National College of Ireland

Detection of Audio Emotional Intelligence Using Machine Learning Algorithms

MSc Research Project
Data Analytics

Tejesh Batapati
x17108811

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Tejesh Batapati
Student ID:	x17108811
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Dr.Catherine Mulwa
Submission Due Date:	13/9/2018
Project Title:	Detection of Audio Emotional Intelligence Using Machine Learning Algorithms
Word Count:	7169

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	12th August 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detection of Audio Emotional Intelligence Using Machine Learning Algorithms

Tejesh Batapati
x17108811
MSc Research Project in Data Analytics

12th August 2018

Abstract

The interaction between humans and machines are increasing day by day and due to advancement in technologies the medium of audio as interaction has been growing exponentially, but unlike text medium, audio interaction needs to be more human-like to make it more effective. To reduce the gap between human and chatbot interaction. In this research, various features such as Mel-frequency cepstral coefficients, root mean square energy, tonnetz and zero crossing rate are extracted and analysed to show which features contribute more to the identification of emotions. In addition, several machine learning models are developed and results are presented. The result of this project will help customers interact with a chatbot effectively. The ensemble model used in this project resulted in a accuracy of 67% with MFCC features which is the highest when compared to other models. After successful identification of emotions, a chatbot framework is presented which can adapt to interactive dialogues with the customer based on the emotion from the speech in the audio. In addition to this, the results of the previous related work and research gaps are discussed. To fill the research gaps feature manipulations and models are developed.

1 Introduction

It is a well known fact that machines are not capable of understanding the emotions of human beings because of a simple fact that machines are machines. But as the emotional state of humans alters the meaning of the same linguistic content, it is necessary for the machines to detect them during interactions to make it more human-like. The idea to make machines detect the emotions from the speech started around 1980's by using statistical techniques by (Bezooijen; 1984; Tolkmitt and Scherer; 1986) and on the other side since the late fifties, there has been a constant research on this by converting speech to words or text El Ayadi et al. (2011). But the demand for this field is exponentially increased as the possibility to solve the problems has grown with the advent of data science, and moreover as the number of fields that are using emotion detection has increased enormously. This research is concentrated on effective detection of emotions from a human speech by extracting features from audio using machine learning techniques and build models to detect the emotion in the speech.

1.1 Motivation and Project Background

Interaction of human beings with machines is done mainly in three forms such as audio, video and textual Zhang et al. (2016). Though textual is widely used it is not as comfortable as video and audio for all the humans to communicate, for example, to communicate with the computers one can write code which is a complicated task when compared to just talking to Siri (Apple's chatbot) via audio. Because of this reason all big and small companies have started using audio medium interactions such as Google assistant by Google to help the users in different assistance activities such as ordering products online using voicemails, Siri by Apple as a personal assistant for the users which communicates with users voice commands and in similar lines Amazon has Alexa and giant companies like IBM and Microsoft are also involved. More importantly, it is not restricted to big scale or big companies, but it is implemented in a lot of small and emerging companies as well. So this demand to implement any procedures that can improve the effectiveness of communication between the humans and machines. The main business motivation here is that it has a potential that can be adapted by many businesses upon successful implementation. For example it is very challenging for the machines to detect the sarcasm of the user during interactions this is because machines cannot understand the hidden meaning under the tone of the user; in scenarios like this detection of emotion based on voice features helps a lot Gidhe and Ragha (2017) and when interaction with blind customers where text medium is not feasible audio emotional intelligence plays a big role. The main motivation for doing this research is not just because of its scarcity but also because it is one particular aspect that can bring value and change in many fields where human and machine interactions are there via voice medium.

Detection of emotions from the speech is very challenging because of the following reasons which speech features are best suited for emotion recognition, in the same utterance there might be more than one emotion, multilingual personalities, speaking styles and cultural differences affect the basic speech features of pitch and energy contours El Ayadi et al. (2011). To detect the emotions in the literature there are two different methods, one is by using the acoustic and spectral features of speech and other is by converting speech to word and applying natural language processing techniques on it. This research is based on the former one i.e. using acoustic and spectral features as this has an advantage over speech to word conversion method in which it overcomes the problems of multilingual personalities, speaking styles and cultural differences.

Linguistics has established over 300 types and subtypes of emotions that encounters in human life Schubiger (1958). classifying such number of emotions are not actually helpful in making a business or personal decision so in literature, researchers decomposed 300 emotions into primary emotions similar to basic colours where any colour is a combination of basic colorsCowie et al. (2001). In this research, the scope is to classify primary emotions namely neutral, calm, anger, disgust, fear, joy, sadness, and surprise.

1.2 Research Question

RQ:*"How can emotions (anger, calm, disgust, fear, happy, neutral, sadness, and surprise) in human speech be identied and detected using machine learning models (Naive Bayes, K-Nearest Neighbour, Neural Network, Support Vector Machine, Random Forest, Ensembles) to help customers during interaction with audio chatbot) ?"*

Emotion detection from human speech is significant i.e., suppose in customer recommendation services where humans interact via voice with audio chatbots like Apple's Siri, Microsoft's Cortana, to get recommendations if it can grasp the emotions in the voice chatbots can recommend or change the recommendations more effectively

Sub RQ: *"How can identified emotions in human speech be applied to Audio chatbot framework ?"* To improve the efficiency of the interaction between human and audio and chatbots.

To solve the research question the following objectives are specified and implemented

1.3 Research Objectives

Obj1: A critical review of literature on human emotions speech detection (2004-2015)
Obj2(a): Extract features(mfcc, rmse, tonnetz, zcr) from speech signals.
Obj2(b): Implement,evaluate and results of Naive Bayes.
Obj2(c): Implement,evaluate and results of K-Nearest Neighbour
Obj2(d): Implement,evaluate and results of Convolution Neural Network.
Obj2(e): Implement,evaluate and results of Support Vector Machine.
Obj2(f): Implement,evaluate and results of Random Forest.
Obj2(g): Implement,evaluate and results of Ensemble.
Obj2(h): Comparision of developed models.
Obj2(i): Comparision of developed models with existing models.

The rest of the technical report is structured as follows, chapter 2 presents an investigation of existing literature in detection of emotions from speech, features of speech in machine learning, usage and implementation of speech emotion detection in chatbots, based of the results of the literature chapter 3 presents scientific methodology followed by feature extraction from speech signals and finally chapter 4 presents implementation evaluation and results.

2 Literature Review on Audio Speech Emotions Recognition (2004-2018)

2.1 Introduction

This literature review investigates detection of human emotions from human speech. This section is divided in to many sub sections i.e. (i) literature review on audio speech emotional intelligence and identified gaps, (ii)investigation of audio chatbots and identified gaps, (iii)an investigation of audio feature extraction, (iv) critique of techniques, models and metrics used in detection of audio emotions and (v) comparision of reviewed techniques used in audio emotion detection

2.2 Literature Review on Audio speech Emotional Intelligence and Identified Gaps

In between humans, the natural and fastest mode of communication happens in the form of speech signals. But when it comes to interactions between humans and machines speech signal communication is not natural nor can be understood. To understand the speech signals from humans, machines should have that kind of intelligence. To inculcate such kind of intelligence in machines many kinds of research are looking at it from various directions. El Ayadi et al. (2011). The various ways to teach audio emotional intelligence in machines can be categorised into two methods linguistic features and acoustic features

Linguistic feature-based methods are done by using the words used in the speech by humans. For example sadness, despair, resignation, disappointment all these words carry the meaning of sadness and when in a human speech these words are used in the sentence the machines will consider the emotion as Sadness. In Devillers and Vidrascu (2006) they tried to detect real-life emotions from speech using linguistic features. But there are some fundamental problems in this concept which are it fails in the detection of sarcasm in the speech, sometimes words carry more than one emotion with the difference in utterance and it fails when the human speaks a different language to the corpus of language the machine has. Though this approach has the mentioned problems it has been widely researched in the field of speech emotion recognition Schuller et al. (2009) Jin et al. (2015) Schmitt et al. (2016) mainly because of its simplicity in logic and as it has shown good results in few scenarios where sarcasm is not involved and in altogether another phase it is being mixed with acoustic feature-based analysis to detect the emotion in the speech.

In between Linguistics and Acoustic based emotion detection, acoustic way is more complex and better one. Because it uses the basic energy in the audio rather than the words. In El Ayadi et al. (2011) authors specified that speech emotions are identified from two dimensions which are activation and valance. Activation, in general, is based on the energy. According to some physiological studies loud emotions such as fear, joy and Anger results in high rates of blood pressure, heartbeat, dryness of the mouth, greater sub-glottal pressure and occasional muscle tremor producing the speech which is fast, high pitched and strong high-frequency energy. In the case of emotions of sadness, the speech is slow, low pitched and little high-frequency energy. And Valance is useful in distinguishing within the high frequency or low-frequency emotions. Therefore researchers have been using the acoustic features of articulation of the speech signal, timing and pitch to make machines understand the emotions in the audio.

2.3 Review of Audio Chatbots and Identified Gaps

In the advent of big data and data science the usage of audio emotion detection is being used in many fields such as education sectors, social networks, e-commerce, tourism, in cars where system can recognize the mental state of the driver and triggers the safety system if necessary, in therapies to aid psychiatrists, in cockpits of aircrafts etc. Dwivedi and Roshni (2017) In all these fields the functionality of audio chatbots are used as interaction between the human and machine. But the main drawback of the interaction is so fundamental that machines cannot understand the emotional state of the user so the responses it gives may not be in the expected areas. As the application of audio chatbots being implemented in many sensitive fields such as therapy, cars safety system, cockpits

etc. it is much more demanding than ever that need for emotion detection should be inculcated in the chatbots. Respecting this the researchers are concentrating in this area to reduce the gap between humans and machines Izquierdo-Reyes et al. (2018) big companies such as Google (Google assistant), Apple (Siri), Microsoft (Cortona), Amazon(Echo) etc are using the audio linked chatbots. So there is a need to increase the efficiency of chatbots with the intelligence of audio emotion detection Ikemoto et al. (2018).

2.4 An Investigation of Audio Feature Extraction

In the context of applying audio or speech as input to the machine learning algorithms, they are not capable of taking a raw audio file as input unless they are transformed in to feature vectors of arrays of frequencies in the time frame. There are mainly two types of features namely 1)temporal features and 2)spectral features. Temporal features are time domain features such as ZCR,STE of the signal etc and spectral features are frequency-based features which are obtained by applying fourier transform of time based features, some of the spectral features are Mel Frequency Cepstral Coefficients (MFCC), Local Discriminant Bases (LDB), Linear Predictive Coding (LPC) etc. Yang and Krishnan (2018)

ZCR attempts to capture the change in sign of the signal or successive samples and also it can be used to capture the rate at which the change in sign happens. The formula used to identify ZCR is as below

$$z_i = \sum_{m=1}^{n-1} sign[x_i(m-1) * x_1(m)]$$

where, n is the number of samples in the ith frame
It is highly suitable to be used in non voice speech recognitions like music, for example the sound of humming where the rate of sign changes from positives to negatives and vice versa are helpful in determination Dutta and Ghosal (2017).

Another frequently used temporal feature is STE. As the name suggests this feature is used to know the transitions in the energy over the signal. The transition in voiced to unvoiced regions in the signal can be used in the intelliglible speech signals.The formula used to identify STE is as follows Dutta and Ghosal (2017)

$$E_i = 1/\mathrm{n} * \sum_{\mathrm{m}=0}^{\mathrm{n}-1} [x_i(m)]^2$$

where, n is the number of samples in the ith frame

Coming to the spectral features the most commonly used feature is MFCC and also this feature is appreciated highly in yielding best results. MFCC is basically the collection of coefficients of the mel frequency cepstrals. This is obtained by applying Fourier transform of the signal(Voice) followed by applying the transformed signal on the mel scale and then at each of the mel frequencies log and discrete cosines are applied. The main usage for using mel frequency transformations is because of this the signal resembles approximately the response in human auditory system.The equation used to obtain MFCC is as follows Dasgupta et al. (2017)

$$\mathrm{M(f)} = 1125 \ln(1 + \mathrm{f}/700)$$

Because of above mentioned nature of extraction in MFCC it is highly applicable in speech/voice recognition techniques.

The research gap here is that in different scenarios or with different data different combinations of audio features gives best results and there is no clear research that confirms which features should be used for voice based speech recognition.

2.5 Critique of Techniques, Models and Metrics Used in Detection of Audio Emotions

In a survey conducted by Chachada and Kuo (2014) reviewed that most of the researchers used MFCC and LBD features to implement K-Nearest Neighbours (KNN) classifier to classify 37 artificial and natural sounds and in this MFCC features yielded better results. Similarly in classifying the Indian ragas (musical sounds) Kumar et al. (2014) used MFCC features and applied it on KNN, Bayes Network(BN), and Support Vector Machine (SVM) to obtain 80% accuracy. In Bormane and Dusane (2013) researchers used a novel technique called as Wavelet Packet Transform (WPT) and used features of MFCC, LPC and ZCR separately to classify the musical instruments based on the sound and here as well better results are obtained from MFCC features. Likewise in many scenarios traditionally MFCC features are yielding better results than the other features. But in Patil et al. (2012) researchers combined MFCC feature with other features to boost the performance of the models and successfully obtained better results. They used MFCC, LPC, ZCR, Short Time Energy(STE), spectral feature(SF) on optimal BN. When they used only MFCC they got 77% accuracy and when they combined all the other features with MFCC they obtained 86% accuracy. In Cakir et al. (2017) article favoured Convolutional recurrent neural networks (CNN) to classify the bird sounds and as these sounds after feature extraction yields non-linear data CNN has opted and it resulted in an area under curve AUC of 88.5% .

Coming to the metrics Cakir et al. (2017) used receiver operating characteristic (ROC) curve related metrics AUC. The more the AUC is the better the performance of the model is. Chachada and Kuo (2014) and Bormane and Dusane (2013)over all accuracy is used to know the performance of the models but Patil et al. (2012) used overall accuracy with 95% confidence intervals.

2.6 Comparision of Reviewed Techniques Used in Audio Emotion Detection

Comparision of literature on data, classifiers, features and results obtained is in (Table 1). Looking at the comparision table Patil et al. (2012) achieved highest accuracy of 86% with Naïve Bayes in classifying male and feamle voices. The next highest accuracy of 70% is achieved by Kumar et al. (2014) to classify 10 Indian ragas. But for this research Bormane and Dusane (2013) is considered as benchmarking metric though it has achieved only 60% of accuracy it classifies 4 classes and which is close to the data that is being used in this research.

Table 1: Comparision of Literature in Audio emotion detection

Features extracted	**Classifiers and Techniques**	**Compared Results**	**Authors**
MFCC, LDB	KNN,SVM	60% is achieved	Chachada and Kuo (2014)
MFCC	KNN,Naïve Bayes , SVM	70% Accuracy was achieved in classifying 10 ragas	Kumar et al. (2014)
MFCC, LPC ZCR	Wavelet Packet Transform (WPT)	60% of accuracy is achieved with MFCC in classifying 4 classes	Bormane and Dusane (2013)
MFCC, LPC, ZCR,STE	Naïve Bayes	86 % is obtained in classifying male or female voice	Patil et al. (2012).

2.7 Conclusion

Based on the reviewed results and identified gaps in the literature there is a clear evidence that there is a need to develop identification of human emotions from human speech and answer research question (sec 1.2) and research objectives (sec 1.3) .In order to address the gap and also to support human and chatbot interaction. The next chapter presents scientific methodology approach used to develop the detection models in order to support the business users that uses chatbots for customer service.

3 Scientific Methodology Approach Used

3.1 Introduction

In data mining related research is usually done in KDD or CRISP DM methodologies but here in my scenario KDD fits best as deployment of models in the business layer is not applicable here and though the motivation for the research is to bring value in the businesses related to human speech emotion detection it is not directly linked to any business so a modified KDD approach is used here Azevedo and Santos (2008) and the design process of the research, which consists of two-tier structure, i.e. client tier and business logic tier is presented.

3.2 Modified Knowledge Discovery and Data Mining Methodology

Modified KDD methodology (refer Figure 1) for human emotion detection consists of following stages, (i) data selection where the data (human voices) from RAVDESS database is collected which are in .wav format (ii) all the data in .wav format are extracted in to feature vectors based on statistical techniques with the help of python library "Librosa" (iii) The feature vectors are transformed by scaling the features (iv) Gaussian Naïve Bayes, SVM, NN, KNN, Random Forest and Ensemble models are trained (v) Models are evaluated and interpreted by using accuracy as main measure.

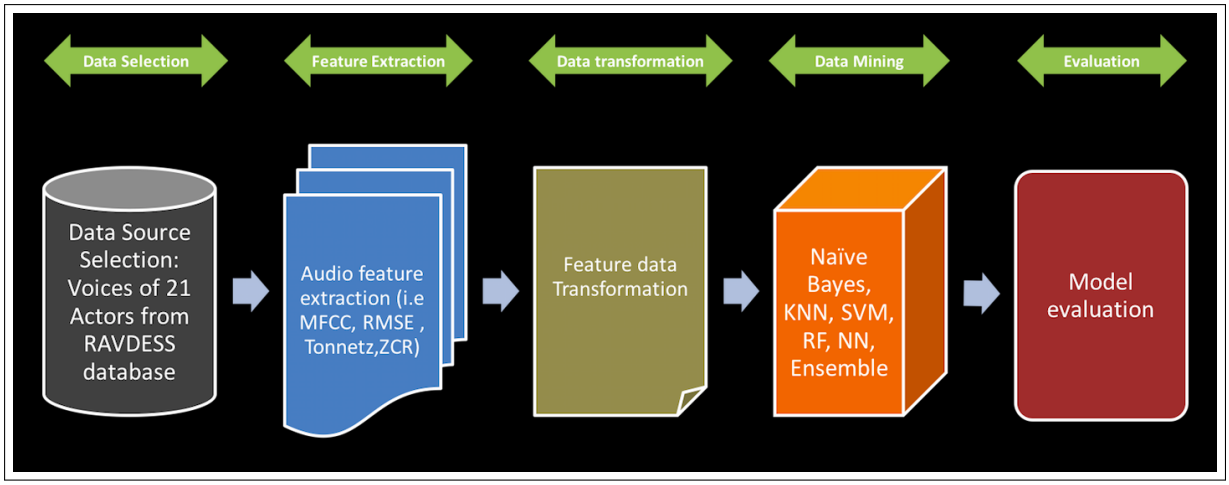

Figure 1: Methodology of human speech emotion detection

3.3 Project Design Process Flow

The project design process (depected in Figure 2) of detection of emotion in human speech consists of (i) client tier and (ii) business logic tier in client tier the interpretations from the classification models and exploratory analysis of the data are represented in visual form using Librosa (python library). In business logic tier data selection, feature extraction, transformation and training of classification models followed by evaluation of models are done.

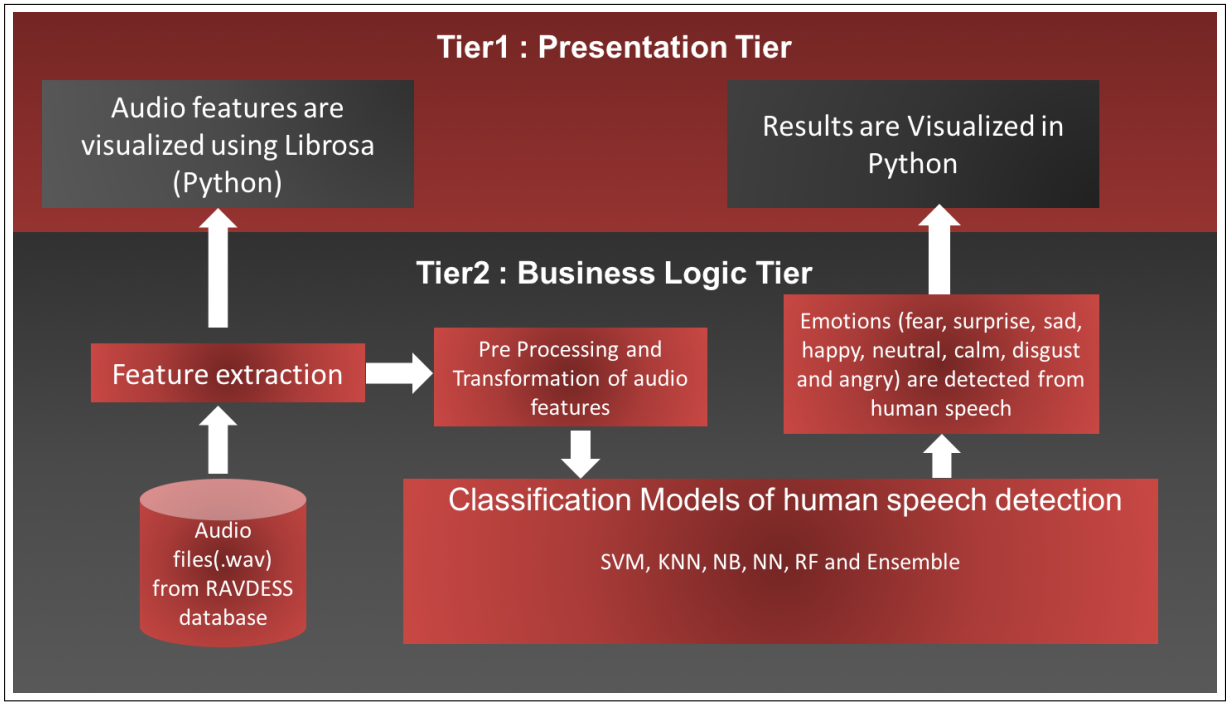

Figure 2: Project Design process of detection of emotion in human speech

3.4 Conclusion

The methodology of KDD is modified for the needs of this project and modified KDD methodology is used for this research. This methodology is adopted to the flow of the

project design process and data is collected from the RAVDESS data source. The 2 tier architecture is used as the project architecture. The implementation, evaluation and results of models to identify the human emotions from audio files is executed in the next section.

4 Implementation, Evaluation and Results of Audio Speech Emotions Detection Models

4.1 Introduction

The implementation, evaluation and results of models used in identifying the emotions from the audio are discussed in this section. Along with the implementation of the extraction of features and its selection is also thoroughly explained in this section. To evaluate the models, accuracy is used as a metric and confusion matrix is used to study the classwise true or false rate prediction model wise. In the last part of this section implemented models are compared and the highest accurate model is selected.

4.2 Extract Features from Speech Signals

Extraction of features is the most important step in this research project and research objective(a) is successfully met with the extraction of MFCC,RMSE,Tonnetz and ZCR features.The analysis of audio signals is necessary to know which features are to be extracted from the audio files to successfully classify the audio emotions. Audio analysis is also necessary to visualze the audio in the form of waves and chromogram so that to identify the change in emotions visually. The Figure 3 shows the waveplot representation of the audio signals in each emotion, it is plotted by using the amplitude of the signal and plotting it along the timeframe. "librosa.display.waveplot" library is used to get the plots.

By analysing the waveplots of different emotions it is found that each emotion is uttered with a specific pattern of amplitude along its timeframe. For example in the waveplot of neutral emotions amplitude is very less when compared to happy emotion's plot at the time 1. Though calm and happy emotions wave is filled at time 1, calm's wave plot is boarder. To analyse more about the pitch differences in the audio signals chromogram is plotted. Chromogram plot can identify 8 different pitch classes and its probability in a particular signal. In the figure 4 the chromogram of pitches (A,B,C,D,E,F,G) are plotted with a degree from 0 to 1. It can be observed that while happy the pitches are all in low values and while disgust most of the pitches are all in high degrees that is more towards 1. So it is concluded that each emotion has some particular feature pattern though it is very minor especially in a short audio signal.

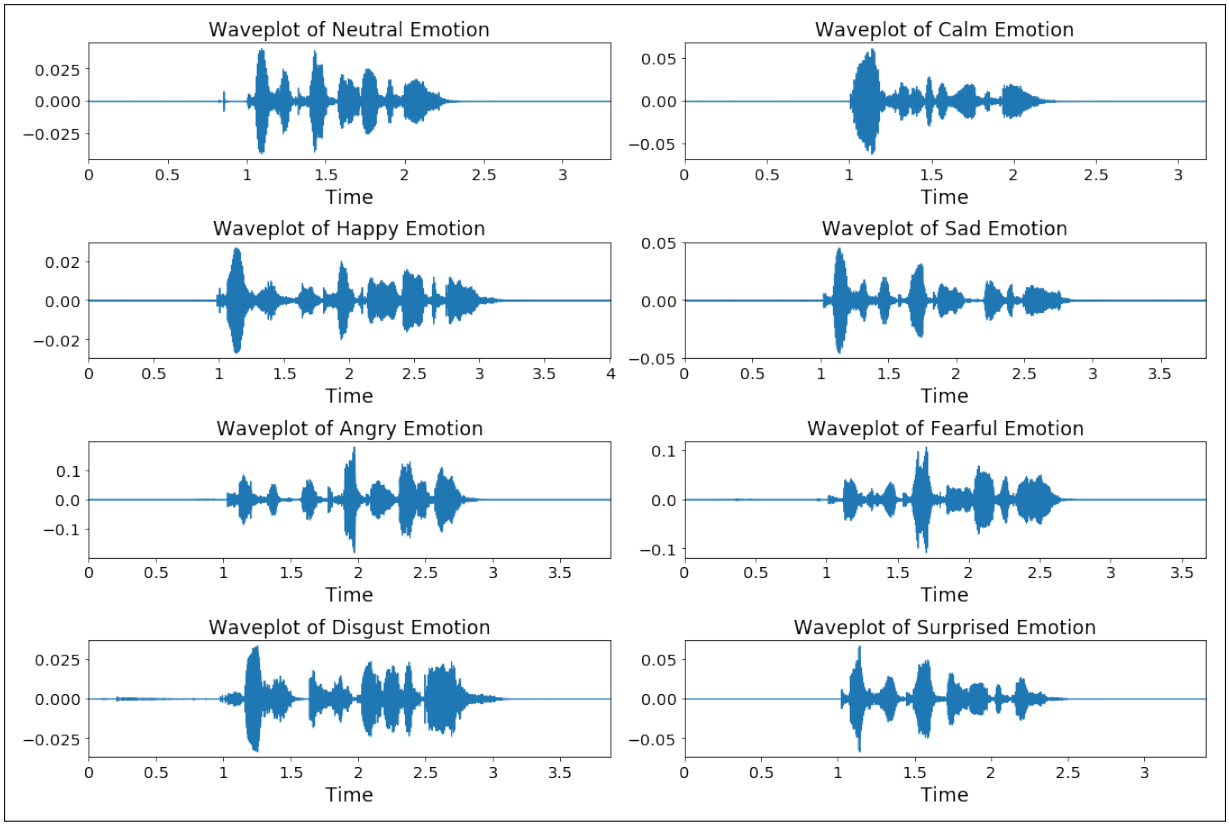

Figure 3: Waveforms of audio in various emotions

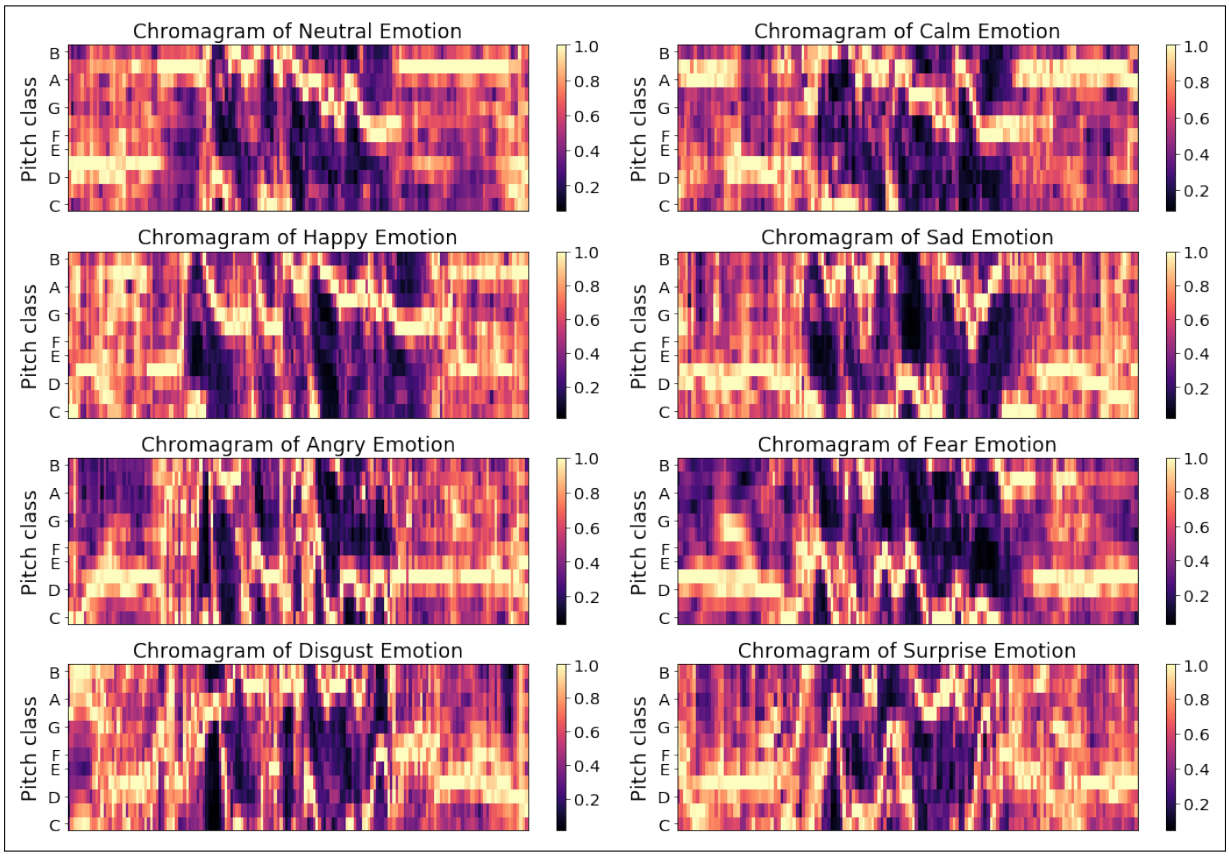

Figure 4: Chromogram of audio in various emotions

Different features such as MFCC, RMSE, Tonnetz and ZCR are extracted from the audio signals to train the models with these features in various combinations to identify which features contribute the most in the identification of emotions from the audio signals.All the features are extracted using "librosa"[1] library provided in python to extract features from music and speech signals

MFCC Extraction: The MFCC features are extracted using "librosa.feature.mfcc" and represented in the form of a matrix as follows [-8.16412911e+02, -8.16412911e+02, -8.16412911e+02, ..., -8.16412911e+02, -8.16412911e+02, -8.16412911e+02], and it is visualized in Figure 5

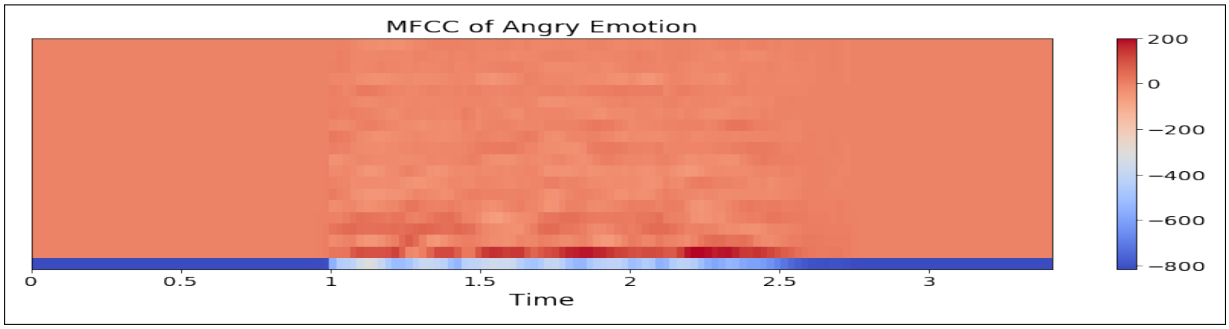

Figure 5: MFCC feature

RMSE Extraction: The RMSE features are extracted using "librosa.feature.mfcc" and represented in the form of an array as follows [1.6270229e-06, 1.8228393e-06, 1.8871159e-06, 1.8401602e-06, 1.6785326e-06, 1.1793761e-06, 1.0207252e-06, 8.5696587e-07,...]

and it can be visualized in Figure 6

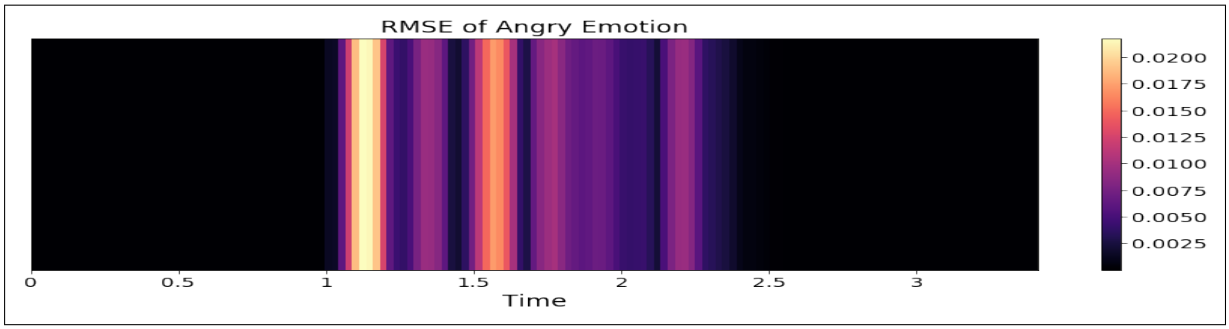

Figure 6: RMSE feature

Tonnetz Extraction: The Tonnetz features are extracted using "librosa.feature.mfcc" and represented in the form of an array as follows [-2.74305230e-02, 1.87691134e-03, 2.54168663e-02, -2.82404297e-02, -3.24596423e-02, -4.06284046e-02, -1.00543391e-01, -6.63241961e-02, -6.34902661e-02,...]

and it can be visualized in Figure 7

[1]http://librosa.github.io/

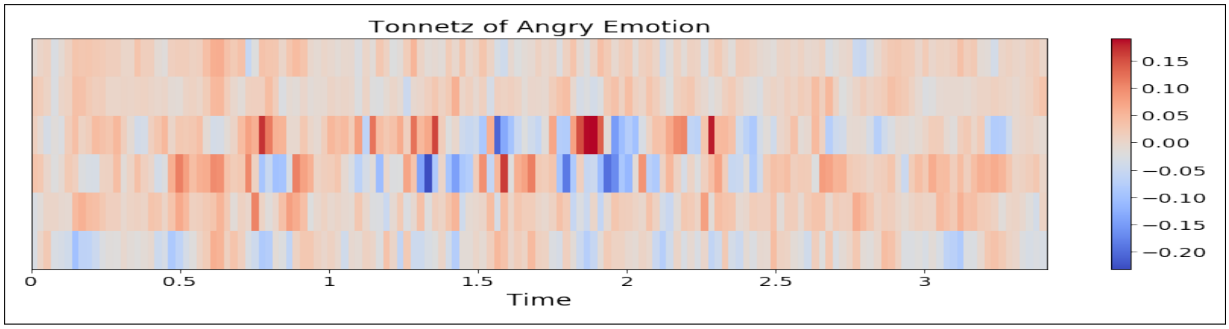

Figure 7: Tonnetz feature

ZCR Extraction: The ZCR features are extracted using "librosa.feature.mfcc" and represented in the form of an array as follows [0.26513672, 0.40185547, 0.53857422, 0.49169922, 0.45898438, 0.32226562, 0.25585938, 0.22998047, 0.21289062, 0.29638672,...]
and it can be visualized in Figure 8

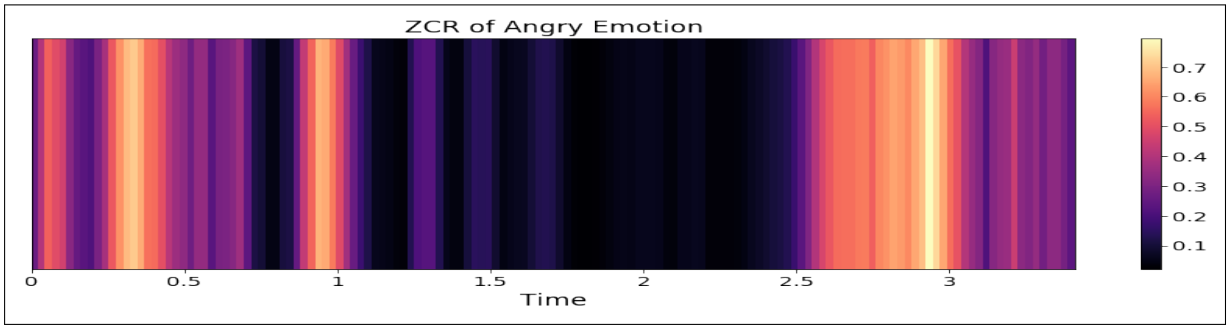

Figure 8: ZCR feature

Combined Features(MFCC+RMSE+Tonnetz+ZCR): To evaluate and compare which features are necessary all the features extracted in the form of array are all combined statistically by transforming each feature along the mean and stacking it one by one vertically Patil et al. (2012) the sample of combination on angry emotion is as follows [-6.60703718e+02, 5.33335932e+01, -7.94904181e-01, 8.58591094e+00, 4.15975216e+00, -2.17486421e+00, -6.37926291e+00, -6.27319129e+00,...]

4.3 Implementation,Evaluation and Results of Naïve Bayes

Naïve Bayes works on the conditional probability theory and since our features are all continuous Gaussian Naïve Bayes is used instead of classical Naïve Bayes which is a ideal model to use when categorical values are there in independent data. Naïve Bayes is executed by splitting the data in to training and testing datasets in 4:1 ratio i.e. 80% of train data and 20% of test data.

Implementation: Gaussian Naïve Bayes is implemented using "sklearn"[2] library in the python. The function used to implement is GaussianNB(). It is trained on the training data. This model is developed in various feature combinations i.e.(1)using MFCC features, 2) using Tonnetz features and 3) using the combination of MFCC+RMSE+Tonnetz +ZCR features)and the Metric accuracy is used here to evaluate the model.

Evaluation and Results:

[2]http://scikit-learn.org/stable/modules/classes.html

Accuracy is calculated based on the following formula

$$Overall\ Accuracy = \frac{Number\ of\ Correctly\ identified\ Audio\ emotional\ instances}{Total\ number\ of\ instances}$$

The accuracy resulted by Gaussian Naïve Bayes with MFCC features is 35% , with Tonnetz features is 16% and with the combination of all the features is 35%. The accuracy percentage is very less in this model and interestingly both the MFCC and MFCC combined with all the features is same which shows other features Tonnetz, RMSE and ZCR are showing very less impact.

It is analysed from the results that Calm and Surprised emotions are the classes that identified in more instances when compared to other classes. Happy emotion being identified as Fearful in most of the scenarios is a poorly identified category and over all this Model resulted in a poor outcome of 0nly 35% accuracy.

4.4 Implementation,Evaluation and Results of K-Nearest Neighbour

KNN works on the nearest neighbor theory. KNN is executed by splitting the data in to training and testing datasets in 4:1 ratio i.e. 80% of train data and 20% of test data. The model is implemented in a grid with different combinations of number of neighbors (1,2,3,4,5,6,7,8) , weights (distance, uniform) with metric minkowski and algorithm (auto, ball tre, kd tree, brute). Finally parameters combination that would give best accuracy is : 'p': 1, 'n neighbors': 1, 'metric': 'minkowski', 'weights': 'distance', 'algorithm': 'auto'

Implementation:

KNN is implemented using "sklearn"[3] library in the python. The function used to implement is KNeighborsClassifier(). It is trained on the training data. This model is developed in various feature combinations i.e.(1)using MFCC features, 2) using Tonnetz features and 3) using the combination of MFCC+RMSE+Tonnetz+ZCR features)and the Metric accuracy is used here to evaluate the model.

Evaluation and Results:

The accuracy resulted by KNN with MFCC features is 63% , with Tonnetz features is 21% and with the combination of all the features is 63%. The accuracy percentage is high in this model and interestingly both the MFCC and MFCC combined with all the features is same which shows other features Tonnetz, RMSE and ZCR are showing very less impact.

As the highest accuracy came with MFCC features it is analysed that this model is efficiently able to differentiate all the classes in more than 60 % of the time in most of the scenarios except 'sad' and 'surprised' which are being identified correctly at 50% of the instances only . Over all this model performance is better than the other models.

4.5 Implementation,Evaluation and results of Deep Neural Network

Deep neural network model is built here with 3 layers, in 2 activation layers 'relu' is used and in the third activation layer (output layer) 'softmax' is used other parameters used here are loss='categorical crossentropy', metrics='accuracy' and optimizer='adam'

Implementation:

[3]http://scikit-learn.org/stable/modules/classes.html

DNN is implemented using "Keras"[4] library in the python. It is trained on the training data. This model is developed in various feature combinations i.e.(1)using MFCC features, 2) using Tonnetz features and 3) using the combination of MFCC+RMSE+Tonnetz +ZCR features)and the Metric accuracy is used here to evaluate the model.

Evaluation and Results:

The accuracy resulted by DNN with MFCC features is 62% , with Tonnetz features is 20% and with the combination of all the features is 6%. The accuracy percentage is high in this model and interestingly both the MFCC and MFCC combined with all the features is same which shows other features Tonnetz, RMSE and ZCR are showing very less impact.

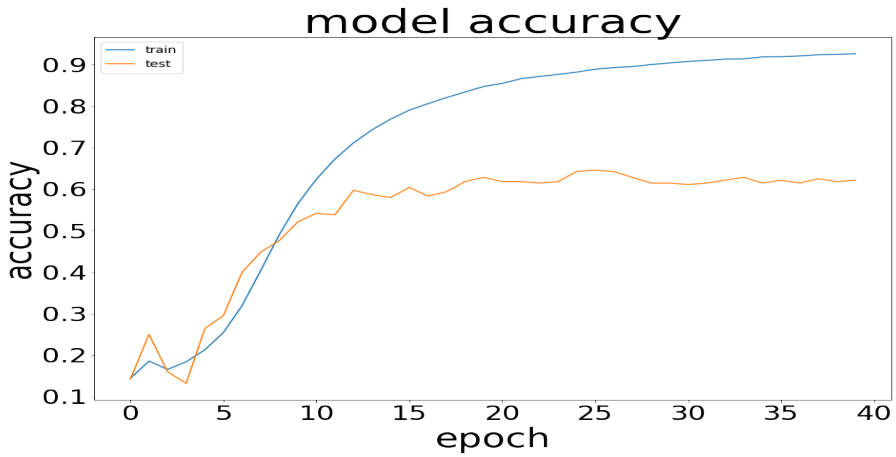

Figure 9: Accuracy across epochs with MFCC feature

In the above figure the model accuracy at each step of epoch is plotted for both the training and testing data and compared. As it is shown in the figure the testing data is resulted on a average of 62 % accuracy. This is a fair model but computationally it has drawbacks than KNN which resulted in 63% accuracy. But there are more chances that this model can beat other models given more training data as is the case in all deep learning models.

4.6 Implementation,Evaluation and Results of Support Vector Machine

SVM works on finding the plane that seperates the classes efficiently. SVM is executed by splitting the data in to training and testing datasets in 4:1 ratio i.e. 80% of train data and 20% of test data. The model is implemented in a grid with different combinations of 'C' value, gamma and kernel (i.e. 'C' (0.001, 0.01, 0.1, 1, 10, 100), 'gamma' (0.0001, 0.001, 0.01, 0.1), 'kernel' (linear and rbf) Finally parameters combination that would give best accuracy is : kernel = 'rbf',C=100 and gamma =0.0001.

Implementation:

SVM is implemented using "sklearn"[5] library in the python. The function used to implement is SVC(). It is trained on the training data. This model is developed in various feature combinations i.e.(1)using MFCC features, 2) Using Tonnetz features and 3) Using

[4]https://keras.io

[5]http://scikit-learn.org/stable/modules/classes.html

the combination of MFCC+RMSE+Tonnetz+ZCR features)and the Metric accuracy is used here to evaluate the model.

Evaluation and Results:

The accuracy resulted by SVM with MFCC features is 57% , with Tonnetz features is 10% and with the combination of all the features is 58%. The accuracy percentage is high in this model and interestingly both the MFCC and MFCC combined with all the features is same which shows other features Tonnetz, RMSE and ZCR are showing very less impact.

As the highest accuracy came with MFCC features it is analysed that this model is efficiently able to differentiate calm, neutral, happy and fearful emotions approximately with 65% though its over all accuracy is below 60%. Over all this Classifier performed fairly well in identifying all the emotions.

4.7 Implementation,Evaluation and Results of Random Forest

Random Forest is an ensemble of tree classifier. Model is executed by splitting the data in to training and testing datasets in 4:1 ratio i.e. 80% of train data and 20% of test data. The model is implemented in a grid with different combinations of estimators (10, 31, 52, 73, 94, 115, 136, 157, 178, 200) and max features (auto, log2) The highest accuracy is obtained with max features': 'auto', and estimators as 178

Implementation: Random Forest is implemented using "sklearn"[6] library in the python. The function used to implement is RandomForestClassifier(). It is trained on the training data. This model is developed in various feature combinations i.e.(1)using MFCC features, 2) Using Tonnetz features and 3) Using the combination of MFCC+RMSE+Tonnetz +ZCR features)and the Metric accuracy is used here to evaluate the model.

Evaluation and Results:

The accuracy resulted by Random Forest with MFCC features is 59% , with Tonnetz features is 18% and with the combination of all the features is 60%. The accuracy percentage is high in this model and interestingly both the MFCC and MFCC combined with all the features is same which shows other features Tonnetz, RMSE and ZCR are showing very less impact.

Though highest accuracy came with combined features the difference is just 1% and as the computational expenses with the combination of all the features is very high when compared to only MFCC features. The later method should be preferred. it is analysed that this model is efficiently able to differentiate calm, disgust and surprised emotions approximately with 65% though its over all accuracy is below 60%. Over all this classifier performed fairly well in identifying all the emotions.

4.8 Implementation,Evaluation and Results of Ensemble

Ensemble is a method of using all the pre developed models and the idea is that as few models are better in identifyng few classes while other models are better in identifying others by ensembling these models better accuracy can be achieved. Here of all the models developed before KNN, SVM and Random Forest are chosen for emsembling where Neural Network and Naïve Bayes are not chosen becasue of computation drawback and less accuracy respectively.

[6]http://scikit-learn.org/stable/modules/classes.html

Implementation: Ensemble is implemented using "sklearn"[7] library in the python. Here the outcomes are considered based on the highest voting from the other pre trained models (KNN, SVM, Random Forest) The function used to implement is VotingClassifier(). It is trained on the training data. This model is developed in various feature combinations i.e.(1)using MFCC features, 2) Using Tonnetz features and 3) Using the combination of MFCC+RMSE+Tonnetz+ZCR features)and the Metric accuracy is used here to evaluate the model.

Evaluation and Results:

The accuracy resulted by Random Forest with MFCC features is 67% , with Tonnetz features is 20% and with the combination of all the features is 67%. The accuracy percentage is high in this model and interestingly both the MFCC and MFCC combined with all the features is same which shows other features Tonnetz, RMSE and ZCR are showing very less impact.

With MFCC and combined features this model resulted in highest accuracy when comapred to other individual models. Confusion matrix is provided below for further analysis of its efficiency in each class identification . It is clear that in more than 80% of the instances it is able to identify happy and sad emotions where all all other features are identified with more than 60% of the insatnces except fear which is onlt 51%. Overall this model resulted in best performance with highest accurate results.

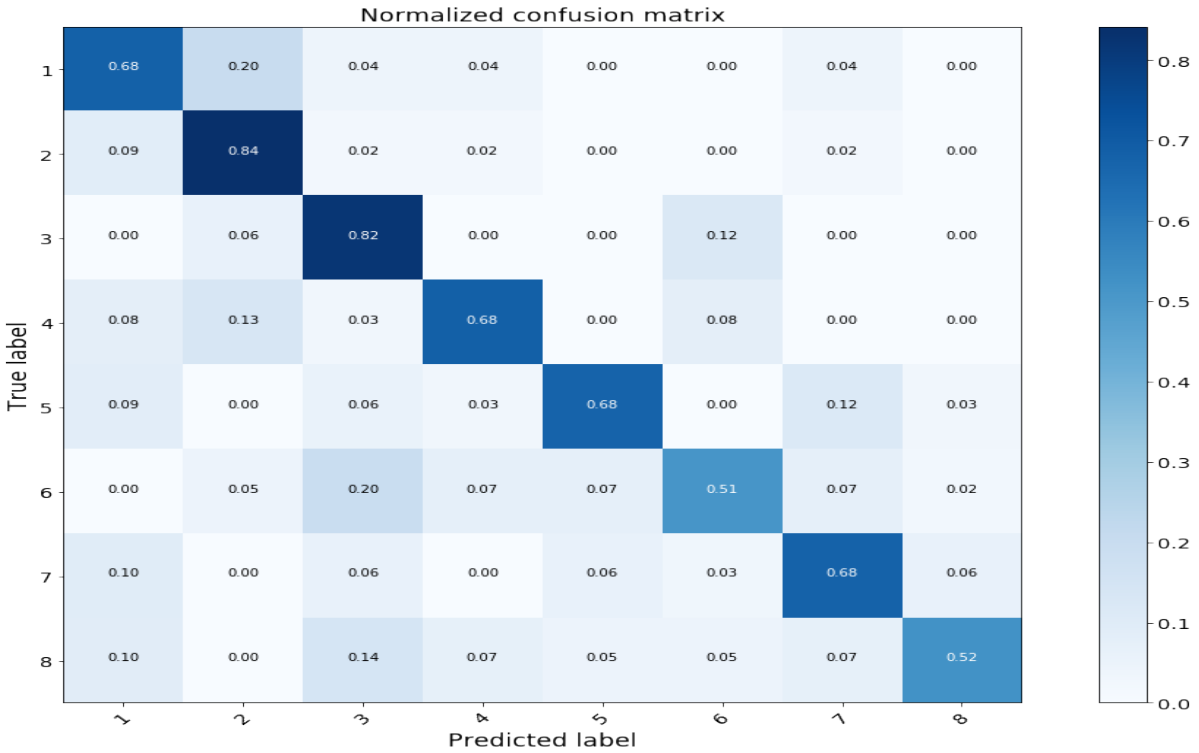

Figure 10: Confusion Matrix of MFCC in Ensemble

4.9 Comparision of Developed Models

All the developed models Naïve Bayes,KNN, DNN, SVM, Random Forest and Ensemble are compared below. It is clear from the figure that Ensemble model has highest accuracy

[7]http://scikit-learn.org/stable/modules/classes.html

in both the MFCC and combined features scenario and also it is noted that MFCC is the main feature that is contributing for the classification in the models as both the accuracies in MFCC and Combined features are not much of a difference and so in order to have better computational value it is suggested to use the ensemble model with only MFCC feature

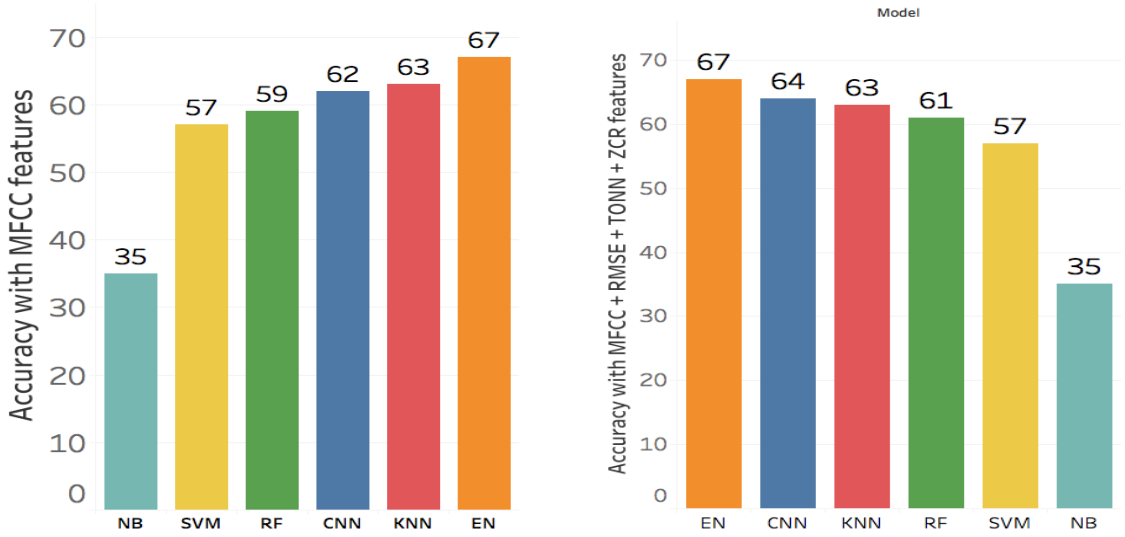

Figure 11: Model comparision with MFCC and combined features

4.10 Comparision of Developed Models with Existing Models

From the comparison table (Table 1) in literature review Kumar et al. (2014) has achived 70% accuracy in classifying the music of 10 indian ragas and Bormane and Dusane (2013) achieved 60% in classifying 4 classes of sound.

Author	Data Classified	Accuracy	Reasons for high/low Accuracy
Kumar et al (2014)	10 Indian music ragas	70%	Classification of emotions is more complex and detailed than music ragas
Bormane and Dusane(2013)	4 music classes	60%	Better feature extraction is executed here
Patil et al (2012)	Male or Female voice	86%	Only 2 classes (male or female voice)

Figure 12: Comparision with previous models

The figure 12 presents the reasons for high/low accuracy achieved compared to this research project.

4.11 Chatbot Framework

The chatbot framework is implemented here to establish how the identified emotions can be used by an audio chatbot in its interactions just by using the audio without depending on the meaning of a text. The motivation here is to make chatbot interaction

more human-like. Sample framework is provided below when the customer tone is angry when speaking to the chatbot and how it can respond.

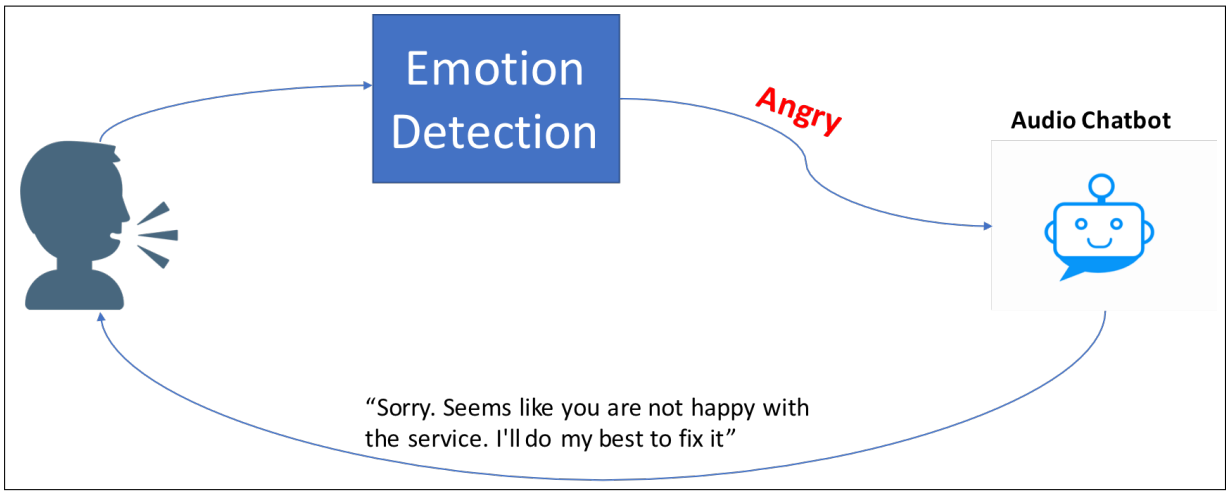

Figure 13: Chatbot framework example

Likewise, the chatbot can change its framing of dialogue interaction depending on the customer interaction. If the customer's voice is identified as "disgust" chatbot can respond as "Apologies for your situation. I will do my best to fix it" if he/ she voice is "fearful" it can ask "Please feel free to let me know if there are any doubts in your mind" and if he/she voice is "happy" it can respond with "I am glad that you are happy with my service".

4.12 Conclusion

Based on the implemented models and produced results the project has fully answered research question in section 1.2. In addition all the research objectives (refer section 1.3) have been tackled and the resulta have been answered. The developed models and results will contribute significantly to the body of knowledge and also improve the field of audio emotional intelligence.

5 Conclusion and Future Work

This research crucial step is in extracting features from the audio and identifying the right features for efficient identification of emotions. MFCC feature is the best feature suitable for classification and several classifiers are used to build the model but however ensemble method is the best with highest accuracy of 67% and it performed better than than the benchmarking value of 60% Bormane and Dusane (2013). This can be applied in businesses where customer support is there, since most of the customer support services should be 365 x 24 hours chatbots with emotional intellogence can be a game changer.Though chatbot framework is used as an example to show the implications and usage of emotion detection it can be applied in many business area such as in car safety alert design for psychological doctors when treating patients e.t.c. The limitations of this project is that it can identify only 8 types of emotions from the audio and when there is external noise in the background of the voice the models may not work effectively.

Future Work:

This research work can be followed up with identifying the emotions from audio with external noise as well using data of audio files with voices in different emotions and with natural background noise. As the implemented project is done in 3 months of time the data collection was a big hurdle. In the future if more amount of data can be used there is chance in increasing the accuracy of the models. As only supervised data is used here with limited conversations in the future unsupervised methods can be used to make use of more data available in the audio databases.

6 Acknowledgement

I would specifically like to extend my thanks to Dr. Catherine Mulwa for her supervision, guidance and support through out the research. I would like to acknowledge my mother, father and brother for their trust in me.

References

Azevedo, A. I. R. L. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview, *IADS-DM* .

Bezooijen, R. v. (1984). *Characteristics and Recognizability of Vocal Expressions of Emotion:*, DE GRUYTER, Berlin, Boston.

Bormane, D. and Dusane, M. (2013). A novel techniques for classification of musical instruments, *Information and Knowledge Management*, Vol. 3, pp. 1–8.

Cakir, E., Adavanne, S., Parascandolo, G., Drossos, K. and Virtanen, T. (2017). Convolutional recurrent neural networks for bird audio detection, *Signal Processing Conference (EUSIPCO), 2017 25th European*, IEEE, pp. 1744–1748.

Chachada, S. and Kuo, C.-C. J. (2014). Environmental sound recognition: a survey, *APSIPA Transactions on Signal and Information Processing* **3**: e14.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. G. (2001). Emotion recognition in human-computer interaction, *IEEE Signal processing magazine* **18**(1): 32–80.

Dasgupta, S., Harisudha, K. and Masunda, S. (2017). Voiceprint analysis for parkinson's disease using mfcc, gmm, and instance based learning and multilayer perceptron, *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, IEEE, pp. 1679–1682.

Devillers, L. and Vidrascu, L. (2006). Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs, *Ninth International Conference on Spoken Language Processing*.

Dutta, S. and Ghosal, A. (2017). A hierarchical approach for silence/speech/music classification, *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, IEEE, pp. 3001–3005.

Dwivedi, S. and Roshni, V. K. (2017). Recommender system for big data in education, *E-Learning & E-Learning Technologies (ELELTECH), 2017 5th National Conference on*, IEEE, pp. 1–4.

El Ayadi, M., Kamel, M. S. and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition* **44**(3): 572–587.

Gidhe, P. and Ragha, L. (2017). Sarcasm detection of non # tagged statements using MLP-BP, *2017 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pp. 1–4.

Ikemoto, Y., Asawavetvutt, V., Kuwabara, K. and Huang, H.-H. (2018). Conversation strategy of a chatbot for interactive recommendations, *Asian Conference on Intelligent Information and Database Systems*, Springer, pp. 117–126.

Izquierdo-Reyes, J., Ramirez-Mendoza, R. A., Bustamante-Bello, M. R., Pons-Rovira, J. L. and Gonzalez-Vargas, J. E. (2018). Emotion recognition for semi-autonomous vehicles framework, *International Journal on Interactive Design and Manufacturing (IJIDeM)* pp. 1–8.

Jin, Q., Li, C., Chen, S. and Wu, H. (2015). Speech emotion recognition with acoustic and lexical features, *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, IEEE, pp. 4749–4753.

Kumar, V., Pandya, H. and Jawahar, C. (2014). Identifying ragas in indian music, *Pattern Recognition (ICPR), 2014 22nd International Conference on*, IEEE, pp. 767–772.

Patil, H. A., Madhavi, M. C., Jain, R. and Jain, A. K. (2012). Combining evidence from temporal and spectral features for person recognition using humming, *Perception and Machine Intelligence*, Springer, pp. 321–328.

Schmitt, M., Ringeval, F. and Schuller, B. W. (2016). At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech., *Interspeech*, pp. 495–499.

Schubiger, M. (1958). *English intonation, its form and function*, M. Niemeyer Verlag.

Schuller, B., Batliner, A., Steidl, S. and Seppi, D. (2009). Emotion recognition from speech: putting asr in the loop, *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, pp. 4585–4588.

Tolkmitt, F. J. and Scherer, K. R. (1986). Effect of experimentally induced stress on vocal parameters, *Journal of Experimental Psychology. Human Perception and Performance* **12**(3): 302–313.

Yang, W. and Krishnan, S. (2018). Sound event detection in real-life audio using joint spectral and temporal features, *Signal, Image and Video Processing* pp. 1–8.

Zhang, W., Wang, H., Ren, K. and Song, J. (2016). Chinese sentence based lexical similarity measure for artificial intelligence chatbot, *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pp. 1–4.