Speech Emotion Classification

Arya Rajiv Chaloli
Computer Science
PES University
Bangalore, India
aryarajivchaloli@gmail.com

Greeshma Karanth
Computer Science
PES University
Bangalore, India
greeshmakaranth.13@gmail.com

Shivangi Gupta
Computer Science
PES University
Bangalore, India
shivangig078@gmail.com

***Abstract*—Audio emotion analysis is trying to extract features from audio clips and classifying the emotion of the speaker. The data-set chosen is extracted from the The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This implementation uses many different approaches to classify speech snippets into eight different emotions - Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised - after pre-processing the extracted data. Analysis of the different classifiers used and their pros and cons is presented in this paper.**

***Index Terms*—speech, audio, emotion, classification, features, comparing classifiers**

I. Introduction

Emotion has always strongly influenced a person's conversation. A message can be conveyed correctly only when the words along with the context of the message is understood. A truism: "It's not what you say, but how you say it". Expressions matter, as do the sentiment behind each encounter and the emotions raised. Emotion is entwined with the literal meaning of words used. With advancements in speech and audio technology, it becomes imperative to make machines understand and interpret the underlying emotions in speech. Extending a machine's capabilities to make it understand the emotions with which the audio clip was produced can help interpret the intended message better and with more efficiency.

However, this task is not easy. It poses a lot of challenges. The first hardship arises from the fact that the concept of emotions is subjective. It is hard to define the notion of emotions as each person would interpret a particular sentence differently. Next, annotating a single record is challenging. Deciding how many emotions should the clips be classified into also defines the accuracy of the results. Smaller number of emotions will give higher accuracy than a larger range of emotions even if the model is not precise. The difficulty arises when the range of emotions increases, since each detail in the audio will impact the classification.

The next obstacle is the collection of data. Most audio clips will have a lot of interference or noise. Finding or collecting data that is noise-free and at the same time strongly indicative of some emotion is difficult. Another aspect of this is that even if the data is manually recorded, a bias gets introduced due to the voice actor speaking the lines.

Lastly, creating the data-set of audio clips and the emotions classified is tasking. It takes a lot of manual effort and requires unbiased predictions, preferable to be made by trained professionals. The result also has to be verified by multiple people to verify the veracity of the classification. This is why, it is preferable to use many of the already available data-sets to first build a steady and successful classifier.

Further sections of this paper are organized as follows. Section 2 discusses related work in this field. Section 3 specifies the problem statement and the data-set description. Section 4 describes the implementation of the speech emotion classifier. Section 4 discusses the evaluation metrics and how the classifiers were tested. Section 5 shows experiments and results. Finally, Section 6 concludes the paper.

II. Literature Survey

Emotion is inferred from audiovisual inputs using decision based multimodal systems in one approach [1]. The inferences are classified into basic emotions such as happy, sad etc. using SVM and Radial Basis functions. While the algorithm yields a 98% classification rate, the accuracy is given mainly by the fusion of audio and video features. With just audio inputs - the classification rate is found to be 53%, which shows major scope for improvement. However, the decision based multimodal approach is proven to work on signals from environments which are highly susceptible to noise.

This approach focuses on emotion recognition from music - combining data from audio, MIDI and lyrics. However, the classification is done for music and not speech, which lack the supplementary features of songs such as instruments and lyrics, and therefore cannot be applied to speech signals [2].

Another method classifies seven emotions found on a Spanish and Berlin database using recurrent neural networks. The features used for classification are MFCC and MSFs and the combination has yielded high accuracy rates greater than 85% for both data-sets. However, the features used in this study are fewer than optimal and increasing the number of

features for study can be looked into [3].

The relationship between emotions and learning was investigated using a device called AutoTutor which has an automated emotion classifier. By applying multiple regression on results on test subjects, conclusions were drawn for comparisons between predicted and actual values. This method is sufficient but it relies on the AutoTutor device for accurate results [4].

The emotion was detected based on speech and facial gestures with PCA and LDA applied to select features and Gaussian classifiers used for classification. With only audio features, the recognition rate received was just 67%. Better results can be gained with other classifiers such as SVM [5].

Another approach talks about using partial least squares models on 2-5 predictors selected from the data-set to predict basic emotions from audio. The project however has been pursued in MATLab and can therefore provide different and improved results when implemented in R [6].

Another project attempted to use deep learning, instead of the commonly used machine learning approaches, as well as image classification in order to recognize emotion and classify them according to the speech signals. Transfer learning is used to train the model and although it reduces the computation cost and training time, it also results in reduced accuracy. Moreover, the data-sets used for each emotion were not enough. This approach does not work on purely audio based data-sets [7].

The main feature attribute considered in the prepared data-set was the peak-to-peak distance obtained from the graphical representation of the speech signals. The data was classified using Weka software. The experiment was repeated for 30 times and the average value was chosen to estimate the standard deviation. However, the results show that the best accuracy is when the training size is small, thus, not proving useful for large data-sets [8].

III. Problem Statement

Having understood the necessity of emotion classification from audio and reviewed the above advances, this paper attempts to overcome the above challenges and try to implement an approach that gives better results.

The problem statement for this project is recognizing and classifying the underlying emotions of given audio clips. Currently, machines are only capable of getting text from an audio clip. To analyse the clip and predict the emotions poses more challenges, the most prominent of which is that emotions are subjective. It is hard to define the concept of emotions, especially in a context that enables machines to learn from it. This is why the clear distinction of emotions needs to be defined before expecting a machine to learn and predict accurately. This demarcation is brought about by defining a moderately large range of emotions that are chosen to classify speech into. Here, eight emotions have been chosen - Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised.

These emotions are identified by analyzing features that can be extracted from audio clips like the MFCC score, tonnetz, chroma, contrast etc. Thus, the data-set used for this project contains numerical features extracted from the set of audio clips along with the emotion classification that the clip corresponds to. There are two more fields in this data-set, which are defined along with the audio clips itself. These two fields are gender of the speaker (male and female) and intensity. This numeric data-set was manufactured from the raw data - which was a set of speech audio clips available through the RAVDESS data-set.

This entire project was implemented in Python and R. There was no noticeable language bias induced by the two different programming languages on the results obtained, except for in one model, which is explained later.

The few assumptions made for this project revolve around the RAVDESS data - mainly that the data is unbiased and noise-free. The other presumption is that the original emotion classifications provided in this data-set are accurate for the audio clips. The final conjecture is that the classification is independent and free from correlation.

The main constraint of this project was the size of the data-set. With a bigger data-set, the accuracy of classification could have been increased. Additionally, the RAVDESS data-set restricts the number of emotions that were classified. With a combination of data-sets, the range of emotions could have been increased.

IV. Methodology

This project was implemented in five stages, as shown in Figure 1. A detailed description of each stage is given below.

A. Data Extraction and Pre-Processing

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a data-set that comprises of 24 actors of different genders. The data-set is comprised of audio files that follow the following file-name nomenclature:

1) Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
2) Vocal channel (01 = speech, 02 = song).
3) Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
4) Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.

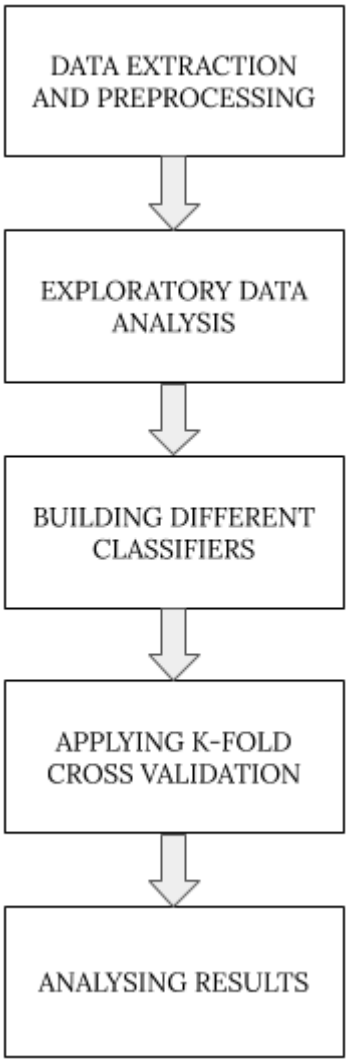

Fig. 1. Process Diagram

5) Statement (01 = ”Kids are talking by the door”, 02 = ”Dogs are sitting by the door”).
6) Repetition (01 = 1st repetition, 02 = 2nd repetition).
7) Actor (01 to 24 - Odd numbered actors are male, even numbered actors are female).

Using this schema, the identified target variables are the Gender, Emotion and the Intensity. This paper highlights the importance and the methods in which the emotion can be predicted, along with the drawbacks associated with them.

The next step involves the extraction of relevant information from the audio files. To do so, it had been identified that the mfccs, chroma, melspectrogram, contrast, and tonnetz provide the most useful information pertaining to the audio signals. Hence, the mean and standard deviations of chroma, contrast, and tonnetz are extracted. As can be seen in the image,

The melspectrogram has significant characteristics over

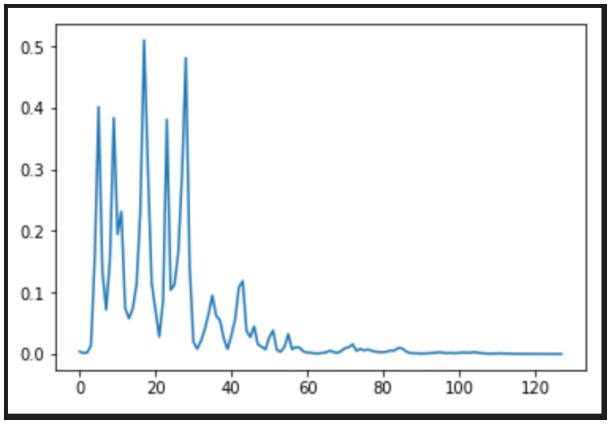

Fig. 2. MelSpectrogram

different time intervals. To capture this variation, the mean an standard deviations of the melspectrogram has been calculated over intervals of 20 samples. This extracted data of all these features are stored in a .csv file for quick access in the future.

The data features are fianlly normalized to prevent unwanted bias to be enforced upon the those having larger values in comparison to the others features.

B. Exploratory Data Analysis

To detect the presence of any highly correlated features, correlograms were plotted and the correlation matrix was visually inspected as seen in the figures below.

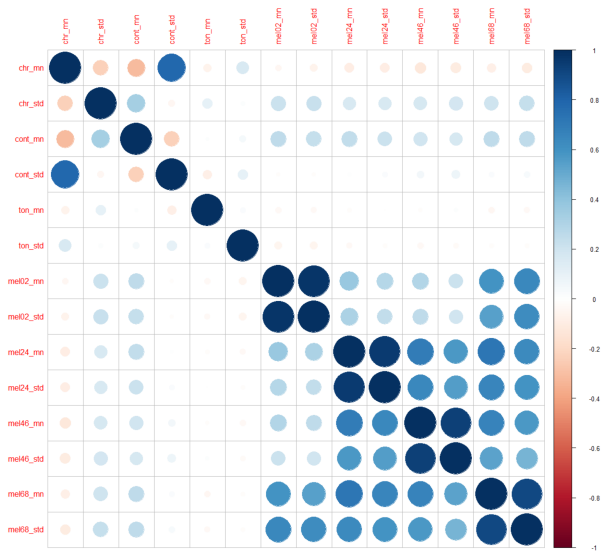

Fig. 3. The visualisation of the Correlation Matrix

From the plots, it can be inferred that, the the mean of the melspectrogram portions are highly correlated to the standard deviations of the corresponding intervals. This is a useful inference that could be drawn. However, though the standard

deviation of the contrast and the mean of the chrome have a high correlation of 0.79, but on further inspection of the inner semantics of the extracted features, it can be convincingly be confirmed that the correlation is spurious.

Further, the relation between the gender of the speaker and the features were analysed as seen below.

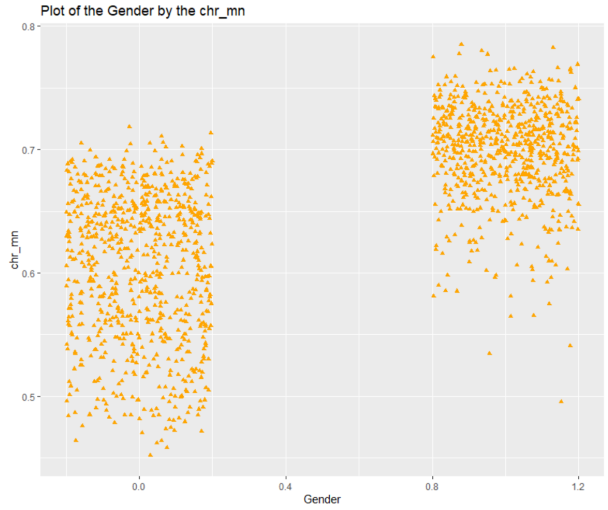

Fig. 4. The relation between the Gender and the Chrome

From these plots it can be observed that, the contrast and the tone do not have a huge role in differentiating the gender. But the chrome of the audio clips have a very clear separation across the genders i.e. an audio clip having a higher chrome value has a greater probability of it being part of the ”Male” category than the ”Female” category.

The last part of the exploratory data analysis involves, finding a relation between the ”Emotion” and the other extracted features.

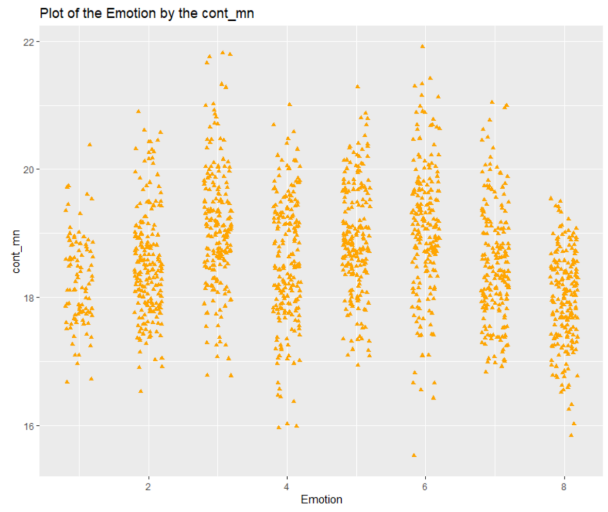

Fig. 5. The relation between the Emotion and the Contrast

From the following plots, it can be, inferred that the chrome and tone do not show any significant change across the emotions. However, the contrast, has an increasing then decreasing trend. The happy, angry and fearful emotions have a significantly higher range compared to the other emotions.

Further, when the following melspectrogram characteristics are plotted against the emotion, it can be observed that the ”anger” and the ”fearful” emotions have a higher spread in their respective distributions.

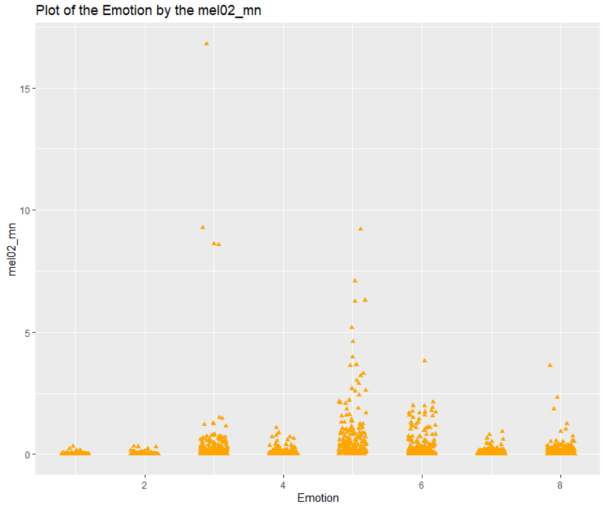

Fig. 6. The relation between the Emotion and the Mel (0-20 interval)

C. Building Different Classifiers

The data was first split into a training and testing data-set to be able to apply models to predict the target variables. A 75-25 split of the original extracted data-set was made to get the two resulting data-sets. This data division is random and changes on every run.

On this training and testing data-set, the different models were built. The models that have been compared in this project are the following - Decision Trees, Random Forests, Extra Trees, AdaBoost Classifier, XGBoost Classifier, Gradient Boosting and Bagging with Decision Trees. The choice of using Decision Trees and its variations was made because using a threshold value to separate the attributes for the different emotions seemed logically sound. Boosting classifiers were also implemented to try and improve the accuracy of the predictions. The models chosen to be compared in this report are chosen based on the comparatively higher accuracies they provide. Nonetheless, other models such as K-Nearest Neighbours, Neural Networks and Support Vector Machines with linear and radial basis function were also implemented, details of which can be found in later sections of this report.

D. Applying k-fold Cross Validation

To evaluate these models, 10-fold cross validation was applied. Cross validation ensures that there is no over-fitting of data or bias towards the training data. In other words, it helps utilize the entire data-set fairly to get the most effective models. The accuracy of these models during one run after applying cross-validation is listed below.

MODEL	MEAN ACCURACY	STANDARD DEVIATION
XGBoost Classifier	33%	4%
Gradient Boosting	32%	3%
Random Forest	31%	2%
Extra Trees	30%	3%
Decision Trees	27%	4%
AdaBoost Classifier	27%	3%
Bagging with Decision Trees	27%	1%

Fig. 7. Accuracy Table

E. Analysing results

Seven different classifier models were used to classify the emotions. Of these, XGBoost classifier had the highest average accuracy of 33% and a standard deviation of 4%. XGBoost is an ensemble technique that performs well due to system optimization and enhanced algorithms. XGBoost model has the best combination of prediction performance and processing time. It gives slightly better results than Gradient Boosting as it optimizes the GBM algorithm by implementing parallelized tree building, tree-pruning using depth-firth approach, cache awareness and out-of-core computing, regularization for avoiding over-fitting and efficient handling of missing data. Gradient boosting is also prone to over-fitting and requires careful tuning of hyper-parameters.

Another intriguing observation was that a single decision tree model performed as good as a bagged model on decision trees and AdaBoost classifier. However, the difference lies in the standard deviation and bagging gives the least amount of deviation, 1%, several subsets of data from training sample are chosen randomly with replacement and each collection of subset data is used to train their decision trees. This gives an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree. This is followed by AdaBoost with deviation 3%, and finally decision trees (4%).

Other models such as random forest also gave low accuracy as the final prediction is based on the mean predictions from subset trees, and thus, this doesn't give precise values for the regression model.

V. Evaluation Metrics

The classifier was evaluated on the testing data-set by using k-fold cross validation. k was chosen to be 10. A higher value of k was taken to induce less bias towards overestimating the true error. Having a moderately large k also gives us more samples to estimate a more accurate confidence interval.

The results gained from this were compared with the actual values using a confusion matrix. This showed that the emotions most accurately classified are 'Calm' and 'Angry'. This also showed that the emotions most commonly mixed up are 'Sad' and 'Calm', 'Angry' and 'Fearful'.

Then, the classification report of the predictions was also analyzed. This showed the f1-score, precision, recall and support. This produced some interesting results.

Comparing the reports for decision trees, random forests and gradient boosting, a few stark differences were observed. While Decision Trees provide the highest precision - of 0.58 - for the emotion 'Calm', Random Forests provide similar precision for four emotions - 'Calm', 'Angry', 'Fearful' and 'Surprised'. On the other hand, the highest precision gained for any emotion is 0.67 for 'Surprised' in the AdaBoost Classifier, followed closely by 0.62 for 'Angry' and it is given by the Gradient Boosting model. The XGBoost classifier gives almost equal precision for both 'Angry' and 'Fearful' emotions whereas bagging using decision trees as well as Extra Trees provides only moderate precision for all.

Coming to the recall and f1-scores, the highest recall is seen for the emotion 'Calm' in almost all classifiers except AdaBoost, whose highest recall is for the emotion 'Angry', which comes a close second in terms of recall in other classifiers. Based on the recall and precision values observed, we can conclude that the f1-scores of the respective classifiers will be in general higher for emotions such as 'Calm' and 'Angry'.

VI. Experiments and Results

A lot of models were used to experiment on this data-set.

First, a Support Vector Machine model with a linear kernel was used. This resulted in only 18% accuracy. Trying to improve on this, the classification was projected into higher dimensions using the radial basis function. However, this caused only a 1% increase in accuracy, that is, a total of only 19% accuracy. One curious thing to note here is that results were found in Python. But, when the SVM with linear

kernel is run in R, it resulted in an accuracy upwards of 30%. Looking into this anomaly, it can be concluded that this may be because of the inbuilt models used in the two languages. In R, the model scales the training data and then keeps the scaling parameters for using in predicting new observations. However, this does not happen in Python, possibly resulting in a lower accuracy. It was observed that with SVM, all target values were getting classified into one or two emotions. One conceivable reason why SVM did not work well with this data-set might be that there are too many emotions with widely varying attributes to define distinct support vectors.

Another approach to classify the data was by using neural networks. First a neural network was developed with only 2 hidden layers, which gave an accuracy of about 7% then using the MLPClassifier in Python with five hidden layers, an accuracy of around 26% was achieved. This was also not a good model and could be due to an incorrect loss function, a user defined loss function might be necessary. It could even be due to class imbalance in the data-set.

Next, the K-Nearest Neighbours model was implemented, taking the 5 nearest neighbours. This resulted in only 29% accuracy. The resulting low accuracy of this classifier could be accounted by several reasons viz., every characteristic of the method has the same result on calculating distance, this can be solved by giving weights. Another reason could be that KNN is the determination of new data classes which is based on a simple vote majority system, where the majority vote system ignores the closeness between data, this is unacceptable when the distance of each nearest neighbor differs greatly against the distance of the test data [9].

Finally, the decision was made to work on the Decision Trees Model. This was made using the already available sklearn module in python. The model resulted in around 31% accuracy, which was better than the previous results. Therefore, to improve this, variations such as Random Forests and Extra Trees were attempted. These methods yielded almost the same accuracy as Decision Trees, with only a slight increase in accuracy. Owing this to the fact that all these models use the same basic approach of division of attributes, boosting was performed to improve the accuracy. Trying out the AdaBoost, Gradient Boosting and XGBoost classifiers, a maximum accuracy of 33% with a standard deviation of 4% was achieved. The analysis of these results are provided in earlier sections of this paper.

VII. Conclusions

After applying several classifiers on the extracted data-set, a maximum of 34% accuracy could be received. Multiple algorithms with various tunings were implemented to improve the accuracy but they failed to raise it by a considerable amount. This might be due to wrong assumptions made about the RAVDESS data-set. The data-set might not have been noise free and the size wasn't sufficient to give accurate results. Moreover, extraction of the data-sets could be done in a different way, by using moving averages on a window of fixed interval.

Contributions

NAME	CONTRIBUTION
Arya Rajiv Chaloli	• Data Extraction and Preprocessing • Exploratory Data Analysis • Decision Tree Baseline Model • Final Report and Presentation
Greeshma Karanth	• Data Preprocessing • Literature Survey Report • KNN Baseline Model • Final Models • Evaluation Metrics • Final Report and Presentation
Shivangi Gupta	• Data Extraction with Moving Averages • Literature Survey Report • SVM and ANN Baseline Models • AIC Evaluation Metric • Final Report and Presentation

Fig. 8. Contributions Table

References

[1] Ntombikayise Banda, Peter Robinson , 'Noise Analysis in Audio-Visual Emotion Recognition', unpublished

[2] R. Panda , R. Malheiro , B. Rocha , A. Oliveira and R. P. Paiva,'Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis', Centre for Informatics and Systems of the University of Coimbra, Portugal

[3] Leila Kerkeni, Youssef Serrestoul, Mohamed Mbarki, Kosai Raoof and Mohamed Ali Mahjoub, 'Speech Emotion Recognition: Methods and Cases Study', 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2, pages 175-182

[4] Arthur GRAESSER, Patrick CHIPMAN, Brandon KING, Bethany McDANIEL, Sidney D'MELLO, 'Emotions and Learning with AutoTutor', unpublished

[5] Sanaul Haq and Philip J.B. Jackson, 'Speaker-Dependent Audio-Visual Emotion Recognition',AVSP 2009 – International Conference on Audio-Visual Speech Processing, University of East Anglia, Norwich, UK September 10–13, 2009

[6] Tuomas Eerola, Olivier Lartillot, Petri Toiviainen, 'Prediction Of Multidimensional Emotional Ratings In Music From Audio Using Multivariate Regression Models', 10th International Society for Music Information Retrieval Conference (ISMIR 2009)

[7] Nithya Roopa S., Prabhakaran M, Betty.P, 'Speech Emotion Recognition using Deep Learning', International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4S, November 2018

[8] Assel Davletcharova, Sherin Sugathan, Bibia Abraham, Alex Pappachen James 'Detection and Analysis of Emotion From Speech Signals' Procedia Computer Science , 2015

[9] K U Syaliman, E B Nababan, O S Sitompul, 'Improving the accuracy of k-nearest neighbor using local mean based and distance weight', 2nd International Conference on Computing and Applied Informatics 2017 IOP Conf. Series: Journal of Physics: Conf. Series 978 (2018) 012047

BIBLIOGRAPHY

1) https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb
2) https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b
3) https://en.wikipedia.org/wiki/Gradient_boosting
4) https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9
5) https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/
6) https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d