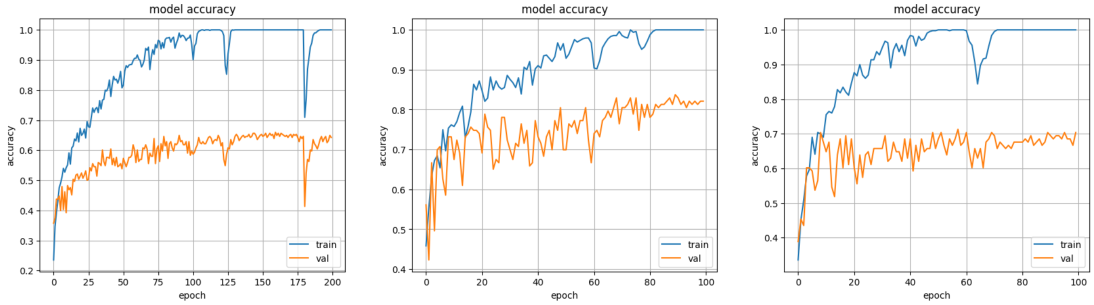


Figure 11: CNN architecture employed, n = No. of output labels

According to the filtering that we perform on the data, n takes the value 7 (for selecting strong emtional intensity files which don't exist for neutral emotion), 8 (if full dataset is taken) and 4 (if only the 4 most separable classes are considered).

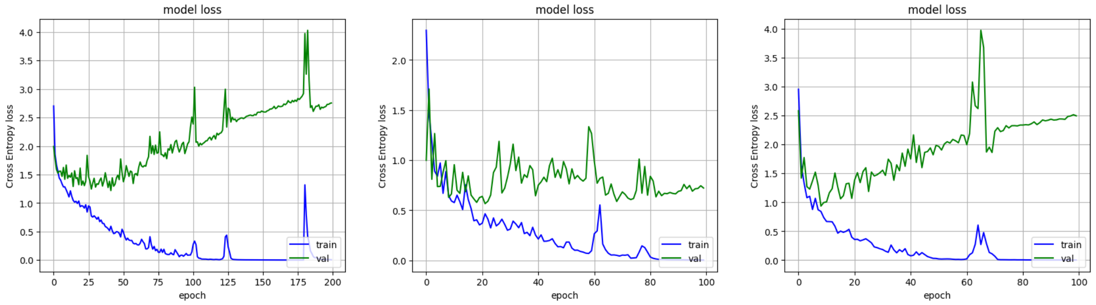
CNN is expected to bring out convolutional correlations among the 180 features extracted, as examined below:



(a) full Data (b) 4 most separable classes (c) High emotional intensity

Figure 12: Accuracy versus epoch for 1D CNN on various version of the data

The model classifies test data with 64.23%, 75.97% and 65.18% accuracies, respectively for the 3 different data versions as mentioned in Figure 12.



(a) full Data (b) 4 most separable classes (c) High emotional intensity

Figure 13: Loss versus epoch for the same model

The above trends demonstrate that for optimum generalisability, training should be stopped around epoch=25. Beyond this point, validation loss starts increasing while the training loss continues to reduce.
Also the accuracies don't fluctuate by a large amount after this checkpoint.

3.3 2D CNN on Short-term spectral feature inputs

Although we have stressed upon time averaged prasodic features, it may be reasonable to argue that time variability should not be ignored. This is something we will explore in this subsection.

For instance, it is observed that speech with neutral or calm tone has constant pitch and amplitude distribution over time, However a sentence spoken with anger or fear may have tapering magnitudes over the ends (See, Fig. 5,6).
This serves as a motivation to explore time variabilities and short term spectral features using CNN.

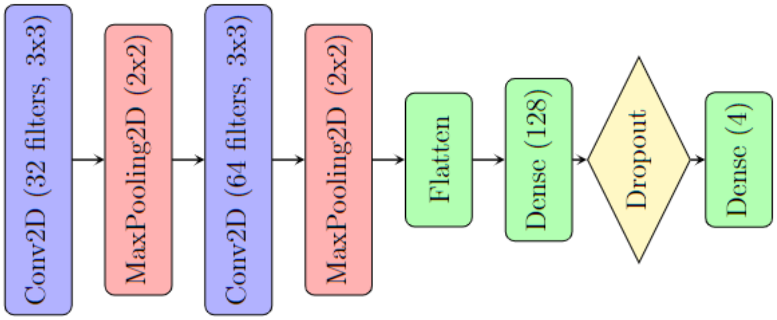
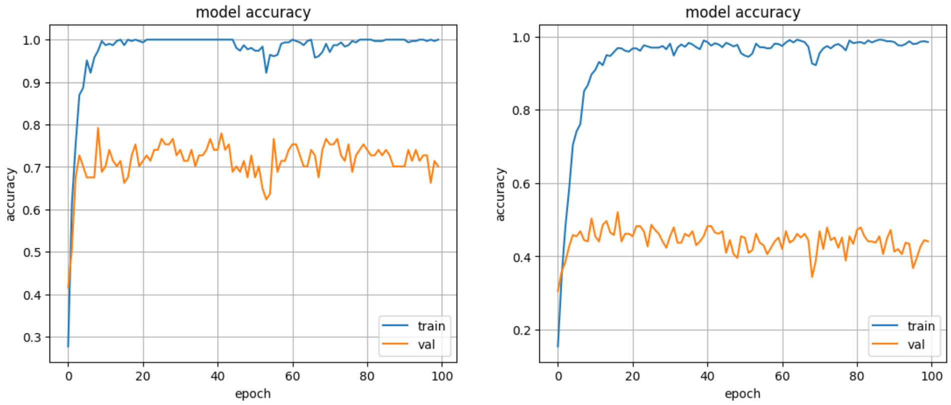


Figure 14: 2D CNN architecture for mel spectrogram input features

Due to expansive nature of the network and large number of trainable parameters, the model converges to 100% training accuracy within a few epochs.

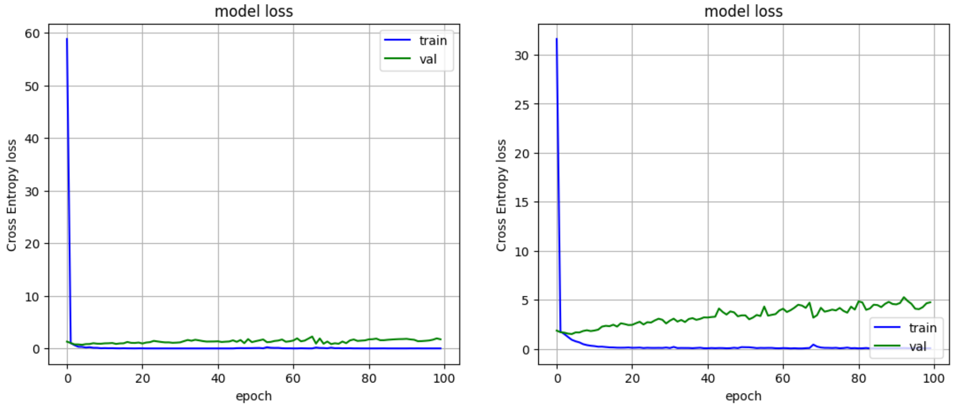


(a) refined dataset

(b) full dateset

Figure 15: Accuracy variation with training process

Accuracies on test data are : 70.12% (refined dataset) and 44.09% (full dataset).



(a) refined dataset

(b) full dateset

Figure 16: Loss variation with training process

Fig. 16 suggests worse performance in terms of avoiding overfit in case of full dataset training. The same is also evident by the separation between the training-validation accuracy curves above.

4 Conclusion

In this project, we have trained multiple models for achieving the required classification task. Based on the class separability of emotional features we observed, we divided the

training tasks into 2 categories, 1) with the full dataset, and 2) with only 4 of the 8 emotions being used for observation.
We also filtered the audio files on the basis of emotion strength (given in the file name itself).
Maximum accuracies were achieved by: **MLP with sigmoid activation - 92.21%** (on time averaged features) in case of 4-label emotion detection tasks; and **1D CNN - 64.23%** (on prasodic inputs again), in case of full data utilisation for training.

Although 1D CNN on prasodic feature inputs provided some improvements to the generalisability of the model, 2D CNN (in our attempt to capture time variability within mel features) failed by a substantial margin.

We conclude that the overall time-averaged behaviour of speech is more expressive in conveying emotions, compared to short term features.
The later is, however, particularly employed in automatic speech recognition (ASR) which is an entirely different domain of audio analysis.

References

[1] Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between, Haytham M. Fayek, 2016

[2] Multimodal Speech Emotion Recognition using audio and text - Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, Proc. SLT2018, Dec 18-21, 2018, Athens, Greece

[3] Identification of emotions from speech using Deep Learning, Abhay Gupta [1] Aditya Karmokar [2] Chennaboina Hemantha Lakshmi [3] and Shivani Goel [4] Bennett University, Greater Noida, India

[4] Speech Recognition using hidden Markov Model, Mikael Nilsson, Marcus Ejnarsson - Master Thesis, MEE-01-27, Blekinge Institute of Technology - March, 2002