e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

SPEECH EMOTION RECOGNITION SYSTEM FOR CLASSIFYING HUMAN AUDIO

Piyush Gawali[*1], Rohit Dahatonde[*2], Zeeshan Bepari[*3], Saransh Pandey[*4], Shivansh Agrawal[*5]

[*1,2,3,4,5]Information Technology, NBN Sinhgad School Of Engineering, Pune, Maharashtra, India.

ABSTRACT

Speech Emotion recognition (SER) system is nothing but series of methodologies that process and classify speech audio to discover the embedded emotions.

It uses the fact that it is the voice that regularly reflects underlying emotion via tone and pitch. The number one goal of SER is to enhance man-machine interface. The main need of Emotion recognition from speech is in the tasks where audio signal is only input for the computer device. This system can be used to check a person's physiological temperament. Librosa is a Python Library for analyzing audio. We are able to use the libraries Librosa, Sound file, and Keras to build a model using MFCC classifier. This will be able to understand emotion from sound files. In this proposed undertaking, we carry out speech data analysis to detect the feelings. Here we are analyzing unique techniques to carry out emotion recognition and speech analysis to perform the task

Keywords: SER, Feature Extraction, Audio Preprocessing, MFCC, CNN.

I. INTRODUCTION

Detecting emotions from voice is a field of study which has many algorithms and methods involved within, but it has vast number of applications. In a human-computer or human-human interaction system, an emotion recognition system can provide an improved service to a user by adapting to the user's emotions.

In a virtual world, emotion recognition helps to simulate more realistic avatar interactions. The scope of work to detect voice emotions is very limited. Recently there is some confusion regarding which algorithm is best for classifying emotions and which emotions are classified. In order to understand a person's attitude and mood through conversation, it is necessary to know who is talking and what they are saying.

Understanding people's moods is often very helpful. For example, a computer with the ability to recognize and respond to human non-lexical communication such as emotions.

In such cases, after recognizing human emotions, the machine can adapt the settings to their needs and preferences.

II. METHODOLOGY

MFCC

Mel frequency cepstral-coefficients was initially recommended for identifying monosyllabic phrases in constantly spoken sentences however no longer for speaker identity. The system is based on the premise that the human ear is an accurate speaker identification system and it replicates the human hearing. MFCC capabilities are rooted within the diagnosed discrepancy of the human ear's important bandwidths with frequency filters spaced linearly at low frequencies and logarithmically at excessive frequencies had been used to retain the phonetically crucial properties of the speech signal. Speech signals usually comprise tones of varying frequencies, each tone with an actual frequency, f (Hz) and the subjective pitch is computed at the Mel scale. MFCC is used to identify different unique numbers by taking audio as input method, as it is one of the effective way out there. Some adjustments have been proposed to the primary MFCC algorithm for better robustness, such as via lifting the log-mel-amplitudes to an appropriate power (round 2 or 3) before applying the DCT and decreasing effects of other low-power parts. The MFCCs are calculated the usage of this equation : $\hat{C}\ n=\sum_{k=1}^{n}(\log \hat{S}\ k)\cos[n(k-12)\pi k]$

III. MODELING AND ANALYSIS

Proposed System

Initially we need to provide input to the system in the form of (Audio File). For data preparation librosa library convert audio file to dataset. Next step is to convert dataset into csv file format and it passes to the feature

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

extraction process. Classifier classify the data and according to dataset parameters the CNN model is create. There are decision algorithms for emotion identification. Finally the expected emotion as a output.

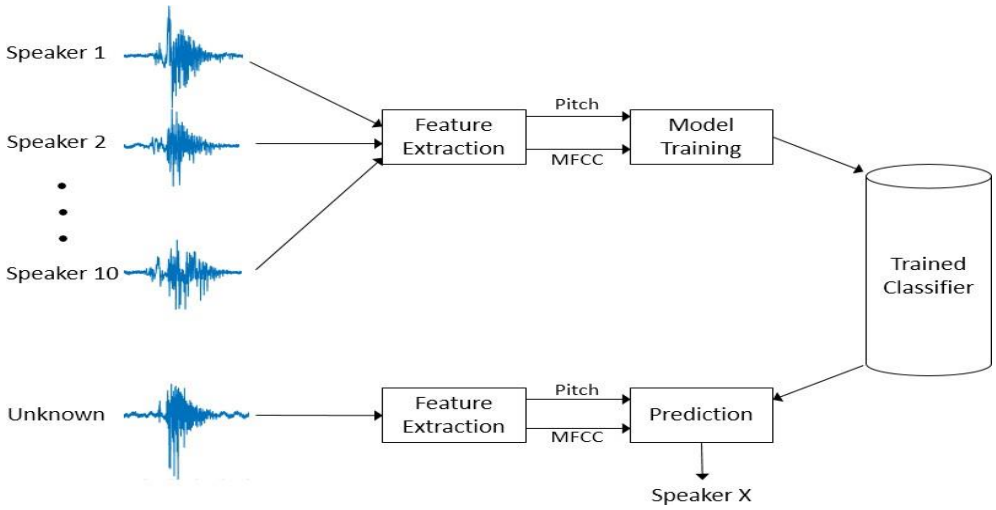

Figure 1: System Architecture

CNN

Artificial Intelligence has been witnessing a huge increase in bridging the space between the talents of humans and machines. The field encompasses a wide range of factors that can produce amazing results, regardless of who is involved. The agenda for this subject is to enable machines to view the world as people do, perceive it in a comparable manner and even use the understanding for a multitude of responsibilities which includes photograph & Video recognition, picture analysis & classification, Media recreation, recommendation systems, natural Language Processing, etc. The advancements in pc vision with Deep learning has been built and perfected with time, often over one precise set of rules a Convolutional Neural network. It computes the significance of input images by learning the weights and biases of the input images.to diverse components/objects inside the image and have the power to differentiate one from another. ConvNet requires significantly less pre-processing compared to other algorithms. Unlike primitive techniques, in which filters were many manually made, they can remember those filters/characteristics with enough training. The structure of ConvNet is mostly like structure of neuron connection patterns that is seen in brain and was stimulated by way of the visual Cortex. In this project 5 layers are used for building CNN model. First 4 layer contains Relu as activation function and last layer is made of Softmax. Max pooling is generally for compaction. Learning rate is kept as small as possible for better gathering of information lr=0.00001. Model contains 370 epoc for repeated learning of audio dataset so as to reduce data loss.

IV. RESULTS AND DISCUSSION

Testing and evaluation takes place simultaneously .For this project epoc for the model is decided to be 370 which is optimum to get maximum accuracy while taking minimum execution time. During each epoc data loses are reduced to by reevaluating same data and learning extracted features more carefully. This eventually result in increasing the accuracy of the train dataset to 83% and overall accuracy of the system increased to 60%. This accuracy is the maximum that can be achieved through this project.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

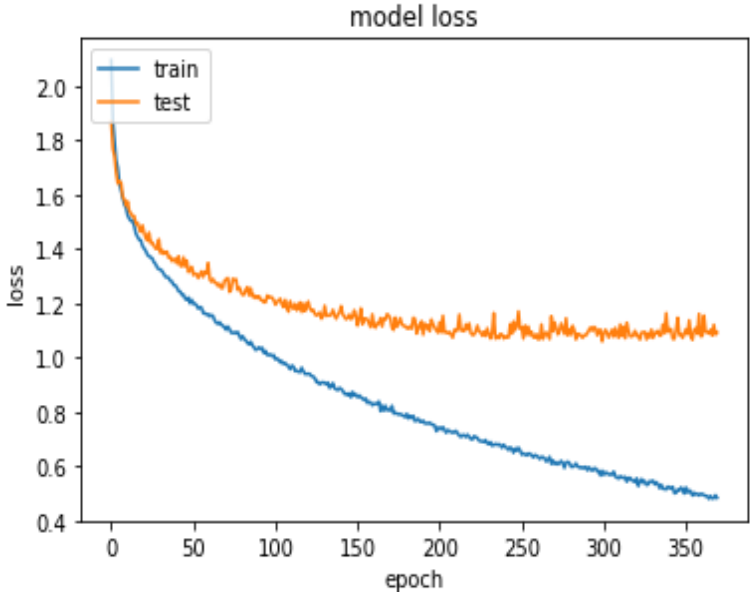

Figure 2: Loss Graph

On the basis of test case results prediction analysis is also done to see how exactly model is creating outputs that are similar or differ to the actual value. As the System accuracy is 60% around 60% of time model will create accurate results. Prediction Analysis outcome can be seen in below diagram.

```
[44] finaldf[130:140]
```

	actualvalues	predictedvalues
130	1	5
131	5	2
132	1	2
133	3	3
134	4	4
135	0	3
136	3	5
137	0	0
138	1	0
139	6	4

Figure 3: Prediction Analysis Outcome.

V. CONCLUSION

Hence above methods present a good way to give the ability to machine to determine the emotion. It will deliver the machine the capability to have a better method towards having a better conversation and seamless verbal exchange like human does. It aims to determine the emotion with the speech of a human and can be extended to integrate with the robot to help it to have a better knowledge of the mood the corresponding human is in, so that you can assist it to have a better conversion. Any e-commerce sites which have an AI based chat bot which recommends the customer to have a good experience our project will determine the customer/s mood and accordingly it can help the chat bot to have and give recommendations.

VI. REFERENCES

[1] Babak Basharirad and Mohammadreza Moradhaseli, “Speech emotion recognition methods: A literature review” AIP Conference Proceedings, vol. 22, no. 6, pp. 1154–1160, Jan. 2017.

[2] Teddy Surya Gunawan, Muhammad Fahreza Alghifari, Malik Arman Morshidi, Mira Kartiwi, “Review on Emotion Recognition Algorithms Using Speech Analysis” Indonesian Journal of Electrical Engineering and Informatics (IJEEI), vol. 6, no. 1, pp. 12–20, Mar. 2018.

[3] Sathit Prasomphan, "Improvement of Speech Emotion Recognition with Neural Network Classifier by Using Speech Spectrogram" IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 10–21, Dec 2015.

[4] Chenchen Huang, Wei Gong, Wenlong Fu, and DongyuFeng, "A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM," Comput. Human Behav., vol. 2014, no. 5, pp. 1–7, Aug. 2014.

[5] Ashish B. Ingale, D. S. Chaudhari, "Speech Emotion Recognition," International Journal of Soft Computing and Engineering (IJSCE), vol. 2, no.1, pp. 2231-2307, Mar. 2012.

[6] Mehmet Berkehan Akcay and Kaya Oguz, "Speech Communication" Elseviers, vol. 116, no. 1, pp. 56–76, Jan. 2020.

[7] A. B. Ingale and D. S. Chaudhari, "Speech emotion recognition," International Journal of Soft Computing and Engineering (IJSCE), vol. 2, pp. 235-238, 2012.

[8] A. Joshi and R. Kaur, "A Study of speech emotion recognition methods," Int. J. Comput. Sci. Mob. Comput.(IJCSMC), vol. 2, pp. 28-31, 2013.

[9] W. Fei, X. Ye, S. Zhaoyu, H. Yujia, Z. Xing, and S. Shengxing, "Research on speech emotion recognition based on deep auto-encoder," in 2016 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 308-312, 2016.

[10] Mengna Gao, Jing Dong, Qiang Zhang and Deyun Yang, "End-to-End Speech Emotion Recognition Based on One-Dimensional Convolutional Neural Network," International Conference on Innovation in Artifical Intelligence, vol. 25, no. 3, pp. 556–570, Mar. 2019.

[11] Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R, Rajesh Kumar Muthu, "Speech Emotion Recognition using Support Vector Machine," School of Electronics Engineering, Vellore Institute of Technology, Jun. 2018.

[12] Russell, J.A. and Mehrabian, A., "Evidence for a three-factor theory of emotions." J. Res. Personal., vol. 11, no. 3, pp. 273–294, 1977.

[13] Nicolaou, M.A. , Gunes, H. , Pantic, M., "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space." IEEE Trans. Affect. Comput.. Vol. 2, no. 2, pp. 92–105, 2011.

[14] J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen, "Segment-based emotion recognition from continuous Mandarin Chinese speech," Comput. Human Behav., vol. 27, no. 5, pp. 1545–1552, Sep. 2011.

[15] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.