# Neural Network Modelling of Speech Emotion Detection

*Y.* Sri Lalitha[1], *Althaf Hussain* Basha Sk.[2], *M. V.* Aditya Nag [3]

[1]Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India
[2] Department of CSE, Chalapathi Institute of Engineering and Technology, Guntur, Andhra Pradesh, India.
[3]Department of Mechanical Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana, India

**Abstract**: In making the Machines Intelligent, and enable them to work as human, Speech recognition is one of the most essential requirement. Human Language conveys various types of information such as the energy, pitch, loudness, rhythm etc., in the sound, the speech and its context such as gender, age and the emotion. Identifying the emotion from a speech pattern is a challenging task and the most useful solution especially in the era of widely developing speech recognition systems with digital assistants. Digital assistants like Bixby, Blackberry assistant are building products that consist of emotion identification and reply the user in step with user point of view. The objective of this work is to improve the accuracy of the speech emotion prediction using deep learning models. Our work experiments with the MLP and CNN classification models on three benchmark datasets with 5700 speech files of 7 emotion categories. The proposed model showed improved accuracy.

**Keywords—**Supervised Learning, Speech Recognition, CNN Classification.

## 1 Introduction

In making the Machines Intelligent, and enabling them to work as human, Speech recognition is one of the essential requirements. Understanding ones Emotions and responding suitably in a human - computer, conversations makes machines more reliable. Determining efficient techniques to identify the emotions in the speech signal has a variety of applications. As we have been using many computer applications in our day-to-day life, recognizing the emotion has a significant influence and has become a demand from markets to medical management. Emotion detection is used in medical field which helps in spotting mental issues by determining Patients Speech patterns[13], in business marketing understanding customer's requirements, enables customized promotion of the products, and in E-Commerce sites such as Amazon or Flipkart, to know the customer feedback of a product need efficient speech emotion recognition systems. Identifying emotion is a challenging work, because emotions are subjective, individuals would draw out them differently. The complexity of SER also includes various other factors such as language, pitch, energy, loudness, rhythm etc, in the sound signal, along with the context such as gender, age, words, time duration of a signal and emotion, all of these will have an influence on the kind of emotion we are determining.

Although there exists wide variety of Probabilistic and Machine Learning techniques such as Hidden Markov Models (HMM), Support Vector Machines (SVM), Gaussian Mixture Models (GMM) in literature that exhibited around 70% of accuracy. Some of the studies proposed earlier showed better results with deep learning models.

## 2 Literature Survey

In the field of Speech Recognition lot of research contributions are available using Machine Learning, Soft Computing, Neural Networks and a combination of methods. In the work proposed in [9], Machine Learning and NN techniques Random Forest, SVM, Naïve Bayes[, KNN, Multi-layer Perception and Logistic Regression methods are combined to determine the emotion. The NN, SVM and KNN methods are joined in [10, 16, 17]. In [11, 12] Deep Belief Networks (DBN) are applied to automatically extract emotional features. Deep belief networks (DBN) can automatically discover the multiple levels of representations in speech signals.

In [1] Random deep belief networks (RDBN) method is proposed where in low level features of the input signal are extracted and applied them to form random subspaces. Each random sub-space is then input to DBN to generate high level features. These features are input to a classifier to determine the Emotion.

In [2] speech and video data are considered together as an input parameter to determine the emotions. The

authors used two datasets. One dataset is large which constituted Video and Speech input signals and an eNTERFACE database. Firstly from the given voice input signal features, they obtained Mel-spectrogram, which was an imaged based spectrum with various signals as parameters. This image spectrum is then input to 2d CNN followed by Ensemble classification models to determine the outputs. Likewise video frames are extracted from video signals; prominent video features are then input to 3d CNN followed by RDBN to determine the fusion of scores. This output is then fed to the final classification modes SVM to determine the emotion label.

In [3] ANN is applied for emotion recognition. Here a High Pass Filter was designed to filter out noise, unwanted signals from speech, and considered high frequency signals. After reading all the frequencies it passes only the frequency which is high in all the reading values for feature extraction process. Using Mfccs feature extraction and ANN classifier they have determined the final emotion for an input speech signal.

In [4] the Direct Modelling of Speech Emotion from Raw Speech was proposed in [4], where parallel Convolution layers with different filters width and LSTM models were combined to determine the emotion, Raw speech signal is processed to extract the features, applied CNN with max-pooling and extracted multiple filter lengths. These are input to LSTM with fully connected layer using softmax activation layer for classification.

The work in [14, 15] introduced Wavenet and DNN based Probabilistic feature extraction methods to extract features for emotion detection.

Scope of this work is to improve the performance of SER systems using DNN to extract the quality features from the audio, including the emotions related to it and deploy it as a Webapp.

# 3 Methodology

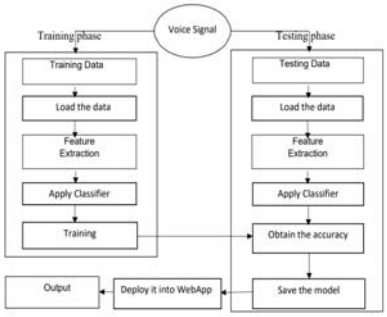


**Fig. 1.** Proposed system

## 3.1 Datasets

We have considered three different datasets namely savee, tess and ravedess. Each of these datasets consist of different persons with various emotions. The dataset is taken from Kaggle. We have considered 5732 speech files. It includes 7 emotions or features like calm, neutral, happy, sad, fearful, disgusting and angry.

## 3.2. Algorithms

### *3.2.1 Multilayer perceptron (Mlp) classifier*

MLP comes under supervised classification that makes utilizes backpropagation. It belongs to the class of artificial neural networks (ANN). It is made up of many perceptrons. It comprises of one output layer, one input layer, and an indefinite number of hidden layers between these input and output layers, depending on the user's requirements. That is, at least three levels should be present: input layer, concealed layer, and output layer. It can identify data that is not linearly separable due to its nonlinear activation.

### *3.2.2. Convolutional Neural Network (CNN) classifier*

It is a subdivision of deep learning approaches for classification that rely entirely on feed-forward architecture. This algorithm is generally used for identifying jobs because they improve data classification. The intake information is processed in shape of responsive province by networks, which has tiny neurons on each layer of the selected model design. CNN consists of three layers namely: input layer, hidden layers, and an output layer The middle layers are known as Hidden layers. Before the final convolution the activation module covers the inputs and outputs. A CNN typically consists of a multiplication or other dot product layer with a ReLU activation function as its activation function. After this layer there are many other layers such as pooling layers.

## 3.3. Description of Various Module

### *3.3.1 Librosa:*

Is a package that is employed and used for music and audio analysis. It uses sound file and audio browse. It primarily provides the significant building block to make audio, music knowledge retrieval systems. At present, the sound file doesn't support MP3, and it'll cause the library to crash. Hence, librosa uses the audio read modules to figure with files like MP3.

### *3.3.2 Keras*

Keras is an open-source neural network library which will operate for high Theano or TensorFlow and is extremely helpful for operating with high-level neural networks in-built Python. It supports convolutional and continual neural networks, in addition as combos of the 2. It's varies in range of neural network building elements, like optimizers, activation functions, and layers to form it rather more easier to figure at the side of image, audio in addition as text and knowledge to

form it easier write deep neural network code. It developed with a spotlight on simplifying models and conjointly enabling quicker experimentation.

### *3.2.3. TensorFlow*

TensorFlow is an open-source toolkit for numerical and large-scale machine learning and conjointly deep learning calculations. It makes use of Python to grant a straightforward front-end API for developing apps, which will be executed in high performance C++. Recurrent neural networks, deep neural networks, and natural language processing tasks could all be trained and operated with TensorFlow. It assists within the creation of dataflow graphs, that represent however knowledge flows across a graph or a series of processing nodes.

### *3.3.4. Sklearn*

Sklearn is a Python machine learning package that is free and open-source machines, Decision Tree Classifier, support vector machines svm neural networks, k-means, and other classification, regression, and clustering methods are supported in Sklearn. It's built to work with NumPy and SciPy, which are two numeric and also scientific libraries.

## 4. Procedure

The training phase and the testing phase. The training phase is on the left, while the testing phase is on the right. To train and test our model, we must first divide our dataset into two steps. After partitioning our dataset, we must load it and perform two processes: extracting the dataset's features and then performing various classifiers to the system to identify the correct emotion from the given input voice audio signal. We must check the correctness of our model after we have completed training and testing by doing feature extraction and then applying the classifier.

## 5. Results

The accuracies of the two models MLP and CNN for the combined three datasets.

**Table 1.** Accuracy of Models

| Accuracy | MLP Model | CNN Model |
|---|---|---|
| Combination of 3 Datasets | 86.41 | 89.01 |

Correlation between the voice inputs in knowledge set helps to search out how strong or weak they're co-related to every alternative within the set. The numeric values from high to low represents vary of the inputs. the high value represents the positive relationship exists between the variables. The numeric low values i.e. zero with darker color offers the more negative relationship between the input voice variables. The below figure 2-3 is the projected model of Mlp.

```
from sklearn.metrics import confusion_matrix
matrix = confusion_matrix(y_test,y_pred)
df_matrix=pd.DataFrame(matrix)
sns.heatmap(df_matrix,annot=True,fmt='')
plt.title('Confusion Matrix',size=25)
plt.xlabel('Predicted Labels',size=14)
plt.ylabel('Actual Labels',size=14)
plt.show()
```

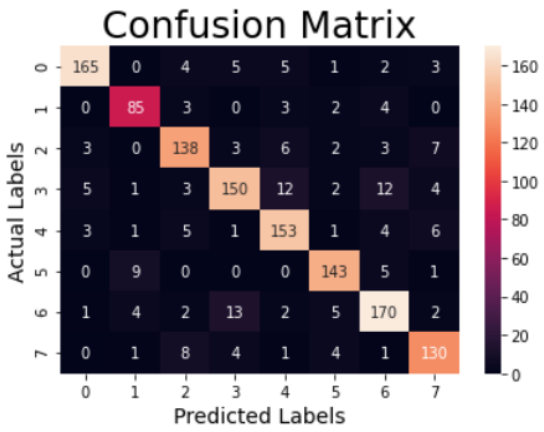


**Fig. 2.** MLP confusion matrix

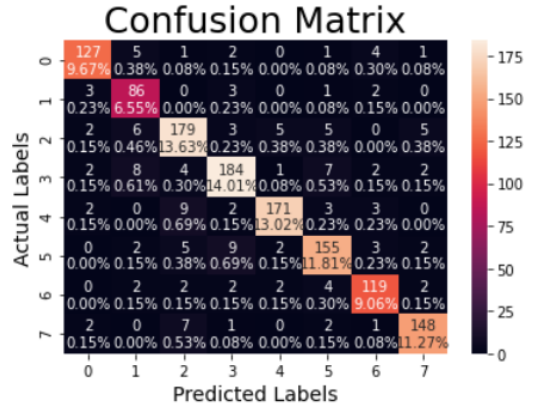


**Fig. 3.** CNN model Confusion Matrix

Prediction of emotion by combining all the three datasets in the MLP model. In MLP model we train the entire data set to the algorithm and extract each feature from the dataset. The gives the prediction of all instances in the data set at a time.

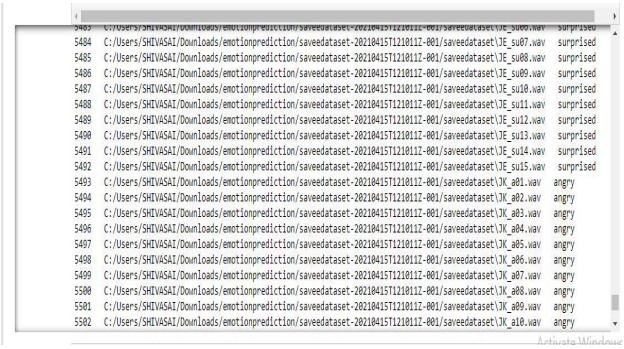


**Fig. 4.** MLP Emotion Prediction Snapshot

Prediction of emotion by combining all the three datasets in the CNN model. After training the model, to test the prediction of the audio we need to send the input to the testing part in the CNN model. Fig 5. shows the prediction of one audio input voice. In Fig 5 we have given a voice to the CNN testing code, and it predicted the emotion as neutral.

```
import keras
import numpy as np
import librosa

class livePredictions:
    """
    Main class of the application.
    """

    def __init__(self, path, file):
        """
        Init method is used to initialize the main parameters.
        """
        self.path = path
        self.file = file

    def load_model(self):
        """
        Method to load the chosen model.
        :param path: path to your h5 model.
        :return: summary of the model with the .summary() function.
        """
        self.loaded_model = keras.models.load_model(self.path)
        return self.loaded_model.summary()
```

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_3 (Conv1D)            (None, 40, 64)            384
_________________________________________________________________
activation_4 (Activation)    (None, 40, 64)            0
_________________________________________________________________
dropout_3 (Dropout)          (None, 40, 64)            0
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 10, 64)            0
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 10, 128)           41088
_________________________________________________________________
activation_5 (Activation)    (None, 10, 128)           0
_________________________________________________________________
dropout_4 (Dropout)          (None, 10, 128)           0
_________________________________________________________________
max_pooling1d_3 (MaxPooling1 (None, 2, 128)            0
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 2, 256)            164096
_________________________________________________________________
activation_6 (Activation)    (None, 2, 256)            0
_________________________________________________________________
dropout_5 (Dropout)          (None, 2, 256)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 8)                 4104
_________________________________________________________________
activation_7 (Activation)    (None, 8)                 0
=================================================================
Total params: 209,672
Trainable params: 209,672
Non-trainable params: 0
_________________________________________________________________
Prediction is   neutral
```

**Fig. 5.** CNN Emotion Prediction Snapshot

The classification report where it gives the precision, recall, f1-score, and support values with avg weighted score. Fig 6. gives the MLP Classification Report with weighted accuracy score 0.86.

```
from sklearn.metrics import classification_report
report = classification_report(y_test, y_pred)
print(report)
# 0 = neutral, 1 = calm, 2 = happy, 3 = sad, 4 = angry,
# 5 = fearful, 6 = disgust, 7 = surprised
```

```
              precision    recall  f1-score   support

           0       0.92      0.90      0.91       141
           1       0.79      0.91      0.84        95
           2       0.86      0.87      0.87       205
           3       0.89      0.88      0.88       210
           4       0.94      0.90      0.92       190
           5       0.87      0.87      0.87       178
           6       0.89      0.89      0.89       133
           7       0.93      0.92      0.92       161

    accuracy                           0.89      1313
   macro avg       0.89      0.89      0.89      1313
weighted avg       0.89      0.89      0.89      1313
```

**Fig. 6.** Classification Report of MLP model

The classification report of the CNN with weighted accuracy 0.89 is depicted in the figure 7.

```
# classification report
from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

       angry       0.93      0.89      0.91       185
        calm       0.84      0.88      0.86        97
     disgust       0.85      0.85      0.85       162
     fearful       0.85      0.79      0.82       189
       happy       0.84      0.88      0.86       174
     neutral       0.89      0.91      0.90       158
         sad       0.85      0.85      0.85       199
   surprised       0.85      0.87      0.86       149

    accuracy                           0.86      1313
   macro avg       0.86      0.87      0.86      1313
weighted avg       0.86      0.86      0.86      1313
```

**Fig. 7**. Classification Report of CNN model

From the data which combined the three data into single unit we counted the number of emotions range the emotions where happy, sad, angry, disgust, clam, neutral, fearful. The range is from 0 to 800 ,we took total 5725 voices as input is shown in fig 8.


**Fig. 8.** Count of Emotion

## 6. Conclusions and Future Work

In this research work, we took voice as an input parameter and detected emotions. For detecting feelings, we used MLP classifier and also CNN set of rules strategies. By using Savee dataset, Tess dataset, Ravdess dataset we trained our models. At last, to boom the training data we've mixed most of these 3 datasets into one. This can further be deployed into any internet site like a customer service internet site or different web sites in which they need to become aware of their feelings and act or reply accordingly. We also want to boom education information via way of means of including a few greater. We also want to improve our model accuracy, so we need to attempt a few different architectures on combined datasets. we ought to make certain that the new datasets we will add in future must be same as the previous datasets we introduced. Additionally to boom the education information, we will carry out a few augmentation strategies. Now we just took voice as an input parameter but in future we try and hit upon feelings via way of means of the use by taking image, text, video, and voice as input parameters.

## References

1. G. Wen, H. Li, J. Huang, D. Li, E. Xun, "*Random Deep Belief Networks for Recognizing Emotions from Speech Signals*", Computational Intelligence and Neuroscience, **2017**, 9 (2017).
2. M. S. Hossain, G. Muhammad, "*Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data,*" Inf. Fus. **49**, 10, (2019).
3. Pawan Kumar Mishra, Arti Rawat, "Emotion Recognition through Speech Using Neural Network", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 5, Issue 5, pp. 422-428, May 2015.
4. S. Latif et al. "*Direct Modelling of Speech Emotion from Raw Speech*", Proc. I. S., **2019**, 5 (2019).
5. M. Xu et al. "*Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation*", arXiv:2102.01813, (2021).
6. S.R. Livingstone, F.A. Russo, "*The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English"*. PLoS ONE **13,** 5, (2018).
7. P. Fuller, M. Kathleen, Dupuis and Kate, , "Toronto Emotional Speech set (TESS)", Scholars Portal Data verse, Version 1.0. https://doi.org/10.5683/SP2/E8H2MF, (2020)
8. Survey Audio-Visual Expressed Emotion (SAVEE) Database (http://kahlan.eps.surrey.ac.uk/savee/Download.html )
9. B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "*Speaker independent speech emotion recognition by ensemble classification*," in Proc. of IEEE Int. Conf. on M.M and Expo, 4 (2005).
10. M. K. Sarker, K.M.R. Alam, M. Arifuzzaman, "*Emotion recognition from speech based on relevant feature and majority voting*,", in Proc. of the Int. Conf. Info., Elec. and Vision 5,(2014).
11. C. Huang, W. Gong, W. Fu, and D. Feng, "*A research of speech emotion recognition based on deep belief network and SVM*", Math. Prob. Engg., **2014**, 7, (2014).
12. C. Huang, W. Gong, W. Fu, and D. Feng, "*A research of speech emotion recognition based on deep belief network and SVM*," Mathematical Problems in Engineering, **2014**, 7 (2014).
13. Y. Sri Lalitha, et. al., "*Efficient Tumor Detection in MRI Brain Images*", Intl. J. O. B. M. Engg., 16, 9, (2020).
14. Hemanta Kumar Palo Mihir Narayan Mohanty, "*Wavelet based feature combination for recognition of emotions*", in A. Shams Engg J. **9**, 7 (2018).
15. A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, "*WaveNet: a generative model for raw audio*", in 9th ISCA Speech Synthesis Workshop 10 (2016).
16. Y. Sri Lalitha, Dr. A. Govardhan, "Semantic Framework for Text Clustering with Neighbor", Proc. of 48th Ann. Conv.of CSI, A.I.S.C., Springer **2**,10, (2013).
17. Y.J. Nagendra Kumar, B.Mani Sai, V. Shailaja, S.Renuka, P.Bharathi,"*Python NLTK Sentiment Inspection using Naive Bayes Classifier*", IJRTE, **8**,4 (2019).
18. B. Dhanalaxmi, G. Apparao Naidu, K. Anuradha, "*Adaptive PSO based association rule mining technique for software defect classification using ANN*", Procedia Computer Science, **46**,11, (2015).