JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 06, 2020

Application of AI in Speech Emotion Detection

Y. Sri Lalitha[1], Althaf Hussain Basha Sk.[2], Y. Gayatri [3]

[1]Department of Information Technology, GRIET, Hyderabad, Telangana, India
[2] Department of CSE, CIET, Guntur, Andhra Pradesh, India.
[3] Department of Humanities, GRIET, Hyderabad, Telangana, India

Abstract:

In making the Machines Intelligent, and enable them to work as human, Speech recognition is one of the most essential requirements. 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 bench mark datasets with 5700 speech files of 7 emotion categories. The proposed model showed improved accuracy.

Keywords—Supervised Learning, Convolution Neural Networks Algorithm, Classification, Plots etc.,

I. 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, in business marketing understanding customer's requirements, promoting the product accordingly, and for E-Commerce sites such as Amazon or Flipkart, to know the customer feedback of a product need efficient speech emotion recognition systems. Identifying emotion has become challenging 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 and emotion, all of these will have an influence 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.

Detection from speech focuses on determining the emotion of certain person from an audio sample. The focus of this work is to interpret the general communication by combining different voice files instead of using traditional equipment to understand words and promote the audience's response accordingly. These audios consist of different parameters like age, gender, emotion. It is important to figure out the meaning of the content and depict the emotion it fetches. Emotion recognition is the procedure of examining the present situation of individual from his/her voice. We have considered many parameters like energy, Pitch, Rhythm, Loudness under sound signal.

II. LITERATURE SURVEY

[1] **Random Deep Belief Networks for Recognizing Emotions from Speech Signals** :"Random Deep Belief Networks for Recognizing Emotions from Speech Signals" was proposed by Eryang Xun, , Huihui Li , Guihua Wen, Jubing Huang and Danyang Li proposed, For recognizing feelings from the speech signals it talks about the learning technique of the Random Deep Belief Networks (RDBN) method. They applied the method of the Random subspaces before extracting low level features of given input speech signal parameter. Each Random

ISSN- 2394-5125 VOL 7, ISSUE 06, 2020

subspace is fed into the input of DBN for extracting higher level features from input speech signal parameter and provided these higher-level features as the input to the base classifier to output a predicted emotion label.

[2] **Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data** "Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data" was proposed by M. Shamim Hossain and Ghulam Muhammad, By using speech and also video as an input parameter it describes how emotions can be detected. They have used two datasets; one is Big Data database which contains both speech and also video input files, and eNTERFACE database.They first extracted the given input speech signals features to obtain a Mel-spectrogram, which is considered an image. This Mel-spectrogram is fed to 2d CNN followed by extreme learning machines (ELMs) for the fusion of scores. Some representative frames from a video segment are extracted and fed to the 3d CNN, followed by the extreme learning machines (ELMs) for the fusion of scores. Output is given for the final classification of the emotions to a support vector machine (SVM) of given input speech and video signals.

[3] **Emotion Recognition through Speech Using Neural Network**:An "Emotion Recognition through Speech Using Neural Network" was proposed by Pawan Kumar Mishra and Arti Rawat. It explains how we can recognize the feelings from the given input speech signals parameter by using the Neural Networks. In this method, firstly, they have proposed High Pass Filter, which is an electronic filter that is used to remove unwanted sounds from the given input speech signal which passes only the frequency which is higher and when there is low frequency cut off frequency occurs, and which passes only the frequency which is higher for further performing feature extraction process. They have used the Mfccs for feature extraction, They used a neural network to classify final emotions from the input speech signal after performing feature extraction.

[4] "Direct Modelling of Speech Emotion from Raw Speech" was proposed by Siddique Latif et al, By combining Convolution Neural Networks (CNNs) and also Long Short-Term Memory (LSTM) for the recognition of emotions. Firstly, for extracting features from raw speech and concatenated all these parallel convolution layers followed by a 2d convolution layer with max-pooling they used parallel convolutional layers with multiple filter lengths. To recognize emotion from the given input signal outputs are fed into LSTM followed by a fully connected layer with the final SoftMax activation function.

III. METHODOLOGY

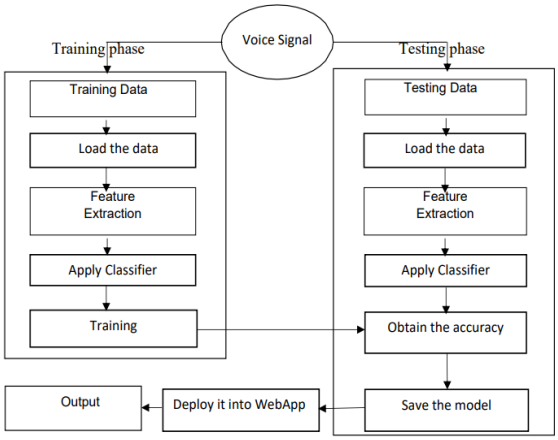

Fig 1: Workflow of proposed system

A. Datasets

For the dataset we have considered three different datasets namely savee, tess and ravedess. Each of these datasets consist of different persons with various emotions. The datasets have been obtained from the Kaggle website. The data set has 5732 instance It includes 7 emotions or features like calm, neutral, happy, sad, fearful, disgusting, angry.

B. Algorithms:

i) **Multilayer perceptron (Mlp) classifier:** A supervised classification technique which uses backpropagation for training is the multilayer perceptron (Mlp) classifier. It belongs to the class of feed-forward artificial neural networks (ANN). It is made up of many perceptrons. It comprises of one output layer, one input

ISSN- 2394-5125 VOL 7, ISSUE 06, 2020

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. With the exception of the input layer, each layer is a neuron with a nonlinear activation function. This is distinguished from a single layer feed forward neural network by its nonlinear activation function and several layers. It can identify data that is not linearly separable due to its nonlinear activation.

ii) **Convolutional Neural Network (CNN) Classification :** CNN is a subset of deep learning approaches for classification that rely entirely on feed-forward architecture. CNNs are frequently employed for recognition jobs because they improve data classification. The input data is processed in the form of receptive fields by these networks, which have small size neurons on every layer of the chosen model architecture. An input layer, hidden layers, and an output layer make up a convolutional neural network (CNN). All middle layers in a feed-forward neural network are referred to as hidden layers. The activation function receives their inputs and outputs, which are masked by the activation function before final convolution. Hidden layers in a convolutional neural network (CNN) include convolutional layers. A CNN typically consists of a multiplication or other dot product layer with a ReLU activation function as its activation function. Following this layer, several convolution layers such as pooling layers, fully connected layers, and normalization layers are applied.

C. Description of Various Module

i) Librosa: Librosa is a package which is used for music and audio analysis. It uses soundfile and audio read. It basically provides the meaningful building block to build audio, music data retrieval systems. At present, the soundfile does not support MP3, and it will cause the library to crash. Hence, librosa uses the audioread modules to work with files like MP3.

ii) Flask: Flask is a micro web framework written in Python. This provides us with some libraries, tools and technologies that allows us to build Web API's. Web API's can also be other web pages, blogs, Wikipedia or some marketing websites. The main advantage of using this framework is, it is light, and very few dependencies to update and to look after for any bugs.

iii) Keras: Keras is an open-source neural network library that can operate on top of Theano or TensorFlow and is highly useful for working with high-level neural networks built in Python. It supports convolutional and recurrent neural networks, as well as combinations of the two. It has a variety of popular neural network building components, such as optimizers, activation functions, and layers to make it much more easier to work along with image, audio as well as text and data to make it easier to write deep neural network code. It developed with a focus on simplifying models and also enabling faster experimentation.

iv) TensorFlow: TensorFlow is an open-source toolkit for numerical and large-scale machine learning and also deep learning calculations. It makes use of Python to give a simple front-end API for developing apps, which are then executed in high-performance C++. Recurrent neural networks, deep neural networks, and natural language processing tasks may all be trained and operated with TensorFlow. It assists in the creation of dataflow graphs, which represent how data flows across a graph or a series of processing nodes.

v) Sklearn: Sklearn is a Python machine learning package that is free and open-source. Random forests, support vector machines, Decision Tree Classifier, 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.

D. 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 applying several classifiers to recognize the exact emotion from the input 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.

ISSN- 2394-5125 VOL 7, ISSUE 06, 2020

IV. RESULTS

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

Datasets	MLP Model	CNN Model
Combination of 3 datasets	86.41	89.01

Correlation between the voice inputs in data set helps us to find out how strong or how weakly they are co-related to each other in the set. The numeric values from high to low represents the range of the inputs .the high value represents the positive relationship exists between the variables. The numeric low values i.e. 0 with darker color gives the more negative relationship between the input voice variables. The below fig is the proposed model of CNN confusion Matrix.

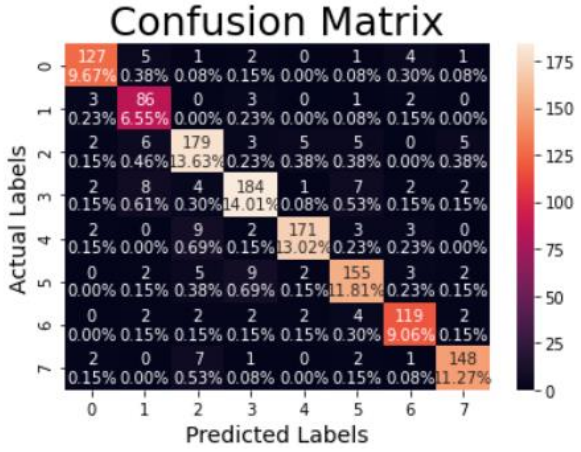

Figure 1 : CNN Model Confusion Matrix

Prediction of emotion by combining all the three datasets in the MLP model. In MLP model we train the entire data set and extract features. This gives the prediction of all instances in the data set at a time.

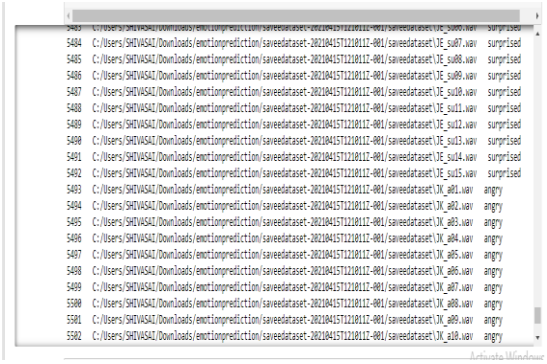

Figure 2: Emotion Prediction Snapshot

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. The below fig shows the prediction of one audio input voice. For the below fig we have given a voice to the cnn testing code and it predicted the emotion as neutral.

ISSN- 2394-5125 VOL 7, ISSUE 06, 2020

```
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
```

Figure 3: CNN Emotion Prediction Snapshot

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

The classification report where it gives the precision,recall,f1-score,and support values with avg weighted score. The below fig gives the CNN Classification Report with weighted accuracy score 0.89.

	CNN				MLP			
	Precision	Recall	f1-score	support	Precision	Recall	f1-score	support
Neutral	0.92	0.90	0.91	141	0.89	0.91	0.90	158
Calm	0.79	0.91	0.84	95	0.84	0.88	0.86	97
happy	0.86	0.87	0.87	205	0.84	0.88	0.86	174
Sad	0.89	0.88	0.88	210	0.85	0.85	0.85	199
angry	0.94	0.9	0.92	190	0.93	0.89	0.91	185
Fearful	0.87	0.87	0.87	178	0.85	0.79	0.82	189
Disgust	0.89	0.89	0.89	133	0.85	0.85	0.85	162
Surprised	0.93	0.93	0.92	161	0.85	0.87	0.86	149
Accuracy			0.89	1313			0.86	1313
Macro Avg	0.89	0.89	0.89	1313	0.86	0.87	0.86	1313
Weighted Avg	0.89	0.89	0.89	1313	0.86	0.86	0.86	1313

Figure 4: Classification Measures of CNN vs MLP

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

ISSN- 2394-5125 VOL 7, ISSUE 06, 2020

Figure 5 Emotions in dataset

CONCLUSION AND FUTURE WORK

In this research paper, 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) 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.
2) M. S. Hossain and G. Muhammad, "Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data," Information Fusion, vol. 49, pp. 69-78, September 2019.
3) 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.
4) Siddique Latif et al. "Direct Modelling of Speech Emotion from Raw Speech", Proc. Interspeech 2019, 3920-3924, September 2019.
5) 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).
6) Livingstone SR, Russo FA, 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): e0196391, May 2018.
7) Pichora-Fuller, M. Kathleen, Dupuis and Kate, 2020, "Toronto emotional speech set (TESS)", Scholars Portal Dataverse, Version 1.0
8) Y. Sri Lalitha et.al "Analysis of Parts of Speech Tagging in Text Clustering", International Journal of Innovative Technology and Exploring Engineering, June 2019, Volume 8, Issue : 8, pp : 2287-2291, ISSN: 2278-3075(online)
9) 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).