International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 8, Issue 3, May - June 2019 ISSN 2278-6856

Speech Emotion Recognition Using Convolutional Neural Network

Dr. B.S. Daga[1], Glaston Dsouza[2], Aadesh Bassi[3], Lionel Lobo[4]

[1]Associate Professor, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Fr. Agnel Ashram, Bandstand, Bandra (West), Mumbai, India - 400050.

[2]Student, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Fr. Agnel Ashram, Bandstand, Bandra (West), Mumbai, India - 400050.

[3]Student, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Fr. Agnel Ashram, Bandstand, Bandra (West), Mumbai, India - 400050.

[4]Student, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Fr. Agnel Ashram, Bandstand, Bandra (West), Mumbai, India - 400050.

Abstract: ***In today's world, Human machine interaction are used nowadays in many applications. Speech is one of the medium of interaction. Detection of emotion from speech is a main challenge. Communicating with emotions is more affective compared otherwise, since they can be expressed and identified in a better way through facial expressions, speech, gestures etc. If machines understand the emotional content they will start behaving in a friendlier way.***

Recognition of emotion is always a difficult problem, particularly if the recognition of emotion is done by using speech signal. Significant research has been done on emotion recognition using speech signal. The primary challenges are choosing the emotion recognition corpora i.e. database, identification of different features related to speech and selecting an appropriate choice of classification model.

Speech Emotion problem is categorized as:

1) Feature extraction from speech: For this MFCC is used. We use 13 MFCC with 13 velocity and 13 acceleration component as features.

2) Feature classification: The features from MFCC are passed to CNN layer where the classification of above features is done.

3) Emotion detection: Depending on the output from CNN the corresponding emotion is detected i.e. happy, sad, angry, calm, fear.

Keywords: CNN-Convolution neural network, MFCC-Mel Frequency Cepstral Coefficient, SAVEE, RAVDESS.

1. Introduction

Nowadays, Human machine interaction is widely used in many applications[6]. Speech is one of the mediums of interaction. The main challenge in human machine interaction is detection of emotion from speech. Emotion can be recognized from different biological signals. When two persons interact with each other they can easily recognize the underlying emotion in the speech, spoken by the other person. The main objective is to mimic the human perception mechanisms. Emotion plays an important role in decision making[9]. An efficient emotion recognition system can be useful in the field of medical science robotics engineering[7], call centre application[9] etc.

Human can easily recognize emotion of speaker. Human first analyses different characteristics of particular speech and then using previous experience or observation he recognizes the emotion of the speaker. There is a need to build a human like system that can detect emotions effectively and efficiently. In this field several systems are proposed for recognizing emotional state of human from speakers voice or speech signal. Some universal emotions include anger, sadness, neutral, happiness, fearful etc.

For the last two decades several intelligent systems are proposed by researchers[1][2][3]. The various techniques used are Segmentation selection[1],Gaussian Model[3][8],Adaboost Algorithm[14],LSTM[16] was also used, the most used algorithm was recurrent neural network[4][10] and then the use of CNN[5][12][17][18] came into picture.There were many other researches as well which had very less accuracy such as random forest[15] and many others [2][7]. Only pitch and MFCC features are used for recognition of emotion.

2. Architecture and working

2.1 Proposed System Architecture

After studying the previous works[5][17][18] in the field of Speech Emotion Recognition and studying the different components involved in Speech Emotion Recognition, we have broken down the entire process of Speech Emotion Recognition into five parts: Speech signal, Preprocessing, Feature Extraction using MFCC, Feature Classification using CNN and Emotion Recognition[12].

The input is in the form of speech signal with .wav extension.Before using MFCC we make some Pre-processing on the data set. All the speech files are with .wav extension; first we compute amplitude values of each file with a sample rate of 44,100 samples per second.
Feature Extraction is the process of identifying various features. MFCC are used for feature extraction. Feature Classification is done using CNN Model[12]. The MFCC are loaded to the CNN Model where classification is done on 5 emotions for Male and Female.
The Output of the CNN Model is the emotion generated on the basis of the input data.

The flow diagram is given below:

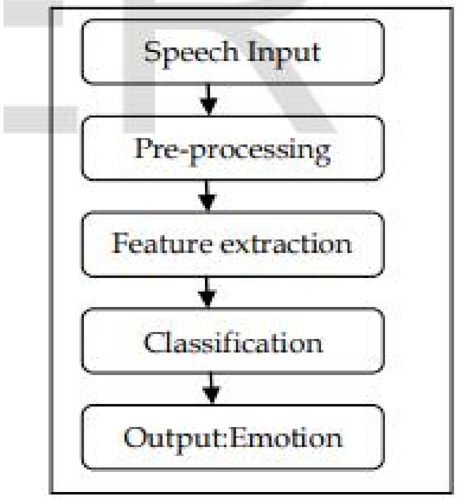


Figure 1 Flow Diagram of System

The dataset used here are SAVEE[6] and RAVDESS[6] having dataset of 1920 audio signals out of which 1200 are used for testing and training. The common emotions which are taken into consideration are happy, sad, angry, fearful and calm. The emotions are further divided into male and female for increasing the accuracy.

The below diagram shows the proposed system architecture which shows the various section such as preprocessing, feature extraction, classification, training, testing and emotion as the output which is male emotion or female emotion as speech.

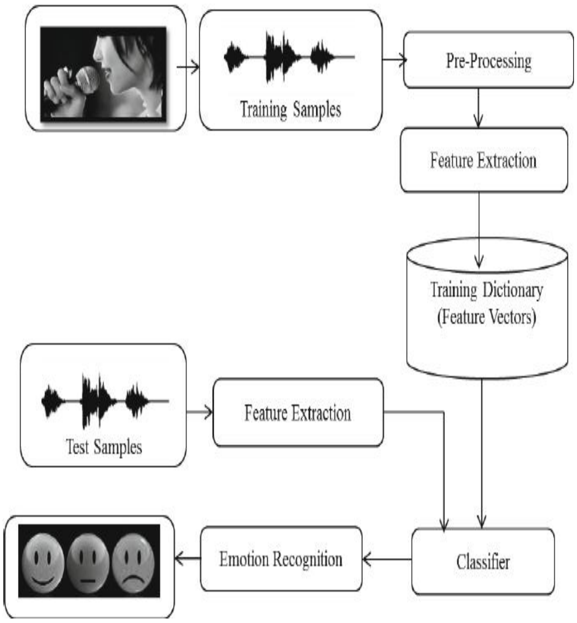


Figure 2 Proposed System Architecture

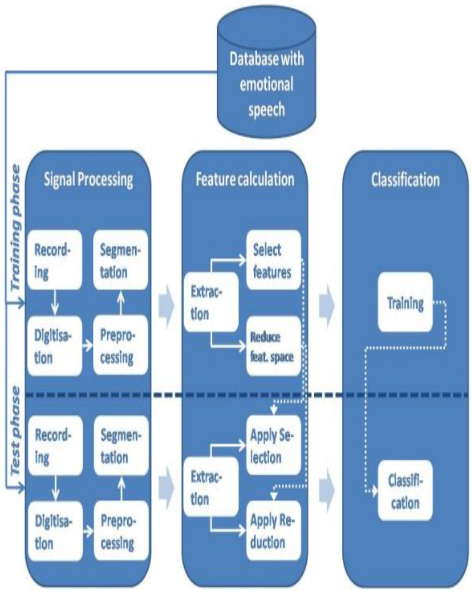


Figure 3 Overview of the System

2.2 Proposed Algorithm for Feature Extraction

The point to understand about speech is that the shape of vocal tract including tongue, teeth etc filter the sounds generated by a human. This shape determines what sound comes out. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope.

MFCC are used for audio signals and speech for feature extraction. In MFCC 13 features are taken into consideration for every frames in a given audio signal.
The block diagram shows different steps used to generate MFCC:

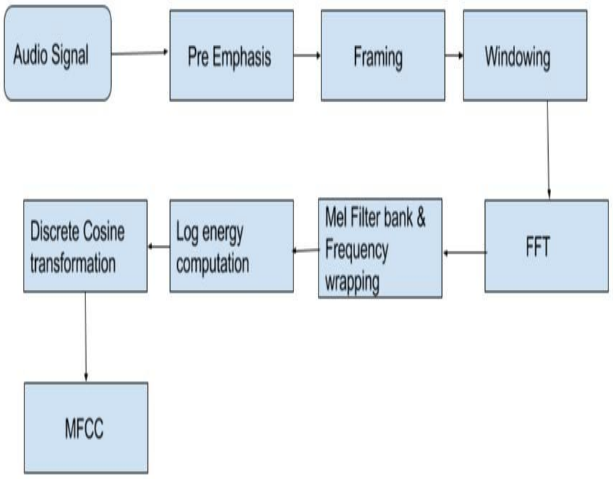


Figure 4 Block Diagram for MFCC

2.3 Proposed Algorithm for Classification

After studying the different approaches for Classifier Construction like Support Vector Machines (SVM), Random Forest in previous researches [3] [7][9] and comparing the results of previously built systems [4][8], we selected Convolution Neural Network for classification.

Convolution neural network or CNN consist of several layer of convolution [12]. In CNN, nonlinear activation function for example rectified linear unit (ReLU) or Sigmoid function are used to the result. In neural networks nodes of input layer are connected with the nodes of hidden layer and those hidden layer nodes are fully connected with nodes of output layer. In CNN, convolution are applied on input layer to generate the output. Each part of input are convoluted by different filters and combining them we get the final output

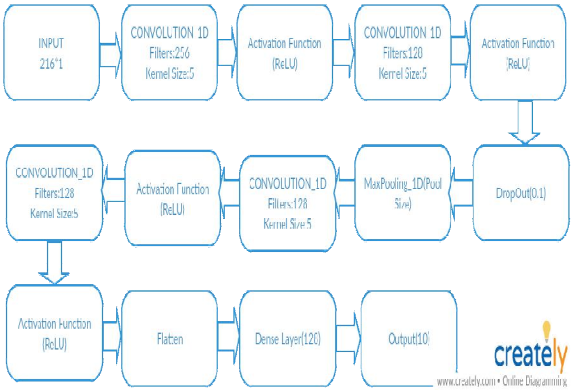

INPUT 216*1
CONVOLUTION 1D Filters:256 Kernel Size:5
Activation Function (ReLU)
CONVOLUTION 1D Filters:128 Kernel Size:5
Activation Function (ReLU)
DropOut(0.1)
MaxPooling_1D(Pool Size)
CONVOLUTION_1D Filters:128 Kernel Size:5
Activation Function (ReLU)
CONVOLUTION_1D Filters:128 Kernel Size:5
Activation Function (ReLU)
Flaten
Dense Layer(128)
Output(10)
creately
www.creately.com • Online Diagramming

Figure 5 CNN Model

2.4 Working

The libraries used are librosa used for generating MFCCs and pandas are used for calculations. The data sets used are SAVEE and RAVDESS which has 1200 audio files

The first step involves pre-processing and then the audio files is framed and each frames have 13 MFCCs and the 4 sec audio file is divided into frames having size 30ms .The MFCCs are then fed to the CNN model which is then classified based on 5 different emotions and differentiate as male and female for more accuracy. Keras and tenserflow is used for modelling and different libraries such as scipy and numpy is used for matrix calculations.

The entire setup has been on the Anaconda IDE used for Python programming, along with Jupyter Notebook which allows programmers to execute code step by step in order to debug with greater ease.

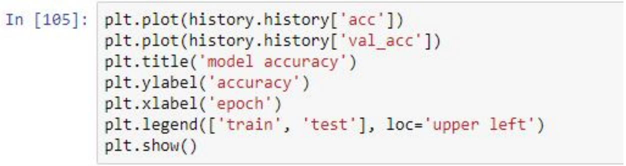
3. TESTING AND RESULTS

3.1 Testing various Sub Approaches

MLP Model (multilayer perceptron):

A multilayer perceptron (MLP)[3][14][5] is a class of feedforward artificial neural network. MLP at least 3 Layers of nodes: An Input, hidden and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes supervised learning techniques such as backpropagation for training . non- linear activation function and multiple layers separate it from linear perceptron.

The MLP model created had a low validation accuracy of around 25% with 8 layers, batch size of 32, softmax function at the output and 550 epochs.



```
In [105]: plt.plot(history.history['acc'])
          plt.plot(history.history['val_acc'])
          plt.title('model accuracy')
          plt.ylabel('accuracy')
          plt.xlabel('epoch')
          plt.legend(['train', 'test'], loc='upper left')
          plt.show()
```

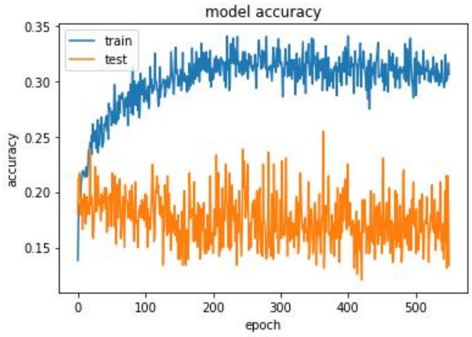
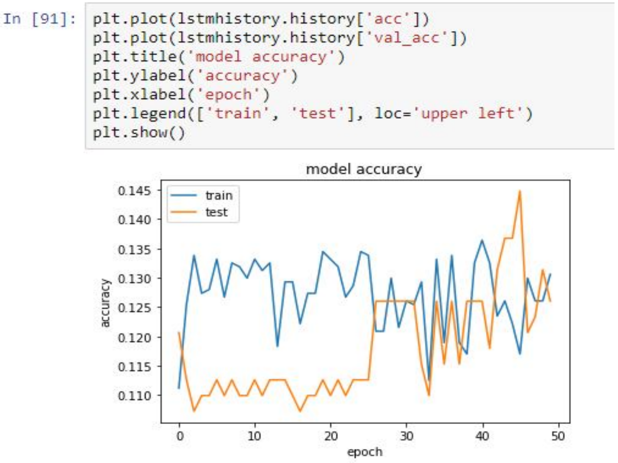


Figure 6 MLP Model Accuracy

LSTM(long short term memory):

Long short-term memory (LSTM)[16] is artificial recurrent neural network (RNN) architecture. LSTM has feedback connection that makes it “general purpose computer”, unlike normal feed forward network neural networks. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. It has three gates that regulate the flow of information in and out of the cell.

The LSTM model had the lowest training accuracy of around 15% with 5 layers, batch size of 32; tan h activation function and 50 epoch

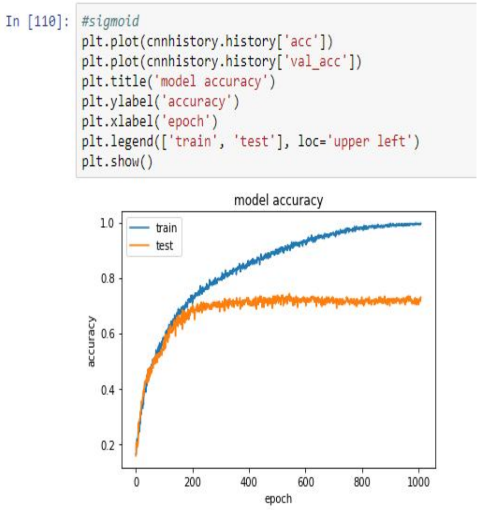


```
In [91]: plt.plot(lstmhistory.history['acc'])
         plt.plot(lstmhistory.history['val_acc'])
         plt.title('model accuracy')
         plt.ylabel('accuracy')
         plt.xlabel('epoch')
         plt.legend(['train', 'test'], loc='upper left')
         plt.show()
```

Figure 7 LSTM Model Accuracy

CNN (Convolutional neural network)

In neural networks, one of the main categories to do images recognition, images classifications is Convolutional neural network (ConvNets or CNNs)[4][17]. Objects detections, recognition face etc., are the areas where CNNs has been used widely. Technically, deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC) and apply Softmax function to classify an object with probabilistic values between 0 and 1.



```
In [110]: #sigmoid
          plt.plot(cnnhistory.history['acc'])
          plt.plot(cnnhistory.history['val_acc'])
          plt.title('model accuracy')
          plt.ylabel('accuracy')
          plt.xlabel('epoch')
          plt.legend(['train', 'test'], loc='upper left')
          plt.show()
```

Figure 8 CNN Model Accuracy

CNN model was the best for our classification problem. After training numerous models we got the best validation accuracy of 92.22% with 18 layers, softmax activation function, rmsprop activation function, batch size of 32 and 1000 epochs.

3.2 Training and Testing

We merged two datasets namely The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[6]

And Surrey Audio-Visual Expressed Emotion (SAVEE) Database[6].

We got audio datasets with around 2000 audio files which were in the wav format from the following websites:

http://neuron.arts.ryerson.ca/ravdess/?f=3,
http://kahlan.eps.surrey.ac.uk/savee/Download.html

The first website contains speech data which is available in three different format.

1. Audio Visual – Video with speech
2. Speech – Audio only
3. Visual – Video only

As we are dealing with finding emotions from speech, we went with the Audio only zip file because. The zip file consisted of 1500 audio files all in wav format.

The second website consisted of 500 audio speeches from four different actors with different emotions.

3.3 Results

1. Audio Recorder output

Audio recorder code is used to output a (.wav) file with a length of 4 seconds with 2 channels and a sampling rate of 44100 Hz.

The file is saved as output10.wav in the current working directory.

2. Separating the male and female voices

We used Librosa library in Python to process and extract features from the audio files. Librosa is a python package used for analysis of music and audio.

It provides all necessary building blocks to create music information retrieval systems. We were also able to extract features i.e. MFCC (Mel Frequency Cepstral Coefficient) using the librosa library. We also separated out the male and females voice by using the identifiers provided on the website It could be because the pitch of the voice affects the results.

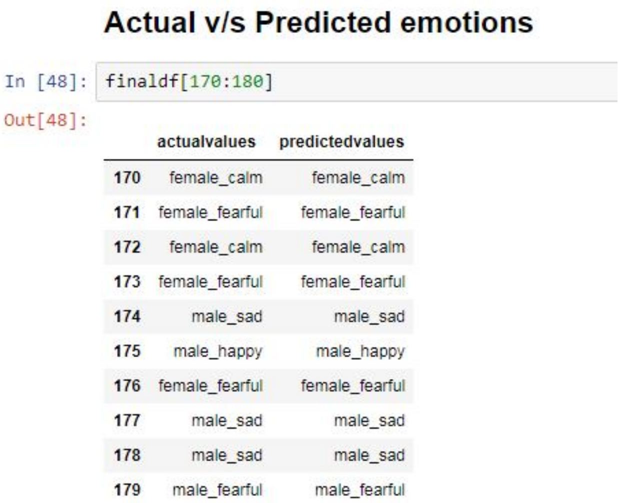
3. Predicting the emotion

We successfully predicted the following 5 emotions for male and female.

1. Calm
2. Happy
3. Sad
4. Angry
5. Fearful

3.4 Final Output

After training the model we had to predict the emotions on our test data. The following picture shows our prediction with the actual values.



Actual v/s Predicted emotions

```
In [48]: finaldf[170:180]
Out[48]:
```

	actualvalues	predictedvalues
170	female_calm	female_calm
171	female_fearful	female_fearful
172	female_calm	female_calm
173	female_fearful	female_fearful
174	male_sad	male_sad
175	male_happy	male_happy
176	female_fearful	female_fearful
177	male_sad	male_sad
178	male_sad	male_sad
179	male_fearful	male_fearful

Figure 9 Final Output

4 Conclusion and future scope

After building numerous different models, we have found our best CNN model for our emotion classification problem[17][5]. We achieved a validation accuracy of 90% with our existing model. Our model could perform better if we have more data to work on. What's more surprised is that the model performed excellent when distinguishing between a males and females voice. We can also see above how the model predicted against the actual values.

In the future we could build a model with CNN and LSTM[16][5] together to increase the input length anything more than 4 seconds.

REFERENCES

[1.] Using Voiced Segment Selection Algorithm[J]. 2016.

[2.] Deng J, Xu X, Zhang Z, et al. Fisher kernels on phase-based features for speech emotion recognition[M]//Dialogues with Social Robots. Springer Singapore, 2017: 195-203.

[3.] Patel P, Chaudhari A, Kale R, et al. EMOTION RECOGNITION FROM SPEECH WITH GAUSSIAN MIXTURE MODELS & VIA BOOSTED GMM[J]. International Journal of Research In Science & Engineering, 2017, 3.

[4.] Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and Recurrent Neural Networks[C]//Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific. IEEE, 2016: 1-4.

[5.] Trigeorgis G, Ringeval F, Brueckner R, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network[C]//Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016: 5200-5204.

[6.] Institute of Automation Chinese Academy of Sciences. The selected Speech Emotion Database of Institute of Automation Chinese Academy of Sciences (CASIA) [DB/OL].2012/5/17.

[7.] Cheng Ming, An approach of speech interaction and software design for humanoid robots, thesis, Changsha, Hunan University, 2016.

[8.] Patel P, Chaudhari A, Kale R, et al. EMOTION RECOGNITION FROM SPEECH WITH GAUSSIAN MIXTURE MODELS & VIA BOOSTED GMM[J]. International Journal of Research In Science & Engineering, 2017, 3.

[9.] F.Noroozi, N.Akrami, and G. Anbarjafari, "Speech-based emotion recognition and next reaction prediction," 2017 25th Signal Process. Commun. Appl. Conf. SIU 2017, no. 1, 2017.

[10.]C.-W. Huang and S. S. Narayanan, "Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition," pp. 1–19, 2017.

[11.]H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," Neural Networks, vol. 92, pp. 60–68, 2017.

[12.] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," 2017 Int. Conf. Platf. Technol. Serv., pp. 1–5, 2017.

[13.]K. Han, D. Yu, I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine", Proceedings of INTERSPEECH ISCA Singapore, pp. 223-227,2014.

[14.]Using Adaboost Algorithm Along with Artificial Neural Networks for Efficient Human Emotion Recognition From Speech Efstathios Stamatatos, "A Survey of Modern Authorship Attribution Methods", Department of Information and Communication Systems Engineering, University of the Aegean.

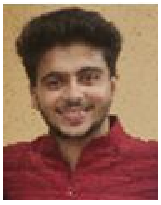
[15.]Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest Li Zheng1, Qiao Li2, Hua Ban1 , Shuhua Liu1*

[16.]SPEECH EMOTION RECOGNITION USING AUTOENCODER BOTTLENECK FEATURES AND LSTM Kun-Yi Huang, Chung-Hsien Wu, Tsung-Hsien Yang, Ming-Hsiang Su, and Jia-Hui Chou

[17.]Z. Huang, M. Dong, Q. Mao, Y. Zhan, "Speech emotion recognition using CNN", Proc. 22nd ACM Int. Conf. Multimedia, 2014.

AUTHORS

Dr. B. S. Daga has received the Bachelor in Computer Engineering from Government Engineering College, Amravati in 1990 and Master in Computer Science & Engineering from National Institute of Technology, Allahabad. He has completed the Ph.D. degree with Department of Computer Engineering, SGB, Amravati University. He has also worked as Coordinator in Entrepreneurship Development Programs with Department of Science & Technology (Government of India). His research interest includes Multimedia Systems, Data Mining, Artificial Intelligence and Machine Learning.



Glaston Dsouza has completed his B.E. in Computer Engineering from Fr, Conceicao Rodrigues College of Engineering, Mumbai. He was also the member of Project Cell of Fr. CRCE, a student body that does many project and go to various competiton in many colleges. He has strong communication skills and he is technically sound. With respect to academics, he is fluent in various areas in the field of Computer Science. His areas of interest include Machine Learning, Artificial Intelligence and Databases.

Lionel Lobo acquired his degree in engineering from Fr. Conceicao Rodrigues College of Engineering in the field of Computers. He is fluent in various programming languages and has worked especially with designing websites and databases. He has exceptional speaking skills

Aadesh Bassi has completed his Bachelor’s degree in Computer Engineering from Fr. Conceicao Rodrigues College of Engineering, which is one of the best institutes in Mumbai, India. He is fluent with many programming language.He has a hobby of teaching .He has interest in Machine Learning and Artificial Intelligence