Speech Emotion Recognition using MLPClassifier

[1]Manikant, [2]Vipin Kumar Yadav, [3]Sudhanshu Gupta, [4]Saket Kumar, [5]Mr. Yuvaraj Subramanian, M.E. (Ph.D.)

[1,2,3,4]Students, [5]Assistant Professor
Computer Science and Engineering
Sri Eshwar College of Engineering
Coimbatore, India.

***Abstract*: Language is the most important communication medium for human beings and speech is the primary medium of communication. Emotion plays an important role in social interaction. Identifying emotions in speech is very important and challenging because here we are dealing with human-machine interactions. Emotions vary: the same person has different feelings from person to person, and everyone expresses it together in different ways. When a person expresses their emotions, everyone has a different energy, the pitch and tone variation of different elements are grouped together. Therefore Speech Emotion Recognition is the future goal of Computer Vision. The goal of our project is to develop a Conventional Neural Network based on the Smart Emotion Recognition Speech. It uses a variety of modules for emotional recognition and classification is used to differentiate emotions such as Happy Sad Angry Surprise. The machine rapidly converts human speech signals and processes its routine and finally it displays emotion. Data Speech Sample and Features Extracted from the speech sample using the Librosa package. We are using the RAVDESS dataset which we use as an experimental dataset. This study shows that all classifications for our dataset achieve 87% accuracy.**

***Keywords*: Emotion, RAVDESS Dataset, Speech Emotion Recognition, MLPClassifier.**

I. INTRODUCTION

In natural human-computer interaction (HCI), speech Emotion recognition (SER) is on the rise Important in various applications. Currently, Speech Sense Identity is an artificially evolving crossing area Intelligence and artificial psychology; In addition, it is popular Signal processing and sample identification research topics. Research is widely applied in human-computer interaction, Interactive teaching, entertainment, security area, etc. Speech Emotion Processing and Recognition System Usually consists of three parts, the first is speech Signal acquisition, followed by feature extraction By emotional recognition. The technology that best suits for Speech recognition is a neural network-based approach. Artificial neural networks, (ANN) are biologically motivated Tools for information processing. Speech Recognition Modeling Prerequisites through Artificial Neural Networks (ANNs) are not required. Speech process and knowledge of this technique quickly Has become an attractive alternative to HMM. RNN can learn Sublunar Relation of Speech - Data and Capacity Modeling time based tone. Conventional nerve There are several types of multi-layer receptor (MLP) networks It is rapidly becoming obsolete for speech recognition and for a variety of purposes Other Speech Processing Applications. Have speech recognition The process of converting the acoustic signal extracted by Microphone or telephone for a set of letters. They can It also serves as an input for further language processing Achieving Speech Comprehension is a topic covered in the section. As We know that speech recognition performs similar functions with The human brain.

II. MLP NEURAL NETWORK MODEL

Of the many neural networks available, feed forward networks monitored by Multi Layer Perceptron (MLP) back propagation are the most popular and most widely used. This class of network uses a gradient decent algorithm that reduces the average square error in neuron output. Training the neural network is a non-linear optimization problem in which the goal is to create a set of weights that reduce cost performance. The cast function, commonly known as the network mapping error function, describes the surface in weight space, often referred to as the error surface. Training algorithms can be seen as methods for finding this surface minimum. The complexity of the search is maintained by the nature of the surface. Typically, a network consists of an input layer, an output layer, and at least one hidden layer. MLP has been shown to enforce arbitrary convex decision boundaries with a hidden layer. Furthermore, it has been shown that this type of network can approximate any non-linear function.

A. Neural Network Initialization

Weight initiating is widely recognized as one of the most effective ways to speed up the training of neural networks. Based on the Kauchi inequality, the network initialization method and the linear algebraic (minimum-squares) method [20] were implemented in our work. When neurons with 11 activation functions have 1 or -1 output, the derivative of the activation function evaluated by this value is zero. Therefore, despite the difference between the target value and the actual output, there is no change in weight. The initial method ensures that the outputs of the neurons are in the active region and increases the convergence rate.

B. Neural Network Training

The training algorithm is gradient decent with positive learning rate backpropagation. The learning rate is related to panic and varies according to the speed of training. Inputs Of Neural Network needs to be normalized to have a range between -1 and +1. In addition, the output of the neural network is converted to the original level.

III. PROPOSED SYSTEM

A. Neural networks

Neural networks are a set of algorithms that are loosely designed. The human brain, which is designed to recognize patterns. The patterns they identify are numeric, in vectors, Contains images, sound, text or all real-world data The time series must be translated. It helps us to cluster and classify. You can think of them as a clustering and taxonomic layer Above is the raw data that you store and maintain. They help Assemble the unlabeled data according to the similarities between Example input, and classify the data when they have a Datasets are labeled for training. Neural networks have evolved into a Fascinating sound modeling approaches at the end of ASR 1980s. Since then, neural networks have been used by many Elements of speech recognition, such as phoneme classification Isolated word recognition, audio-visual speech recognition. Make less explicit assumptions about neural networks HMM and has statistical features compared to most of them Features make them attractive identification patterns for speech Identity.

B. Deep feedforward and recurrent neural networks

The deep feed-forward neural network is an artificial nerve A network with multiple hidden layers of units between inputs And output layers. The DNN complex can model non-linear Relationships. Whose structures prepare composition patterns, Additional layers here start the composition of features from the bottom Layers, providing massive learning potential and thereby potential Created complex models of speech data. One The basic principle of deep learning is removal To use hand-crafted feature engineering and raw features. This This principle was first successfully explored in architecture Deep auto-encoder on "Raw" spectrogram or linear filter bank features showing its superiority over Mel-septrol Features that have specific stages of change From the spectrogram. The true "raw" features of speech, Waves, recently shown to produce Excellent large-scale speech recognition results.

C. Mel-Frequency Cepstral Coefficients (MFCC)

One of the most popular audio features is the male-frequency septic coefficient (MFCC). This is a representation of the speech signals that characterize the window's sepstream.

The short-time signal is taken from the FFT of that signal. The signal is then converted to the frequency axis of the melfrequency scale using a log based transformer, which is then decorated with a modified isolated cosine. The steps to capture MFCC features include pre-emphasis, frame blocking and Windows, FFT size, mail filter, bank, log energy and DCT. MFCC uses the Mel-scale, which is designed for the frequency response of the human ear. Because of this, MFCC has proven to be invaluable in the field of speech recognition and efforts have been made to integrate it with emotion recognition. According to spectral audio features e.g. MFCC is best suited for n-way classification.

D. Multilayer Perceptrons Classifier (MLP Classifier)

Work done with multilayer perceptrans has shown that they have the ability to predict the XOR operator as well as many other non-linear functions. Multilevel comprehension is often applied to supervised learning problems. They train on a set of input-output pairs and learn to model the correlation (or dependence) between those inputs and outputs. The network therefore has a simple description as the input-output model, weights and thresholds (bias) are the free parameters of the model. Important issues in MLP design include the number of hidden layers and the specification of the number of units in these layers. Number of hidden units

Usage is not very clear. It is advisable to use the hidden layer as any starting point with a number of units equal to half the number of input and output units.

IV. SPEECH EMOTION RECOGNITION USING MLP CLASSIFIER

Audio files are input into the Speech Emotion Recognition System (SER). The data set travels through several blocks of processes that can be executed to assist in the analysis of speech parameters. The data is pre-processed for conversion to the appropriate format and the features related to the audio files are extracted using various steps such as framing, hammering, window making etc. This process helps to break down audio files.

Numeric values that indicate frequency, time, amplitude, or other parameters that aid in the analysis of audio files. After the necessary extraction Features from audio files, model trained. We have used RAVDESS and TESS dataset of audio files containing the speeches of 26 people with various parameters. For training, We store numerical values of emotions and their related properties in separate ranges. These ranges are given as inputs to the MLP classification Enabled. The classifier identifies the different categories in the dataset and classifies them as different emotions. The model is now able to understand the range of values of speech parameters that fall into specific emotions. To test the model performance, if we enter an unknown test dataset as input, it retrieves the parameters and evaluates the sentiment according to the training dataset values. System accuracy is displayed as a percentage, which is the end result of our project.

V. DATASETS DESCRIPTION

We used two datasets throughout our project. One of them, the RAVDESS dataset, contains the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, which contains 1440 audio-speech files of 24 professional actors, 12 of whom are men and 12 are women, 60 of them. There are a total of 1440 audio-speech files for and 1012 audio-song files of 24 professional actors, including 12 men and 12 women, 44 trails with a total of 1012 audio-song files each. Audio speech files contain expressions such as sadness, calm, anger, joy, fear, disgust, and surprise, and audio-song files contain sadness, calm, anger, fear, and happy emotions. The second dataset we considered was the Toronto Emotional Speech Set (TESS) dataset.

In the TESS dataset, 2800 special speech files were recorded by two actresses aged 26 and 64, both of whom spoke a set of 200 target words in the carrier phrase "speak the _____ word". The two actresses are from the Toronto area, speak English as their first language, go to university and take music lessons. There are about 1401 special speech files for a 26 year old and about 1399 special speech files for a 64 year old. Combining these two speech files, we have 2800 unique speech files and each of them has seven different emotions that are anger, disgust, fear, joy, pleasant surprise, sadness and neutral.

We combined these two datasets and observed eight different emotions for our project. They are neutral, calm, happy, sad, angry, fearful, disgusted and surprised.

VI. EXPERIEMENTAL SETUP

This section describes the experimental setup and the libraries used for in-depth learning to help identify emotions.

A. SYSTEM SETUP

For testing we have used a system setup with a Core i7 6th generation 3.7 GHz processor, SSD with 512GB of memory space, NVIDIA GeForce GT 730 2GB GPU card with Ubuntu 16.04 installed. For Deep Learning I used Tensor Flow 1.5 to implement the Inception Net model and the Tensor Board to visualize learning, graphs, histograms etc.

B. TRAINING METHOD

All images labeled with relevant emotion are designed to train the model. The proposed MLPClassifer model was implemented using sklearn liberary. Spectrogram images created from IEMOCAP resized to 500 x 300. More than 400 spectrograms are generated from all the audio files in the dataset. For each emotion, approximately 500 image thresholds are collected for each class emotion from the corpus database. The training process is set to 100 batch size for 20 ages. The initial learning rate was set at 0.01 with a decrease of 0.1 after every 10 ages. The training data model is made on the Nvidia GeForce GT 730 with 2GB of onboard memory. The training took about 35 minutes and the best accuracy was achieved after 28 ages. There was a loss of 0.71 in the training set and a loss of 0.95 in the test set. 35.95% accuracy was achieved for each spectrogram. It is important to note here that the overall accuracy is very low. These may be due to transfer learning and less datasets used for each class of emotions.

VII.DISCUSSION OF RESULT

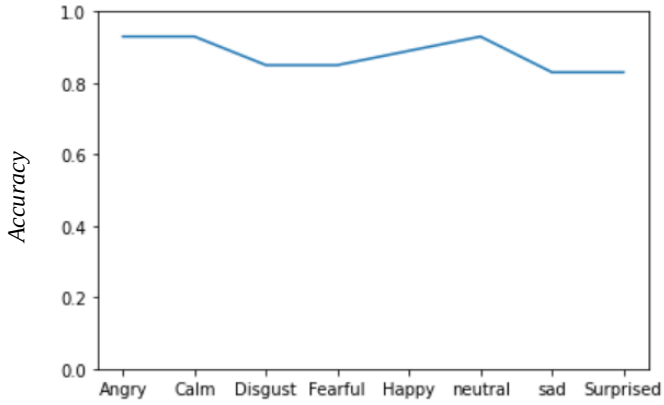
The evaluation phase's findings demonstrate the model's effectiveness when compared to baselines and the state of the art on the RAVDESS dataset. The precision, recall, and F1 values obtained for each of the emotional classifications are shown in the table. These results show that precision and recall are very well balanced, allowing us to achieve F1 values for almost all classes that are dispersed around the value 0.88. The low variability of F1 findings demonstrates the model's resilience, which successfully classifies emotions into eight separate categories.

The model is less accurate in the classes "Sad" and "Surprised," but this is not surprising because it is well recognized in the literature that these are the most difficult classes to determine not only by speech but also by observing facial expressions or examining written text.

RESULTS OF THE MODEL ON THE TEST SET PER EACH CLASS

Emotion	Precision	Recall	F1-Score	Support
Angry	0.93	0.86	0.89	188
Calm	0.93	0.79	0.85	104
Disgust	0.85	0.90	0.88	149
Fearful	0.85	0.83	0.84	185
Happy	0.89	0.87	0.88	173
neutral	0.93	0.91	0.92	165
sad	0.83	0.90	0.87	201
Surprised	0.83	0.93	0.88	148
Accuracy			0.88	1313
Macro avg	0.88	0.87	0.87	1313
Weighted avg	0.88	0.88	0.88	1313

EMOTION ACCURACY IN DATASET



VIII. CONCLUSIONS AND FUTURE SCOPE

In this project, we MLPClassifier to identify emotions by taking voice as input. We have trained and evaluated our model using two datasets namely RAVDESS Dataset, TESS Dataset. Finally, we combined these two datasets into one to enhance the training data and evaluate our model using this combined dataset.

This model can be used as a web application, it can be added to any website like customer support website or other websites where they can identify their feelings and act accordingly or respond accordingly. In the future, we would like to increase the training data by adding a few more datasets like these two datasets. However we need to make sure that the naming process in the new dataset is the same as what we added to the first two datasets and at the same time we can make some improvement methods to increase the training data. Additionally, we would like to try some other architectures in this combined dataset to improve the accuracy of our model. This proposed model could simply detect emotions using voice as an input; In addition, we want to try to identify emotion by adding video, image, text and these four video, image, text and voice as inputs so that it can be useful in applications that require acting or feedback. It is necessary to find the feeling depending on the feelings of the particular person. In the future, we would like to try multiple emotions from a single input file, because in real time, people have different emotions when speaking. Currently, our model can only detect one emotion from a given input speech signal.

REFERENCES

1. S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, “Speech-to-Text and Speech-to-Speech Summarization,” vol. 12, no. 4, pp. 401–408, 2004.
2. S. G. Koolagudi and S. R. Krothapalli, “Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features,” Int. J. Speech Technol., vol. 15, no. 4, pp. 495–511, 2012.
3. J. Rong, G. Li, and Y. P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, 2009
4. Guihua Wen, Huihui Li, Jubing Huang, Danyang Li, and Eryang Xun, “Random Deep Belief Networks for Recognizing Emotions from Speech Signals”, Computational Intelligence and Neuroscience, Volume 2017, Article ID 1945630, 9 pages, March 2017.
5. Pawan Kumar Mishra and 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.
6. Abhay Kumar et al. “Speech Mel Frequency Cepstral Coefficient feature classification using multi level support vector machine”, In 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON). IEEE, pp. 134-138, October 2017.
7. D. Bharti and P. Kukana, “A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals”, In 2020 International Conference on Smart Electronics and Communication (ICOSEC). IEEE, pp. 491-496, September 2020.
8. Mandeep Singh, Yuan Fang, “Emotion Recognition in Audio and Video Using Deep Neural Networks”, arXiv:2006.08129, June 2020.
9. D.R. Hush, B.G. Horne, "Progress in Supervised Neural Networks – What's New Since Lippmann"," IEEE Signal Processing Magazine, pp. 8-39, January 1992.
10. https://www.sciencedirect.com/science/article/pii/S111086 6512000345
11. https://en.wikipedia.org/wiki/Deep_learning
12. Awni Hannun, Ann Lee, Qjantong Xu and Ronan Collobert, Sequence to sequence speech recognition with time-depth deperable convolutions, interspeech 2019, Sep 2019.
13. K. R. Scherer, "What are emotions? And how can they be measured?", Social Sci. Inf., vol. 44, no. 4, pp. 695-729, 2005
14. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, et al., " Emotion recognition in human-computer interaction", IEEE Signal Process. Mag., vol. 18, no. 1, pp. 32-80, Jan. 2001.
15. A. D. Dileep and C. C. Sekhar, " GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines", IEEE Trans. neural Netw. Learn. Syst., vol. 25, no. 8, pp. 1421-1432, Aug. 2014.
16. M. S. Hossain and G. Muhammad, " Emotion recognition using deep learning approach from audio–visual emotional big data", Inf. Fusion, vol. 49, pp. 69-78, Sep. 2019
17. A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech recognition using deep neural networks: A systematic review", IEEE Access, vol. 7, pp. 19143-19165, 2019.
18. L. Deng and D. Yu, "Deep learning: Methods and applications", Found. Trends Signal Process., vol. 7, no. 3, pp. 197-387, Jun. 2014