2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) | 978-1-6654-7413-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSES55317.2022.9914336

Human Voice Emotion Recognition Using Multilayer Perceptron

D.Femi and S.Thylashri

Department of Computer Science and Engineering Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology Chennai
E-mail : dfemi20@gmail.com thylashri93@gmail.com

Abstract- The Human Voice Emotion recognition system needs to extract features from speech signals. Time, frequency, and Cepstral domain features are among these features. Because the algorithm might affect the classification accuracy, choosing an effective regularization approach to generalize features is tricky. In our work we present the effect of the use of different regularization methods on language features to identify emotions. The dataset is collected from Ryerson Audio-visual Database where the emotional voices are the actors were filmed in eight different moods. Our research divides emotions into eight categories: neutral, calm, happy, sad, angry, afraid, disgusted, and startled. Speech indicators had been analyzed to classes Spectral assessment and characteristic extraction algorithms were used to extract the prosodic capabilities of speech formant frequencies, entropy, variance, minima, median, Linear Prediction Corrector (LPC). The emotional speech parameters are then extracted using Multilayer Perceptron Neural Networks (MLPNNs).

Keywords— Human Computer Interaction, Multilayer Perceptron

I.INTRODUCTION

In the field of computer science, speech emotion recognition is one of the hottest study subjects. Emotion is a means of expressing how a person feels and their mental state. Emotions are vital in sensitive jobs like that of a surgeon, a military commander, and many others where one must keep their emotions under control. Predicting emotions is difficult because everyone speaks in a different tone and intonation. Neutral, calm, happy, sad, angry, surprise is some of the feelings elicited. To classify these emotions using the best suited approach from a given voice sample.

Emotions play a significant part in human existence. It very well may be recognized from the speech uttered by the person; it is a medium of expression of one's point of view or sentiments to other people. This cycle could be accumulated alongside video that is, the face recognition by utilizing its own APIs converged with voice and can be created as a software application in future specialized industry, thus that it will portray the state of the individual with whom we are collaborating with. It will be more and more useful for person abroad with their relatives and parents in their hometown. Voice signal acquisition, feature extraction, and emotion recognition were the three aspects of a speech emotion processing and recognition system.

Textual and emotional information are both present in speech. The computer should automatically understand the emotional content of voice signals to establish a pleasant human-computer interaction experience. In a customer service system, for example, the voice may detect a customer's emotions. When combined with an educational assistant system, it has been demonstrated to improve children's social-emotional and intellectual abilities. Parents and educators are capable of quickly resolving problems. Speech can be used to detect emotions by the operating system. The device will give an early warning if the driver's mood is overly tense or enraged. This could reduce the likelihood of traffic collisions. Speech emotion recognition software has a wide range of applications in a variety of industries.

Emotional displays transmit a lot of information about a person's mental health. This has spawned a new branch of study known as automatic emotion recognition, whose primary goal is to comprehend and recall desired feelings. Speech recognition and speaker identification techniques allow machines to recognize "what is said" and "who said it." Machines can also understand "how it's spoken" if they are equipped with emotion recognition techniques. Aside from facial emotions and gestures, speech is a potent channel for communicating with emotional intelligence in the field of human computer

interaction (HCI).

The extraction of information from the voice signal and conversion to an appropriate format for simpler processing are required for speech emotion recognition. These features include Time, Frequency, and Cepstral-domain features. Selecting a normalization algorithm that is efficient in voiceemotion recognition is a difficult task because the algorithm will affect the performance of the emotion discrimination classifier by removing the emotion discrimination's effectiveness This study uses the Ryerson Audio-visual collection of emotional speech and songs to investigate the impact of various normalisation techniques on the categorization process (RAVDESS).

The aim of Human Voice Emotion Recognition is to improve man-machine interface. It can be utilized to monitor the psychological condition of an individual in lie detectors. In ongoing time, speech emotion recognition also discovers its applications in Medication, Forensics and Google assistant.

II. Literature Review

In this paper [1] the author proposed the Multilayer perceptron algorithm for the recognition of different emotional categories from the standard speech RAVDESS database. The comparison of the two algorithms for performance analysis which is supported by the confusion matrix and accuracy.

In this paper [2] the author proposed Support Vector machine for auto-regressive model for training and classification. The main objective behind using such a model is to extract dependency among extracted speech feature vectors as well as the multi-modality in their distribution. In this paper [3] the author explores audio sounds and categorizes them into four main motion states. To this end we studied various high-energy ZCR (Zero Crossover) MFCC (Mel Frequency Cepstral Coefficient) sounds for representations of sensor data.

In this paper [4] the author describes in-depth learning techniques and multi-layered perceptual algorithms for predicting output emotions and confusion metrics. The main objective of this work is to train the model between the classifier and the derived speech.

III. Existing System

Human Voice Emotion Recognition has changed in the era of using computer-generated gestures to extract emotional characteristics and analyse characteristic parameters and the resulting emotional changes. Here the reliability and integration of all emotions are used. The translated speech should be a symbol of emotional state so it is necessary to identify the emotions contained in the source speech and integrate the equivalent emotions in the target speech. The words in the speech are grouped but in rare cases.

IV. Proposed System

A Multi-layered Perceptron is a machine learning algorithm consisting of a neural network as a three-tiered sub-algorithm. Input Layer, Hidden Layer and Output Layer Converted Virtual Recording can be input as modified audio file. Compare layer weights with pre-processed data to avoid layer confusion in input layer output records. The RAVDESS database for model devices is sent to the MLP classification which divides the database by 75:25 for training test data set. The dataset includes sound samples from 24 expert artists with North American voices. Eight styles of emotion are included. Classifications are used mainly because they are very convenient for a fully oriented time series for emotionally predictable speech in this context.

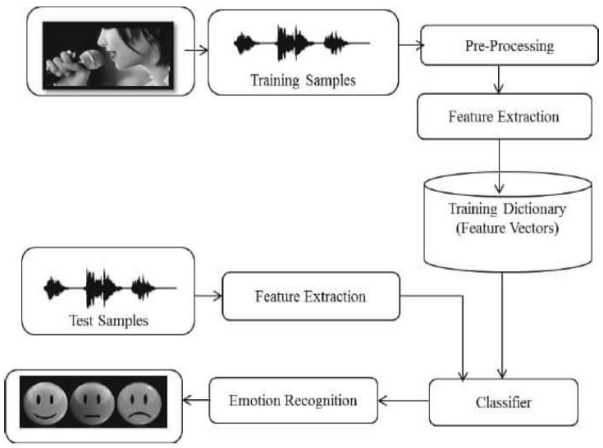


Fig 1: System Architecture

The emotion recognition system consists of six modules such as Speech input, pre-processing, spectral analysis, feature extraction, neural network training, and detected emotions are some of the techniques used. The solution to recognizing

emotions depends on the features of the voice. This can be seen in the system architecture fig 1 where we take sound as a training sample then pre-process to extract the sound features and then put it into the training group. This feature is used to form a classifier to get the decision emotion.

A. Dataset

The RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song) has 7356 files that include videos and audios of talks and songs.

B. Pre-processing

Data visualisation approaches are used to study data to do pre-processing. The balancing is examined, as well as the amount of data.

C. Feature Extraction

Extraction of features the librosa, pandas, and NumPy libraries are used in the feature extraction method. Images of Mel spectrums is got as the output and that is applied on convolutional neural network, but the results were insufficient to process them. One-dimensional characteristics are extracted instead of processing spectrum images.

D. Training and Classification

A feed forward neural network augmentation is the multilayer perceptron (MLP). It has three layers: an input layer, an output layer, and a concealed layer. The input layer receives and processes the input signal. The output layer is responsible for tasks like prediction and categorization. Any number of hidden layers are introduced between the input and output layers in the present MLP computation engine. In MLP the feed forward like a network forwarding from the data input to the output layer.

In MLP, Neurons are trained using a back-propagation learning algorithm. MLP is designed to predict continuous tasks and can solve problems that cannot be separated linearly. The primary use case format for MLP is classification, recognition and prediction.

The MLP consists of three or more layers of non-linear active nodes (putting a layer in one layer or more hidden layers). In supervised learning issues, multi-layered preceptors are frequently used. They learn to predict the correlation (or dependence) between input and output by training the classifier of input and output pairs. Emotional traits such as happy, sad, angry when there are clear data such as male and female traits. This can be converted into a representation of the actual encoding values.

This training helps the model to understand whose emotional features are. The invisible data can then be connected to the above-mentioned input mode, it will be able to correlate and predict the emotion. Once a neural network has been trained it can be used to make various predictions. Multilayer artificial neural networks analyse how to adjust and reduce the cost function of a neural network.

Then, utilising each feature extracted to the known emotion, we train the system and test it, using 75% of the data for training and 25% for testing by partitioning the data.. We know that the RAVDEES database is emotionally based on the latter because we first know how to download data from a folder that can be run through the Python Library Globe and use the OS library to get its name. The statement X and Y represents emotion. The next step is to model the sentiment classification used in MLP Classifier.

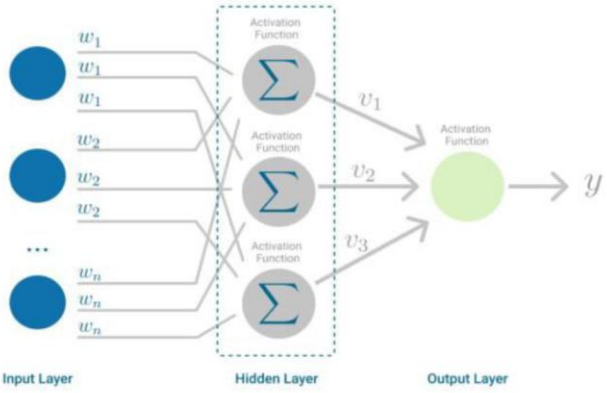


Fig 2: Multi-layered Perceptron

The trained model will classify the emotions:

- Neutral
- Calm
- Happy
- Sad
- Angry
- Surprise

V. Results And Discussion

We are using Visual studio to make use of obtained output to detect the emotion using trained model classifier and then print the output. The amount of

training samples, testing samples, and identified features may all be seen. We derive the accuracy and confusion matrix by comparing predictions after training. Fig 2 describes the number of training samples used and how many features obtained.

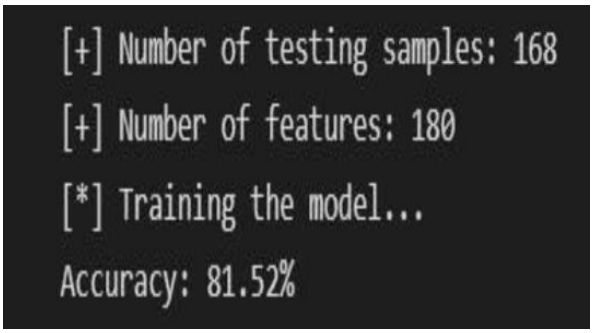
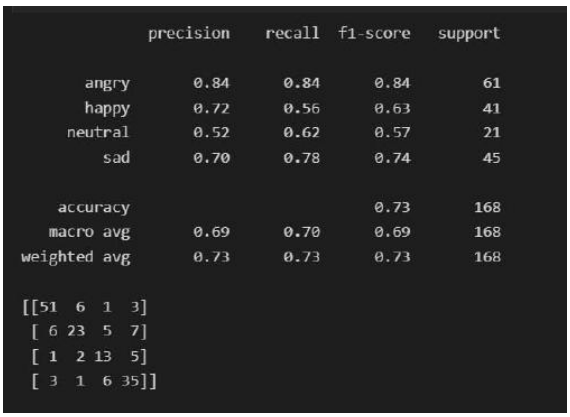


Fig 3: Training the data

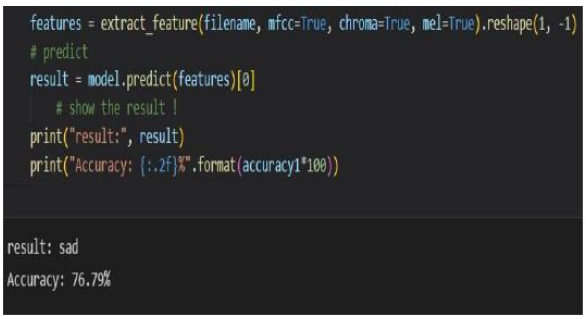
Fig 4 and 5 shows the 81.52% percent of accuracy which is obtained from confusion matrix. which tells that our model works well. With this we are trying to correct the false prediction by extracting some more features.



	precision	recall	f1-score	support
angry	0.84	0.84	0.84	61
happy	0.72	0.56	0.63	41
neutral	0.52	0.62	0.57	21
sad	0.70	0.78	0.74	45
accuracy			0.73	168
macro avg	0.69	0.70	0.69	168
weighted avg	0.73	0.73	0.73	168

```
[[51  6  1  3]
 [ 6 23  5  7]
 [ 1  2 13  5]
 [ 3  1  6 35]]
```

Fig 4: Confusion Matrix



```
features = extract_feature(filename, mfcc=True, chroma=True, mel=True).reshape(1, -1)
# predict
result = model.predict(features)[0]
    # show the result !
print("result:", result)
print("Accuracy: {:.2f}%".format(accuracy1*100))
```

```
result: sad
Accuracy: 76.79%
```

Fig 5: Emotion and Accuracy

VI. Conclusion

The studies over this idea fetches us the knowledge, that this technique is yet play the vital position in scientific and technical field. Librosa is used in our work to derive the features of emotion recognition. We use Pyaudio to record audio. The Matplotlib module is used in our work to draw waves and provide them for future use. We used a classifier model to categorize emotions.

VII. Future Enhancement

We have shown how to use machine learning to gain basic knowledge of speech and basic emotions from human emotions. The system can be used for processes such as call center structure or digital assistant based on voice or chatbot linguistics. The correct use of speech rhythm can be checked to see if it can explain many of the weaknesses of the model. Find a way to clear the silence from the occasional audio clip. Browse the various sound features of the audio data to make sure it applies to the language recognition area.

References

[1] Sardar, A. A. M., Islam, S., Bhuiyan, T. (2021). A Review on Automatic Speech Emotion Recognition with an Experiment Using Multilayer Perceptron classifier. In Soft Computing Techniques and Applications (pp. 381-388). Springer, Singapore.

[2] Ho, N. H., Yang, H. J., Kim, S. H., Lee, G. (2020). Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. IEEE Access, 8, 61672-61686.

[3] Sefara, T. J. (2019, November). The effects of normalisation methods on speech emotion recognition. In 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC) (pp. 1-8). IEEE.

[4] Berkane, M., Belhouchette, K., Belhadef, H. (2019). Emotion recognition approach using multilayer perceptron network and motion estimation. International Journal of Synthetic Emotions (IJSE), 10(1), 38-53.

[5] Navya Damodar, Vani H Y, Anusuya M A. Voice Emotion Recognition using CNN and Decision Tree. International Journal of Innovative Technology and Exploring Engineering (IJITEE), October 2019.

[6] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomedical Signal Processing and Control, vol. 47, pp. 312-323, 2019.

[7] A. S. Popova, A. G. Rassadin, and A. A. Ponomarenko, "Emotion recognition in sound," in International Conference on Neuroinformatics, 2017: Springer, pp. 117-124