**AUDIO EMOTION RECOGNITION SYSTEM USING DEEP LEARNING**

**Dissertation submitted in part fulfilment of the requirements for the degree of**

**MSc. in Data Analytics**

**at**



**Kaustubh Tikar**

**10526351**

**MSc. Data Analytics August, 2020**

**DECLARATION**

I, Kaustubh Tikar, a student of Dublin Business School. Studying Masters of Data Analytics, declare that this research is my original work and it has never been presented to any institution or university for the award of Degree or Diploma. In addition, I have referenced correctly all literature and sources used in this thesis and this work is fully compliant with the Dublin Business School’s academic honesty policy. Steps have been taken to adhere to GDPR policies.

Signed: Kaustubh Tikar

Date: 25th August 2020.

**ACKNOWLEDGMENTS**

I would like to take this opportunity and thank Professor Abhishek Kaushik, my Supervisor. Thank you, Sir, for leading me on this daunting journey to finish my master's thesis. For helping me always to gain the knowledge necessary to work on this topic of dissertation and to answer even my rudimentary questions and doubts. Also in Dublin Business School I would like to thank everyone for their kind support. All my lecturers and academic staff were a tremendous help in getting me to this point in the process of the master's thesis. That journey would not have been possible had it not been for your support. At last I would also like to thank all my friends and family for who stood up with me during all this process.

**ABSTRACT**

In order to the increase the recognition accuracy of speech emotion recognition and the characterization ability of speech, an Audio Emotion Recognition (AER) based on deep neural network is proposed. Later, research methods involving pre-processing of the data and feature extraction techniques are presented. For the utterance of speech, Mel-frequency Cepstral Coefficient is used further. Then we applied convolution Neural Network (CNN) and Long Short Term Memory (LSTM) for building our model where CNN outperformed LSTM. The following research is carried out on two files from benchmark datasets i.e. RAVDESS which includes speech and song files. Further recognition of eight emotions neutral, calm, happy, sad, angry, fearful, disgust, surprised, anger, fear, joy, can be accomplished. A remarkable accuracy of 76% was gained after evaluating the proposed model

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| CHAPTER 1: INTRODUCTION | 9 |
| 1.1 Background of the Dissertation Topic | 9 |
| 1.2 Detailing the problem or Defining the problem | 10 |
| 1.3 Research Question | 10 |
| 1.4 Dissertation Roadmap | 11 |
| CHAPTER 2: LITERATURE REVIEW | 13 |
| 2.1 Deep neural networks for emotion classification using speech | 13 |
| 2.2 Machine learning methods for emotion recognition using speech/audio | 15 |
| 2.3 Feature extraction techniques for audio/speech | 16 |
| 2.4 Conclusion | 18 |
| CHAPTER 3: RESEARCH METHODOLOGY | 19 |
| 3.1 Introduction | 19 |
| 3.2 Research Philosophy | 20 |
| 3.3 Research Method | 20 |
| 3.3.1 Business Understanding | 21 |
| 3.3.2 Data Understanding | 21 |
| 3.3.3 Data preparation | 22 |
| 3.3.3.1 Pre-processing and feature extraction for audio | 23 |
| 3.3.4 Design specification | 24 |
| 3.3.4.1 Design specification for CNN model | 24 |
| 3.3.4.2 Design specification for LSTM model | 25 |
| 3.3.5 Model building | 25 |
| 3.3.5.1 Building CNN model for audio modality | 25 |
| 3.3.5.2 Building LSTM model for audio modality | 27 |
| 3.3.6 Evaluation | 27 |
| 3.3.6.1 Hyperparameter tuning for CNN model | 28 |
| 3.3.6.2 Hyperparameter tuning for LSTM model | 28 |
| CHAPTER 4: ARTEFACT DESIGN & DEVELOPMENT | 29 |
| 4.1 Introduction | 29 |
| 4.2 Hardware setup | 29 |
| 4.3 Software Information | 29 |
| 4.4 Data preparation | 31 |
| 4.5 Implementation | 32 |
| 4.6 Keras for deep learning | 35 |
| CHAPTER 5 : DATA FINDINGS AND ANALYSIS | 36 |
| 5.1 Results of CNN model | 36 |
| 5.2 Results of LSTM model | 37 |
| CHAPTER 6: DISCUSSION | 39 |
| CHAPTER 7: CONCLUSION AND FUTURE WORK | 40 |
| References | 41 |

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Fig. 1 Dissertation Roadmap | 11 |
| Fig.2 Industry Process for Data Mining (CRISP-DM) | 19 |
| Fig.3 Block diagram of Mel Frequency Cepstral Coefficients (Source: Created by Author) | 23 |
| Fig.4 Mel-frequency cepstral coefficient of a sample audio signal (Source: Created by Author) | 23 |
| Fig.5 Waveform and Mel-frequency cepstral coefficient of a sample audio signal (Source: Created by Author) | 24 |
| Fig.6 Emotion recognition system architecture for CNN model (Source: Created by Author) | 25 |
| Fig.7 Emotion recognition system architecture for LSTM model (Source: Created by Author) | 25 |
| Fig.8 CNN model for audio (Source: Created by Author) | 26 |
| Fig.9 LSTM model for audio (Source: Created by Author) | 27 |
| Fig.10 Computer Specification (Source: Created by Author) | 29 |
| Fig.11 Anaconda Software for Python (Source: Created by Author) | 30 |
| Fig.12 Anaconda Installer (Source: Created by Author) | 30 |
| Fig.13 Jupyter Environment (Source: Created by Author) | 31 |
| Fig. 14 Extracting features and MFCC coefficients | 31 |
| Fig. 15 Building CNN model | 32 |
| Fig. 16 Building LSTM model | 32 |
| Fig. 17 Building Support vector model | 33 |
| Fig. 18 Building decision tree classifier | 33 |
| Fig. 19 Building random forest classifier | 34 |
| Fig.20 Adaboost ensemble technique | 34 |
| Fig. 21 Keras Documentation | 35 |
| Fig. 22 Graph plots of accuracy and loss function for train and test data. (Source: Created by Author) | 36 |
| Fig.23 Confusion Matrix for CNN model (Source: Created by Author) | 37 |
| Fig.24 Confusion Matrix for LSTM model (Source: Created by Author) | 37 |

**LIST OF TABLES**

|  |  |
| --- | --- |
| Table.1 Software Requirement Specifications (Source: Created by Author) | 20 |
| Table.2 Data description | 21 |

**Chapter Six**: This is the chapter of discussion, here is where the whole thesis is reviewed and scrutinized. This is where we will be implementing the findings. In this chapter we will answer the research questions and give discussion on the subject findings.

**Chapter Seven:** Is the Theme Conclusion and sets out the final results of the study and concludes the subject.

**Chapter Eight**: In this study the References used. It is important to have this topic, so that we can recognize our fellow academics and researchers' work. It also contains details of the various sites and reading material that were used to understand the research question.

**Chapter Six**: This is the chapter of discussion, here is where the whole thesis is reviewed and scrutinized. This is where we will be implementing the findings. In this chapter we will answer the research questions and give discussion on the subject findings.

**Chapter Seven:** Is the Theme Conclusion and sets out the final results of the study and concludes the subject.

**Chapter Eight**: In this study the References used. It is important to have this topic, so that we can recognize our fellow academics and researchers' work. It also contains details of the various sites and reading material that were used to understand the research question.

**Chapter Six**: This is the chapter of discussion, here is where the whole thesis is reviewed and scrutinized. This is where we will be implementing the findings. In this chapter we will answer the research questions and give discussion on the subject findings.

**Chapter Seven:** Is the Theme Conclusion and sets out the final results of the study and concludes the subject.

**Chapter Eight**: In this study the References used. It is important to have this topic, so that we can recognize our fellow academics and researchers' work. It also contains details of the various sites and reading material that were used to understand the research question.

**CHAPTER 1**

**INTRODUCTION**

In recent year's quick movement in the field of innovation expands fascination of brilliant gadgets in our day by day life. As there is exceptional development in clients of cell phones, speech has gotten one of the most well-known methods in which human machine correspondence is built up. Comprehension of the human emotion by a machine has benefits in segments like mechanical autonomy, computer generated reality and games. Any player's misery in feelings can be perceived in computer generated reality and games. Discourse is one of a kind for every speaker and the feelings in it can't be covered up with the exception of if the speaker attempts to imagine it to do as such. What's more, consequently hypothetically feelings of the speaker can be assessed. So as to make PC to do so a computational strategy is utilized to recognize human emotions from speech sound. One of the model is the examination of feelings on voice mails where speech methodology assumes an imperative job to catch human feelings in which audio features could be pulled out from a speech.

**1.1 Background of the Dissertation Topic**

Speech is the most ordinary way of expressing ourselves as humans. It is just characteristic at that point to stretch out this correspondence medium to PC applications. We describe speech emotion recognition (SER) systems as assembly of methodologies that process and categorize speech signals to distinguish the rooted emotions. Speech Recognition has become progressively 'glib' nowadays: Alexa, Cortana, Siri, and a lot more discourse frameworks have hit the purchaser advertise on a more extensive premise than any time in recent years, yet do any of them really notice our feelings and respond to them like a human conversational accomplice would?

In the past emotion recognition from audio signals has been a challenging and active field of research. Verbal expression together with human emotion makes emotional speech processing more complex and calculation based. (Mohan Ghai, 2017)Previous work done in this area consists of use of various classifiers like SVM, Neural Network, Bayes classifier. (Bagus Tris Atmaja, 2019) Recognizing human emotions by machine gives a big advantage in areas such as games, virtual reality, robotics and call center. Furthermore, emotion recognition is useful in inquisition process.

Speech emotion recognition (SER) intends to perceive the basic enthusiastic condition of a speaker from her voice. The territory has gotten expanding research intrigue through all these years. There are numerous uses of recognizing the feeling of th e people like in the interface with robots, sound observation, online E-learning, business applications, clinical investigations, diversion, banking, call focuses, cardboard frameworks, PC games, and so on.

Verbal expression together with human emotion makes emotional speech processing more complex and calculation based.

**1.2 Detailing the Problem or Defining the Problem**

Even though a remarkable progress is made in the field of speech emotion recognition, some problems still are into existence. For various speech emotion datasets, distinctive feeling models might be utilized by the various analysts which makes selection of model more challenging for emotion computation. We are still a long way from normally communicating with the machines. A SER system need a classifier, a supervised learning concept that will be trained to identify from new speech signals and such supervised system also requires labeled data consisting emotions involved in it. Also the data requires preprocessing before extracting the features and features are mandatory for classification process which helps in reducing the data to its more characteristics importance.

(Mohan Ghai, 2017)A single utterance may have different emotions where each part may represent some part of the utterance and boundary between these parts is not easy to determine hence making it difficult for predicting a single emotion from the utterance. One of the issue is the type of emotion which depends on speaker, culture and environment. Also feature learning is main factor behind improving the performance of speech emotion recognition it’s problematic to extract effective features by using machine learning algorithms.

Obviously, there are mixed emotions. While these are captivating turns of events, obstacles in selection remain. Emotion detection and analysis triggers a few privacy and security issues. Are clients ready to have their feelings being analyzed? Is assent required? It will require some investment to correctly choose cases and to decide the best datasets to catch while ensuring the data can be adequately predicted to improve client experience.

**1.3 Research Question**

1. How the human emotion can be accurately classified using Convolutional Neural Network.

**1.4 Dissertation Roadmap**



**Fig. 1 Dissertation Roadmap**

This dissertation is divided into the following chapters:

**Chapter One**: This chapter includes in depth of background of this master’s research. It states the problem with its sources. Also an emphasis is put forward on the research question, and at end what will be obtained as an output. The objectives and aims, the limitations of this research, plus how the research is lagging and to overcome it what improvements can be made.

**Chapter Two**: Delivers the literature review of the proposed topic. It also appraises the presence of each topic by doing an in depth study of the topic.

**Chapter Three**: The following chapter will expose the proposed research methodology and methods of leading the research topic. It shows how the research is been conducted and the methods combined during this research.

**Chapter Four**: This chapter involves the artefact design and it’s through development. It shows how the artefact was made and how is the working of the prototype. It goes deep into the design of the bot and its databases.

**Chapter Five**: This chapter Discusses about the Data collected and the evaluation of data. It reveals the findings of the model evaluation study. It discloses the details of the study.

**Chapter Six**: This is the chapter of discussion, here is where the whole thesis is reviewed and scrutinized. This is where we will be implementing the findings. In this chapter we will answer the research questions and give discussion on the subject findings.

**Chapter Seven:** Is the Theme Conclusion and sets out the final results of the study and concludes the subject.

**Chapter Eight**: In this study the References used. It is important to have this topic, so that we can recognize our fellow academics and researchers' work. It also contains details of the various sites and reading material that were used to understand the research question.

**List Of Key Words: CNN, LSTM, RNN, MFCC, SVM, CRISP-DM**

**Chapter Seven:** Is the Theme Conclusion and sets out the final results of the study and concludes the subject.

**Chapter Eight**: In this study the References used. It is important to have this topic, so that we can recognize our fellow academics and researchers' work. It also contains details of the various sites and reading material that were used to understand the research question.

**CHAPTER 2**

**LITERATURE REVIEW**

**2.1 Deep neural networks for emotion classification using speech**

The study (Haiqing Zheng, 2019)puts emphasis on recognizing the musical/song by using deep neural network. The design of emotion recognition consists of four networks. At the very beginning is the ID CNN layer followed by TD-FC layer. Then comes BiGRU layer for mining the needed temporal for emotion recognition. And finally a system is built by risking a CNN layer on time time-distributed layer producing an accuracy of 48%.But the accuracy appears to be quite low as compared with alternate standard model.

In (Richard Orjesek, 2018) use of multiple restricted Boltzmann Machine (RBM) is done to on audio signals under Deep Belief Machine. For differentiating emotions speed of speech is used and finally a 24 dimensional MFCCs are used for extracting the time frequency features. Further this extracted features are provided as input to RBM network, building a DBN network which comprises of multiple layers of RBMs.AT the end the gained accuracy on training data was 86% and on testing data was 78%.The reconstruction error of DBN model is condensed by 5.3% due to use of ReLu activation function.

In (Jillur Rahman Saurav, 2018) various Gaussian Mixture Model consisting Hidden Markov Model (GMM-HMM) based and Deep Neural Network (DNN-HMM) based models are being analyzed for the purpose of speech recognition in Bangla language for a search engine. Then acoustic model was applied to predict the phonemes and language modeling technique was also applied in order to improve the accuracy. Further for feature extraction author applied MFCC and then LDA transforms and MLLT estimation was also applied which Kaldi already has for feature transformation. Further author also applied SAT and boosted MMI to obtain better accuracy. At the end, lowest WER was recorded as compared to other work done on Bangla language.

A SER method built on spectrogram and phoneme sequence is anticipated in (Promod Yenigalla, 2018). Three different network architecture were considered in search of better accuracy on a benchmark IEMOCAP dataset. First was CNN model with phoneme input where sequence phoneme is being extracted from spectrum input speech which helps in identifying words or set of words hidden in speech. Second model is CNN for spectrogram features in which spectrogram is given as input to 2D CNN and the spectrogram is generated by STFT. Third model used is Multi-channel CNN model with phoneme and spectrogram features which consists of combination of both phoneme and spectrogram. In the end it was observed that the model formed by combination provides 4% more accuracy then the existing method

In (Caroline Etienne, 2018) a neural network is designed for recognizing emotions in speech. Further to overcome with class imbalance and small size of dataset, data augmentation is carried out by vocal tract length perturbation and over sampling the least represented class of dataset. Experiments were conducted using various depths of convolution and Bi-LSTM modules. Different combinations were implemented like “CNN + deep Bi-LSTM”, “deep CNN + shallow Bi-LSTM” and “deep CNN + deep Bi-LSTM”. At end the best result were achieved with choice of 4 convolution and 1 Bi-LSTM layers. The obtained weighted accuracy was 64.5% and unweighted was 61.7% respectively.

(Sandeep Kumar Pandey, 2018) Puts emphasize on catching and categorizing state of emotion from speech utterances by using CNN and LSTM. A CNN model involving 3 convolution layers with input shape of spectrogram image 129 × 251 and to lessen size of feature maps max-pooling size of 2 × 2 is used further implemented. To capture the temporal dependencies in a better way, a bi-directional LSTM model with cell size 128 is applied. Plus a combination of CNN and LSTM is proposed next to achieve end-to-end emotion recognition where CNN is for feature extraction and LSTM to gain class probabilities. Then the model is optimized by Adam optimizer and RELUS as an activation function which lowers likelihood of vanishing gradient. And by fluctuating Number of cells in LSTM and by varying number of convolution layers and filters in CNN, A remarkable accuracy of 82.35 % is achieved at the end.

In (Jun Deng, 2017), focus is on exploring deep learning technique on anger detection on speech. German voice data was used and was labeled with three labelers, two men and one women and then two classification tasks including three-class and two-class tasks. Then a LSTM RNN architecture consisting one self-connected linear memory cell with three multiplicative gate which has one input gate, a forget gate and an output gate. Further, a network with two LSTM and two convolution layers is proposed with 200 hidden units per LSTM layers. The BLSTM model has obtained a maximum accuracy UARs of 80.1 % and 79.4 % for ComParE feature and log-mel filter bank and it was observed that deep neural network performs quite better.

(Md Amaan Haque, 2018)Consists of four DNN architecture i.e. triangular, rectangular, modified triangular and modified rectangular and comparison is made between them on the basis of training point and accuracy. The number of input and output layer neurons in all the architecture is set aside to be 1500 and 10. And also in total 6 hidden layers were added in every model. All the architecture vary by just fluctuating the number of neurons in hidden layer as compared with input layer. Then these models are further evaluated for recognition of speech by using TIDIGITS database. Then by implementing MFCC technique 1500 feature vectors are extracted. Finally 99.31 was the highest accuracy obtained and it was for modified triangular architecture. On the other hand minimum training time of 49.72 was recorded for triangular architecture

The study under (Aharon Satt, 2017) presents emotion recognition from para-lingual content in speech carried out by deep neural network which is then applied to spectrograms. IEMOCAP a very common dataset is used which consists of scripted improvised dialogue for the purpose of testing. In the part of preprocessing, unwanted noise was removed from the logapectrogram and noise filtering is carried out further. Then topologies were evaluated with one and two LSTM layer and one to six convolution layers. It was noticed that one of the network tested topologies reached 67.3% and 62% vs. precious research 63.9% and 62.8% for overall accuracy and class accuracy accordingly.

Study (JUN DENG, 2017)states how emotion recognition models can be trained on unlabeled whispered speech data based on acoustic feature learning. Here three feature transfer learning methods denoising autoencoders, shared-hidden-layer autoencoders, and extreme learning machines autoencoders are used. INTERSPEECH 2009 Emotion challenge was chosen, which includes 12 functions applied to 2 X 16 acoustic LLDs counting their first order delta regression coefficients. While comparing the performances of all auto encoder based methods, it was observed that this method is highly influenced by fluctuating the number of hidden units in a specific range. For acoustic features, finally, the proposed method drastically improved the performance in UAR and it also was beneficial for normal phonated speech recognition to some extent.

In (Michael Neumann, 2019), it is seen that how recognition accuracy can be increased by adding representations learnt by unsupervised autoencoder in an emotion classifier based on CNN model. Angry, sad, happy, and neutral classes are used from IEMOCAP for SER. ACNN together with multi-view is used for SER and openSMILE toolkit for the purpose of feature extraction. In order to extract, autoencode training and for creating spectrogram representations, auDeep toolkit is used. Then various combinations of uni-directional and bidirectional encoders were tested and it was clear that using both of them was beneficial. Maximum accuracy was seen for sad, neutral and angry when ACNN and AE model are combined together. Further discriminative strength of the representations was found in t-SNE visualization.

**2.2 Machine learning methods for emotion recognition using speech/audio**

(Fatemeh Noroozi, 2017) Comprises of vocal-based emotion recognition method by using random forest. RF decision making algorithm is applied to speech signal which consist of emotion categories. Further by using multi-class classifier, every voice signal is assigned with an emotion label. Then speech signal consisting set of emotions which are to be analyzed are produced. Further variation in intensity and speech are evaluated by avoiding linguistic data. Then to classify and recognize the emotions, RF decision trees are used. At last, on the basis of majority of voting from the forest of the trees class labels are assigned to input vector. For choosing different partitions of training and validation sets, k-fold cross validation is also used. By using sample for the test and letting other for training, the obtained average recognition rate was 66.28% and happiness had highest recognition performance i.e. 78%.

In (Salsabil Besbes, 2019)for the extraction of cepstral feature from speech under stress, a multitaper technique is proposed. Speech samples from SUSAS database are used to gain results obtained from multitaper method. To carry out extraction multitaper MFCC extraction is implemented. Further kernel SVM based classifier is used for the purpose of classification. One Against One (OAO), One Against All (OAA) and Oneclass SVM (OC-SVM) are the three approaches selected for multiclass support vector machines classification. As a result, best recognition rate was produced by combination of OC-SVM and Multipeak MMFCC which was 99.46%.

In (Kun-Yi Huang, 2019) both verbal and nonverbal sounds in utterances were taken into consideration for the recognition of emotions using deep neural network. The work was then further divided into three sub categories. Initially verbal and non-verbal sound segments were extracted by Prosodic Phases auto tagger and PRAAT silence detector. Further for classifying the verbal and non-verbal sounds, SVM model was used which are then provided as input to Bi-LSTM model. The obtained accuracy was 63 %, 8 % increase in the previously studied researches. Plus the model also extracted feature for noise in the background, which was an error while categorizing the human emotion.

**2.3 Feature extraction techniques for audio/speech**

(AlifBinAbdulQayyum, 2019) Consists of a CNN network based emotion recognition system. Here SAVEE dataset is used and for feature extraction, MFCC and modulation spectral features (MSF) are considered. The MFCC values are obtained by the first 12 discrete cosine transform (DCT) of Mel log energies for the purpose of classification. And by removing the spectro-temporal processing and considering regular acoustic frequency jointly with modulation frequency, which are performed in human auditory system. Then to select the best or remove the worst feature a model (e.g. linear regression or SVM) is used accordingly. To remove the mean feature and normalize them into unit variance Speaker normalization is carried out further. And at last 7 types of emotions are classified by great accuracy.

An emergency parking instruction recognition system which is completely based on speech emotion recognition is presented in (Tian Kexin, 2019). MFCC is used further for feature extraction in which firstly pre-emphasis is done on the signal to improve SNR and then 6dB octave pre-emphasis filter is used with frame length 25ms.Then spectral characteristics are gained from transforming signal from time domain to frequency domain by using FTT. Then statistical characteristics calculated by MFCC feature are applied as input to SVM. To separate sentimental samples and speech emotion recognition into 5 parts, 5-fold cross validation is used. Then DWT algorithm is implemented for matching parking command distance where the lowest sum of matching distance is used, thus giving an accuracy of 95% for the speech recognition.

Whereas in (Mohan Ghai, 2017) SER is carried out to gain emotion such as anger, boredom, disgust, anxiety, sadness, happiness and neutral. For this purpose Berlin database is used. Then some of the crucial features like MFCC and energy of speech are extracted from the audio signals and then these features are summarized in order to apply supervised learning algorithm. Feature extraction was further done by combining 25 frames in segments and using this segments as data point’s leads to obtain increased features and averaging effect. Then relationship between actual measured frequency and supposed frequency of pure tone is obtained from Mel scale. Further energy is extracted from intensity of speech and then by combining MFCC feature and energy, clear view of emotion of speech are determined. Further SVM is applied followed by gradient boosting and Random Forest algorithm. Finally it was seen that Random forest gained maximum accuracy among all which was 81.05%. Whereas the samples of anger class received the highest accuracy and it was minimum for happiness.

In (Novita Belinda Wunarso, 2017) SER is carried out on Indonesian language by extracting simple features from 3420 voice data samples collected from total of 38 participants. Then speech data is further classified into happy, neutral and sad by providing it to neural network and SVM. Further feature extraction is carried out in which three features were extracted namely the average of amplitude, the length of speech signal and average of approximation coefficients derived from one level decomposition of DWT using Daubechies wavelet db1 to db4. Overall SVM gives more accuracy which was 76.84% from which 60% was for non-emotional and 85% for emotional speech. It was observer that emotions in speech can be withdrawn from speech duration which can be very beneficial for Indonesian speech-emotion automatic recognition (I-SpEAR) system.

Whereas, in the study (Bagus Tris Atmaja, 2019) identical work was done on IEMOCAP dataset by audio modality. Here the speech segments are exposed to silence removal filter by minimum number of samples and threshold value. Audio features are then extracted from these filtered data and the obtained output is in the form of vectors. Attention layer is inserted with LSTM layer to deal with unnecessary extracted features. Hence the overall training period gets shortened and only those speech segments which are valuable in emotional recognition are allowed to get through as an input to output layer. The accuracy gained was 76 % which outperformed other models.

(Yu Tian, 2018)Puts emphasis on accuracy of children speech in noisy and clean condition. For the purpose of feature extraction FFT based MFCC is used, but MFCC based recognition didn’t work accordingly due to robustness and hence to overcome RSA and RFC are engaged further. For expectation maximization (EM), Baum-Welch algorithm is taken into consideration which is present in hidden Markov model. Then to eliminate the spikes in spectrogram obtained due to recognition error, RSA is implemented. Without noise the obtained accuracy was 99% and decreases further if noise is added. It was noticeable that in white noise, accuracy was increased by about 40 % due to RSA and 50 % due to RSF concluding that RSF overcomes RSA. Also in children isolated speech recognition system, RSA and RSF improves robustness.

**2.4 Conclusion**

Literature review is conducted to gain perspective into the world of Speech emotion recognition. Best of the researches are compared in overall to gain a better insight on CNN based models. Also a remarkable knowledge of Feature extraction was gain on MFCC coefficients extraction process. Also broader research done on LSTM model will be supportive in building the system. The techniques used in Speech emotion recognition in the researches are beneficial in generating artifact for our proposed methodology.

The Bi-LSTM model is originally built using the default parameters. A ReLu activation function is used to activate a BiLSTM layer to avoid problems related to the vanishing gradients. In Bi-LSTM and LSTM layers, a recurrent dropout layer is added to avoid the Modell overfitting. The model is further assessed by tuning the hyperparameters such as optimizer, loss function, and applying criteria for early stoppage. Initially, for model building, an Adam optimizer is used, followed by attempting to optimize Adagrad and RMSprop. Then, taking sparse categorical cross-entropy and categorical cross-entropy tunes the loss function. Third, the model is subject to early stop criteria, for which accuracy and validation accuracy are chosen for monitoring the model's early stoppage. In addition, the model uses the 14 confusion matrix to assess the predicted outcomes.

The Bi-LSTM model is originally built using the default parameters. A ReLu activation function is used to activate a BiLSTM layer to avoid problems related to the vanishing gradients. In Bi-LSTM and LSTM layers, a recurrent dropout layer is added to avoid the Modell overfitting. The model is further assessed by tuning the hyperparameters such as optimizer, loss function, and applying criteria for early stoppage. Initially, for model building, an Adam optimizer is used, followed by attempting to optimize Adagrad and RMSprop. Then, taking sparse categorical cross-entropy and categorical cross-entropy tunes the loss function. Third, the model is subject to early stop criteria, for which accuracy and validation accuracy are chosen for monitoring the model's early stoppage. In addition, the model uses the 14 confusion matrix to assess the predicted outcomes.

**CHAPTER 3**

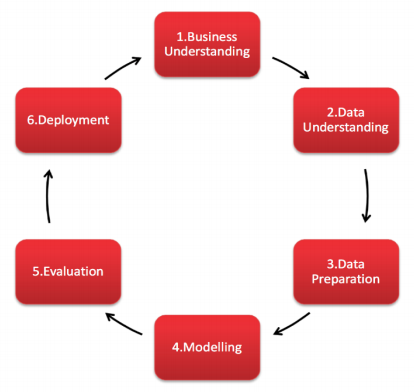
**RESEARCH METHODOLOGY**

**3.1 Introduction**

The research methodology applied for this artifact is the exchange among deliberations and strategies obtained during the conceptualization of the framework. Deliberation in this edge of reference is the calculations, libraries and datasets used. Whereas the methods referred here are for creating functions and deep analysis of some critical areas. Specific libraries are summed-up in order to get accurate results. This section is basic in characterizing a way to accomplish the objective of the thesis. A research method is an efficient and sorted out examination of an issue planned for comprehending it. In this master’s thesis we comprise of a novel methodology in the field of speech emotion recognition i.e. CRISP-DM which will provide a boost to the previous proposed systems. Increasing the accuracy, robustness and gaining accurate emotions in this methodology can be more advantageous in building more realistic SER models.

CRISP-DM is broad data mining methodology and process model that equip anyone with actual information for operating a data mining project. CRISP-DM is the standard methodology when you are handling data-centric projects as it is robust and simultaneously provides flexibility and customization The CRISP-DM model summarizes the steps involved in performing data science activities from business need to deployment but more importantly defines a framework that allows communication through all the phases.

This model consists of six major phases defined as follows: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.



**Fig.2 Industry Process for Data Mining (CRISP-DM)**

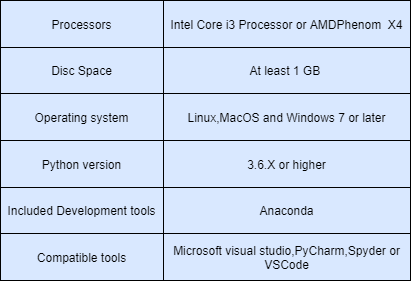
**3.2 Research Philosophy**

Reason it is still conveyed and utilized in this day, as a result of its six stages which levels an amateur individual to fabricate a business project, from initially understanding the Business Requirements what are the key segments and need for setting up business venture at that point Understanding Data how it tends to be gathered and be utilized for building up the business project, in light of the understandings in like manner set up the Data(Data Preparation) which would be utilized in information digging apparatuses for applying Modeling Techniques to perform tests on the removed business information whether it meets the necessary quality and how well it predicts on the noteworthy information before any future expectations, in view of this test it is additionally Evaluated (Evaluation) in view of the discoveries and model yield at that point decide the following stage, were we sum up the entire procedure to characterize the business measures, regardless of whether information gathered was adequate and precise to take right choices, before definite Deployment of the Business Project in which we give the game plan, last report how the task would execute and what steps are required for the execution of the Business Project. Extra it gives an advantage, which permits hi006D to move or bounce to and fro do numerous cycles between various stages at whatever point required and roll out essential improvements according to the necessities.

Having such points of interest oppressed above, CRISP-DM approach has a few disadvantages as stages are not been fused or missed which are available in the upgraded variant of CRISP-DM and different Methodologies such TDSP, SEMMA and so on

**3.3 Research Method**

In this research approach the initial step was to gather the prerequisites for the frameworks. The necessities for the framework are expressed underneath in the accompanying table:

****

**Table.1 Software Requirement Specifications (Source: Created by Author)**

**3.3.1 Business Understanding**

In this stage, the essential target of the business is controlled by characterizing the related provision of the difficult case. Information given by emotion examination is multifaceted and can give data on each part of the connection at every snapshot of the scene which can help in tracking emotional reactions over time. Emotion recognition can be utilized to get on a client's manner of speaking and state of mind, and to characterize the call with the correct need to the correct operator which will ultimately lead to emotion based call routing. At the point when an operator is in line with a client's emotions, the discussion can be custom fitted to guarantee sympathy, thus powering customer personalization. For instance, specialists giving directions to setting up a smart TV can see confusion enlisted on a client's face, empowering them to repeat or improve the steps. Voice analytics may enable an operator to identify significant levels of dissatisfaction and offer customized support that tends to the client's particular issue. When there's a language obstruction or a noisy environment, a voice-to-message application will empower specialists to profit by feeling investigation, giving experiences into a client's temperament when discourse or facial examination isn’t possible.

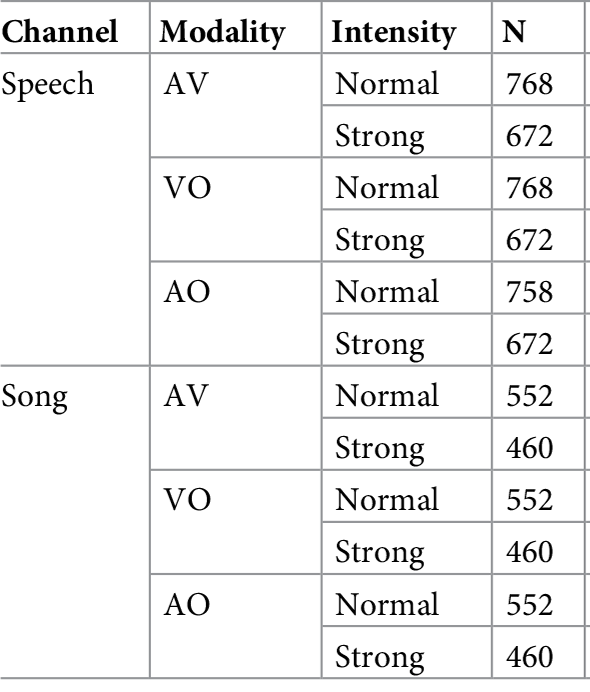
**3.3.2 Data Understanding**

This stage involves initial understanding of raw data and class distribution of target variables. For this research work, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) containing 1440 speech files and 1012 Song files is been used. Speech file (Audio\_Speech\_Actors\_01-24.zip, 215 MB) consists of 1440 files: 60 trials per actor x 24 actors = 1440. Song file (Audio\_Song\_Actors\_01-24.zip, 198 MB) includes 1012 files: 44 trials per actor x 23 actors = 1012. The database contains 24 expert entertainers (12 female, 12 male), vocalizing two lexically-coordinated articulations in an unbiased North American intonation. Speech consists of calm, sad, happy, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions.

Each of the 7356 RAVDESS files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 02- 01-06-01-02-01-12.mp4). These identifiers define the stimulus characteristics:

* Modality (01 = full-AV, 02 = video-only, 03 = audioonly);
* Vocal channel (01 = speech, 02 = song);
* Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised);
* Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the ’neutral’ emotion;
* Statement (01 = ”Kids are talking by the door”, 02 = ”Dogs are sitting by the door”);
* Repetition (01 = 1st repetition, 02 = 2nd repetition);
* Actor (01 to 24. Odd numbered actors are male, even numbered actors are female)

The emotion substance of the dataset was assessed by 350 human members, every utterance getting 10 appraisals from 10 unique members. The members were approached to recognize the emotion communicated by the performer from the arrangement of the objective feelings, or show none is right. An understanding rate extending from 0 to 1 was evaluated for every utterance. If there is total agreement between target emotion and evaluator is meaning of score 1 and if scores 0 it means complete lack of agreement. In order to avoid use of exaggerated performance only utterances with normal intensity were used. Further, the dataset was decreased for two concerns: (1) performer having only has speaking data and (2) there are two emotions to speaking. Following table shows the description of data for sung and spoken expression across modality, channel and emotional intensity and since neutral has no intensity manipulation thus neutral scores were merged into normal category.

****

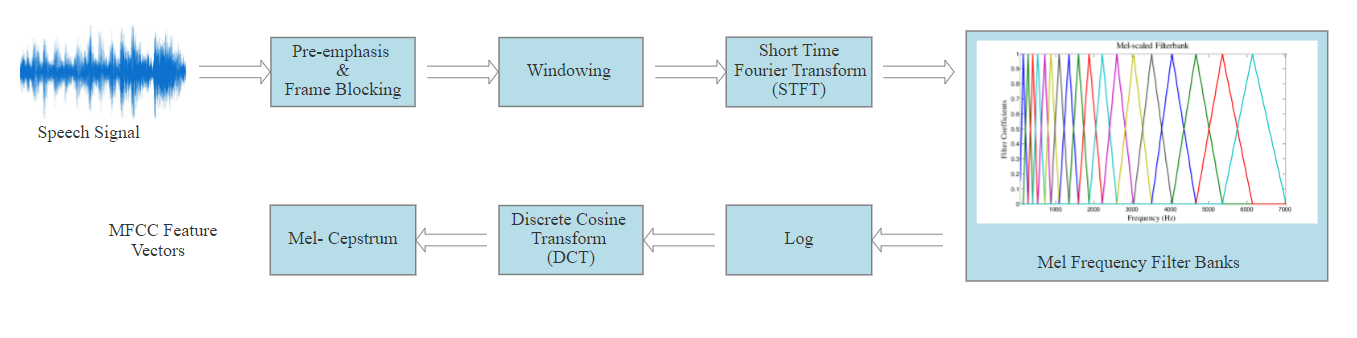
**Table.2 Data description**

**3.3.3 Data preparation**

To train a deep learning model and to obtain valuable insights from data we need to transform data. In this stage selected data is first cleaned and then features are being extracted from it.

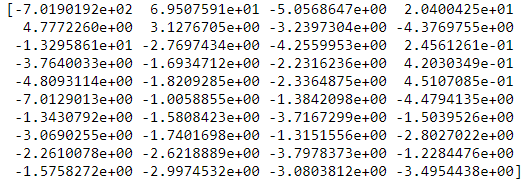
**3.3.3.1 Pre-processing and feature extraction for audio**

Speech is a waveform consisting of continuously changing signal. The model of classification of emotions proposed here is based on deep learning and built on dense layer and convolution neural network. Thus main idea here is considering Mel-frequency cepstral coefficients (Mohan Ghai, 2017), which is commonly mentioned as the “spectrum of spectrum”, which is the only feature used to train the model. MFCC is an alternate translation of the Mel-frequency cepstrum (MFC), and it has been shown to be the best in class of sound formalization in automatic speech recognition task. The MFC coefficients have essentially been utilized has the result of their ability to speak to the sufficiency range of the sound wave in a minimal vectorial structure. The block diagram shows the process of extracting MFCC.



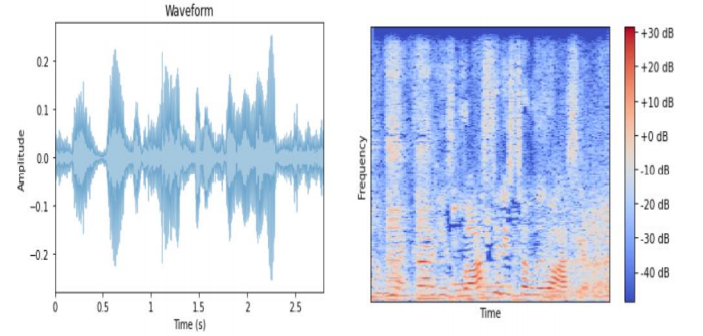
**Fig.3 Block diagram of Mel Frequency Cepstral Coefficients (Source: Created by Author)**

In this research librosa library is used in which has inbuilt function to extract these MFCC features for us, thus saving the time. As described in (AlifBinAbdulQayyum, 2019), the file record is divided into frames, for the most part utilizing a fixed window size, so as to get statistically fixed waves. The Discrete Fourier Transformation is applied on the acquired small frames, and just the amplitude spectrum's logarithm is kept. The extraction of MFCC parameters is also subject to FFT, which helps in reducing the amount of calculation of this part. The amplitude range is standardized with a decrease of the "Mel" frequency scale. This activity is performed for identifying frequency more important for a critical recreation of the wave as the human auditory framework can see. 40 features have been extracted for each audio file.

****

**Fig.4 Mel-frequency cepstral coefficient of a sample audio signal (Source: Created by Author)**

After converting every audio file to a time series of floating point, the features have been generated which can be seen in **Fig.4** and a sequence of MFCC has been created through the time series. The fig.5 shows the waveform plotted using amplitude and time for a sample signal.



**Fig.5 Waveform and Mel-frequency cepstral coefficient of a sample audio signal (Source: Created by Author)**

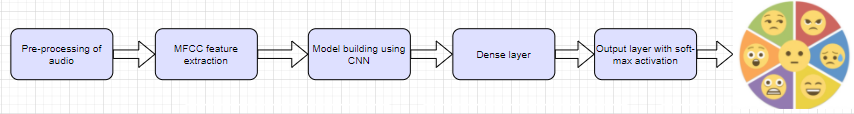
Humans are better at differentiating slight changes at low level of frequencies and the wrapping of frequency provides speech a better demonstration. But when dealing with robustness, performance of MFCC based system is not as expected (Yu Tian, 2018).

**3.3.4 Design specification**

In this part we will be manipulating data according to project need and thus making predictions for further study.

**3.3.4.1 Design specification for CNN model**

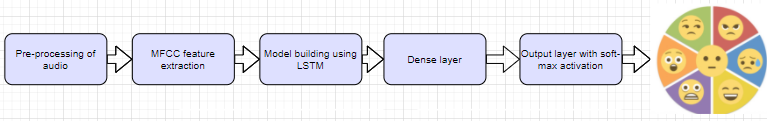
In this section design architecture of our SER is been illustrated. The CNN model comprises of all the steps involved in the fig BELOW. At the initial stage Pre-processing of raw data is carried out in order obtain a precise output. Then later feature extraction is been done by extracting MFCC features. Accordingly CNN model is built further for the main purpose of decision making.



**Fig.6 Emotion recognition system architecture for CNN model** **(Source: Created by Author)**

**3.3.4.2 Design specification for LSTM model**

The LSTM model consists of all the steps involved in the FIGURE. At the underlying stage Pre-preparing of raw data is completed all together acquire an exact output. Afterward feature extraction is been finished by separating MFCC highlights. And then a LSTM model is formed followed by dense layer and output layer to gain results.



**Fig.7 Emotion recognition system architecture for LSTM model** **(Source: Created by Author)**

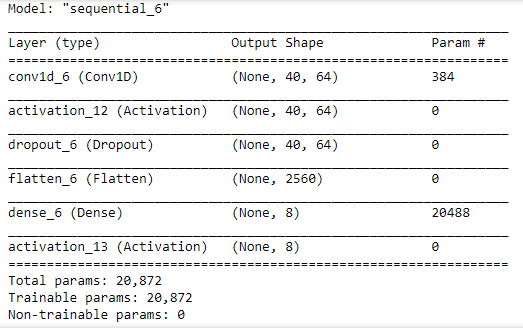
**3.3.5 Model building**

As indicated by CRISP-DM, Modeling stage is iterative and recursively returns back to the data arrangement stage. In this section we will be building our model for speech emotion recognition. We have built two models for audio modality: 1. CNN model for audio modality and 2. LSTM model for audio modality which are discussed below.

**3.3.5.1 Building CNN model for audio modality**

In the feed-forward design, CNN is a quite popular. In Deep neural Network, to substitute the affine transformation, Convolution operation is utilized by CNN (Sandeep Kumar Pandey, 2018). In general, CNN comprises of a convolution layer consisting of sets of convolutional filters in order to extract local patterns, so that many feature maps can be produced at local region in input space (Sandeep Kumar Pandey, 2018). A CNN is a Deep Learning algorithm which will take an input and will provide importance to different aspects and objects in the input and finally will differentiate them from each other.

The deep neural network intended for the classification task is accounted to be operational in FIGURE BELOW.

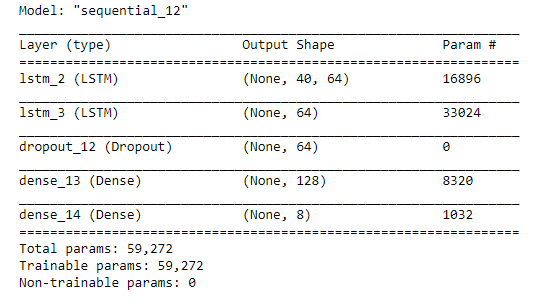
****

**Fig.8 CNN model for audio (Source: Created by Author)**

The network is functional to work on vectors of 40 features for every audio file given as input. The compact numerical form of the audio frame of 2s length is signified by 40 value. Thus, we provide as input a of size < number of training files > x 40 x 1 on which we executed one round of a 1D CNN consisting of ReLu activation function with dropout of 20% and a max-pooling function 2 x 2. . The rectified linear unit (ReLu) can be dignified as g(z) = max{0, z}, and it permits us to obtain a big value in case of activation by applying this function as a decent choice to signify hidden units. .In this case, Pooling can help the model to keep focus only on crucial characteristics of every part of data which takes *tem* invariant according to their position. Further we ran the described process once again by fluctuating the kernel size. Then next another dropout was applied which flattered the output by compiling it with the next layer. Finally, we added one Dense layer ( fully connected layer) together with a softmax activation function, changing the size of output from 640 components to 8 and accepting the likelihood conveyance for every one of the classes appropriately encoded (0=Neutral; 1= Clam; 2= Happy; Sad=3; Angry=4; Fearful= 5; Disgust=6; Surprised=7).

**3.3.5.2 Building LSTM model for audio modality**

Next version of Recurrent neural network (RNN) is Long Short Term Memory (LSTM) which uses gates to control data flow and outperforms RNN’s problem of exploding and vanishing since RNN used previous time frames. The design of LSTM itself is constructed is such a way that it can read the long term dependencies of sequence. And this provides assistance while capturing important and valuable information even through utterances. LSTMs are expressly intended to maintain a strategic distance from long term dependencies issue. Recalling data for extensive amount of time is their only soul purpose , All intermittent neural systems have a chain of continuously rehashing modules of neural system, while in regular RNN a simple structure is used by repeating module. The figure BELOW illustrates the model for LSTM built for our research.



**Fig.9 LSTM model for audio (Source: Created by Author)**

**3.3.6 Evaluation**

In this segment, the executed models are assessed to get an ideal answer for emotion recognition. The evaluation of the proposed models have been carried so that we can investigate if the model can produce results with an accuracy that are good and sufficient to yield interesting reflections .These reflections can be used for work in future on the subject which can involve speeches in real noisy domains. Various models of classification are been evaluated beyond that plan which makes it possible to generate baselines for the received results. This section mainly consists of

**3.3.6.1 Hyperparameter tuning for CNN model**

Initially, the CNN model is built by implementing default parameters. For activation of a convolution layer, a ReLu activation function is used in order to neglect the difficulties caused due to vanishing gradients. Plus, a recurrent dropout layer is places in addition to escape the overfitting of the proposed model. Then further the model is evaluated by tuning the hyperparameters like loss function, optimizers, and implementing early stopping criteria. At very first, an Adam optimizer is used for building model followed by making an attempt for Adagrad and RMSprop as an optimizer. Then, tuning of loss function is carried out by taking sparse categorical cross-entropy and categorical cross-entropy. At last, the model is exposed to early stopping criteria, in which validation accuracy and accuracy are selected for checking the early stopping of the model. In addition, the model uses confusion matrix to access the predicted outcomes.

**3.3.6.2 Hyperparameter tuning for LSTM model**

The Bi-LSTM model is originally built using the default parameters. A ReLu activation function is used to activate a BiLSTM layer to avoid problems related to the vanishing gradients. In Bi-LSTM and LSTM layers, a recurrent dropout layer is added to avoid the Modell overfitting. The model is further assessed by tuning the hyperparameters such as optimizer, loss function, and applying criteria for early stoppage. Initially, for model building, an Adam optimizer is used, followed by attempting to optimize Adagrad and RMSprop. Then, taking sparse categorical cross-entropy and categorical cross-entropy tunes the loss function. Third, the model is subject to early stop criteria, for which accuracy and validation accuracy are chosen for monitoring the model's early stoppage. In addition, the model uses the 14 confusion matrix to assess the predicted outcomes.

The Bi-LSTM model is originally built using the default parameters. A ReLu activation function is used to activate a BiLSTM layer to avoid problems related to the vanishing gradients. In Bi-LSTM and LSTM layers, a recurrent dropout layer is added to avoid the Modell overfitting. The model is further assessed by tuning the hyperparameters such as optimizer, loss function, and applying criteria for early stoppage. Initially, for model building, an Adam optimizer is used, followed by attempting to optimize Adagrad and RMSprop. Then, taking sparse categorical cross-entropy and categorical cross-entropy tunes the loss function. Third, the model is subject to early stop criteria, for which accuracy and validation accuracy are chosen for monitoring the model's early stoppage. In addition, the model uses the 14 confusion matrix to assess the predicted outcomes.

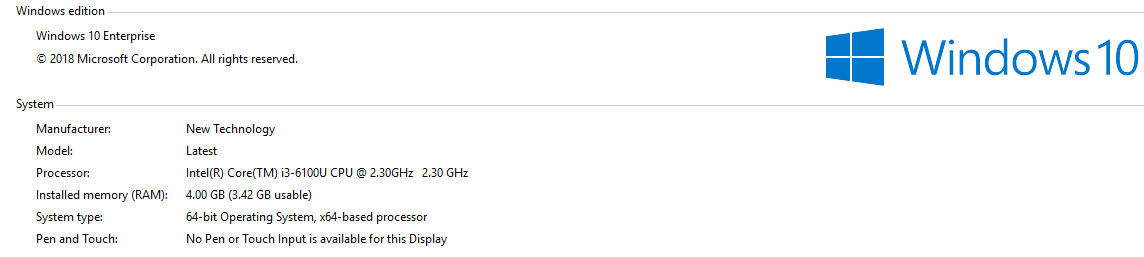
**CHAPTER 4**

**ARTEFACT DESIGN & DEVELOPMENT**

**4.1 Introduction**

A graphical demonstration of a prototype system is nothing but the artefact design. It consists of all the work done in the implementation part in Jupiter notebook(Anaconda). The actual components used in order to build a prototype are illustrated in the upcoming figures. Here the main focus of presenting this artifact is to give user a better insight of the prototype.

**4.2 Hardware setup**

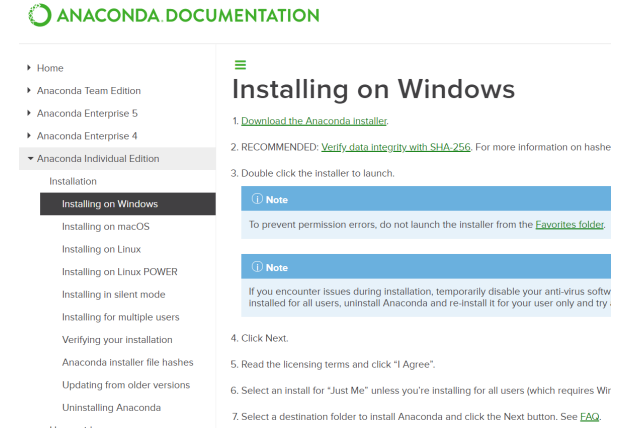
****

**Fig.10 Computer Specification (Source: Created by Author)**

The machine specifications used in this research project are illustrated in figure above. It Has 4 GB RAM installed with 64-bit Operating System and Windows 10 installed Configure. The machine's processor is Intel ® Core TM i3-6100U CPU with Intel ® UHD Graphics620.

**4.3 Software Information**

For this research project, an Anaconda software is installed in our machine for accessing the Jupyter environment. All of the prototype work is carried out in this environment.

****

**Fig.11 Anaconda Software for Python (Source: Created by Author)**

The anaconda can therefore be installed with proper functioning for each operating system. A windows 64-bit anaconda was installed for this work.



**Fig.12 Anaconda Installer (Source: Created by Author)**

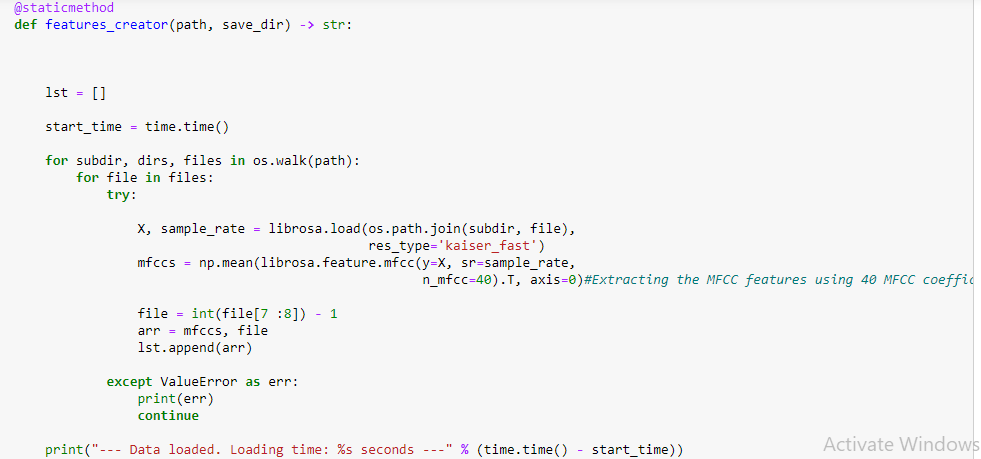
We created Jyputer notebook by clicking New on the python 3 option as demonstrated in the figure below. Consequently, in this work we have created it.



**Fig.13 Jupyter Environment (Source: Created by Author)**

**4.4 Data preparation**

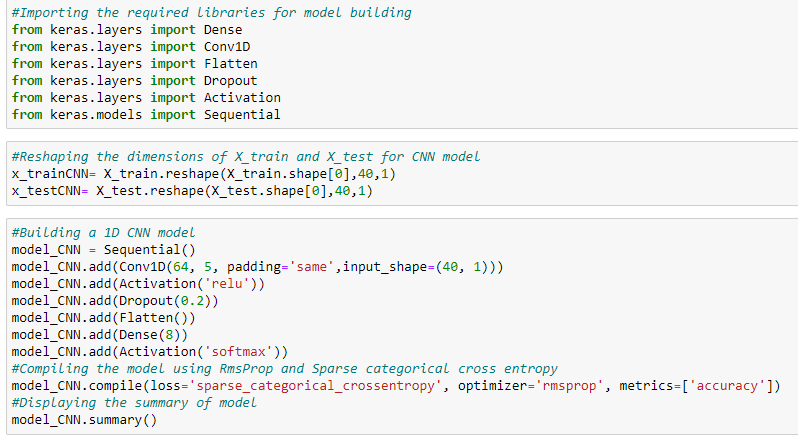
The figure below states the steps involved in feature extraction i.e. MFCC features and appending the labels.

****

**Fig. 14 Extracting features and MFCC coefficients**

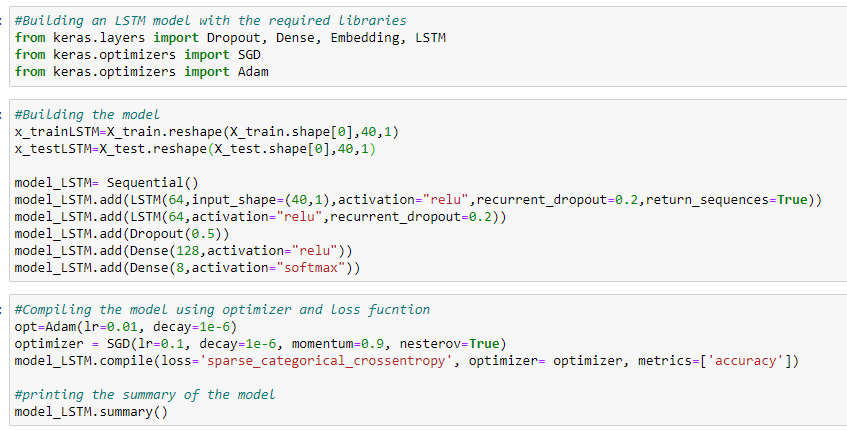
**4.5 Implementation**

In figure below CNN model is applied as proposed in our work.

****

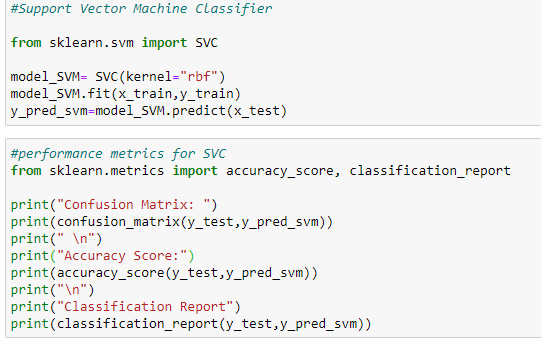
**Fig. 15 Building CNN model**

Further we build LSTM model which can be seen in next diagram to check if we get better performance but an issue of over fitting occurred.

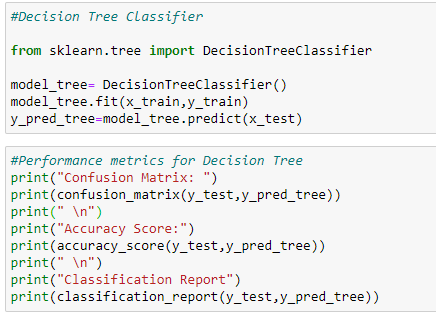
****

**Fig. 16 Building LSTM model**

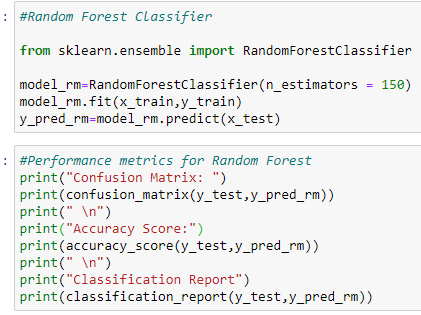
Further we applied SVM, decision tree classifier, random forest classifier and Adaboost ensemble technique in order to check if they can do well but it was observed that CNN outperforms all of them.

****

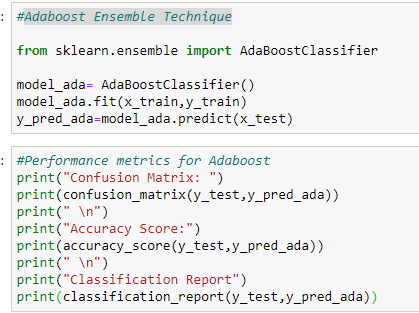
**Fig. 17 Building Support vector model**

****

**Fig. 18 Building decision tree classifier**

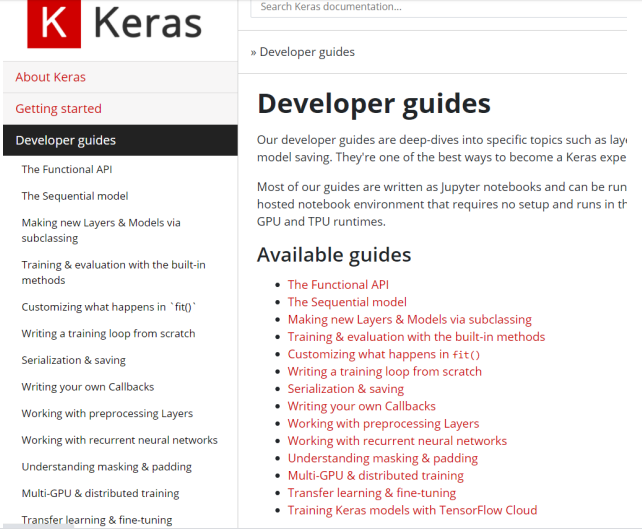
****

**Fig. 19 Building random forest classifier**

****

**Fig.20 Adaboost ensemble technique**

**4.6 Keras for deep learning**

****

**Fig. 21 Keras Documentation**

The Keras model is constructed by following the guidelines given in the Keras documentation Official Website. The documentation covers both the CNN, and LSTM model construction Templates along with hyperparameters and information optimizer.

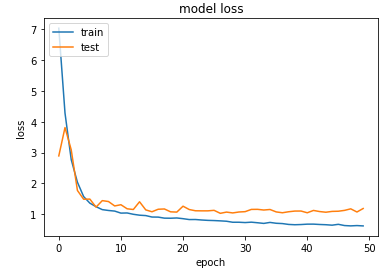
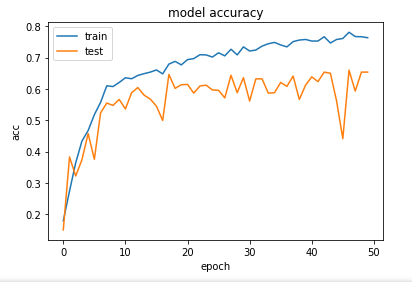
**CHAPTER 5**

**DATA FINDINGS AND ANALYSIS**

This section consists of all the results obtained after implementation of the SER model. Also an analysis is carried out further to compare the outcomes so that at last we can conclude with best suitable.

**5.1 Results of CNN model**

In this part outputs of CNN model obtained and evaluated. It is observed that the model accuracy of the CNN model is around 76% which is slightly higher in terms of model accuracy of other baseline models. SER's best-chosen hyperparameters are SGD for optimizing, categorical cross-entropy as a loss function, and early stopping using precision as a monitoring parameter. The figure below portrays the model’s accuracy on validation data and training data

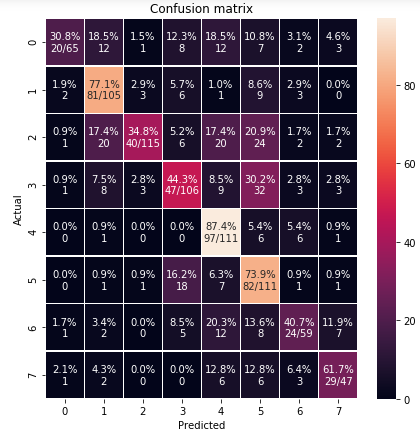
****

**Fig. 22 Graph plots of accuracy and loss function for train and test data. (Source: Created by Author)**

Also we obtained confusion matrix which can be observed below for our proposed CNN model. These figures consist of the confusion matrix for the predicted test values, accuracy graphs for the validation, training, and testing data and loss function graph for train and test data.

Also we obtained confusion matrix which can be observed below for our proposed CNN model. These figures consist of the confusion matrix for the predicted test values, accuracy graphs for the validation, training, and testing data and loss function graph for train and test data.

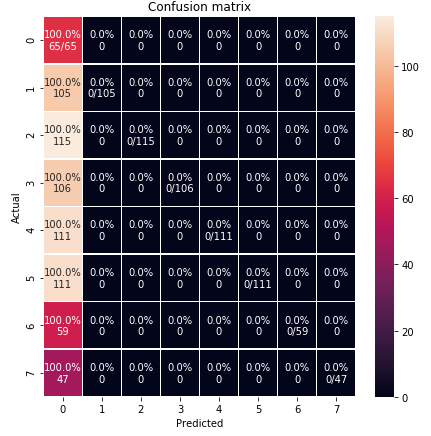
Also we obtained confusion matrix which can be observed below for our proposed CNN model. These figures consist of the confusion matrix for the predicted test values, accuracy graphs for the validation, training, and testing data and loss function graph for train and test data.

****

**Fig.23 Confusion Matrix for CNN model (Source: Created by Author)**

**5.2 Results of LSTM model**

In this research work, as two different neural networks were implemented in classifying the audio emotions, the CNN model outperformed the LSTM model with an accuracy of 76%. The best-chosen hyperparameters for audio analysis are RMSprop for optimizer, categorical cross-entropy as loss function, and early stopping using accuracy as a monitoring parameter. LSTM model was overfit, thus making CNN as best for evaluation for our research.



**Fig.24 Confusion Matrix for LSTM model (Source: Created by Author)**

Finally, the performance of our final LSTM model is evaluated using a confusion matrix. Further we also some machine learning models like SVM, Random forest classifier, Adaboost Ensemble Technique and Decision tree and evaluated them by obtaining confusion matrix for each of them. But finally it was observed that CNN outperforms all of them. .

In this section, the important findings are critically discussed. For this research work, the modal model performances of text and audio are been compared. The CNN model outperformed the other implemented models with an accuracy of 76%, whereas the LSTM model was overfit because it only predicted one emotion. Data preparation for audio is carried out by extracting MFCC features and appending the labels for them. The dataset contained files sorted according to actors and each actor has recorded his voice in terms of songs and speech. Each file has all the emotion files in it thus making it difficult to apply our model. So firstly I arranged data according to our eight proposed emotions. Eight folders are created with all of the emotion from 01 to 08 and then exported the files using OS library. Then further the obtained results were shocking .we performed CNN model first and then LSTM model so that a comparison can be made. Then CNN model received a benchmark accuracy of 76%. But the LSTM model seemed to be over fit since it was able to predict only one of the eight emotions. When the models were evaluated using confusion matrix it was detected that ‘Angry’ emotion is the mostly accurate predicted class. On the other hand ‘Neutral class gained less accuracy for prediction purpose. Thus considering the research question, the performed experiments conclude that Human emotions can be classified by using Convolution Neural Network with a precise accuracy.

In this section, the important findings are critically discussed. For this research work, the modal model performances of text and audio are been compared. The CNN model outperformed the other implemented models with an accuracy of 76%, whereas the LSTM model was overfit because it only predicted one emotion. Data preparation for audio is carried out by extracting MFCC features and appending the labels for them. The dataset contained files sorted according to actors and each actor has recorded his voice in terms of songs and speech. Each file has all the emotion files in it thus making it difficult to apply our model. So firstly I arranged data according to our eight proposed emotions. Eight folders are created with all of the emotion from 01 to 08 and then exported the files using OS library. Then further the obtained results were shocking .we performed CNN model first and then LSTM model so that a comparison can be made. Then CNN model received a benchmark accuracy of 76%. But the LSTM model seemed to be over fit since it was able to predict only one of the eight emotions. When the models were evaluated using confusion matrix it was detected that ‘Angry’ emotion is the mostly accurate predicted class. On the other hand ‘Neutral class gained less accuracy for prediction purpose. Thus considering the research question, the performed experiments conclude that Human emotions can be classified by using Convolution Neural Network with a precise accuracy.

**CHAPTER 6**

**DISCUSSION**

In this section, the important findings are critically discussed. For this research work, the modal model performances of text and audio are been compared. The CNN model outperformed the other implemented models with an accuracy of 76%, whereas the LSTM model was overfit because it only predicted one emotion. Data preparation for audio is carried out by extracting MFCC features and appending the labels for them. The dataset contained files sorted according to actors and each actor has recorded his voice in terms of songs and speech. Each file has all the emotion files in it thus making it difficult to apply our model. So firstly I arranged data according to our eight proposed emotions. Eight folders are created with all of the emotion from 01 to 08 and then exported the files using OS library. Then further the obtained results were shocking .we performed CNN model first and then LSTM model so that a comparison can be made. Then CNN model received a benchmark accuracy of 76%. But the LSTM model seemed to be over fit since it was able to predict only one of the eight emotions. When the models were evaluated using confusion matrix it was detected that ‘Angry’ emotion is the mostly accurate predicted class. On the other hand ‘Neutral class gained less accuracy for prediction purpose. Thus considering the research question, the performed experiments conclude that Human emotions can be classified by using Convolution Neural Network with a precise accuracy.

**CHAPTER 7**

**CONCLUSION AND FUTURE WORK**

After working on this research a strategy for their implementation is provided in this section. Here the operational side of our business is improved by making predictive analytics. The proposed SER system has surely increased the performance and reliability of the system. The experiments conducted thus conclude that human emotions can be classified with precise accuracy using Convolution Neural Network, considering the research question. The improvements gained over previous models will make the model more effective. Due to lack of resources and time we can carry out deployment in near future.

In this research work, two models are built namely CNN and LSTM for the crucial purpose of speech emotion recognition. Accuracy of recognition for both accent-based recognition and also the emotion-based recognition system for speech can be increased or improved by increasing the number of speakers. Further standardization of these samples of speech carried out. In future multiple modality can be used for recognizing the human emotions and accordingly a fusion methodology can be implemented. Thus, it becomes necessary to allow more researchers to take innovative efforts in this area.

# **References**

Aharon Satt, S. R. R. H., 2017. Efficient Emotion Recognition from Speech Using Deep Learning on. *INTERSPEECH 2017,* p. 5.

AlifBinAbdulQayyum, A. S., 2019. Convolutional Neural Network (CNN) Based speech emotion recognition. *2019IEEE International Conference onSignal Processing, Information, Communication & Systems(SPICSCON),* p. 4.

Anaconda Documentation, 2020. *Anaconda Documentation.* [Online]   
Available at: https://docs.anaconda.com/anaconda/install/windows/  
[Accessed 15 07 2020].

Bagus Tris Atmaja, M. A., 2019. Speech Emotion Recognition Based on Speech. *2019 IEEE International Conference on Signals and Systems (ICSigSys),* p. 5.

Caroline Etienne, G. F. ,. A. P., 2018. CNN+LSTM Architecture for Speech Emotion Recognition with Data.

Fatemeh Noroozi, T. S. ,. D. K., 2017. Vocal-based emotion recognition using random forests and decision tree. *Int J Speech Technol,* p. 9.

Haiqing Zheng, Y. Y., 2019. An Improved Speech Emotion Recognition. *2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS),* p. 5.

Jillur Rahman Saurav, S. A. S. K. M. S. R., 2018. Bangla Speech Recognition for Voice Search. *International Conference on Bangla Speech and Language Processing(ICBSLP), 21-22 September, 2018,* p. 4.

Jun Deng, F. E. B. S. F. B., 2017. Deep Neural Networks for Anger Detection from Real Life Speech Data. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW),* p. 6.

JUN DENG, S. F. Z. Z., 2017. Recognizing Emotions From Whispered Speech. *IEEE,* Volume VOLUME 5, 2017.

KERAS, n.d. *Developer Guides.* [Online]   
Available at: https://keras.io/guides/  
[Accessed 15 07 2020].

Kun-Yi Huang, C.-H. W. Q.-B. H. M.-H. S. a. Y.-H. C., 2019. SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORK CONSIDERING. *ICASSP 2019,* p. 5.

Md Amaan Haque, J. S. R. A. V., 2018. Evaluation of Modified Deep Neural Network. *2018 International Conference on Intelligent and Advanced System (ICIAS),* p. 5.

Michael Neumann, N. T. V., 2019. IMPROVING SPEECH EMOTION RECOGNITION WITH UNSUPERVISED. Volume ICASSP 2019.

Mohan Ghai, S. L. S. D. l. a. S. M., 2017. Emotion Recognition On Speech signals using machine learning. p. 6.

Novita Belinda Wunarso, Y. E. S., 2017. Towards Indonesian Speech-Emotion Automatic Recognition. *2017 4th International Conference on New Media Studies,* p. 4.

Promod Yenigalla, A. K. S. T. C. S. S. K. J. V., 2018. Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. *Interspeech 2018,* p. 5.

Richard Orjesek, R. J. M. C. M. K., 2018. DNN Based Music Emotion Recognition from Raw. p. 4.

Salsabil Besbes, Z. L., 2019. Stressed speech recognition using MMFCC. *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD'19),* p. 6.

Sandeep Kumar Pandey, H. S., 2018. Deep Learning Techniques for Speech Emotion. *IEEE,* p. 6.

Tian Kexin, H. Y. G. L., 2019. Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition system. *IEEE,* p. 5.

Yu Tian, J. T. X. J. H. T. a. Y. M., 2018. Accuracy on Children’s Speech Recognition under Noisy circumstances. *The 18th International Symposium on Communications and Information Technologies (ISCIT 2018),* p. 4.