ISSN 2347 - 3983
Volume 8. No. 5, May 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter34852020.pdf
https://doi.org/10.30534/ijeter/2020/34852020

MLP Model for Emotion Recognition using Acoustic Features

Suhail MSK[1],Guna Veerendra Kumar J[2], Mahesh Varma U[3],
Hari Kiran Vege[4], Swarna Kuchibhotla[5]
[1,2,3]B. Tech, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,Vaddeswaram, Guntur, A.P, India. msksuhail24@gmail.com,gunaveerendra@gmail.com,maheshvarma509@gmail.com
[4,5]Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India. hari.vege@kluniversity.in, drkswarna@kluniversity.in

ABSTRACT

Emotion recognition is an important topic of research lately. There are already a few methods that can predict emotion but, in this paper, we im to make a unique model that is not only lightweight but also fast and accurate. We are currently focussing on predicting 4 emotions like angry, sad, neutral and happy through frequencyanalysis. To achieve this, we first needto extract7 features including 195 sub-features from an audio file. The features like MFCC, Root mean square, Spectral contrast, Tonnetz, Zero-crossing rate, Mel spectrogram frequency and Chroma combined are useful in determining emotion from the audio frequencies accurately. We have trained the model using RAVDESS and TESS which are both open source databases for audio emotions. We made an intelligent python program that can analyze the frequency pattern of each emotion using MLP classifier and predict the emotion with an accuracy of 83%.Moreover, we can predict emotion either from a file or from the microphone instantly.

Key words: Dataset fusion, MLP classifier, RAVDESS, TESS.

1. INTRODUCTION

Emotions play an important role in our day-to-day life and they are generated by humans naturally. In computer science, a lot of research been done especially since finding an emotion from the text or voice is not that simple and it depends on the current state of the speaker and also the culture of the speaker. If we take this simple sentence “Why don’t you ever text me!” we can either arbitrate as an angry or sad emotion. Since there are no facial expressions, the detection of emotion from voice should be a challenging problem.

Emotions are basic human traits and have been studied by researchers in the fields of psychology, sociology, medicine, computer science, etc. for the past several years. Some of the prominent working understanding and categorizing emotions include Ekman’s six lass categorization and Plutchik’s “Wheel of Emotion” which issuggested as eight primary bipolar emotions. Given the vast nature of the study in this field, there is naturally no consensus on the granularity of emotion classes [1],[2]. In this paper, we consider 4 emotion classes, Happy, Sad, Angry and Neutral and classify the voice recording into one of the above four categories.

The major use of this application is since emotions are important to determine the state of the sentence (In real-world) sometimes judging the emotion from text/semantic meaning analysis from recording could be ambiguous. So, for accurate emotion analysis, we go with frequency analysis. It tells the state of the speaker. And, it can be used alongside the text recognition.

In our project, we made use of MLP classifier to detect emotions from the audio files which are given as input in .wav format.Our project holds four types of emotions Happy, Sad, Angry andNeutral.

The multilayer perceptron is mainly used for supervised learning problems, which train on a set of input-output pairs and learns to model the correlation between those inputs and outputs. This type of training adjusts the parameters, or the weights and biases, of the model to minimize the errors. It is a classifier that maps a set of input data to desired output dataThediagrammatic representation is shown in figure 1.

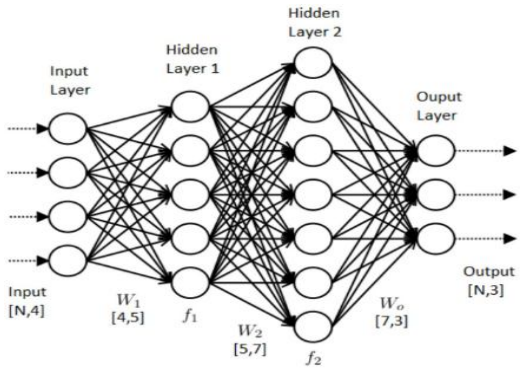

Figure 1: Typical neural network diagram

We have developed a model to define emotions from the speech which has undergone training and testing process.

2. DATASETS USED

RAVDESS and TESS are two open-source emotion audio sample databases that contain several audio samples with different emotions. There are 8 emotions in RAVDESS and 7 emotions in TESS to be precise. Now since we wanted to stress on 4 emotions i.e angry, sad, neutral and

happy, we took only 4 emotion files from the databases which total 2272 files (672 files from RAVDESS and 1600 files from TESS) for training and testing[6].

2.1 TESS

Ithas very high-quality audio samples.Most of the other dataset is skewed towards male speakers and thus brings about a slight imbalance representation. So, because of that, this dataset would serve a very good training dataset for the emotion classifier in terms of generalization.

There are a set of 200 target words were spoken in the carrier phraseand recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral).

There are 2800 data points (audio files) in total.
The dataset is organized such that each of their emotions is contained within its folder. And within that, all 200 target words audio files can be found. Refer to table 1 for more details.

2.2 RAVDESS

The RAVDESS is a validated multimodal database of emotional speech.Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and the song contains calm, happy, sad, angry, and fearful emotions.
Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in voice-only formats.

The set of 7356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. A further set of 72 participants provided test-retest data.
High levels of emotional validity and test-retest reliability were reported. Corrected accuracy and composite "goodness" measures are presented to assist researchers in the selection of stimuli. Refer to table 1 for comparison.

Table 1: No. of files comparison

Files used	RAVDESS	TESS
Angry	192	400
Sad	192	400
Neutral	96	400
Happy	192	400
Total	**672**	**1600**

3. FEATURE EXTRACTION

A machine cannot understand the data provided in the audio files. For analysing the audio files, we extract features. Feature extraction is the basic step for classification, prediction, and to define the best algorithms[10].

There are two types associated with wav files one is the bit depth and the other is Sample frequency. The sample frequency states the number of times per second the computer stab in and go.

With the help of the Sample rate and Sample data, we can perform many techniques to extract features from audio files.

Features extracted in our project include

1. Zero-Crossing Rate
2. MelFrequency Cepstral Coefficients (MFCC)
3. Chroma Vector
4. Tonnetz
5. Root Mean Square (RMS)
6. Contrast
7. MEL Spectrogram Frequency (mel)

Each feature holds its importance, and we are using every feature for the classification of emotions[7],[8],[9].

3.1 Zero-Crossing Rate

The measure of the signal value that crosses the zero line. The Zero-Crossing Rate (ZCR) of an audio frame predicts the count of signal change value, from positive to negative and vice versa, divided by the length of the frame.

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} \mid sgn[x_i(n)] - sgn[x_i(n-1)]$$

(1) is ZCR Equation where sgn $(\cdot)$ is the sign function, i.e.given in equation (2)

$$sgn[x_i(n)] = \begin{cases} 1, & x_i(n) \geq 0, \\ -1, & x_i(n) < 0. \end{cases}$$

3.2 Mfcc

MFCC is widely used in speech processing applications. It can capture the phonetical important characteristics of speech. A mel is a measurement based on the human ear's discerned frequency. Variation of the human ear's critical bandwidths with frequency is the basis for MFCC.Filters spaced linearly atlogarithmically at high frequencies and low frequencies are used to capture the phonetics which are important for determining the characteristics of speech.

MFCC uses in the mel-frequency scale for measurement, which is logarithmic frequency spacing above 1kHz and linear frequency spacing below 1 kHz.

3.3 Chroma

Chroma feature is used to represent the vocal content in audio files. It describes the angle of pitch rotation as it traverses the helix
• Two octave-related pitches share common angle in a chroma circle
For analysing the western tonal music we quantize its angle into 12 pitch sets or positions. (3) represents Chroma equation.

$$f_c(k_{lf}) = f_{\min} \times 2^{\frac{k_{lf}}{\beta}} \quad (3)$$

From equation (3), Fmin, is the minimum frequency obtained on analysis.

Klf , refers to integer filter index ∈ [0,(β × Z) – 1] β = bins per octave Z = number of octaves.

Features of chroma:

- Filterbank of the overlapping window.
- It acts as a Center frequency of windows

All the available windows are normalized to a unit sum (Short Term Feature extraction).

This method splits the input file into small windows and computes various features on each frame. This process leads to a sequence of feature vectors for the complete audio file.

3.4 Tonnetz

Tonnetz also is known as Tonal Centroids consists of harmonic content present in each audio signal.It arranges all sounds according to the pitch relationships into inter-dependent spatial and temporal structures.

Motifs characterizing chords, melody, keys, largely depends on understanding these structures. Tonnetz can make changes in a special relationship that exists between octave intervals.

3.5 Root Mean Square

RMS describes the average power output of the speakers over a long period. RMS is an accurate measurement of the power output of an audio signal. (4) is RMS formula.

$$\sqrt{\frac{\sum_{i=1}^{n} a_i^2}{n}} \qquad (4)$$

3.6 Contrast

Contrast Analysis analyses two different selections in an audio track to determine the RMS difference between foreground (speech) and background such as noise.

Each frame of a spectrogram S is divided into sub-bands. For each sub-band, the energy contrast is estimated by comparing the mean energy in the top quantile (peak energy) to that of the bottom quantile (valley energy).

High contrast values are clear, narrow-band signals, while low contrast values correspond to Noise.

3.7 Mel Spectrogram Frequency

It is the representation of the short-term power spectrum of a sound signal, based on a linear cosine transform of a log power spectrum on a nonlinear Mel-scale of frequency.It's the cepstral representation of the audio clip.

The difference between the cepstrum and the mel-frequency cepstrum is the frequency bands that are equally spaced on the mel scale of MFCs, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping allows a better representation of sound. For example, in audio compression.

4. METHODOLOGY

We aim to make a novel methodology that can predict emotion without compromising accuracy. So, there are several ways we can achieve it. We have classifiers like Support Vector Classifier (SVC), Random ForestClassifier, Gradient Boosting Classifier, K Neighbour Classifier, MLP Classifier, Bagging Classifier, Recurrent Neural Networks (RNN).
According to the study for audio, RNN and MLP Classifier gave the best results for speech synthesis. The first choice for speech emotion analysis is RNN but we have decided to go with MLP Classifierin this paper because we want to see how MLP performs and the results are surprising.Figure 2 is the flow chart of the process we have done.

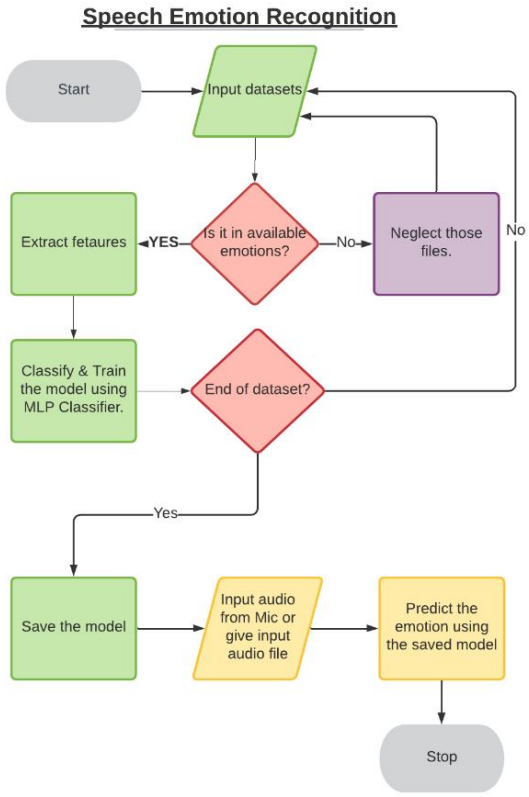

Figure 2: Flow chart of the methodology

We have taken RAVDESS and TESS databases as our inputs and we are only taking the files of angry, sad, neutral and happy and we are not processing other emotion files. Before processing the audio files (.wav file) to mono and set the sample rate to 16000Hz which means that it samples an average of 16000 times per second to make a discrete signal from a continuous signal[4],[5].

So, to observe the frequency pattern of each emotion file of the database, we have to extract these 7 spectral features using the librosa python package.The audio files should be in the mono format and have a ".wav" extension.
Zero-Crossing Rate (zcr), Contrast (contrast), Chroma (chroma), MFCC (mfcc), Tonnetz (tonnetz), Root Mean Square (rms), MEL Spectrogram Frequency (mel).

Among them, for the first 5 features, we are applying "mean" for optimized analysis. For more details refer

Introduction part. For each feature, there are other sub-features. The mean values then stacked to the result. Based on the results, MLP classifier trains itself.

>Zero-Crossing Rate (ZCR) - 1
> Contrast - 7
> Chroma - 12
> MFCC - 40
>Tonnetz - 6
> Root Mean Square - 1
> Mel Spectrogram Frequency - 128

Together we are extracting 195features.

MLP vs RNN **:** Given a sequence of length T, an RNN is recursively (hence the name) unrolled T times, once for each time instance, and the copies are connected through the memory cell (aka hidden or recurrent state) as shown in the figure 3.

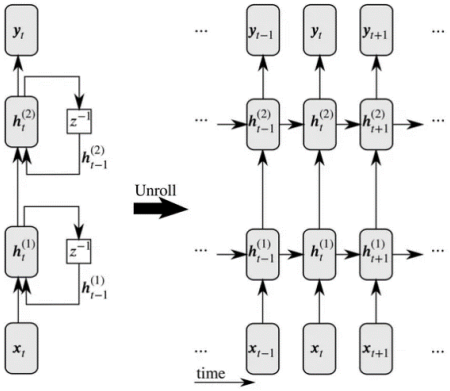

Figure 3: RNN memory cell representation

In contrast, the input layer to an MLP's has a fixed size and cannot handle variable-length inputs.

Also, they have differences in parameter tying. When the RNN is unrolled into those T copies, the sameparameters are used for all the copies[9]. This can greatly reduce the number of parameters compared to MLP. To make the parameter sharing obvious, sometimes we draw RNNs [11]as shown in figure 4.

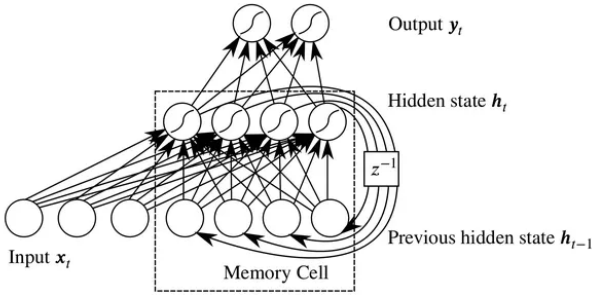

Figure 4: TypicalRNN diagram

The important thing to note is that the same network, with the same connection weights, is used for every time step[2],[3]. The only thing that varies over time is the inputs, the outputs, and the hidden state passed forwards from the previous time step. Finally, an MLP taking the entire sequence would be able to "see" the entire raw sequence at once. In contrast, subsequent time steps are only connected through the hidden states in RNNs. The memory cell usually has size much smaller than the entire input sequence, so the RNN has to learn a representation that fits in the memory cell and summarizes the semantically meaningful aspects of the sequence up until that point.These differences make RNNs a useful tool for mapping variable-length input sequences into a fixed-length output which summarizes meaningful information about the input, and the reduction in the number of parameters due to parameter sharing makes training tractable. Unfortunately, the lack of a global view of the entire input also contributes to criticisms that RNNs have difficulty capturing global structures and long-range dependencies.

Now, for a multi-layer perception (MLP) feeding from the *n*time steps. It will contain *n*hidden units feeding from a concatenated new input vector.

The dimensionality of *x*^is *n* times the dimensionality of the normal *x* single time step input. So, you need *n* hidden units for *n* time steps and thus *n* independent weight vectors for each hidden unit. The weight vectors should have the dimensionality of *x*^but with the majority of the weights set to zero for the other time steps.

So, in conclusion, both have advantages and disadvantages. MLP can have an unnecessarily large number of parameters to optimize if the number of time steps *n*is large while the RNN will have just the same number of parameters, making the MLP cumbersome to work with. Also, the RNN design means that it can handle variable time steps. Now for the MLP, consider that you now wish to process (*n*+1) time steps, you will have to train a new MLP unfortunately.The plus side for the MLP is that if the time step *n*is sure to be constant throughout the application then training it will not suffer from exploding/vanishing gradient problem.

Now we are using **MLP Classifier**, we are observing these 195 features for each file and classifying them accordingly into 4 emotions internally. So, to reduce the system effort in every run, we are saving the model in a folder so that it can be used for further runs.

We implemented grid search over random search because it methodically creates and evaluates model in the best combination of the algorithm parameters which were specified in the grid. So, the best model can be achieved by the grid search and theirparameters are set as follows:

a. Alpha value: 0.005
b. Batch size : 256
c. Epsilon value : 1e-08
d. No. of hidden layers : 500
e. Learning rate : Adaptive
f. Max iterations : 500

We have done algorithm tuning or Hyperparameter optimization with the adaptive learning rate because it gave us the best result and learning isstopped when the performance of the model starts degrading which negatively impacts the prediction rate and accuracy.

We can test this program in 2 ways. Either by giving an audio file or taking input from the microphone. We made a program in such a way that it takes input until we speak i.e starts recording when the voice is registered and ends recording until we stop speaking. Even in this process, the silence is recorded. To remove this, we normalize audio and trim the silence parts at the start and the end then pads 0.5 seconds of blank audio at the start and the end of our recording. We did this because to make sure that our recording does not have chopping.

5. RESULTS

We have calculated 2 types of accuracies. One testing with its database files and other with custom audio files which include the samples collected from the internet and recorded through the microphone. The latter's accuracy is more important and is instances of real-world scenarios. Anyways we have included both accuracies here.

While testing the files from its database, the program divides splits the data in the ratio of 3:1 where 3 parts are for training and 1 part is for testing. We make sure there are no common files for both training and testing.

5.1 Results tested with RAVDESS database files

There are total of 672 files of angry, sad, neutral and happy emotion audio files in the RAVDESS dataset. Out of those, we have divided 504 files for training and 168 files for testing. Accuracy table is shown in table 2.

Table 2: Actual vs Predicted on RAVDESS as test data

RAVDESS Testing Data	Actual	Predicted correctly	Accuracy
Angry	48	40	83.33%
Sad	46	37	80.43%
Neutral	24	20	83.33%
Happy	50	39	78%
Total	**168**	**136**	**80.95%**

A. Results tested with the custom files

Analysis of report for each emotion in custom files and its comparison is given in table 3.

Table 3: Actual vs Predicted on custom data

Custom files	Actual	Predicted correctly	Accuracy
Angry	25	19	76%
Sad	25	20	80%
Neutral	25	17	68%
Happy	25	17	68%
Total	**100**	**73**	**73%**

From table 3 we can make a confusion matrix as shown in figure 5.

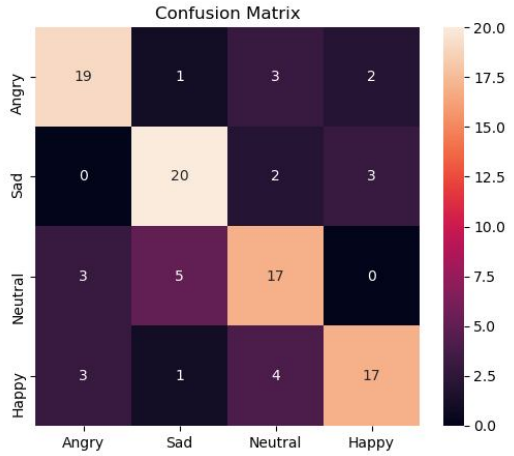

Figure 5: Confusion matrix of custom files trained with RAVDESS dataset

5.2 Results tested with TESS database files

There are a total of 1200 files of angry, sad, neutral and happy emotion audio files in the TESS dataset. Out of those, we have divided 800 files for training and 400 files for testing. Accuracy table is shown in table 4.

Table 4: Actual vs Predicted on TESS as test data

TESSTesting Data	Actual	Predicted correctly	Accuracy
Angry	112	112	100%
Sad	96	96	100%
Neutral	108	108	100%
Happy	84	84	100%
Total	**400**	**400**	**100%**

A. Results tested with custom files

Analysis of report for each emotion in custom files and its comparison is given in table 5.

Table 5: Actual vs Predicted on custom data

Custom files	Actual	Predicted correctly	Accuracy
Angry	25	20	80%
Sad	25	17	68%
Neutral	25	16	64%
Happy	25	22	88%
Total	**100**	**75**	**75%**

From table 5 we can make a confusion matrix as shown in figure 6.

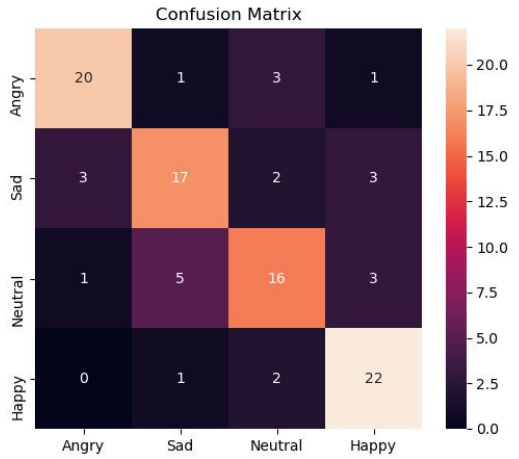

Figure 6: Confusion matrix of custom files trained with TESS dataset

5.3 Model trained with Dataset fusion (RAVDESS and TESS)

There are a total of 2272 files of angry, sad, neutral and happy emotion audio files in both the RAVDESS & TESS datasets. Out of those, we have divided 1704 files for training and 568 files for testing. Accuracy table is shown in table 6.

Table 6: Actual vs Predicted on both RAVDESS and TESS as test data

Fusion Testing Data	Actual	Predicted correctly	Accuracy
Angry	147	137	93.19%
Sad	132	121	91.66%
Neutral	131	125	95.41%
Happy	158	136	86.07%
Total	**568**	**519**	**91.37%**

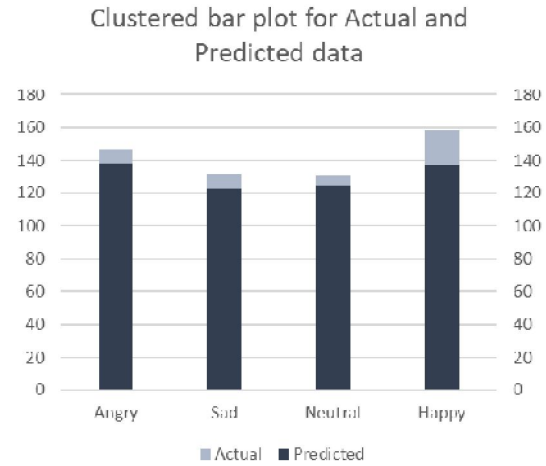

Figure 7: Clustered bar plot of Actual vs Predicted on both RAVDESS and TESS as test data

Figure 7 shows the Actual vs Predicted stats.

A. Results tested with the custom files

Analysis of report for each emotion in custom files and its comparison is given in table 7.

Table7: Actual vs Predicted on custom data

Custom files	Actual	Predicted correctly	Accuracy
Angry	25	23	92%
Sad	25	21	84%
Neutral	25	18	72%
Happy	25	21	84%
Total	**100**	**83**	**83%**

From table 7 we can make a confusion matrix as shown in figure 8.

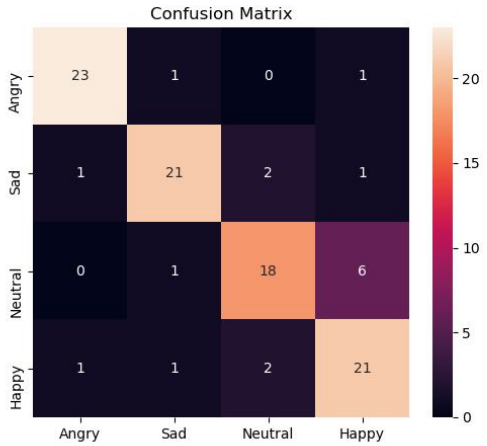

Figure 8: Confusion matrix of custom files trained with Dataset fusion (RAVDESS and TESS combined)

The accuracy is approximately 83% for the custom audio files (both recorded from the mic and files) and should be the same for real-world scenarios too.

5.4 Comparative analysis of results

When we compare all the results and the compare accuracies of the 3 types i.e. RAVDESS, TESS and Dataset fusion trained models (Refer table 8 and figure 9), we observe that among the emotions, neutral emotion prediction is less accurate. It may be due to the training data. Neutral audio files are mostly predicted as "Happy". This may be due to external factors like recording environment, mic quality, etc. which can affect the prediction.

Table 8: Accuracy comparisons

Custom files	RAVDESS	TESS	Data Fusion
Angry	76%	80%	92%
Sad	80%	68%	84%
Neutral	68%	64%	72%
Happy	68%	88%	84%

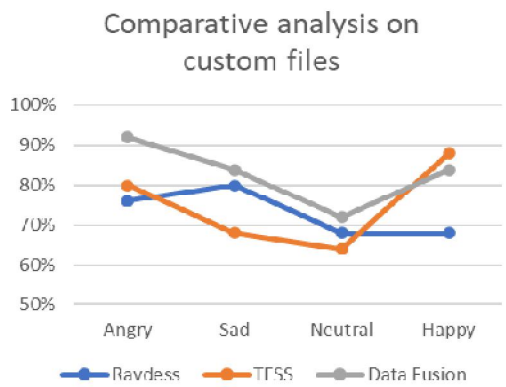

Figure 9: Comparative analysis of custom files

But our MLP classifier model works as expected and it creates less load on CPU with even 500 hidden neural networks. The estimated training time is from 45 seconds to 3 minutes (depends on CPU threads and IPC). So, it can run not only inthe server but also in the user (local) PC effectively.

6. CONCLUSION AND FUTURE SCOPE

The methodology that we have used can predict the emotion instantaneously because we have saved the trained model and is being used whenever the inputs are passed. It gotan accuracy of about 73% and 75% when trained with the RAVDESS and TESS datasets individually. Upon fusing them, we got an accuracy of 83% which is a significant increase from the previous prediction rate.We still have scope to make it even better. We can polish our model even more by giving several other databases. Not only 4 emotions, but we can also expand this methodology to predict 10+ emotions with few changes in the program. Since we have used 2000+ files for training, still we need more files to optimize this program. There is a lot of scope in the future.We can implement this methodology in various applications like automated systems, virtual assistants, etc.

Generally, digital assistants record audio and they analyze the queries based on several factors like emotion, previous query, and current situation to reply accurately. Emotion plays an important role in determining the true basis of the query. So, in such systems, our model can be implemented to predict emotion from audio. It can be applied in various fields - the possibilities are endless.

ACKNOWLEDGMENT

The success in this research would not have been possible but for the timely help and guidance rendered by many people at K L University. We express our sincere thanks to all those who have assisted us in one way or the other for the completion of our research.

EDITORIAL POLICY

The author and all co-authors had given valuable time and had worked hard to complete this research on their own. We own the rights of this research work. If anybody is interested to use this content/work we encourage them to cite this paper and give credits.

REFERENCES

1. AnvarjonTursunov ,Soonil Kwon , and Hee-Suk Pang **Discriminating Emotions in the Valence Dimension from Speech Using Timbre, Features.** *Applied Sciences*, 9(12):2470, June 2019. https://doi.org/10.3390/app9122470
2. ZhengweiHuangy, Ming Dongz, Qirong Maoy, Yongzhao Zhany. **Speech Emotion Recognition Using CNN in** *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 801–804, November 2014. https://doi.org/10.1145/2647868.2654984
3. W. Lim, D. Jang and T. Lee. **Speech emotion recognition using convolutional and Recurrent Neural Networks**, *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju*, pp. 1-4, December 2016. https://doi.org/10.1109/APSIPA.2016.7820699
4. **Emotion Detection Motion from Speech Computer Science Tripos Part II** Gonville & Caius College 2009-2010.
5. Kolla, B.P., Dorairangaswamy, M.A. & Rajaraman, A. 2010, **A neuron model for documents containing multilingual Indian texts**, *International Conference on Computer and Communication Technology*, ICCCT-2010, pp. 451, 2010. https://doi.org/10.1109/ICCCT.2010.5640489
6. Kolla, B.P. & Raman, A.R. **Data Engineered Content Extraction Studies for Indian Web Pages, Advances in Intelligent Systems and Computing**, Volume 711, 2019, Pages 505-512, 2019. https://doi.org/10.1007/978-981-10-8055-5_45
7. Swarna Kuchibhotla, HD Vankayalapati and KR Anne, **Anoptimal two stage feature selection for speech emotion recognition using acoustic features**. *International Journal of Speech Technology*, pages 1-11, 2016. https://doi.org/10.1007/s10772-016-9358-0
8. Swarna Kuchibhotla, Niranjan M.S.R. **Emotional classification of acoustic information with optimal feature subset selection methods**, *International Journal of Engineering & Technology*, Vol 7, No 2.32, Special Issue 32, 2018. https://doi.org/10.14419/ijet.v7i2.32.13521
9. Anila, M. & Pradeepini, G. **Study of prediction algorithms for selecting appropriate classifier in machine learning**, *Journal of Advanced Research in Dynamical and Control Systems*, vol. 9, no. Special Issue 18, pp. 257-268, 2017.
10. El Ayadi, M. M., Kamel, M. S., &Karray, F. **Speech emotion recognition using gaussian mixture vector autoregressive models. In: Acoustics, Speech and Signal Processing**. *IEEE International Conference on, IEEE*, vol. 4, pp. IV-957, 2007. https://doi.org/10.1109/ICASSP.2007.367230
11. Ji, S., & Ye, J. **Generalized linear discriminant analysis: A unified framework and efficient model selection**. *IEEE Transactions on Neural Networks*, 19(10), 1768–1782, 2008. https://doi.org/10.1109/TNN.2008.2002078