iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.36973

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Sentiment Perception inArticulation

Ganesh NagaVenkata Sai MohanKancherla[1], Bhargava Sai Sathvik Gontla[2]
[1]*School of Electronics and Communication Engineering Vellore Institute of Technology, Vellore India*
[2]*School of Computer Science andEngineering Vellore Institute of Technology, Vellore India*

Abstract: Emotion is quite prevalent aspect in daily life. Every individual has a inequity levels of anxiety in the finding the concealed emotion present in a speech or talk. So we had decided to procreate a new methodology in which every emotion which is present in a speech can be detected. The system we developed can detect any emotion with a great extent of efficiency. Any type of emotion will be detected using Machine learning algorithms in a effective way. We will utilize Multi-Layer Perceptron in the initial stage and then we will compare this with working model of Convolution Neural Networks. We want to develop an Artificial Intelligence perception system which leads to detection of emotion in any articulation

Keywords: Mel Frequency Cepstral Coefficient and Classifier, LSTM,CNN,MLP, recurrent neural network (RNN) SAVEE (Surrey Audio-visual Expressed Emotion)

I. INTRODUCTION

Nowadays Identifying feelings is quite possibly the main promoting methodologies. Any individual can personalize various emotions explicitly to relevant situation. For this reason, we had chosen one task where we could distinguish an individual's emotions simply by tone that would allow to inspect various AI applications.A small number of models can be counting call focuses to play music when one is furious on the call. In subsequent method, mixed system was proposed [1]. Trisubsystems (LLD-DNN, LLD-SVM, DNN-SVM) arerequired to detect the feelings of a clip. Another could bea smart vehicle easing back down when one is furious orunfortunate. As a result, this kind of operation had a lot potential on the planet that would profitorganizations andeven give wellbeing to buyers. The thought behind making this undertaking is to assemble an AI model that can recognize feelings from the discourse we have with one another constantly. These days personalization issomething that is required in every one of the things we experience. So why not have a feeling locator that will check your feelings and, later on, suggest you various things dependent on your disposition. The types of tones present in the speechare classified in different pieces : phonetic features and prosodic features [2] This can be utilized bynumerous businesses to offer various administrations likepromoting organization proposing you purchase items dependent on your feelings, auto industry can identify the people feelings and change the speed of self-sufficient vehicles as needed to stay away from any crashes etc. The attributes of human voice like the pitch, tone, commotion, and tone makehuman voice a flexible to impart. It very well may be seen that people can likewise communicate their feelings by shifting the expressed attributes. In any situation if the tone is happy, angry or scared [3]This takes into account distinguishing human feeling by analysing speech. With various feelings and temperaments, not exclusively does the apparent quality shift, however the related discourse designs change as well. [4]In speech feature extraction, there are ways like prosodic feature, sound quality feature and spectral feature For example, individuals may will in general talk in uproarious voices when scared furthermore, utilize piercing orshrill voices when in a frightened or froze passionate state. A few group will in general meander aimlessly when they get energized or apprehensive. Despite what might be expected, when in a meditative enthusiastic state,individuals will in general talk gradually and make longer stops, subsequently showing an increment in time separating between successive expressions of their discourse. Having a counterfeit specialist comprehend crude human feeling will add to upgrading current condition of virtual specialists. Aside from causing a counterfeit specialist tocomprehend human emotion, speech assumption investigation can likewise be utilized in making people more mindful of the feeling of the individual conversing with them. Sound attributes of discourse can be utilized in situations where face to face correspondence isn't possible or where there are a language limitation and legitimate model for dictionary-based discourse investigation notpromptly accessible. Following is such situations where discourse attributes can fill in as an instrument for distinguishing human feeling [5], [6]and they had created history in the state-of-the-art presentation in time-series based classification tasks:

1) *Playing music and changing the surrounding room's lighting according to the tone of the discussion.*
2) *Execution in sociology research*

Customers administration focuses can accumulate experiences on their consumer loyalty by essentially investigating the discourse of their clients. Additionally, the scores got as a piece of this examination can be utilized to evaluate the general assessment of an organization/item/administrations. There are various algorithms which are effectively implemented in an allured way but there are plethora of implementationproblems while using vast datasets SER[7]-[11] mechanisms

©IJRASET: All Rights are Reserved

II. LITERATURE SURVEY

The main objective of the project is to build a model thatis very well trained to distinguish between male andfemale voices and should distinguish with 100% accuracy. The model should also be tuned to detect emotions with 100% accuracy. For the identification of the certain model for this paper we had done a brief literature survey and we identified different gaps from that papers.

Zamil [12] used the framework used is Sound Processing Technique: Mel Frequency Cepstral Coefficient and Classifier: Logistic Model Tree. Using this famework they had detected all kinds of signals present in a speech. They had implemented Voting mechanism on Classified frames. There are two datasets Berlin Database of Emotional Speech (Emo-DB) and Ryerson Audio Visual Database of Emotional Speech and Song (RAVDES S) used as source of data in this paper. The relevant finding from this paper is Extraction of required speech features using Mel Frequency Cepstral Coefficient (MFCC). The gap identified from this paper is the quality of the systemcan possibly be increased by using contextual informationalong with the audio clip characteristics for classification.

Zhuang [13] used the main theoretical model used is Mel frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstral Coefficient (LPCC) Classifier: Cubic spine Support Vector Machine (SVM) classifier model. They had executed Cubic SVM Classifier Based FeatureExtraction. The main concept we can extract from this paper is Extraction of required speech features using Mel Frequency Cepstral Coefficient (MFCC). The dataset thathad utilized in this paper is YouTube speech recordings of 10 different actors and 10 different actresses of Bollywood. Main limitation we identified from this paper is the system could be made very efficient and quick in addition of help of a proper selection of characteristics.

Xiaoyan [14] then main concept used for emotion recognition is the recurrent neural network (RNN) algorithm. In this paper they also allured to the new optimized model of Mel Frequency Cepstral Coefficient (MFCC) and used it. They had proposed a hybrid model which is the combination of both Recurrent neural network (RNN) and Mel Frequency Cepstral Coefficient(MFCC) for the emotion detection. The dataset used in this paper is this paper is Chinese speech emotion database which contains equal number of male and female speeches and five emotions are taken into action. The main limitation identified from this paper isimproving the concussion of remaining inequity featureson professor feelings.

Zhang [15] had the main idea of the paper is to overcomethe fundamental issues present in the field of machine learning algorithms such as data representations, herein represented as deep spectrum signs. They hadimplemented a model to overcome these problems by showing weaknesses in a parallel combination of attention-based bidirectional long short-term memory recurrent neural networks with attention-based fully convolutional networks (FCN). The datasets used in this paper are IEMOCAP which includes the conversations of dual actors in five classes of time and FAU which consists of various speech samples in German language. The main limitation observed from this paper is they couldn't establish their techniques in any other speech recognitiontasks.

Lotfian [16] they had implemented a synthetic speech signal that makes understand the same language data andit is pointed to the same targeted sentence that is present in the vast database. Initially they had used the SENAINEdatabase. Then they annoyingly used FEELTRACE records. And they used SEMAINE database for the second round of experiments. The main limitation identified from this paper is the emotions detected at the end of this paper are arranged in an inappropriate way likethe output is correct but the data present in the output is displayed in an ununiform way.

III. OVERVIEW

A. *Motivation*

Now a days many people are hiding their emotions and speaking to others in an artificial way. We are very allured to the emotions which are concealed and started researching about finding these secret emotions which are hidden viciously deep inside the words and their pitch while speaking. Wecome to know that there are plethora of ways in which ones original emotion can be detected. We had a deep interest in neural networks field which can predict the most accurate answer for any question or situation. In many sectors our paper plays a vital role suchas interrogations, political speeches etc. This technology is required badly in our day to day life. In today's world our motive makes a drastic change in any social segmentspresent in present day to day life.

B. *Dataset*

Use of two specific datasets:

1) RAVDESS (Ryerson Audio-visual Database of Emotional Speech and song)The dataset comprises of 1500 audio records from 24 kinds of artist people. each input data report has aunique identifier on the 6th element of the file name which could be used to find the emotion the audio report is composed. we've got 5 distinct feelings in our database. Calm(22.6%), glad(18.7%), sad(17.4%), angry(8.78%), nervous(13%) and the rest were neutral

©IJRASET: All Rights are Reserved

2) SAVEE (Surrey Audio-visual Expressed Emotion) This datasheet consists of 500 audio files recorded by four one of a kind male artist the primary two characters of thefile name corresponding to the distinct emotions that theyexhibit. We used Librosa library in Python to make and dissolve functions from the audio files. Librosa is a python bundle for songs and audio analysis. It offers the building blocks important to create track informationretrieval structures. using the librosa library we have beenable to take out functions that is MFCC. The MFCCs were characteristic broadly utilized in computerized speech and speaker reputation. We also separated out thewomen and men voice by the means of usage of recognition tags furnished on the website. because as working, we determined out that setting apart male and girl voices multiplied through 15%. it is able to be because ,the voice intensity became affecting the outcomes results and efficiency and average accuracy is mentioned by the final observations made(refer table 1)

C. Pre-processing

All the null values and non-relative data were removed and an outliers were also made to find data that falls out of range instead of losing the data the outliers were replaced instead with the mean value irrespective of including them so that efficiency could be attained and we have split the data into 2 parts first the training part, comprising 70% of the whole dataset available and the remaining dataset to make evaluation of the model after multiple trails we ended up with this ratio so that the efficiency was reasonable and admirable

Initially we use the MFCC for making the pre-processingwith the available dataset. And the audio document comprising of .wav tag, initially we calculate the amplitude magnitude with the every given audio document with a change in pattern of charge 16000 sample adequately on ration to second . Now we calculatethe weighted average in step with respect to duration of media files then convert them identical via giving null values for minimal document to equate them with the common duration record and make a record of all the bigfiles for the same purpose. At the end by following thesemethodology, every documents have become identical inpoint of length. Data standardization (normalization)[17] processing is a general most duty for the data mining After the raw statistics is made, every index is inside the same kind of value, that could be suitable for complete comparative evaluation. each measurement is processed to keep away an effect on various dimensions on distance calculation. In classifiers or clustering algorithms, when distance is used to find and evaluate the likeliness, or when PCA is used to lessen measurement, the Z-score standardization appears to carry out better

$Y_i=(X_i - \text{mean}(X))/\sigma$

This could standardizes the mean value and standarddeviation of the initial data .the stage processed data along with standard normal distributionwhich is the mean value's zero and the standard deviation is 1

IV. PROPOSED MODELS

A. MLP

Multi-Layer Perceptron (MLP) is a community made up of perceptron. Fig(1)It has an input layer that gets the given signal ,an outcome layer that makes predictions or selections for a particular individual requirement, and thelayers found in among the given layer and outcome layer is known as invisible layer. There may be many invisible layers, the variety of invisible layers may be modified as in keeping with the needed. For the proposed technique for Speech Emotion Recognition, the Multi- Layer Perceptron network may have one given layer, of three hundred and more than 40,000 invisible layers and 1 desired layer. The input layer will take as given entry forthe 5 features, which are extracted from the audio document. The extracted 5 capabilities were likely, Mel Spectrogram Frequency, Chroma, Mel Frequency Cepstral Coefficients , Contrast and Tonnetz. The invisible layer uses an trigger function to act upon the enter statistics and to system the information. The Fig(1)triggering characteristic used is logistic activation function. The output layer brings out the facts found out by the community as output ,this layer segregates and returns output of the expected emotion, in respect with the computation accomplished through the invisible layer.

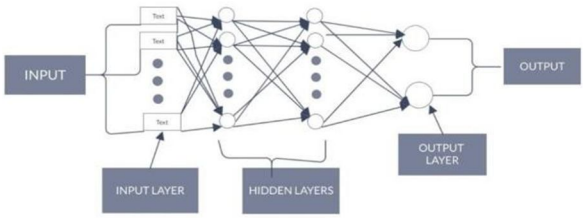


Fig.1 Architecture of MLP classifier

©IJRASET: All Rights are Reserved

Multilayer perceptron is implemented for supervised learningsort of problems. The multi-layer perceptron is taken into needfor the sake of classification. The MLP is made to understandthe subsequent output rises and conditions from the data sheetit takes initially. The training section permits the MLP to studythe mutual dependencies between the set of inputs and outputs.Duringtrain period the MLP adapts model parameters along with weights and biases which will minimise the mistakes in the final outputs. The MLP makes use of Backpropagation, to have weight and bias modifications relative to the error. the error could be calculated in lots of approaches too.

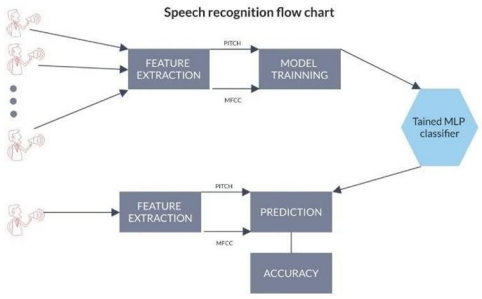


Fig.2 flow chart of MLP model

1) Constructions of MLP Classifier

a) Start the program MLP classifier by mentioning the needed inputs and params Fig(2)
b) Sheet of data will be provided to the Neural network to train it
c) Trained network are used for making the prediction of output
d) And finally calculating the efficiency achieved in the predicted output

2) Working of the MLP: Multi-layer perception classifier depends and comes under the neural network to make and analyze the classification outcome. The MLP executes the multi-layer perception algorithm and instructs the neural network using backpropagation algorithm

3) Feature extraction[18]

Capabilities constitute the traits of human vocal channels and listening systems. Because the feature extraction is a complicated system, economical feature-extraction Fig(2) is difficult assignment in an sentiment identity model uprooting required attribute is the high priories points of sentiment identity model. Linear Prediction Coefficient is highly utilized capabilities for the two speech and sentiment identification. The fundamental thought of an LPC model is as samples of voice at atime t, s(t) could be estimated by the additive amalgamation of the old audio samples.
It is denoted in terms of mathematically as

$$e(n)=s(n)-\sum_{k=1}^{p} a(k)s(n-k)$$

a) Mel Frequency Cepstral Coefficients: (MFCC)[19] is used to retrieve the sound from the inserted audio file by means of using distinct hop period and HTK-styles mel frequencies.Pitch of 1 kHz tone and 38 dB over the perceptual discernible area is defined as one thousand mels, utilised as an indexpoint. TheMFCC simulates a Discrete Cosine trade (DCT) of a proper logarithm of the transient vitality confirmed at the Mel recurrence scale

Mel(f)=2595 log10(1 +freq/700)
f:frequency of the given input signal

b) MEL: The Mel scale relates obvious repeat, or pitch, of an unamplified tone to its actual assessed recurrence. peopleare pretty improved at perceiving little changes in pitch at low frequencies than they're at excessive frequencies. Solidifying this scale makes our capabilities set up even extra eagerly what individuals concentrate.

M(f)=1125Loge(1 + freq/700)

©IJRASET: All Rights are Reserved

c) *Chroma:* The Chroma [20]module is an analyser, which speaks to the tonal substance of melodic sound indicators in a consolidated form. consequently, Chroma spotlight can be taken as good-sized crucial for multiplied stage semantic examination, similar to harmony acknowledgement or consonant nearness estimation. A higher nature of the extricated chroma consists of empowers a good deal higher outcomes in these improved level assignments. The chroma is figured via consisting of the log-repeat size range throughout the octaves. the approach about the plan of chroma vectors can besaid as chroma-gram

$$Cf(b)=\sum_{z=0}^{z=1} X[f(b+z\beta)]]$$

d) *Tonnetz:* The Tonnetz is a pitch area characterised by means of thesystem of connections among melodic makecontributions just inflection .Close symphonies connections are displayed as brief separations on a tremendous Euclidean plane.

B. LSTM

RNN (Recurrent neural [21]networks) has an ability to analyze and respond to the occasion with out modifying the slow fashioned weighs, incorporating short-term triggers to the latest activities. this option is useful with the instance of programs that execution time is an important feature, As they may be trained with the help of back Propagation via Time, mistakes indicators moving backwards with respect to schedule could both turn out to be larger and bigger or disappear relying on the capacity of weights, and could create swing in weights or manipulate the entire network to have a gradual in training and converging in order of comprising the quick-time period model of RNNs and keep away from the problems. they may be incorporating a gradient-primarilybased set of rules enforcing consistent errors float throughcharacter units, specifically designed to deal with the fast-time period; as a consequence, they are able to trim the gradient calculations at a finite factor with out disturbingthe long-term activations. In cutting-edge years, LSTM networks have been turning into in the centre of a hobby for many applications associated with time-collection events. Speech-Processing and mainly speech sentiment identification are of those applications.

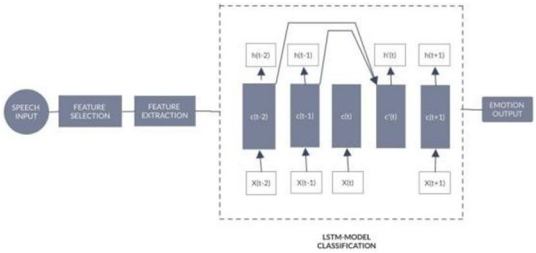


Fig.3 LSTM classifier

1) *Feature Extraction:* Voice is Fig (3) form of nonlinear time collection sign, textual content records are carefully associated with a temporary context, and all of them are associated with time. consequently, apart from in LSTM, there may be nointermediate nonlinear part hidden layer that reasons the growth in difference inside the hidden state factors. In short, the version skills related to CNN and LSTM are each limited. CNN and LSTM networks had been resorted in for speech sentiment popularity, thus turned out as an improvement technique frequently utilized in lots of contemporary research.
2) *Algorithm:* Considering the fact[22] that there is the distinction among the anticipated emotion tag together with the self-tagged emotion Fig(4)profile, these results normalizes the self-tagged emotion profile a good way acquiring identical scale with the detected emotion profile. Initially,the depth of every emotion for each segment in the EP- DB is calculated with the aid of the contributors the use of the size of from zero to five For every section, the pitchfor individual emotion changed into normalize by a scaling characteristic, described as below

$$EE^{p}{}_{i}=Sigmum(EE^{p}{}_{i}) \quad \frac{1}{7} \quad I=1$$

- $EE^{p}{}_{i}$:level of intensity of emotion
- sigmod function $1/1+e^{-x}$

Diff is the difference among the self tagged emotion intensity and intensity of predicted outputemotion that can be mentioned as follows

$$Diff^{p}{}_{i}=PE^{p}{}_{i}-EE^{p}{}_{i}$$

The value of diff from the above equation can be related with gaussian distribution with mean same as per the gaussian distribution and can be utilized to simulate the (GWN)gaussian white noise too

©IJRASET: All Rights are Reserved

Then, every individual function vector is loaded to the sentiment prediction model to acquire the sentiment result. The expected sentiment profile is then modified into the expressed sentiment result with help of the personality-based transfer feature. eventually, the expressed sentiment profile succession is utilized for emotion identification of the use of long short-term memory model appearing the complete matrix

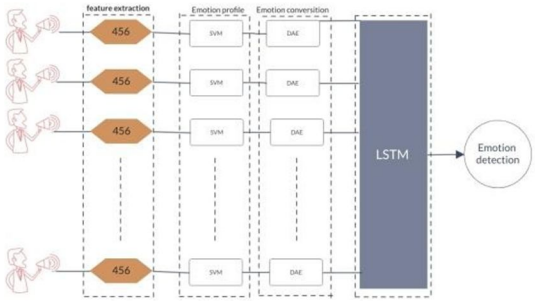


Fig.4 Flow chart of LSTM model

C. CNN

Convolutional neural network encompass several layer ofconvolution .In CNN, non-linear activation characteristicas an specimen of rectified linear unit (ReLU) or Sigmoidfunction were utilized for the final output. In neural networks nodes of the given layer were interlinked with the nodes of hidden layer and people hidden layer nodes were completely linked with nodes of return layer. Within Fig(5) CNN, convolution were implemented on given layer to generate the return. Each and every fragment of given were convoluted using special filters and concatenating them we get the final result. Within CNN, there could be a pooling layer and reason of these layer issub sampling the given against a particular filter. A non- unusual pooling approach is max pooling. In max poolinga max fee is identified from every clear out. Pooling layerlessen the capacity of given input. Max pooling layer could also carryout above a window as opposed to

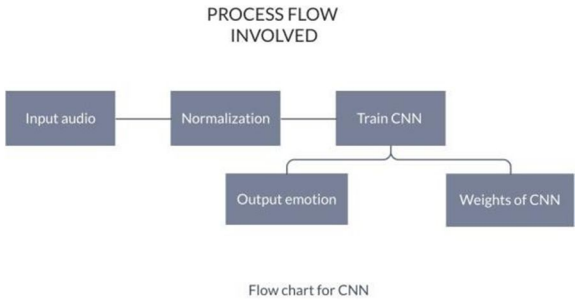


Fig.5 flow chart of CNN model

Statistical data is fetched to the device which is composedthe expression label and Weight train is also given for thatnetwork. Audio is given for the model, Thereafter, depth normalisation is carried out over the audio. Normalised audio is used to train the Convolutional network, this is performed tomake sure that the impact of presentation collection of the examples doesn't affect the schooling overall performance. The collections of weights pop out as an outcome of this training methodology and it acquires the great consequences with this mastering statistics. at the same time as testing, the dataset fetches the device with pitch and strength, and primarily based on the final network it offers the decided emotion. The output is represented in a numerical cost every corresponds to bothof the 5 expressions. There are three feelings that are being detected based on the person's bpmcount, those are Calm, joy or enjoyment, Fear or Anger.

1) Algorithm

a) Step 1: The input is given for the respective audio signal

b) Step 2: plotting the spectogram and waveform from thegiven input

c) Step 3: the use of the LIBROSA, we extract the MFCC(Mel Frequency Cepstral Coefficient) usually approximately 10–20.

d) Step 4: Remixing the records, dividing it in trainingand testing and there after building aCNN version and its successive layers to train the data provided.

e) Step 5: Predicting the output for the given audio signal

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

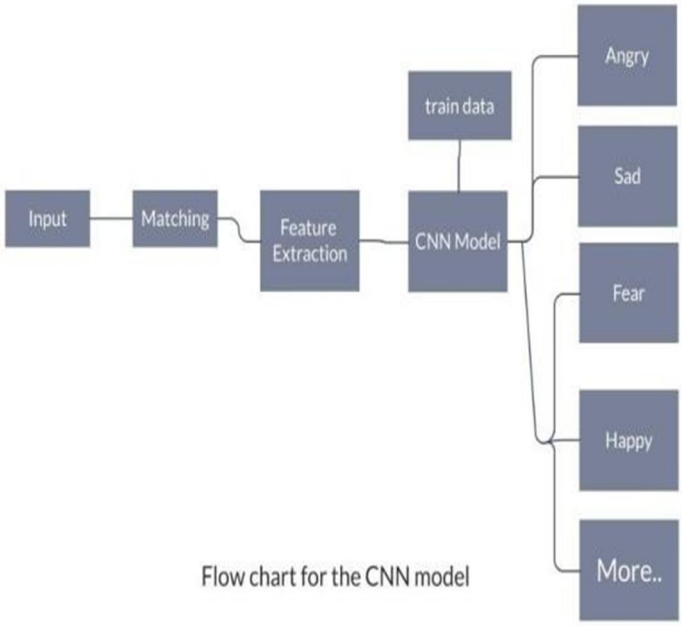


Fig.6 Final output flow of CNN model

2) CNN Layers

a) Audio characteristic Extraction and Visualizations: Characteristics extraction is required for classification and showcase. The audio sign is a 3-d input signal wherein three axes suggest time, amplitude and frequency. Fig(6)we use Libros to analyze and extract thetraits of any audio sign. (.load) feature pulls an audio file and decrypts it right into a 1D array that's of time collection x, and SR is absolutely sampling rateof x. by using default SR is 22 kHz. right here I can display one audio file show with the use of the (IPython.display) function. Librosa. show is vital to represent the audio documents in diverse paperwork i.e. wave plot, spectrogram and colourmap

b) Train the Model: Inside this module we train the version for accuracyestimations. 1st, import important modules. Then load the dataset. we are able to get hold of the sampling runtime Fig(6) with librosa programs and mfcc characteristic function. Thereafter this fee holds other variables. Now audio files and mfcc hold a variable therefore it will adda listing. Then zip the list and keep two variables x & y.Then we've got represented (x, y) form values with theusage of numpy package

c) Implementation of CNN Model: Speech converted into image format in 3 layers, on usage with CNN we have to consider the firsttwo derivatives of image with respect to time and frequency CNN will predict and analyse the speech data and CNN will understand and train itself from speeches and words

V. CLASSIFICATION OF SPEECH EMOTION

Here we expect the output form the CNN model and labelling them with respective to names and labels and taglines and the Fig(6) data output could be displayed with good user understandable interface Comparative analysis among the three algorithms used in our method

A. MLP

Contains either one or greater than one layer of neurons, and the data is loaded into the first inputlayer, and it maybe more than one or maybe one hidden layer which provides the level of abstraction and the required predictions were depicted in the output layer called as a visible layer, suitable for predictions made assigned on the tags and label encoding part, data provided in a CSV format to the feed input, since they were flexible it couldbe applied to other datasets too higher level of predictioncould not be expected in case of audio and image datatypes

©IJRASET: All Rights are Reserved

B. LSTM

Provides the best accuracy in case of time series computations because of the short term sequence of memory and since the dataset is large LSTM may not perform well and in our literature review only a few strongly supported that LSTM provides much accuracy and we have observed a contradiction to it and LSTM networks always follow iterative gradient methodologies and most in prior they use the backpropagation idea, thatalways modulates the weight in respect to time and directly varying with error derivative for minimizing the net error on the total of testing and trainingdata

C. CNN

They were used to map data from input to output datatype, could be used to any data type irrespective of the data type entry of the training data set, they have an ability to convert and interpret any data type into images of 2D dimension ,they generally show abrupt results of thedatatype those have spatial relationship ,Mostly the CNNmodels were used to take the data type of2-dimension butstill they could be converted into 1-dimension array or lists internal representation, spatial relationship could be in our work was the converted audio to image signal and their relationship

VI. RESULTS

The experiment was held throughout on a windows platform and the data comprising of RAVDESS and SAVEE containing 1500 and 500 data entriesrespectively collected from 2 different websitescontaining 2 different parameters and values and the finaloutputs were expected to be as 5 classes Calm(22.6%), glad(18.7%), sad(17.4%), angry(8.78%), nervous(13%) and the rest were neutral All the inputs were the audio files and pre-processing was done for cleaning the data and feature extraction Is made according to the model need and params .

The personal computer is of version CORE I5 and 10th generation, processor 2.50GHz 2.50 GHz, 64-bit operating system graphic card 16GB and theram of 8GB GTX 1650 TI .working library include LibROSA library working with Python, pyAudioAnalysis library working with Python, WavePad Audio Editor and the additional tools used were SciPy(provides fast n-dimensional array Manipulations), TensorFlow (collection of workflows),Keras(deep learning framework), SKlearn(predictive data analysis), NumPy(data structure, n-dimensional array), pandas(Analysing and interpreting the data) and the training batch size was always kept at 32 sizes and the epochs were varied accordingly before loading them intotrain models and upon execution, the results were depicted below in the Tabel1

A. Findings

The data size loaded and the batch sizes were same to allthe experiments made ,As shown in the table 1 the resultsare interpreted with various conditions and models for achieving the maximum efficiency in MLP it is observedthat the efficiency is out as 25% and the data sensitivity is almost high and total of 8 layers were being used and the number of epochs that giving the maximum in MLP were 500 and all the parameters were given are utilized by the model for training and activation function used wasSoftMax, on moving to the LSTM it provided with an efficiency of 15% and data sensitivity was also high ,total 5 layers were used namely embedding,lstm1,dense layer,lstm2 and again a dense layer and a output layer.

Epoch were observed 50 which could be giving the suitable decent accuracy .activation layer as tan h function layer and all the data feed given is utilized and no parameter is left without training on moving to CNN here conditions were reasonably giving good accuracy so 2 different epoch were considered one giving 58.2 percent efficiency and the other with reduced epoch and modified 1d layer giving the output as 81.18% efficiency in detail 18 layers were observed convolution layer(6 in number), activation layer(7 in number) ,droputlayer(2 in number),max-pooling layer(1 in number), flatten layer(1 in number), dense layer(1 in number) excluding the output layer total number of layers were 18 ,2 activation functions were altered SoftMax activation function and rmsprop activation function and all the trainable parameters passed to the feed were used no were leftover and comprising of highest sensitivity 98.7%

©IJRASET: All Rights are Reserved

Heads/Heads	Sensitivity	Efficiency	both dataset s)	Batch size	Total layers	epochs	Activation Function	Trainable paramete rs	Non-Trainable parameter s
(MLP) Multi layer perceptron	98.16%	25%	(1500,500)	32	8	500	softmax	235,897	0
(LSTM) Long Short-term memory	97.8%	15%	(1500,500)	32	5	50	tan h	1,757,481	0
(CNN) Convolution neural network	98.7%	58.25 %	(1500,500)	32	18	700	softmax , rmsprop	445,578	0
(CNN) Convolution neural network	98.7%	81.18%	(1500,500)	32	18	700	softmax , rmsprop	445,578	0

TABLE 1

B. Future work

The quality of the system can be increased if one can usea different dataset rather than the used one. There is chance of increasing the efficiency of the model that hadbeen implemented in this paper. Can build a sequence model to generate voice based on different emotions. Forexample we can automatically simulate the output tagbased on the emotion.

©IJRASET: All Rights are Reserved

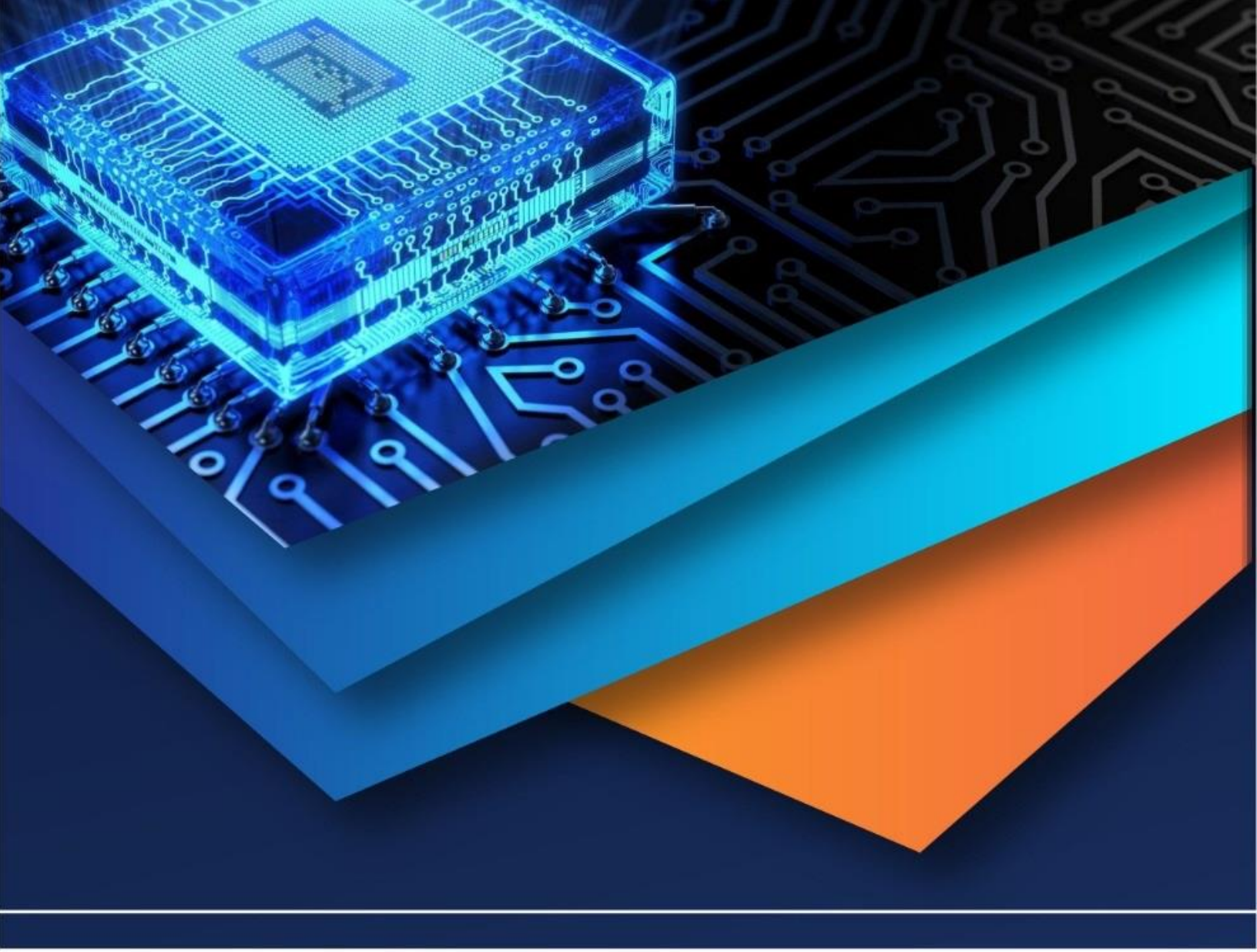
VII. CONCLUSION

After building the models and varying for the different cases in order to yield higher efficiency we ended up with an efficiency of 81.18% which is the maximum out of ourobserved experimental results from the CNN model comprising the specifications of epochs 700 and modified 1d convolution layer, and the final CNN model could still be more efficient if the provided data was much more accurate andin large in size. An noticeable impressive point was it wasable to distinguish between male and female voices and the tag of emotion recognised attached at the end to it

REFERENCES

[1] R. Chen, Y. Zhou, and Y. Qian, "Emotion recognitionusing support vector machine and deep neural network," in National Conference on Man-Machine Speech Communication. Springer, 2017, pp. 122–131
[2] Liu, Z. T., Wu, M., Cao, W. H., Mao, J. W., Xu, J. P.,& Tan, G. Z. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing, 273, 271-280.
[3] Johar, S. (2016). Psychology of Voice. In Emotion, Affect and Personality in Speech (pp. 9-15). Springer, Cham.
[4] LI Hong, XU Xiaoli, WU Guoxin, et al. Research on the extraction of speech emotion feature based on MFCC[J]. Journal of Electronic Measurement and Instrumentation, 2017,31(3):448-453.
[5] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, ''Attention based fully convolutional network for speechemotion recognition,'' in Proc. Asia– Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPAASC),Honolulu, HI, USA, Nov. 2018, pp. 1771–1775.
[6] F. Karim, S. Majumdar, H. Darabi, and S. Chen, ''LSTM fully convolutional networks for time series classification,'' IEEE Access, vol. 6,pp. 1662–1669, 2018.
[7] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, ''End-to-end multimodal emotion recognition using deep neural networks,'' IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1301– 1309, Dec. 2017.
[8] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, ''Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNsand FCNs for speech emotion recognition,'' in Proc.INTERSPEECH, Hyderabad, India, May 2018, pp. 272–276
[9] S. Mirsamadi, E. Barsoum, and C. Zhang, ''Automatic speech emotion recognition using Recurrent Neural Networks with local attention,'' in Proc. IEEE Int. Conf.Acoust., Speech Signal Process. (ICASSP), New Orleans,LA, USA, Mar. 2017, pp. 2227–2231.
[10] M. Neumann and N. T. Vu, ''Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,'' in Proc. INTERSPEECH, Stockholm, Sweden,Aug. 2017, pp. 1263–1267
[11] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, ''An attention pooling based representation learning method for speech emotion recognition,'' in Proc. INTERSPEECH, Hyderabad, India, Sep. 2018, pp. 3087–3091
[12] Zamil, A. A. A., Hasan, S., Baki, S. M. J., Adam, J. M., & Zaman, I. (2019, January). Emotion detection from speech signals using voting mechanism on classified frames. In 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) (pp. 281-285). IEEE.
[13] Jain, U., Nathani, K., Ruban, N., Raj, A. N. J., Zhuang, Z., & Mahesh, V. G. (2018, October). Cubic SVM classifier based feature extraction and emotion detection from speech signals. In 2018 International Conference on Sensor Networks and Signal Processing (SNSP) (pp. 386-391). IEEE.
[14] Jie, L., Xiaoyan, Z., & Zhaohui, Z. (2020, August). Speech Emotion Recognition of Teachers in Classroom Teaching. In 2020 Chinese Control And Decision Conference (CCDC) (pp. 5045-5050). IEEE.
[15] Zhao, Z., Bao, Z., Zhao, Y., Zhang, Z., Cummins, N., Ren, Z., & Schuller, B. (2019). Exploring deep spectrumrepresentations via attention-based recurrent and convolutional neural networks for speech emotion recognition. IEEE Access, 7, 97515-97525.
[16] Lotfian,R., & Busso, C. (2019). Lexical Dependent EmotionDetection Using Synthetic Speech Reference. IEEE Access, 7, 22071-22085.
[17] Cao, G., Ma, Y., Meng, X., Gao, Y., & Meng, M. (2019, July). Emotion recognition based on CNN. In 2019Chinese Control Conference (CCC) (pp. 8627-8630). IEEE.
[18] Palo, H. K., Mohanty, M. N., & Chandra, M. (2015). Use of different features for emotion recognition using MLP network. In Computational Vision and Robotics (pp. 7-15). Springer, New Delhi.
[19] Basu, S., Chakraborty, J., & Aftabuddin, M. (2017, October). Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. In 2017 2nd InternationalConference on Communication and Electronics Systems (ICCES) (pp. 333-336). IEEE.
[20] Jerry Joy1 , Aparna Kannan2 , Shreya Ram3 , S. Rama "Speech Emotion Recognition using Neural Network and MLP Classifier" IJESC, April 2020
[21] Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. Sensors, 21(4), 1249.
[22] Huang, K. Y., Wu, C. H., Su, M. H., & Fu, H. C. (2017, March). Mood detection from daily conversational speech using denoising autoencoder and LSTM. In 2017 IEEE International Conference on Acoustics, Speech andSignal Processing (ICASSP) (pp. 5125-5129). IEEE.
[23] Harini Murugan "Speech Emotion Recognition Using CNN" International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020 ISSN: 1475-7192

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)