International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Gender based Real Time Vocal Emotion Detection

Anusha Anchan*[1], Manasa G. R.[2], Joylin Priya Pinto[3]

Submitted: 22/08/2023 **Revised**: 09/10/2023 **Accepted**: 22/10/2023

Abstract*:* Emotion Detection (ED) is recognizing the emotional facet of speech regardless of its semantics. Although human beings are able to perform this task efficaciously, automatically conducting it, utilizing programming devices or techniques is still a subject of research. This work proposes, several classifier algorithms have been implemented and compared based on the accuracy and emotion category. MLP, SVM, CNN and DNN with LSTM layer have been trained. At first the classifiers are training with initial datasets. The datasets that are being used are CREMA-D, TESS, SAVEE and RAVDESS. Subsequently, with pre-processing of data and data augmentations using four parameters namely Noise, Stretch, Speed and Pitch. Zero Crossed Rate, Energy – Root Mean Square and MFCC are extracted from the data. Models have witnessed improvement in accuracy under the post processing. Models thrived and succeeded in achieving accuracy of 66.61% (MLP), 53% (SVM), 88.35% (CNN) and 91.30% (DNN). This work also proposes a Real time Speech Detection System acts as artificial intelligence which classifies emotions by analyzing the audio files via the trained model based on gender.

Keywords: *Face Recognition, Computer Vision, Face Detection, Uncontrolled Environment*.

1. Introduction

Communication is one the effectual ways to transfer information. The current developments in technology clearly identifies two way communication between humans and machines, in which Emotion Detection (ED) plays a very important part, since it ensures seamless interactivity between a human being and machine. ED is a way of identifying the emotions of a speaker from voice signals. Practical applications of ED can be found in applications used in the field of education, entertainment, audio-video surveillance, text to speech etc.

The major inspiration for this work is underlying study carried out. Work proposed by Manju D. Pawar et al. in [2] uses CNN with MFCC. In [7] Aparna Kannan et al., Panuwit Nantsri in [13] have utilized MFCC to obtain highest accuracy model. Works carried out in [9], [15] and [16] shows positive effect of that data augmentation. The previous works by various scholars have been concentrating on emotion detection with the help of numerous machine learning and non machine learning techniques as well. But consideration of gender in emotion detection is one of the areas in this field which requires further contribution. Hence presented work aims at implementing various machine learning & deep learning algorithms to detect the emotions of speaker from an audio files considering gender into consideration. Then these models are compared based on accuracy score to deploy Real Time Emotion Detection System.

Objectives of the proposed work are:

- Build several different classifier models to detect the emotions based on gender and compare them.
- Explore on feature extraction techniques and data augmentation.
- Deploy Real Time Emotion Detection System where it predicts the probability of the different emotions in the given input audio file.

Current work does evaluation of ED, based on features extracted using Mel Frequency Cepstral Coefficients (MFCC), Mel Spectrogram, Chromagram, Energy - Root Mean Square (RMS), Zero Cross Rating (ZRC) and other parameters or combination of parameters to study the emotions. This work covers utilization of algorithms like Support Vector Machine (SVM), Multi Layer Perceptron (MLP), Convolutional Neural Network (CNN) and Deep Neural Network (DNN) evaluating these trained models to obtain a better performing classifier. The chosen classifier is then deployed for real time recognition system.

2. Related Work

The work carried out by Anusha Koduru et al. in [1] revolves around preprocessing of audio samples. Noise is removed using filter, then in later Mel Frequency Cepstral Coefficients (MFCC), Discrete Wavelet Transform (DWT), pitch, energy and Zero crossing rate (ZCR) algorithms were used. Five models were used which are based on Convolution Neural Network (CNN) by Manju D. Pawar et.al [2] to recognize emotion through signals of speech. Considering seven emotions, Feature extraction was carried out by Mel Energy Spectrum Dynamic Coefficients (MEDC), Pitch and Energy and Mel Frequency Cepstral Coefficients (MFCC). By Aparna Kannan et.al in [7] MLP classifier is used in classifying emotions. MFCC has the

[1,2,3]NITTE (Deemed to be University), Dept. of Computer Science and Engineering, NMAM Institute of Technology, Nitte - 574110, Karnataka, India
** Corresponding Author Email: anusha@nitte.edu.in*

highest accuracy using the logistic function. Enkhtogtokh Togootogtokh et.al [8] proposed speech recognition framework, called DeepEMO having two parts, a) pre-processing inorder to extract efficient features. b) Deep transfer learning model inorder to train then recognize and with 42 epochs, better accuracy was achieved. Shilandari et. Al [9] focuses on netwrok called cycle generative adversarial for data augmentation. This network was built for every five emotions which was of interest. Augmenting artificial data here would help SVM to recognize metadata better and classify metada with better performance. U. Kumaran et. al [11] in their work use, Deep Convolutional-Recurrent Neural Network (Deep C-RNN) to classify the effectiveness of learning emotion variations. It used fusion of Mel-Gammatone filter in convolutional layers which t extract high-level spectral features then recurrent layers is acquired to learn the long-term temporal context from high-level features. The obtained accuracy is morehan 80% and have less loss. The proposed fusion method by Zengwei Yao et.al [12] would integrate three sub classifiers's power. With this WA of 57.1% was achieved. UA of 58.3% was achieved. The obtained results were higher compared to individual classifier. Panuwit Nantasri et.al in [13] used Average values of MFCC's concatenated with delta and delta-delta coefficients for an artificial neural network model (ANN) in classification. The proposed model by Soonil Kwon et al. in [14] is real time SER model with basis of one-dimensional dilated convolutional neural network (DCNN). This model uses a multi-learning strategy in order to parallely extract spatial salient emotional features and learn long term contextual dependencies from the speech signals.In the work proposed by Nikolaos Vryzas et.al in [15] Recognition is performed on successive time frames of continuous speech, while data augmentation techniques are applied as well. The proposed model outperforms SVM with the margin of 8.4% in accuracy. Study carried by Arash Study by Raghavendra Pappagari et.al in [16] proposed a novel augmentation procedure for Speech emotion recognition (SER). It was assumed that performance of SER could be improved utilizing concatenated utterances in model training. Observation indicates that three Copy-Paste schemes help improve SER performance with dataset used MSP-Podcast, Crema-D, and IEMOCAP. Additionally, Copy-Paste performs better compared to noise augmentation.

3. System Design

The proposed model follows the given design in Fig 3.1. The datasets that are being used are CREMA-D, TESS, SAVEE and RAVDESS. These datasets are then pre-processed and cleaned. Appropriate features are selected and extracted to train the model. Four models are trained and then evaluated with test data for emotions prediction.

4. System Implemenation

4.1 Dataset

The dataset used in this project are as follows CREMA-D, TESS, SAVEE, RAVDESS and Common Voice. CREMA-D data set consists of 7,442 original clips having voices of 91 actors belonging to variety of races and ethnicity with 6 various emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad), 4 emotion levels (Low, Medium, High, and Unspecified) of both male and female. In Toronto emotional speech set (TESS), there are a set of 200 target words with seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). It contains 2800 data points (audio files). Surrey Audio-Visual Expressed Emotion (SAVEE), this database comprises of 480 British English utterances. These are voices of 4 male actors exhibiting 7 emotions. Another dataset RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) have 1440 files. The following website (http://voice.mozilla.org/) has data Common Voice is a corpus.

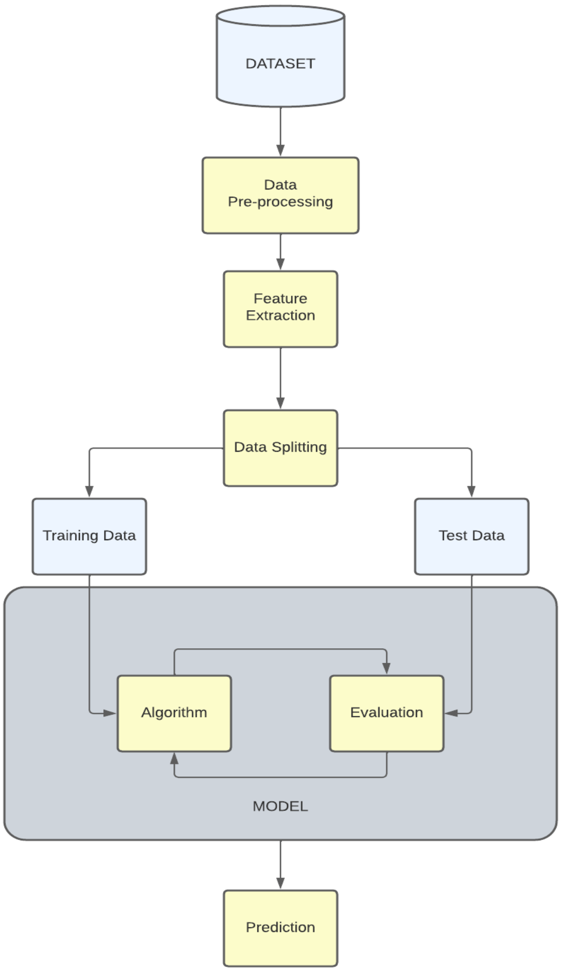


Fig. 3.1 System Design

4.2 Methodology

There were four algorithms to be developed namely SVM, MLP, CNN and DNN. These models followed the general proposed system design with some variations to acquire better accuracy. The development of project can be divided into following stages:

1. Load the dataset
2. Data visualisation and pre-processing
3. Feature extraction and selection
4. Data splitting
5. Build the model
6. Evaluate the model
7. Real time implementation

4.3 Multi Layer Perceptron

MLP Classifier was trained with the original datasets and augmented datasets namely RAVDESS, CREMA-D and SAVEE. The dataset has already been pre-processed and feature extraction of Mel spectrogram has been taken place. The dataset is split into train and test with test size=20% of the dataset. MLP Classifier is built with parameters alpha=0.839903176695813, batch size=150, hidden layer sizes=100, learning rate=adaptive, maximum iterations=100000, solver= Stochastic Gradient Descent. MLP Classifier is trained for both original and augmented datasets. It is also used to detect the gender of the speaker. MLP Classifier with original dataset gave 42% accuracy for emotion detection and 62.41% accuracy for gender detection. Whereas for augmented dataset, 46.33% for emotion detection and 67% for gender detection. Figure 4.1 and 4.2 shows the confusion matrix for gender and emotions respectively.

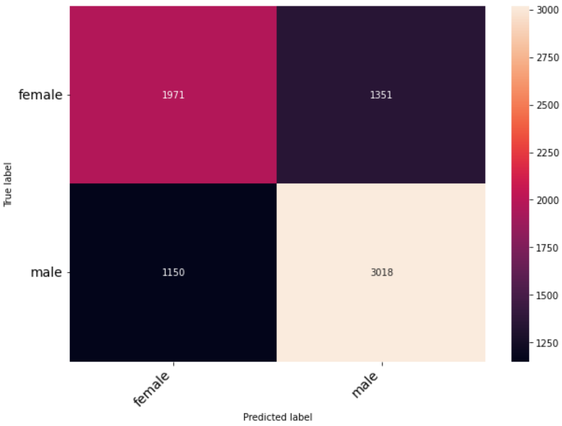


Fig. 4.1 MLP confusion matrix for gender

4.4 Support Vector Machine

The dataset that is used here is RAVDESS dataset. Here, Mel spectrogram, Chromagram and MFCC are extracted. The dataset is split into train and test with test size = 25% of the dataset. Training and Evaluation of SVM model gave an accuracy of 53% for emotion detection. It showed highest accuracy of 69% for 'calm' emotion. Figure 4.3 shows the precision, recall, f1-score and support for various emotions.

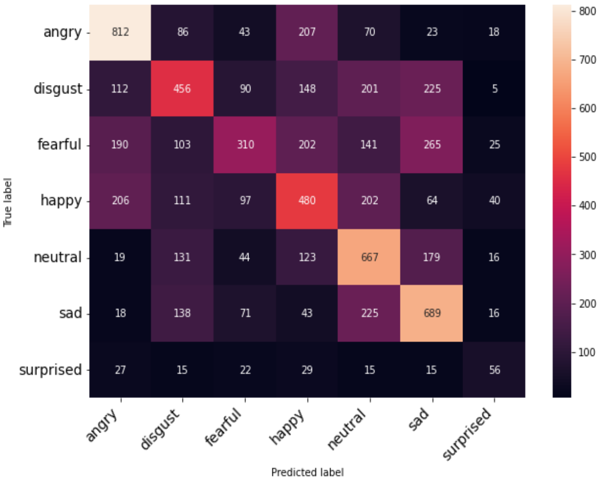
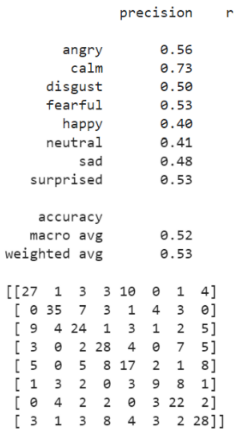


Fig. 4.2 MLP confusion matrix for emotion



```
              precision    recall  f1-score   support

       angry       0.56      0.55      0.56        49
        calm       0.73      0.66      0.69        53
     disgust       0.50      0.49      0.49        49
     fearful       0.53      0.57      0.55        49
       happy       0.40      0.37      0.39        46
     neutral       0.41      0.33      0.37        27
         sad       0.48      0.63      0.54        35
   surprised       0.53      0.54      0.53        52

    accuracy                           0.53       360
   macro avg       0.52      0.52      0.52       360
weighted avg       0.53      0.53      0.53       360

[[27  1  3  3 10  0  1  4]
 [ 0 35  7  3  1  4  3  0]
 [ 9  4 24  1  3  1  2  5]
 [ 3  0  2 28  4  0  7  5]
 [ 5  0  5  8 17  2  1  8]
 [ 1  3  2  0  3  9  8  1]
 [ 0  4  2  2  0  3 22  2]
 [ 3  1  3  8  4  3  2 28]]
```

Fig. 4.3 SVM Classification Report

4.5 Convolutional Neural Network

The datasets used are CREMA-D, RAVDESS, TESS and SAVEE. Data augmentation is done using four parameters namely Noise, Stretch, Speed and Pitch. Mel spectrogram is the extracted feature. This feature is also extracted for each parameter of the augmented datasets. Test dataset size is 25% where train is 75%. This model was built with 9 Layers consisting of 8 Conv1D layer and 1 Dense layer, with following parameters Batch size of 16, 100 Epochs, Learning Rate = 0.00001, Decay = 1 e^{-6}, Loss function is categorical cross entropy and the evaluation metric is accuracy. Figure 4.4 gives the model loss. This model was trained with the original dataset which gave an accuracy of 30% for gender-emotion detection, 65% accuracy for gender detection and 45% accuracy for emotion detection. Whereas for augmented dataset, resulted in accuracy of 76% for gender-emotion detection, 88.35% accuracy for gender detection and 80.3% accuracy for emotion detection. Figures 4.5, 4.6 gives us confusion matrix showing classification of emotions into different categories with gender into consideration and Confusion matrix for only emotions respectively.

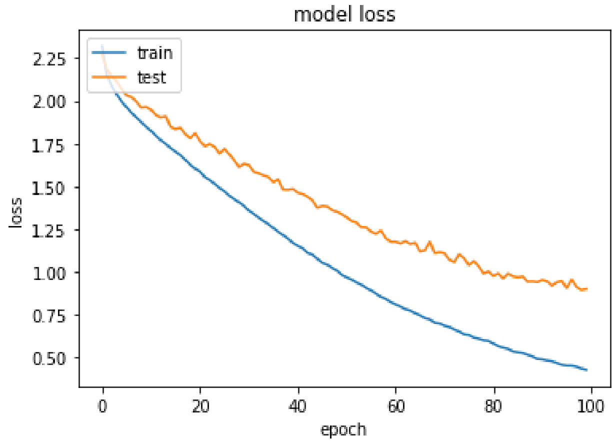


Fig.4.4 Model loss

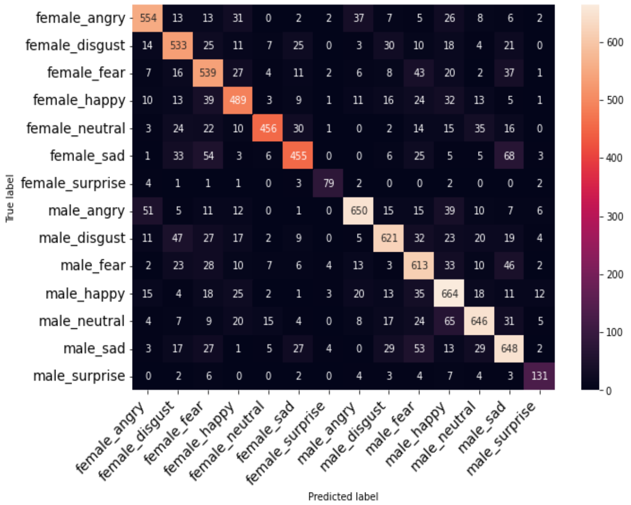


Fig. 4.5 CNN – confusion matrix for gender-emotion

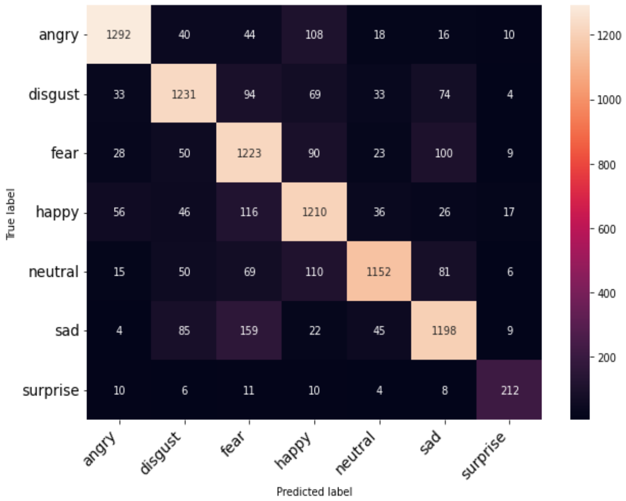


Fig. 4.6 CNN-confusion matrix for emotion

4.6 Deep Neural Network with LSTM Layers

The datasets used are RAVDESS and TESS. Once the dataset is loaded, it is normalized to +5.0 dBFS, transformed to arrays and trimmed. Padding is done to have equal audio length, followed by noise reduction. Zero Crossed Rate, Energy – Root Mean Square and MFCC are extracted from the data. The dataset is divided into train, test and validation category, with train size=87.5%. The remaining 12.5% is split as validation and test data with test size of 30.4% of split data i.e., 3.8% of the dataset. Fig. 4.7 and 4.8 shows the confusion matrix for validation and test data.

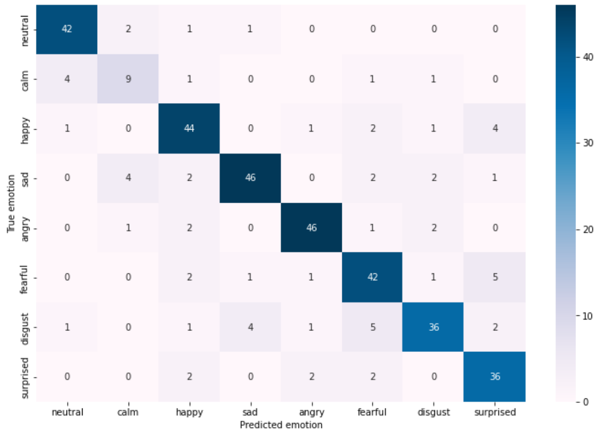


Fig. 4.7 DNN confusion matrix for validation dataset

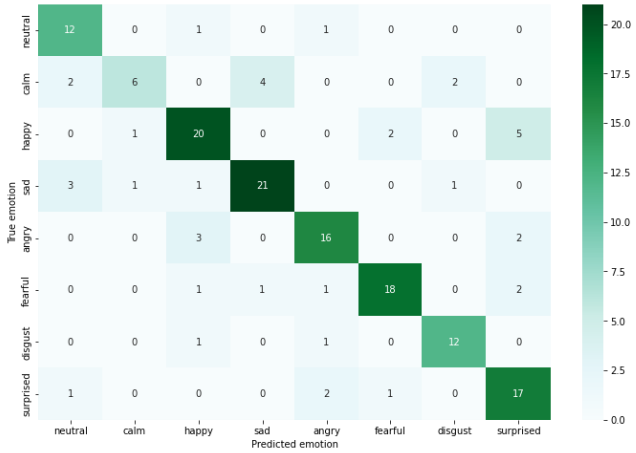


Fig. 4.8 DNN confusion matrix for test dataset

The model is executed with keras library, using 2 hidden LSTM layers with 64 nodes, and an output (dense) layer with 8 nodes, each for one emotion using the 'softmax' activation. The optimizer that led to the best results was 'RMSProp' with default parameters and the batch size chosen is 23 and 350 epochs. Figure 4.9 shows the DNN Model architecture.The figures 4.10 and 4.11 below shows the result of DNN in terms of Loss and Accuracy.

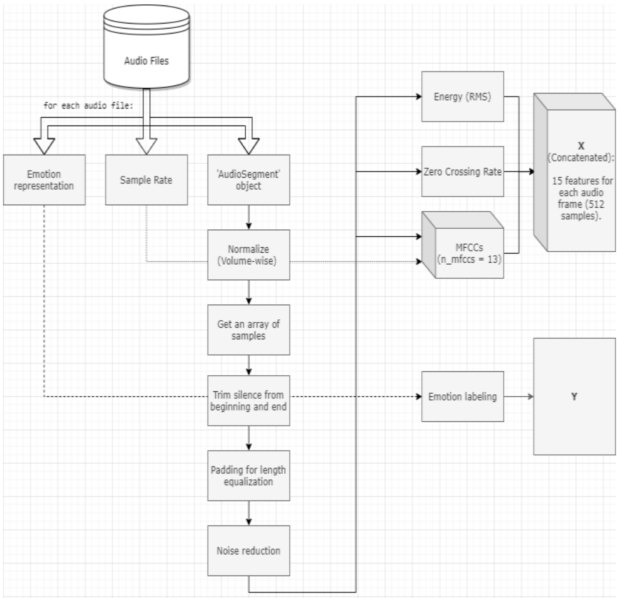


Fig. 4.9. DNN model

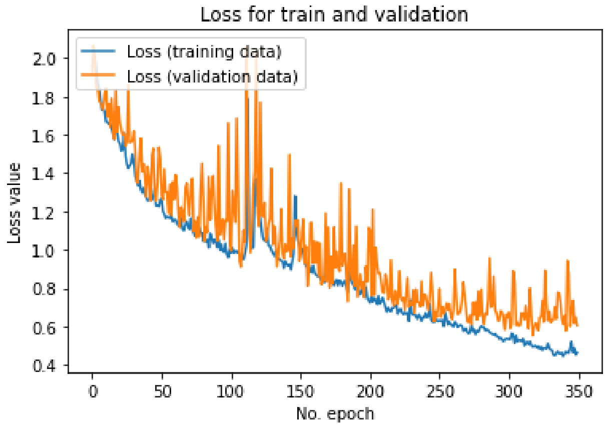


Fig. 4.10 DNN Model loss

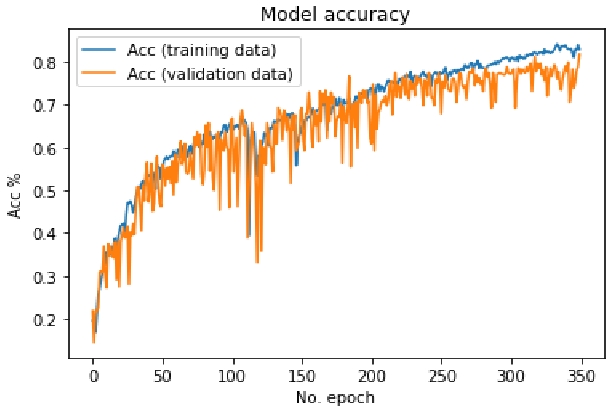


Fig. 4.11 DNN Model Accuracy

4.7 Real Time Implementation

One of the objectives is not only to construct various machine learning models, but to implement the best system based on the results obtained and build the Real Time Emotion Detection system. The following are the steps followed to do the same.

Step1: Load the speech emotion recognition LSTM Model and weights

Step2: Data Preprocessing

An audio input .wav file is processed similarly to the model's preprocess in the following order:

- Sample rate: It is the number of audio samples extracted per second by librosa.
- 'Audiosegment' instance: the audio is loaded to an object by the audiosegment module of pydub.
- Normalization: the 'audiosegment' object is normalized to + 5.0 dbfs, by effects module of pydub.
- Transforming the object to an array of samples by numpy & audiosegment.
- Noise reduction is being performed by noisereduce.

Speech features are extracted as well:

- Energy - root mean square (rms)
- Zero crossed rate (zcr)
- Mel-frequency cepstral coefficients (mfccs)

With frame_length = 2048, hop_length = 512, assuring equally

sequential length. The features are concatenated to an 'x' variable,

adjusted to fit the expected shape of the model: (batch, timesteps,

feature). The function returns 'x_3d' variable.

Step 3: Additional Setup

- An **emotion list** is defined to translate the model prediction output to a readable form.
- Is_silent function is executed as a boolean state if silence of sequential audio was found. Is_silent returns true when the maximum signal within the sequence is less than the threshold value defined.

This implementation of an LSTM Speech Emotion Recognition model carries out a real-time emotion prediction of an audio input, recorded from the soundcard of the platform. The process includes the following:

1. Session start, opening a connection with the input channel by pyaudio.

2. If **not silent**, the input signals will be recorded to a .wav file, by pyaudio and wave.

2.1 After 7.1 seconds, the recording will stumble in order to send the last .wav file to the rest of the process before start recording a new one.

2.2 The .wav file is preprocessed, in preprocess function.

2.3 Model.predict is executed, an array of 8 emotion probabilities is returned. E.g. Predictions = [array([p_neutral, p_calm, p_happy, p_sad, p_angry, p_feaful, p_disgust, p_suprised], dtype=float32)]

2.4 Predictions are transformed to a compact representation (without 'array' and 'dtype' statements) and saved in a list.

2.5 A visualization of predictions is shown by matplotlib.

3. Else, if silence is identified within the last 2 seconds of a .wav file:

3.1 Break; end of the session; close connections.

3.2 Visualize a summary of the session: Mean value of the predictions.

3.3 State the overall session time.

Variables:

RATE = Sample rate = 24414 which is the sample rate of most of

the model's train data. CHUNK = A batch of sequential samples

to process at once. Similar to 'hop length' by librosa, defined 512.

FORMAT = Sampling size and format, 32bit as in the model.

CHANNELS = 1 for mono, a standard of audio recording in PC /

cell phones.

5. Result Analysis and Discussion

The accuracy comparison is carried out between original dataset and preprocessed dataset. Multi Layer Perceptron is the first model built and Table 5.1 shows the resulted accuracy of original and augmented dataset with increase in accuracy by 4.33% and 4.2% with respected emotion and Gender respectively while data is augmented.

Table 5.1 MLP accuracy

MLP	Original	Augmented
Emotions	42%	46.33%
Gender	62.41%	66.61%

Table 5.2 convey us the accuracy of SVM model in each category of emotions taken into consideration. Overall accuracy of 53% was observed with the same model.

Table 5.2 SVM accuracy

Angry	0.56	Happy	0.39
Calm	0.69	Neutral	0.37
Disgust	0.49	Sad	0.54
Fearful	0.55	Surprised	0.53

CNN model's accuracy is compared with respect to original dataset and pre-processed dataset shown in Table 5.3 there is an substantial proliferation in accuracy with percentage of 35% for Emotions, 46% for Gender-Emotions, 23.35% for Gender.

Table 5.3 CNN accuracy

CNN	Original	Augmented
Emotions	45%	80.30%
Gender-Emotions	30%	76%
Gender	65%	88.35%

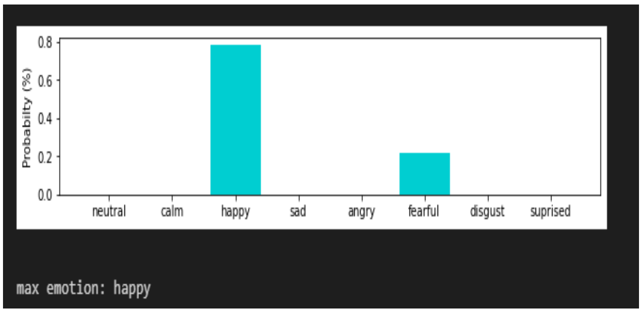
Table 5.4 represents the accuracy for varied emotions of Validation dataset and Test dataset. Its clear form below table that model projects better performance with validation dataset asserting the fact of better training.

Table 5.4 DNN accuracy

DNN	Validation	Test
Neutral	0.9130	0.8571
Calm	0.5625	0.4286
Happy	0.8302	0.7143
Sad	0.8070	0.7778
Angry	0.8846	0.7619
Fearful	0.8077	0.7826
Disgust	0.7200	0.8571
Surprised	0.8571	0.8095

Real time emotion detection system result

The results of three input audios has been displayed here in this section. The system is experimented with three different audio signals, which were recorded by users. The results are portrayed in the form of probabilities of various emotions which occur in an statement of a user. Figure 5.1 shows the probabilities of emotions happy (~75%) and fear full (~25%) captured in first audio input.



```
max emotion: happy
```

Fig. 5.1 Emotion recognition – 1

Figure 5.2 displays the result of second audio input to the system, where model could identify emotions like Angry, disgust and surprised with various probabilities. Its clear from the result that surprised has high probability compared to other two emotions detected.

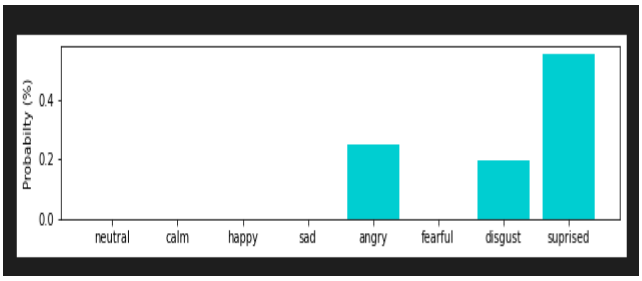
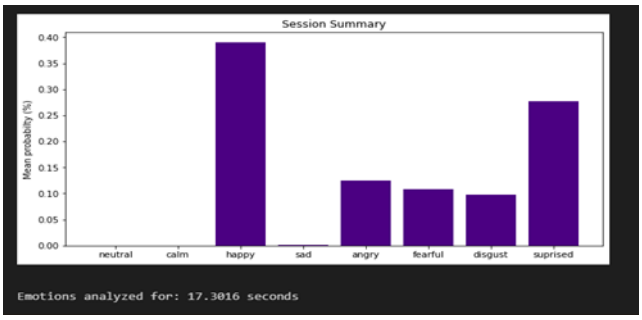


Fig 5.2 Emotion recognition – 2

The result of third trail is shown in Figure 5.3, where model was able to identify diverse emotions such as Happy, angry, fearful, disgust and surprised with respective probabilities. Here we can see Happy dominating the other emotions in the current input.



```
Emotions analyzed for: 17.3016 seconds
```

Fig. 5.3 Emotion recognition – 3

6. Conclusion

This work intended in considering Gender into consideration while detecting emotion of a given audio file. Method proposed here uses MFCC, Mel Spectrogram, Chromagram. Along with data augmentation and pre-processing, in later stage project covers implementing algorithms like SVM, MLP, CNN and DNN. It is evident from the result that DNN out performs other mentioned algorithms in the present work with accuracy of 91.3% ,as compared to 88.3% of that of CNN , 53% with SVM and MLP with accuracy of 66.61%. The Real time application is implemented with DNN model.

References

[1] J. Andersson, L. Baresi, N. Bencomo, R. Lemos, A. Gorla, P. Inverardi and T. Vogel, "Software Engineering Processes for Self-Adaptive Systems," *in Software Engineering for Self-Adaptive Systems II, Springer,* 2013, pp. 51-75.

[2] Anusha Koduru, Hima Bindu Valiveti, Anil Kumar Budati , "Feature extraction algorithms to improve the speech emotion recognition rate" *International Journal of Speech Technology,2020*

[3] Manju D. Pawar & Rajendra D. Kokate , "Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients", *Springer*, 2021

[4] Koo, H., Jeong, S., Yoon, S., & Kim, W., "Development of speech emotion recognition algorithm using MFCC and prosody" *International Conference on Electronics, Information, and Communication (ICEIC)* (pp. 1-4), IEEE.2020

[5] N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas and G. Kalliris, "Continuous PAPERS Speech Emotion Recognition with Convolutional Neural Networks",2020

[6] Mustaqeem, Kwon, S, "MLT-D Net: Speech Emotion Recognition Using 10 Dilated CNN Based on Multi-Learning Trick Approach, *Expert Systems with Applications*,2020

[7] Kudakwashe Zvarevashe and Oludayo Olugbara , "Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition",2020

[8] Jerry Joy, Aparna Kannan, Shreya Ram, S. Rama, "Speech Emotion Recognition using Neural Network and MLP Classifier", *IJESC,2020*

[9] Enkhtogtokh Togootogtokh, Christian Klasen, "Deep-EMO: Deep Learning for Speech Emotion Recognition",2020

[10] Arash Shilandari, Hossein Marvi and Hossein Khosravi "Speech Emotion Recognition using Data Augmentation Method by Cycle-Generative Adversarial Networks",2021

[11] Stavros Ntalampiras , "Speech emotion recognition via learning analogies",2021

[12] U. Kumaran, S. Radha Rammohan, Senthil Murugan Nagarajan, A. Prathik , "Fusion of me/ and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN”,2020

[13] Zengwei Yao, Zihao Wang, Weihuang Liu, Yaqian Liu, Jiahui Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers HSF-DNN, MS-CNN and LLD-RNN”,2020

[14] Panwit Nantasri, Ekachai, Jesseda karnjana, Surasak Boonkla,” ,” A Light-Weight Artificial Neural Network For Speech Emotion Recognition Using Average Values Of Mfccs And Their Derivatives”, *17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, June 2020, DOI:10.1109/ECTI-CON49241.2020.9158221, 2020

[15] Mustaqeem, Soonil Kwon, “MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach”, *Expert systems with Application*, Volume 167, 1 April 2021, 114177, https://doi.org/10.1016/j.eswa.2020.114177

[16] Nikolaos Vryzas, Rigas Kotsakis, Aikaterini Liatsou, “, Speech Emotion Recognition for Performance Interaction”, June 2018, Journal of the Audio Engineering Society. Audio Engineering Society 66(6):457-467 10.17743/jaes.2018.0036.

[17] Verma, D. N. . (2022). Access Control-Based Cloud Storage Using Role-Fully Homomorphic Encryption Scheme. Research Journal of Computer Systems and Engineering, 3(1), 78–83. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/46

[18] Saxena, K. ., & Gupta, Y. K. . (2023). Analysis of Image Processing Strategies Dedicated to Underwater Scenarios. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3s), 253–258. https://doi.org/10.17762/ijritcc.v11i3s.6232