JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SPEECH EMOTION RECOGNITION USING MACHINE LEARNING

Dr. K. Kranthi Kumar (Associative Professor), Dept of IT SNIST,Hyderabad, India

B. Hema Kumari (Assistant Professor), Dept of IT SNIST,Hyderabad, India

Mr. K. Madhan Sesha Sai, Dept of IT, SNIST, Hyderabad, India

Mr. D. Sri Laksh, Dept of IT, SNIST, Hyderabad,India

Abstract- **Pitch, timbre, loudness, and vocal tone are just a few characteristics that may be used to describe the human voice. Many times, it has been noted that as individuals produce speech, their vocal characteristics change to reflect their emotions. The algorithm method for emotion recognition in humans using speech is presented in this research. This paper's main goal is to identify speech feelings and divide them into 6 emotion output classes: frustrated, terror, dissatisfaction, joyous, sorrowful, and silent. The suggested method relies on the Crema-D database of emotional speech using Mel Frequency Cepstral coefficients (MFCC). Data Augmentation is performed on input data audio file, such as Noise, High Speed, Low Speed etc. are added, thus more the varied data is available to the model better the model understands. Feature extraction is done using MFCC and then the extracted features are Normalized (for Independent Variable), Label Encoding (for Dependent Variable (SVM, RF)), One Hot Encoding(for Dependent Variable(for CNN)) is done. After this the dataset is divided into Train, Test and given to different models such as CNN, SVM for Emotion prediction. For each of the several tests, we provide accuracy, f-score, precision, and recall. The results showed that CNN had the best accuracy and correctly identified emotion 88.21% of the time.**

Keywords: - Mel Frequency Cepstral coefficients, Machine Learning.

1. INTRODUCTION

Speech emotion recognition is the process of gathering emotional elements from computer speech signals, comparing them, and analyzing. parameter values and related emotional changes. Feature extraction and classifier training are required for emotion detection from audio sources. A feature vector is used to train a classifier model to identify a specific mood more accurately. It is made up of audio signal components that describe the prominent aspects of a speaker (pitch, pitch, energy, etc.). By way of Twitter, message boards, Facebook, blogs, user forums, and other platforms, social media and the internet have access to a vast amount of opinion data. People's decision-making processes are influenced by ideas posted on the Internet by a variety of thought leaders and regular people. People may share their thoughts and opinions about items and socially relevant products through text-based reviews. Audio and video are two more common ways to communicate ideas. Numerous videos exist that discuss political and social analyses, product and movie reviews, and product unpacking. On YouTube, you may find discussions on these issues and perspectives. There are several audio platforms. internet platform for personal expression. Because there is more information about the speaker's viewpoint in voice, it is frequently more interesting than text. This enormous resource is largely underutilized, and data analysis may benefit greatly by gathering public sentiment/opinion on specific topics as well as popular opinion on social or political concerns. The study of the mind is still a young discipline.

2.LITERATURE SURVEY

1. Vincius Maran et al. assert that learning speech is a laborious process in which the infant's processing of criteria is highlighted by their irrationality on the path to the contemporary production of ambient language segments and structures.

2. G. Tsontzos et al. focused on the emotions that improve our knowledge of one another, therefore it makes sense that this understanding would also benefit computers.

3. Industrial control and robotics applications are the main uses of neural networks, according to Mehmet Berkehan Akçay et al.

4. Discriminative testing has been used for speech detection and recognition for a longer length of time, according to Y. Wu et al.

5. Varghese et al. claim that there are several methods to infer sentiments from shape

3.OVERVIEW OF THE SYSTEM

3.1. EXISTING SYSTEM

According to existing research, most of the current research in this field focuses on lexical analysis of emotion recognition. This is a language-based strategy for classifying emotions into her three groups. H. Anger, joy, and indifference are all emotions. The highest degree of correlation between test and training audio files is used as the key parameter for identifying specific emotion types, and the strongest cross-correlations between discrete-time sequences of audio signals are recorded. The second strategy uses SVM cubic classifiers to extract discriminators and recognize only anger, joy, and neutral emotion segments.
less and majorly all of them are carried out on social networks such as twitter, LinkedIn and Facebook. There are no major studies carried out on Instagram social network.

3.2. PROPOSED METHOD

3.2.1 INTRODUCTION

The suggested system's MFCC function was utilized to divide audio data into several emotion categories using artificial neural networks. Neural networks offer the benefit of being able to categories many sorts of emotions present in variable-length audio inputs in a real-time setting. This approach presents a reasonable balance between the precision of execution and the complexity of real-time approaches.

3.2.2 BENEFITS OF THE PS

- Any machine that supports the Python programming language can use it.
- Quickly processes audio and is simple to use.
- The system is capable of handling audio files of various sizes.

3.2.3 FEASIBILITY STUDY

Project viability is considered at this phase, and commercial concepts, a project plan, and a cost estimate are all offered. A feasibility analysis of the suggested system will be performed as part of the system analysis. This is to make sure the organization won't face major financial difficulties because of the suggested remedy. Basic knowledge of important system requirements is necessary for a feasibility study. When evaluating feasibility, three important elements need to be considered.

1. Economy

2. Technical feasibility

3. Social viability

3.2.4. Economy:

The purpose of this analysis is to examine the economic impact of the system on the company. Long-Term Corporate Profitability Companies have limited ability to invest in system research and development.

3.2.5. Technical feasibility:

The technological viability and system requirements are determined by this investigation. The proposed system shouldn't put an undue strain on the technological resources that are already accessible. This puts a heavy strain on available technical resources. This puts a lot of pressure on the customer. Only minor adjustments or changes are required to run this system, so the system created should meet realistic standards.

3.2.6. Social Feasibility:

Checking user consent for the system is part of the investigation. This includes educating users on how to get the most out of technology. User acceptance is determined solely by how the user is introduced and educated about the system. To give constructive criticism, you must build trust. This is very helpful as a system user.

4. PROPOSED ALGORITHM SYSTEM

4.1 CNN

A deep learning system called a convolutional neural network (ConvNet/CNN) can take an input age and recognize various attributes and objects while giving them a priority using learnable weights and biases. In comparison to other classification techniques, CNN requires a lot less preprocessing. While the fundamental approach necessitates the manual building of filters, CNNs can: With the right training, become familiar with these filters/properties. The organization of the visual cortex parallels the neural connectivity patterns of the brain. The

human brain used to create CNNs. Individual neurons can only respond to internal stimuli the receptive field, which is a small part of the visual field. Several similar fields can be placed on top of each other to cover a larger area. A convolution operation is performed to extract high-level input features such as edges. CNNs require more than just convolutional layers. The first, her ConvLayer, is traditionally responsible for collecting low-level data such as edges, colors, and gradient directions. The architecture also supports top-level quality by adding layers, giving the network the ability to perceive the photos in the dataset as we do. Convolutional neural networks have ushered in a new age of artificial intelligence despite their flaws. CNNs are being employed in computer vision applications such as augmented reality, face recognition, and picture retrieval and editing.

Our findings are unexpectedly useful in the context of convolutional neural networks' ongoing development. We still have a long way to go before we can reproduce the fundamental elements of human intellect, as a works has shown. A convective neural network is made up of several artificial neuronal layers. A mathematical function called an artificial neuron, like its biological counterpart, computes the weighted sum of a set of inputs and outputs an activation value. Different activation functions generated by CNN layers are sent to the following layer.

1. As input, a sample audio file is offered.

2. Audio files are used to draw spectrograms and waveforms.

3. Extract the MFCC (Mel-Frequency Cepstral Coefficients) using the LIBROSA Python package.

4. The data is rearranged, divided into training and test groups, the results are examined, and the data set is trained using the CNN model and its layers.

5. METHODOLOGY

In this study, we thoroughly contrast several methods for speech-based emotion identification systems. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) audio recordings were used for the analysis. After the raw audio recordings had been pre-processed, characteristics such the Log-Mel Spectrogram, Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy were considered. Application of techniques like Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs), Hidden Markov Models (HMMs), and Deep Neural Networks (DNNs) allowed researchers to compare the importance of different features for emotion categorization.

Convolutional neural networks are used in the voice emotion recognition application's implementation. The system's architecture is as follows:

5.1 Models for training and testing:

The system receives training data that includes the expression label and For that network, weight training is furthermore available. The input is an audio file. The audio is then subjected to intensity normalisation. To prevent the influence of the presenting order of the samples from affecting the training performance, the Convolutional Network is trained using a normalised audio. When combined with the learning data, the weight collections created as a result of this training approach yield the greatest results. Based on the final network weights learnt and the system's pitch and energy during testing, the dataset returns the emotion during identification. Based on the individual's bpm value, three emotions—Relaxed/Calm, Joy/Amusement, and Fear/Anger—are identified. Based on the theories of "colour psychology" and "shape psychology," the colours and forms used in the created art are like the emotions observed.

5.2 Speech Archive:

The suggested approaches for voice emotion recognition in this study are validated using several speech databases. Berlin and AIBO are the two datasets that are most frequently utilised. German-language actors recorded Burkhardt et al. Technical University Berlin's Department of Technical Acoustics served as the location of record. Five male and five female German actors each read one of the selected lines as part of the dataset contribution. Various documented emotions include disgust, indifference, fear, rage, happiness, and sadness. Aibo, a robot created by Sony that is controlled by a human operator, was used to play with and interact with 51 youngsters to gather data for another emotional database. The five gathered emotions in AIBO are emphatic, angry, positive, neutral, and positive.

5.3 Extracting Features:

The next step is to extract the characteristics from the audio files that will aid our model in differentiating between them. One of the libraries used for audio analysis is the Librosa package in Python, which we utilise for feature extraction. Since its invention in the 1980s, MFCCs have been the cutting-edge feature while doing Speech Recognition jobs.

5.4 Spectrogram and Waveform:

By charting the waveform and spectrogram of one of the audio files, we examined it to learn more about its characteristics.

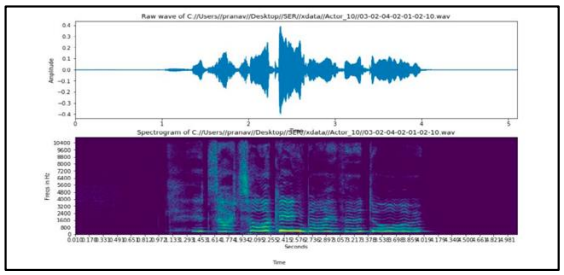

FIG **1** WAVEFORM

We would make use of MFCCs as our input method. If you wish to comprehend MFCCs in depth. Using the Python tool librosa, audio data can be loaded and converted to MFCCs format with ease.

5.5 DATA SET RAVDESS

The voice or singing renditions of 12 actors and his 12 actresses may be found in the RAVDESS.

There are no Song Version details for Actor #18. The data for song version does not contain the emotions of disgust, neutrality, or surprise.

The audio-only zip file was my choice because it discusses extracting emotions from audio. There were roughly 1500 WAV audio files in the zip file. A second website has 500 audio remarks spoken by his four actors at various emotional points. Organising your audio files comes next. A distinctive identifier that may be used to pinpoint the emotions that make up each audio recording can be found in the sixth place of the filename. The dataset contains 5 different emotions: calm, happy, sad, furious, and afraid.

In order to analyse and extract functions from an audio file, I utilised Python's Librosa package. A Python library for analysing music and audio is called Librosa. It offers the components required to develop a music data retrieval system. With the help of the librosa library, I was able to extract the functionality. Mel-Frequency Cepstrum Coefficient (B. eMFCC). A common feature in speech and automated speaker recognition is MFCC. Using the identifiers offered on the website, we also distinguished between the male and female voices. Because in our testing, we discovered a 15% increase in the distance between male and female voices. It can be because your voice pitch has an impact on the outcomes.

Data Set

Table 6.2 RAVDESS Dataset Actors Voice

Actors	Actor Recorded Voise Count
Actor 1	60
Actor 2	60
Actor 3	60
Actor 4	60
Actor 5	60
Actor 6	60
Actor 7	60
Actor 8	60
Actor 9	60
Actor 10	60
Actor 11	60
Actor 12	60
Actor 13	60
Actor 14	60
Actor 15	60
Actor 16	60
Actor 17	60
Actor 18	60
Actor 19	60
Actor 20	60
Actor 21	60
Actor 22	60
Actor 23	60
Actor 24	60

FIG 2 RAVDESS Dataset Actors Voice

6. SYSTEM ARCHITECTURE

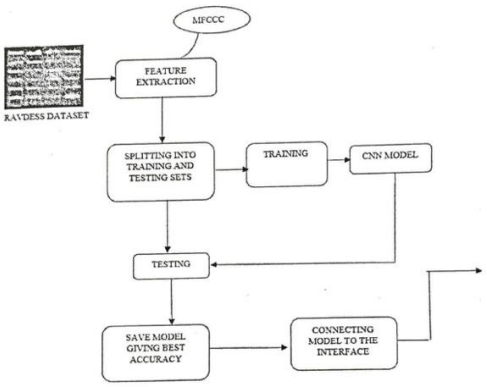

Fig 3. System Architecture

7. RESULTS SCREENSHOTS

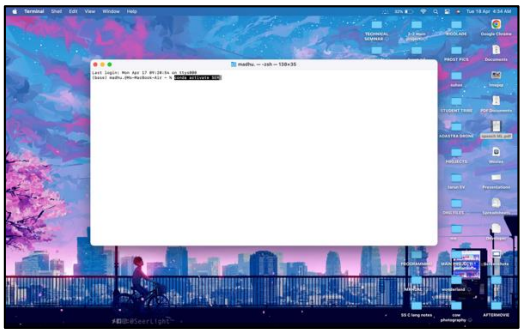

fig.4. Activating Conda

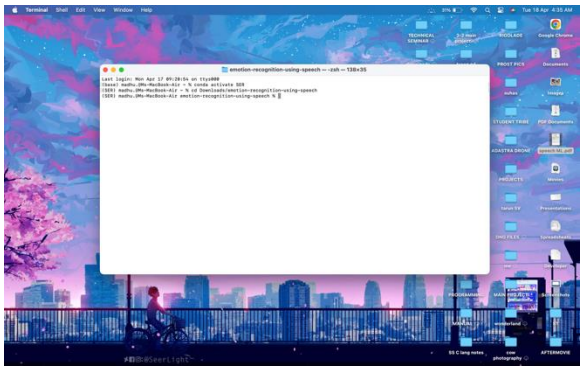

Fig 5. Locating File

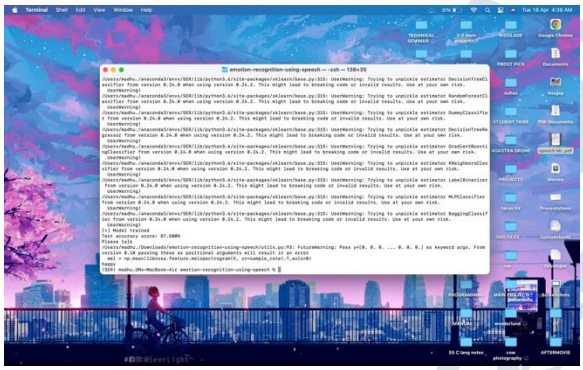

fig.6. Providing Voice

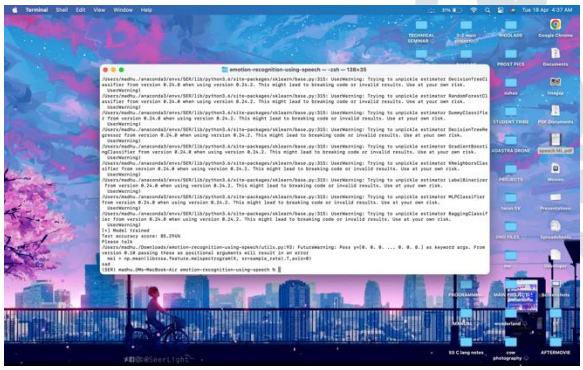

fig.7. Predicted Output

8.CONCLUSION

We trained his CNN model to be able to predict human emotions based on language. Speech-based emotion recognition has been successfully analyzed, created, and implemented as a whole. The results were also tested and found to be correct. This approach uses a CNN algorithm to classify emotions based on human language. This model outperforms all previously trained models at predicting human emotions from speech. The trained model estimates the F1 score to be 91.04 points. In this project, five emotions—happy, sad, anxious, peaceful, and angry—are expressed. It is possible to employ this language-based emotion identification for interpretation. Voice input allows you to convey ideas and comments to the model. For instance, their opinions on a certain brand or political perspective. This method may be used with other music applications to offer song suggestions based on the user's mood. Additionally, this may be used to enhance the product recommendations sent to users of online shopping apps such Amazon.

9. FUTURE ENHANCEMENT

Increasing the system's precision. A few additional feelings, such as disgust, surprise, and others, can be added to it. Connecting the system to several platforms.

10.REFERENCES

[1] Mohammed AbdelwahabAnd Carlos Busso, Multimodal Signal Processing (MSP) Laboratory, Erik Jonsson School OfEngineering &Computer Science, University Of Texas At Dallas, Richardson, Texas 75083, U.S.A.

[2] Voice Based Emotion Recognition WithConvolutional Neural Networks for Companion Robots Eduard FRANT ¸ I1, 2, Ioan ISPAS1, Voichita DRAGOMIR3, Monica DASCA˘ LU1, 3, Elteto ZOLTAN1, And Ioan Cristian STOICA4,

[3] MULTIMODAL SPEECH EMOTION RECOGNITION USING AUDIO AND TEXT Seunghyun Yoon, Seokhyun Byun, And Kyomin Jung Dept. Of Electrical AndComputer Engineering, Seoul National University, Seoul, Korea Fmysmilesh, Byuns9334, Kjungg@Snu.Ac.Kr

[4] Speech Emotion Recognition Using CNN, Article InInternational Journal Of Psychosocial Rehabilitation · June 2020, Harini Murugan

[5] Speech Emotion Recognition Methods: A Literature Review Babak Basharirad, And Mohammadreza Moradhaseli

[6] SPEECH EMOTION RECOGNITION Darshan K.A1, Dr. B.N. Veerappa2 1U.B.D.T. College OfEngineering, Davanagere, Karnataka, India

○ 2Dr. B.N. Veerappa, Department Of Studies In Computer Science And Engineering, U.B.D.T. College Of Engineering, Davanagere.

[7] SPEECH EMOTION RECOGNITION WITH MULTISCALE AREA ATTENTION AND DATA AUGMENTATION Mingke Xu1, Fan

o Zhang2, Xiaodong Cui3, Wei Zhang3 1Nanjing Tech University, China 2IBM Data And AI, USA 3IBM Research AI, USA

[8] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Towards real-time speech emotion recognition using deep neural networks. In

o Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on, pages 1–5. IEEE, 2015.

[9] Koteswara Rao Anne, Swarna Kuchibhotla, Acoustic Modeling for Emotion Recognition, Studies in Speech Signal Processing, Natural

a. Language Understanding, and Machine Learning, Springer Briefs in Speech Technology 2015.

[10] M. E. Ayadi, M. S. Kamel, F. Karray, ―Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases‖, Pattern

b. Recognition 44, PP.572-587, 2011.

[11] A. Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, ―Speech Emotion Recognition Using Hidden Markov Model‖, Eurospeech, 2001.

[12] C. M. Lee, S. S. Narayanan, ―Towards detecting emotions in spoken dialogs‖, IEEE transactions on speech and audio processing, Vol. 13, No. 2,

c. March 2005.

[13] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, ―Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models, ‖ in 2007

d. IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ‘07, 2007, vol. 4, pp. IV–957–IV–960.

[14] S. S. Narayanan, ―Toward detecting emotions in spoken dialogs,‖ IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2005.