Exploring Application of Machine Learning Models For Audio Based Emotion Recognition

Parichit Sharma, Surbhi Paithankar, Apurva Gupta

***Abstract*—This work focuses on audio signal-based emotion classification by utilizing hybrid data, a combination of spectral and prosodic features [1]. We studied and identified features to train multiple neural network architectures on the RAVDESS [2] dataset which consists of audio and video data files. Based on our experiments, we leveraged empirical selection of features to best distinguish emotions and contrast performance of DNN, CNN, SVM against a baseline trivial model. The feature engineering part is motivated by two factors, firstly, the small size of the dataset requires us to inform the network of significant features (for a large dataset the network would have learned by itself). Secondly, it gives us a chance to explore the significance of features on the RAVDESS dataset which is relatively recent and have not been explored intensively. This dataset consists of song and speech signal of different actors. In our work, we have contrasted the performances of the models using them individually as well as in combination. We further analyzed the results in terms of misclassification rate and found synonymous results with [3] that an emotion such as NEUTRAL is more likely to be misclassified compared to other emotions which are easy to capture such as HAPPY, SAD (having high activation function).**

Key words: audio-based emotion recognition, prosodic features, spectral features, machine learning, artificial intelligence

I. BACKGROUND

ONE of the reasons why people listen to music is because of the particular emotions that the music can generate. Audio based emotion recognition deals with identifying specific emotions that are generated in response to audio clips (music, speech, communication etc.). In this context, the emotional response is a qualitative measure of the signals characteristics. This is distinct from quantitative measure such as volume, pitch, genre, frequency etc. that can be directly captured and studied. Emotional assessment of an audio signal generally happens at the listener end and can be affected by several features such as age, gender, individual preferences, cultural background to name a few. Recognizing emotion based only on the quantifiable characteristics is difficult as have been reported in the literature. Consequentially, there has been significant ongoing research in this field that uses both domain specific theories (from sociology, psychology etc.) [4] and computational algorithms to advance this field [5] [6] [7] [8].

A. A brief survey of state of the art in audio-based emotion recognition

In this work we primarily focused on two aspects. First, we studied the improvement in the prediction accuracy of the model by feeding in a combination of features. This was compared with the case where models were trained using only specific features to emphasize the rationale behind feature engineering. Secondly, we did an extensive study of which spectral or prosodic features show significant discriminatory power across the emotions on this particular dataset. This leads to a set of specific features that were used to train the models.

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 [5] [6] [7] [8].

An advanced approach is proposed by *Ooi et al.* [6] where they have employed RBF kernel based nodes in the neural network for better classification. However, the authors have relied only on the spectral features i.e. the MFCC matrix for training the network and have not considered the prosodic features which can also be used for emotion classification. We have tried to use the combination of prosodic and spectral features for emotional classification.

Vog et al. [7] used HMM based classifiers for identifying each type of emotion. The HMM classifier are powerful models because they can capture the temporal dynamics of an audio signal.

In a similar work, Lee, mower et al. [8] have designed a multi-layer model where each layer uses a SVM for binary classification.

B. Description of Audio Features

The audio data can be characterized on a high level by three types of features namely vocal tract,prosodic and spectral features. In this work we have focused only on prosodic and spectral features.

a) Spectral features: They are calculated from the magnitude spectra of an audio signal. Some examples are chroma, spectral centroid and spectral contrast etc. One of the most widely used spectral feature is MFCC (Mel Frequency Cepstral Coefficients) that have been used in several studies [1]. We leveraged the fact that there are standardized libraries such as librosa that can extract many types of spectral features easily.

b) Prosodic features: They Characterize the auditory features of an audio signal. Some examples are loudness, total energy of the signal, zero cross rate, pitch or fundamental frequency. For the calculation of prosodic features, we implemented our own methods in python by referring to the work published in [1].

1) Pitch: The pitch or the fundamental frequency of the signal is the quality to distinguish sounds as high and low. Pitch has been identified as a strong prosodic feature that can distinguish between different types of emotions.

2) Log Energy: The log energy of the audio signal is the sum of squared amplitude of the time domain signal, in other words it also corresponds to the volume of the signal or the normalized power. Happy, sad and anger emotions have relatively lower log energy compared with others.

3) Zero Rate Crossing: This indicates the frequency of oscillation of the signal in a given signal. In other words, it is the weighted sum of the number of times a signal changes its sign from positive to negative and vice versa. It measures the quanta of vibration in a signal.

C. Application(s)

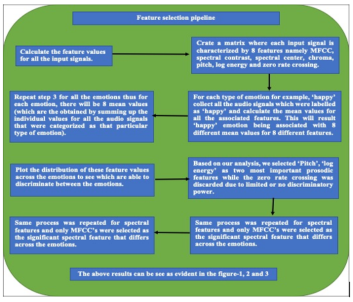
From a commercial point of view, an emotion recognition system can have many applications. Usage in recommender systems, emotion assessment during diagnosis or stress condition, integrating emotional responses in intelligent assistants are some of the examples. Though audio-based emotion recognition is an active field of research but there is limited usage of emotion metric in popular recommendation systems [2] . From application context, we think that a qualitative metric like emotion may be considered for inclusion in recommendation systems and other areas due to its ability to analyze emotions.

II. Dataset Description

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. It contains 7356 files (audio and video format). We have only used the audio data in favor of resources viz. training time and memory requirement. The data is generated synthetically by 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. We selected 2452 audio files (out 7356 files) pertaining to speech and song data. There are a total of 8 emotions including neutral, calm, happy, sad, angry, fearful, disgust and surprise. 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. The data was randomly split into training, validation and test set in the ratio of 0.8, 0.1 and 0.1 respectively.

III. Feature Engineering

The motivation for feature engineering was two-fold. Primarily, we wanted to ensure that the network is able to learn from the input data and not be limited by the size of the dataset. Therefore, we intensively studied a set of features to be used for training the network. This we did so that the automatic feature engineering which is done by the deep networks by leveraging huge datasets should not become a limiting factor in the performance of our models. Consequently, the models were able to learn significant weights to produce a reasonable prediction accuracy. However, this does not pacify the lack of data, more data could have added to the classification prowess of the models, but we were limited by the moderate number of files since we wanted to do audio-based recognition.



Feature selection pipeline

Calculate the feature values for all the input signals.

Crate a matrix where each input signal is characterized by 8 features namely MFCC, spectral contrast, spectral center, chroma, pitch, log energy and zero rate crossing.

For each type of emotion for example, ‘happy’ collect all the audio signals which were labelled as ‘happy’ and calculate the mean values for all the associated features. This will result ‘happy’ emotion being associated with 8 different mean values for 8 different features.

Repeat step 3 for all the emotions thus for each emotion, there will be 8 mean values (which are the obtained by summing up the individual values for all the audio signals that were categorized as that particular type of emotion).

Plot the distribution of these feature values across the emotions to see which are able to discriminate between the emotions.

Based on our analysis, we selected ‘Pitch’, ‘log energy’ as two most important prosodic features while the zero rate crossing was discarded due to limited or no discriminatory power.

Same process was repeated for spectral features and only MFCC’s were selected as the significant spectral feature that differs across the emotions.

Same process was repeated for spectral features and only MFCC’s were selected as the significant spectral feature that differs across the emotions.

The above results can be see as evident in the figure-1, 2 and 3

Fig. 1. Feature Selection Process

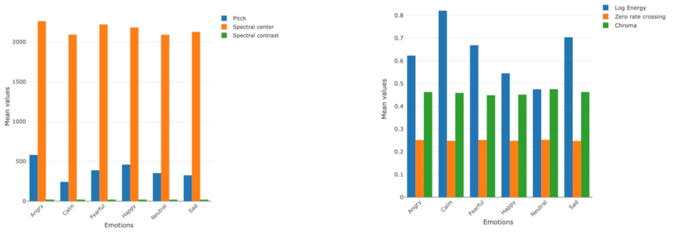


Fig. 2. Distribution of Features across emotions

Secondly, it gave us the chance to study the discriminatory power of features across the emotions in the dataset. Not all features are equally capable of distinguishing between the emotions, some are more likely to identify emotions then others. Also, this is emotion specific as can be seen in the Figure 2 where certain features are sensitive to a specific type of emotion.

A. Design of feature engineering experiment

The design of our feature engineering experiment was motivated by data driven feature selection that can be empirically validated. We decided to use a host of spectral and prosodic features namely, MFCC, spectral contrast, spectral center, chroma, pitch, log energy and zero rate crossing to characterize the signals. Our choice of prosodic features is motivated by published literature [1].

B. Selection of features

An outline of the feature selection process as shown in Figure 1.In figure 2, we can observe that only log energy and pitch have effect on emotions. Rest of the audio features are extremely close to each other. Therefore, we considered log energy and pitch along with MFCCs to build our Machine learning model.

C. Data Preparation

1) Output labels: The dataset contains 8 different emotions viz. neutral, calm, happy, sad, angry, fearful, disgust and surprise. These correspond to 8 labels that were used to

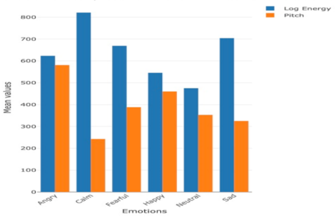


Fig. 3. A distribution of log energy and pitch across the emotions for whole dataset. Both pitch and zero rate crossing have strong biases for specific type of emotions.

calculate the BP error of the network during the training. These 8 labels were integrated into one hot vector encoding for network training.

2) Input features: The dimensions of the MFCC matrix vary between the audio clips due to the difference in the duration of speech. To overcome this, the MFCC matrices were padded with 0s to ensure exact same dimensions for all the input matrices. Alternatively, we could have clipped the matrices to match the dimensions of the smallest one but that could have resulted in loss of actual signal values which we did not want so we chose to expand the matrices.

Thereafter, the prosodic feature values for log energy and pitch were appended to the MFCC matrix to combine the spectral and prosodic features in the same matrix.

The combined matrix thus obtained is used for training the network.

IV. Experiments and Methodology

A. Input to the network

To incorporate all the features in the machine learning model, we standardized all the features in order to keep them on the same scale and improving the predictive power. After the data preparation, we pulled them together to feed into different machine learning models. We predicted the accuracies using a trivial classifier, SVM, Neural networks and convolutional networks.

1) Random Chance: This is a very naive model which classifies any given voice sample to a class with 1/8th probability. This trivial classifier works poorly because it does not incorporate any knowledge or attributes about the voice sample.

2) SVM: SVMs have always been powerful learning models that perform categorization using hyperplanes and support vectors. We used Sklearns linear SVC for implementing the model that performed one vs one reduction. This technique helps in multiclass classification using binary classifier where one trains $\frac{k(k-1)}{2}$ binary classifiers in K-way problem. In our problem, we had to distinguish between 8 different classes for which we used the radial basis function (rbf) kernel. We performed the hyper parameter tuning by using grid search library, that gave us the best gamma value of 0.05.

3) Dense Neural Network: The third model we implemented was dense neural network with two hidden layers and an output layer. The input to this network was an array of size N*F*T, where N is number of training examples, F is the number of features in each example and T is the total number of frames after the feature extraction. The input is passed to fully connected layers with two hidden layers using ReLu activation function followed by an output layer using the Softmax function. The hidden layers and final output layer consisted of 1024,256 and 8 nodes. We used Adam optimizer for our implementation. The prediction is made by choosing the label with the highest Softmax probability for each dimension. In order to optimize our results we performed parameter tuning for adjusting the number of epochs, learning rate, drop outs and batch sizes. The combinations of parameter that yielded the best results were chosen.

4) Convolutional Neural Network: In this last model, we used the convolutional neural networks for emotion classification. In this approach we tested the results for two different kinds of inputs. Firstly, we created a single channel input matrix with dimensions N*F*T*1. On the other hand, we also engineered double channel input matrix with an additional channel comprising of delta values of the first channel. These delta values are local estimates of the derivative of the input data along the selected column axis.The 2-channeled input matrix outperformed the single channeled method by 5 %. In both the cases, we used two convolutional layers of 128 and 256 nodes respectively. After flattening the outputs from these convolutional layers, they were fed to dense output layer with 8 output nodes. Like the basic neural networks, in this model we used the ReLu function for intermediate nodes, and a Softmax function for the final output layer. Adam optimizer was used for training the network. We chose the best combination of parameters by looking at the learning curve for validation loss and accuracy to avoid overfitting of the data.

V. Results

Since our dataset is synthetically generated and has well balanced classes, we computed the accuracies of predicting correct results across all the classes.

RAVDESS dataset consists of both audio and video clips for actors who are communicating their emotions using songs as well as speech. As per literature, the audio signals conveying the emotions in the continuous format of song have proved to give better results than regular speech signals. In our work we have first used the songs and speech signals individually and also combined them together for enlarging the dataset. We have experimented with both the kinds of audio formats to understand how the tonal difference can change the learning of emotional intelligence significantly.

In the table above we repeated same experiments for five times and averaged over their results for all kinds of input data. We used the model loss and accuracy plots for choosing the number of epochs and validating the train vs test data results.

Finally, after the model tuning, we tested our model using our own live voices. These were the voices that were completely new for the model. To implement the live demo feature, we followed all the steps of feature extraction again and fed them to the network. In the next section we have discussed about the results obtained while building the model as well as the live demo in detail.

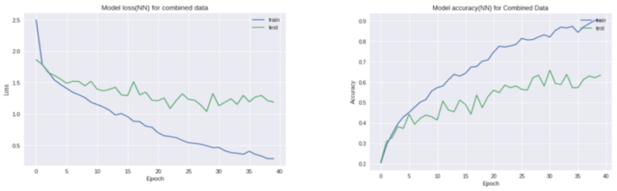


Fig. 4. Learning Curve

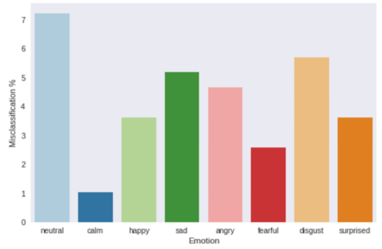


Fig. 5. Overview of Misclassification for Different Emotions

A. Discussion

Amongst all the models that we implemented for creating the emotional intelligence application, CNN outperformed all the others in every case. This can be attributed to the temporal invariance of CNN. While DNN learns through multiple hidden layers and their activations, CNN codes all of these features compactly through convolutions, no matter when they occurred in the input. The learning capabilities of CNN insinuates the great learning potential that the deep learning methodology has when corroborated with large amount of data. Though inferior to the deep learning frameworks, SVM also performed fairly well for this multiclass classification problem.

Furthermore, we recorded that interestingly, the song signal performed very well with 81 % of best accuracy whereas speech signals were the worst amongst all. As per Steven R. Livingstone in [2], the RAVDESS data set has lexically-matched set of emotional songs that could be a distinguishing feature. When we used all the available audio clips in our model training, we could attain an intermediary result of 74 %. However, if given enough data, we can train the model with a greater power.

Going ahead, we also analyzed the misclassification of the trained CNN model to understand the cases where the model did not perform well.

In the figure 5, we note that the neutral has been highly misclassified class. In addition, we see that sad class is also amongst the top misclassifications. This can be attributed to the fact that neutral and sad are emotions that have very low activation energy. Therefore, the model confuses them with each other. Since the tonal patterns in the neutral class are very steady, even a slight variation captured causes the model to misclassify the model into other class. In addition, owing to the subjectivity of disgust and surprised, the model often misclassifies them. Our training set do not contain these two classes for the song data. Therefore, if we can have additional audio clips with these emotions, we can decrease the error rate.

In the live demo, where we performed the feed forward, we observed that the strong intensity productions were identified more accurately like happiness, anger, calmness. On the other hand, the sentences spoken in a normal tone were often confused with sadness or happiness. These results are in line with research which has shown that strongly intense displays are identified more accurately in faces and voices [2]. This topic warrants further study to resolve these pattern of confusions.

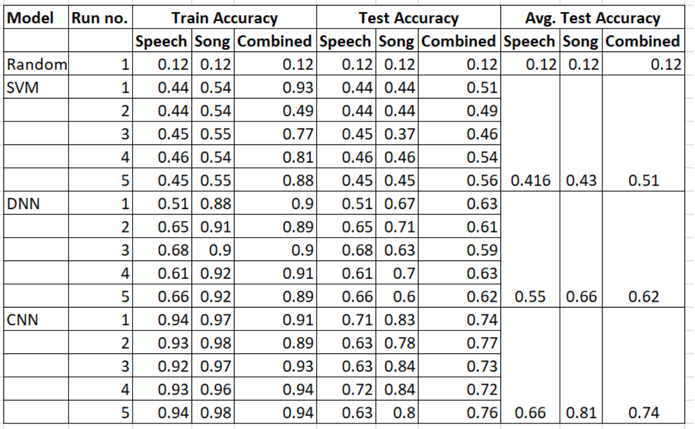
Finally, we tried our hands on other deep learning framework, LSTM. However, due to limited time and resources we could not run it to completion. It would be interesting to implement them and formulate a model with higher complexity to observe if the model develops to be robust enough to understand the intricacies of the confusing emotions.

VI. Conclusion

Machine emotional intelligence is a prospering frontier that could have huge consequences in not only advertising, but in new startups, healthcare, wearables, education, and more. The emotion detection market is expanding tremendously and was recently estimated that the market will grow from upto 3x by the year 2020. In this project we experimented with different machine learning methods by employing spectral and prosodic features together for emotion detection. We used speech, song and their combined data to understand the intricacies of sentiment analysis. We explored various prosodic features like pitch, zero cross rating and log energy in which pitch and log energy showed direct correlation with the variety of emotions. After the feature selection, we built the machine learning models like SVM, NN and CNN. In all the cases, convolutional neural networks gave us the best results with 75 % accuracies for the speech and song data. We also fed forward the network with our own test voices which the model could predict the true emotions for the sentiment with high activations like Happy, sad, surprised, Angry. However, the results showed most of the misclassification for confusing emotions like disgust, calm, neutral. The model needs to learn these intricacies by training on more data. In conclusion, this project aimed at exploring the signal processing technique for emotion detection which is attributed as an interesting blend of psychology and technology.

VII. Future Work

This work typically focuses on using audio features of a given speech and song signal. However, we can gather much higher information for building a stronger model using visual cues. This dataset includes over 4000 clips along with the videos of actors. We can leverage various computer vision methods to make better predictions. Also, since we have temporal data, LSTM have been widely known to produce substantial results. Given enough time and resources, we would like to experiment with this method. Finally, there has been a lot of work done that uses multimodality that uses different methods used in conjunction. Therefore, for the emotion detection purposes, we could use linguistic information along with audio signals to help the machine learning models



Model	Run no.	Train Accuracy			Test Accuracy			Avg. Test Accuracy		
		Speech	Song	Combined	Speech	Song	Combined	Speech	Song	Combined
Random	1	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
SVM	1	0.44	0.54	0.93	0.44	0.44	0.51	0.416	0.43	0.51
	2	0.44	0.54	0.49	0.44	0.44	0.49			
	3	0.45	0.55	0.77	0.45	0.37	0.46			
	4	0.46	0.54	0.81	0.46	0.46	0.54			
	5	0.45	0.55	0.88	0.45	0.45	0.56			
DNN	1	0.51	0.88	0.9	0.51	0.67	0.63	0.55	0.66	0.62
	2	0.65	0.91	0.89	0.65	0.71	0.61			
	3	0.68	0.9	0.9	0.68	0.63	0.59			
	4	0.61	0.92	0.91	0.61	0.7	0.63			
	5	0.66	0.92	0.89	0.66	0.6	0.62			
CNN	1	0.94	0.97	0.91	0.71	0.83	0.74	0.66	0.81	0.74
	2	0.93	0.98	0.89	0.63	0.78	0.77			
	3	0.92	0.97	0.93	0.63	0.84	0.73			
	4	0.93	0.96	0.94	0.72	0.84	0.72			
	5	0.94	0.98	0.94	0.63	0.8	0.76			

Fig. 6. Result Summary

learn and understand the subtleties of the true emotions being conveyed.

VIII. Acknowledgment

The success and outcome of this project required a lot of guidance and assistance from many people and we are extremely fortunate to have got all this during out project work. We would like to express our sincere gratitude to Prof. Minje Kim for providing us all the support and guidance which helped us complete the project on time. We would also like to appreciate the efforts taken by the teaching assistants by presenting their timely guidance and ideas.

References

[1] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, Li Wern Chew, A new approach of audio emotion recognition.

[2] Steven R. Livingstone, Frank 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.

[3] Biqiao Zhang, Georg Essl, Emily Mower Provost, ”Recognizing Emotion from Singing and Speaking Using Shared Models” Computer Science and Engineering, University of Michigan, Ann Arbor

[4] Marcel Zentner, Didier Grandjean and Klaus R. Scherer, Emotions Evoked by the Sound of Music: Characterization, Classification,and Measurement

[5] Y. Wang, L. Guan, and A. N. Venetsanopoulos. 2012. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recogni- tion. IEEE Transactions on Multimedia 14(3):597 607. https://doi.org/10.1109/TMM.2012.2189550.

[6] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. 2014. A new approach of audio emotion recognition. Expert Sys- tems with Applications 41(13):5858 5869. https://doi.org/https://doi.org/10.1016/j.eswa.2014.03.026.

[7] Thurid Vogt, Elisabeth Andre , and Johannes Wagner. 2008. Automatic recognition of emotions from speech: a review of the literature and recommenda- tions for practical realisation. In Affect and emotion in human-computer interaction, Springer, pages 75 91.

[8] Interspeech 2005, Lissabon, Portugal. Busso, C., Lee, S., Narayanan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Transactions on Audio, Speech, and Language Processing, 17(4), 582596.