Exploring Application of Machine Learning Models For Audio Based Emotion Recognition

Surbhi Paithankar, Apurva Gupta, Parichit Sharma
School of Informatics, Computing & Engineering

Abstract

This work focuses on audio signal-based emotion classification by utilizing hybrid data, a combination of spectral and prosodic features [1]. We identified features (spectral and prosodic) to train multiple neural network architectures that provides comparable accuracy on the RAVDESS [2], consisting both speech and song data.

We have tried to leverage empirical selection of features to best distinguish emotions and contrast performance of DNN, CNN against a baseline SVM model.

Background

A crucial aspect of the audio data (speech or music) is the emotional affect it evokes in the listener. Emotions are qualitative in nature since they can't be measured like pitch, average amplitude or other quantitative features. Studies mainly rely on real time input from the listener to classify emotions into various categories [1,5]. However, people also listen to songs due to specific emotions it invokes for example, happiness, calmness or even anger.

To this end, many published studies have explored the application of different machine learning models across datasets, both real time and synthesized data for multi-class and multi-label classification [1,3,4]. Recently, there has been an upsurge in the usage of deep learning models and various architectures including RNN, CNN, DNN [1,8] have been studied and applied to characterize the emotions.

In the context of industry applications, many recommendation systems rely on quantitative features for suggesting songs to their users [3,4]. Though audio-based emotion recognition is an active field of research but there is limited usage of emotion metric in popular recommendation systems [6]. From application context, we think that a qualitative metric like emotion may be considered for inclusion in recommendation systems due to its ability to capture emotional requirement of a user indirectly.

Objective

This work focus on emotion classification based on both speech and song audio signals by comparing with other similar work and trying to improve it. Additionally, we have studied the discrimination power of features for distinguishing specific emotions. We also aim to present the improvement in accuracy based on pure spectral versus spectral and prosodic features.

In this work, we have used the - RAVDESS [2] dataset - 'The Ryerson Audio-Visual Database of Emotional Speech and Song' compiled by Livingstone et al. that contains 7356 files in total of both speech and song type.

Dataset

In this work, we have used the RAVDESS [2] dataset. It consist of 7356 files (audio and video format). We have only used the audio data in favor of resources viz. training time and memory requirement. Speech data was classified into calm, neutral happy, sad, angry, fearful, surprise, and disgust expressions. The song audio signals were classified into calm, happy, sad, angry, and fearful emotions..

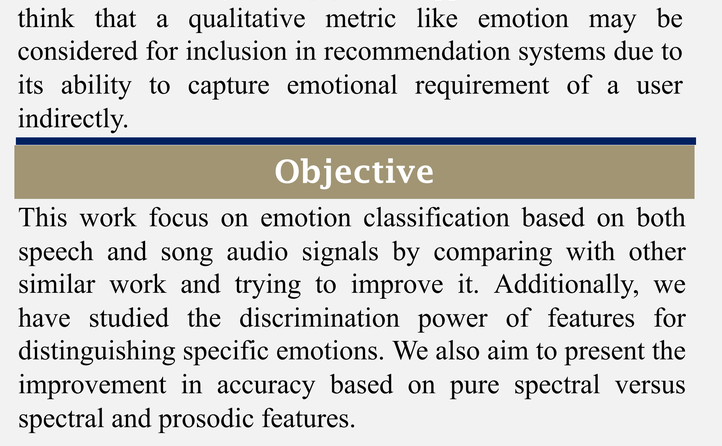


Figure-1: Time domain signals for 'happy' and 'sad' emotions looks different from visual inspection.

Preparation of the training dataset

The RAVDESS dataset It contains audio-visual recordings of 24 adult untrained participants (12 male, 12 female) speaking and singing the same sentences with different emotions at normal and strong emotional intensity, each with two repetitions. There are a total of 8 emotions including neutral, calm, happy, sad, angry, fearful, disgust and surprise. We selected 2452 audio files (out 7356 files) pertaining to speech and song data. The data was randomly split into training, validation and test set in the ratio of 0.8, 0.1 and 0.1 respectively.

Feature Selection & Engineering

To identify specific prosodic features (in addition to the MFCC matrix with reasonable discriminatory power across emotions, we plotted the distribution of these features over different emotions and selected 'Log energy' and 'Pitch' as the ones that vary significantly between emotions. Other prosodic features were not used for training due to lack of their discriminatory power.

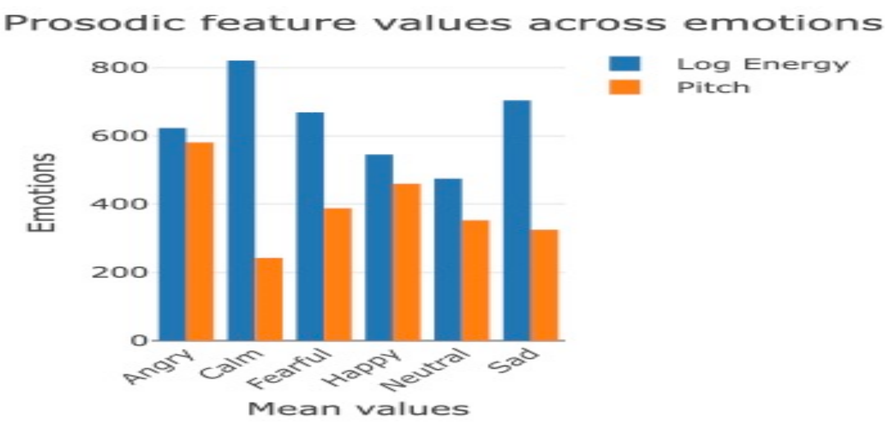


Figure-2: Distribution of mean log energy and pitch across emotions.

Experiments

- After feature engineering, the MFCC matrix of dimension 20 * 200 was combined with log energy and pitch values to obtain a 22 * 200 matrix.
- This was done for all audio signals leading to a matrix of 2452 * 22 * 200.
- For CNN an additional channel containing the delta values of the MFCC matrix was appended to the data thus resulting in a 2452 * 22 * 20 * 2 matrix.
- For testing the feasibility of our feature set, we first trained and tested the model only on the speech audio signal data and tuned the model for optimal performance only on the speech data.
- This was followed by training and testing only on the song data audio signals and further tuning to optimize the performance of the model only on the song data.
- Finally, we merged the speech and song audio signals to increase the size of the dataset and tuned the model to perform optimally on this dataset.
- During all these experiments, the Convolutional NN model performed best followed by Dense NN and a baseline SVM model in terms of classification accuracy.

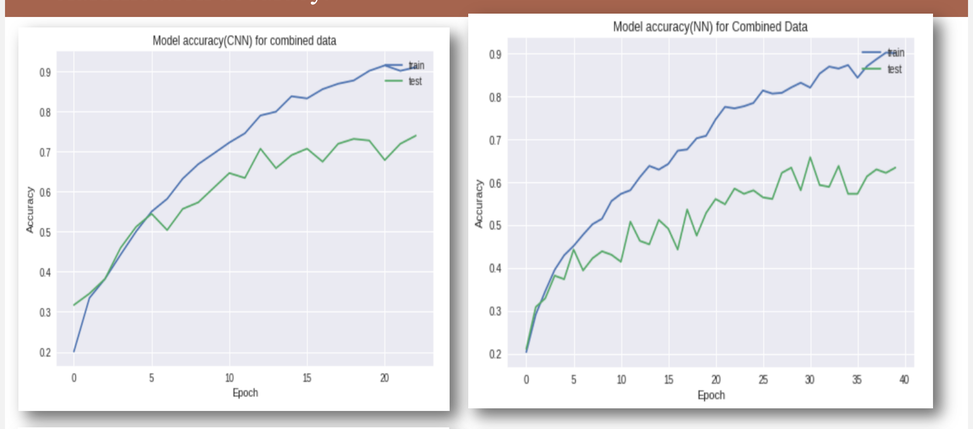


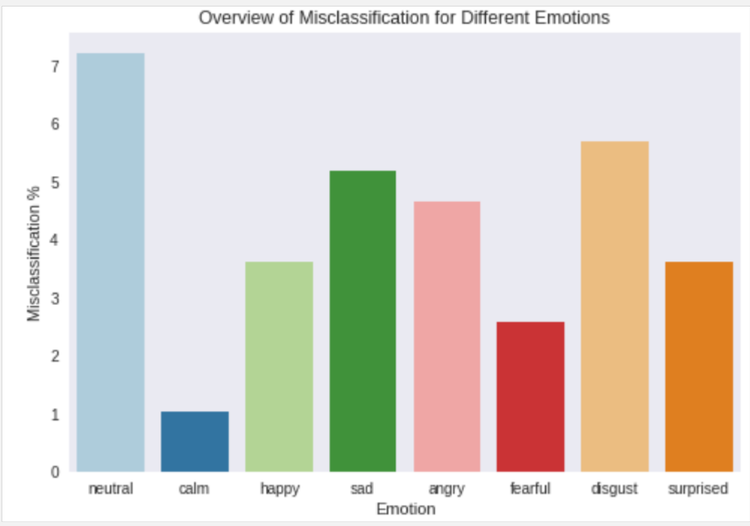
Figure-3: Top-left: Plot of Convolutional NN accuracy with epochs. Top–right: Plot of dense NN accuracy with epochs. Bottom-left: Plot of Convolutional NN loss with epochs. Bottom-right: Plot of Dense NN loss with epochs.

Results

- Evaluation Results: We obtained the following classification results averaged out of five independent runs by using the combination of spectral and prosodic features:
- The CNN based model produced the best classification accuracy of ~70%.
- The DNN based model provided an accuracy of ~60%.
- The baseline SVM models provided an accuracy of ~30%.
- A random guess based classifier resulted in the accuracy of ~12%.

Dataset	Model	Run no.	Train Accuracy	Test Accuracy	Average Accuracy
Combined	Random Guess				0.12
Speech + Song audio signals	SVM gamma=0.01, kernel='rbf' decision_function_ shape='ovo'	1	0.93	0.36	0.35
		2	0.49	0.34	
		3	0.77	0.36	
		4	0.81	0.34	
		5	0.88	0.36	
Speech + Song audio signals	Dense NN 2 hidden layers, Activations: relu, softmax adam, dropout=0.5, epochs=40	1	0.9	0.63	0.62
		2	0.89	0.61	
		3	0.9	0.59	
		4	0.91	0.63	
		5	0.89	0.62	
Speech + Song audio signals	Convolutional NN 3 conv 2 d with max pooling 1 dense 128 nodes, Activations: relu, softmax, adam dropout = 0.5, batch size 128 epochs=23	1	0.91	0.74	0.74
		2	0.89	0.77	
		3	0.93	0.73	
		4	0.94	0.72	
		5	0.94	0.76	

- Table-1: Accuracy of models on the train and test set.



- Figure-4: Mis-classification rate of the CNN model across emotions.

Conclusion and Future Work

- The CNN model performs fairly well. However, the mis-classification rate is higher for emotions (ex.- neutral) and low for easy to recognize emotions..
- This work can be extended to other audio datasets and can include complementary visual features for better classification.

References

[1],https://www.sciencedirect.com/science/article/pii/S0957417414001638 [2], The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English [3] https://www.sciencedirect.com/science/article/pii/S1051200417302439[4],https://www.sciencedirect.com/science/article/pii/S1877050915031841[5],http://web.stanford.edu/class/cs224s/reports/Anusha_Balakrishnan.pdf [6],http://ceur-ws.org/Vol-1462/paper6.pdf[7],https://web.eecs.umich.edu/~emilykmp/EmilyPapers/2015_Zhang_ACII.pdf
[8]https://librosa.github.io/

Acknowledgement
We are thankful to Professor Minje Kim and the entire course team for their valuable feedback on the project proposal and for in-depth lectures on neural networks and relevant ML techniques.