A Clinical Application for Speech Emotion Recognition

|  |  |  |
| --- | --- | --- |
| Pushkar Kane  *Information Technology*  *Pune Institute of Computer Technology*  Pune, India  kanepushkar09@gmail.com | Pratik Mathe  *Information Technology*  *Pune Institute of Computer Technology*  Pune, India  pratikmathe21@gmail.com | Yash Waghumbare  *Information Technology*  *Pune Institute of Computer Technology*  Pune, India  Waghumbareyash5@gmail.com |

***Abstract*—In this age of artificial intelligence, speech-emotion recognition is crucial. AI can free up a doctor’s time to devote to providing patients with individualized care and medical services rather than transactional activities. Analysis and classification of speech signals are used in speech emotion recognition to find the underlying emotions. However, due to the complexity of emotions, which makes them challenging to perceive, the existing SER models still do not reliably understand human emotions. Despite the fact that this profession has been present for a while, new developments have brought it back into the spotlight. centered on use in medical care, which can be utilized to better the medical treatment by relating the relationship between illness and interaction and the sensations that patients and doctors are now experiencing during a visit. Using Deep Learning techniques, we will create a web application for clinical use that can identify and categorize emotions from human speech in speech and aid in better patient interpretation to improve medical treatment by tying illness and interaction together.**

***Keywords—Speech Emotion Recognition, Convolutional Neural Network, Deep Learning, MFCC***

1. INTRODUCTION

Currently, artificial intelligence (AI) is empowering a variety of medical applications, including precision medicine, breast cancer imaging diagnostics, and healthcare. There have been efforts made to help machines understand human emotions because they are an essential component of human interactions and help individuals understand one another better. AI can relieve doctors of the burden of comprehending their patients’ emotional states and move the emphasis from transactional chores to individualized medical treatment and service. However, it necessitates that computers cleverly deduce human speech and comprehend it on a semantic level. Systems for detecting the embedded emotions in voice signals are referred to as Speech Emotion Recognition (SER) systems. The primary goal of SER systems is to identify certain speaker voice traits under various emotional states. centered on use in medical care, which can be utilized to better the medical treatment by relating the relationship between illness and interaction and the sensations that patients and doctors are now experiencing during a visit. SER is the quickest means of communication and information exchange between people and computers, and it has a wide range of practical applications in the field of humancomputer interaction. It results in the expanding study area of Speech Emotion Recognition (SER), where many developments could result in enhancements in a number of fields, including spontaneous translation systems, speech-to-text synthesizer, machine-human interaction, etc.

# II. LITERATURE SURVEY

# Before the era of deep learning, for SER, researchers mostly use complex hand-crafted features and traditional machine learning methods (such as HMM, SVM, etc.) [1]. In 2014, K. Han et al.[2] proposed the first end-to-end deep learning SER model. In this they use deep neural networks (DNNs) to extract high level features from data and show that they are efficient for speech emotion recognition(SER). In this MFCC is used to extract features and audio file is classified into 5 emotions with. In[3], Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier(2016), author has considered different have considered different statistical features of pitch, energy, and ZCR (Zero Crossing Rate) MFCC (Mel frequency cepstral coefficient) from 2000 utterances of the created audio signal database. In that, Pitch feature is extracted by AMDF. Naive bayes is used to classify audio signal. In[4] Automatic speech emotion recognition using recurrent neural networks with local attention author used RNN on IEMOCAP dataset and used an raw and emotional LLD’s for feature extraction, had an accuracy of an 58 Fatemeh Noroozi [5] proposed an Vocal based emotion recognition(2017) using random forests and decision tree The average recognition accuracy rate was 66.28 In 2017, Speech Emotion Recognition[6] from Spectrograms with Deep Convolutional Neural Network worked on CNN to have The proposed CNN model consists of three convolutional layers, three fully connected layers. In [7], authors have explored deep learning model that combines temporal and spatial features They have quadrupled the RAVDESS dataset using AWGN (Additive Gaussian White Noise) for 5760 audio samples. They have built two parallel convolutional neural networks (CNN) to extract spatial features and extract temporal features, classifying emotions from one of 8 classes MFCC is widely used to analyze any speech signal and had performed well for speechbased emotion recognition systems compared to other features. In 2020 ,Mustaqeem [9] proposed CNN model with deep bidirectional LSTM that used MFCC and produced better result than co-existing model on IEMOCAP dataset. Two experiments were carried out by author to check the effectiveness of the state-of-art model of CNN and DSCNN [10] and achieved an accuracy 79.4

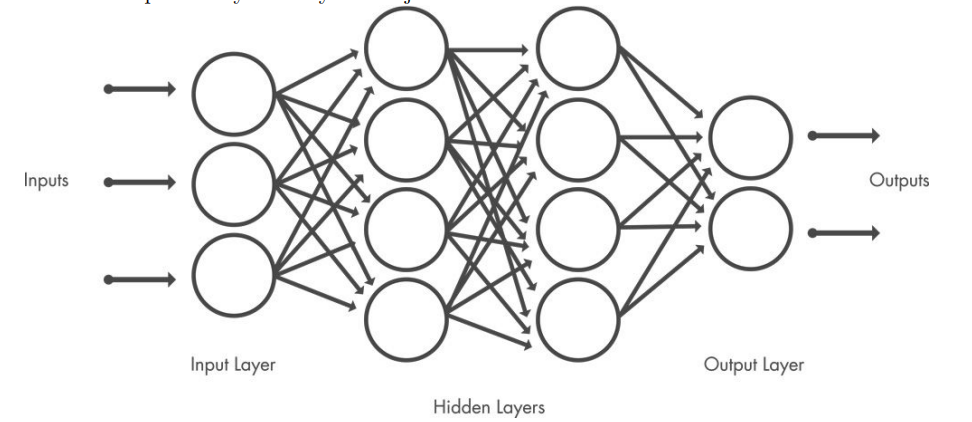
# III. PROPOSED METHODOLOGY

## A. Dataset

This dataset includes around 1500 audio file input from 24 different actors. 12 male and 12 female where these actors record short audios in 8 different emotions i.e., 1=neutral, 2 = calm, 3 = happy, 4 = sad, 5 = angry, 6 = fearful, 7=disgust, 8=surprised. Each audio file is named in such a way that the 7th character is consistent with the different emotions that they represent.

## B. CNN

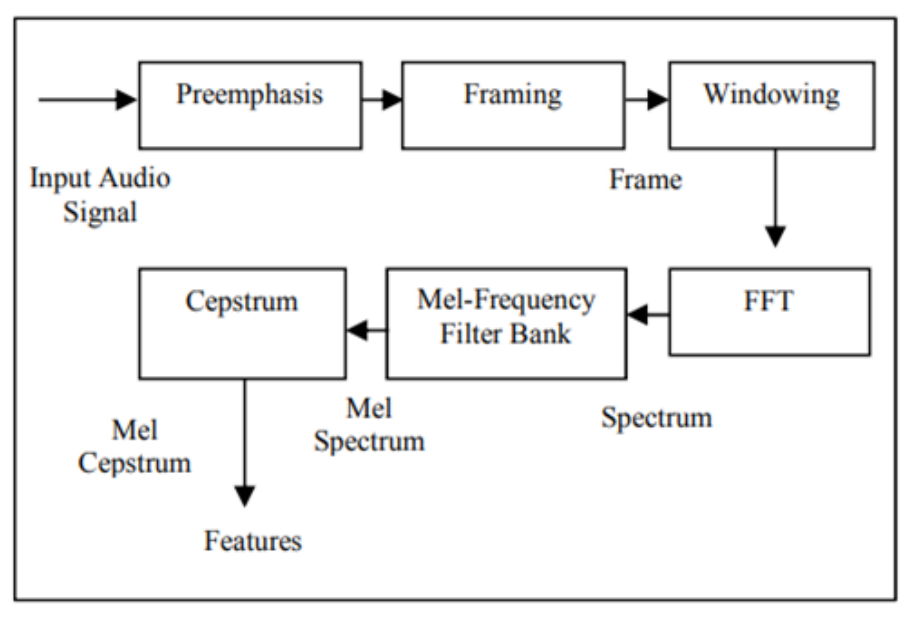
## A deep learning network architecture known as a convolutional neural network (CNN or ConvNet) learns directly from data, doing away with the requirement for human feature extraction. CNNs are very helpful for recognizing objects, faces, and scenes in photos by looking for patterns in the images. For categorizing non-image data, such as audio, time series, and signal data, they can be highly useful. Tens or even hundreds of layers can be present in a convolutional neural network, and each layer can be trained to recognize various aspects of an image. Each training image is subjected to filters at various resolutions, and the result of each convolved image is utilized as the input to the following layer. Beginning with relatively basic properties like brightness and borders, the filters can get more complicated until they reach characteristics that specifically identify the object. A CNN is made up of an input layer, an output layer, and numerous hidden layers in between, similar to other neural networks. These layers carry out operations on the data in order to discover characteristics unique to the data. Convolution, activation or ReLU, and pooling are three of the most used layers.



## 

## C. MFCC

MFCC reduces the frequency information of speech signal into the small number of coefficients which is easy to compute and extract the features Pre-emphasis is the process of the Signal with a pre-emphasis filter applied for a smoother spectral appearance. In the frame blocking procedure, the sound signal is divided into a number of short, overlapped frames with a frame size of 20 ms and a step of 20 ms between each frame. For the analysis of a segment of long signals, windowing is a necessary step. This process removes the aliasing.



# V. CONCLUSION AND FUTURE SCOPE

Successfully conducted the literature survey for selecting the most suitable algorithms for better result accuracy in the predictive analysis and proposed a methodology for the defined scope. Our model can be further extended to provide insights to clinical physicians with the field in which the necessary improvements can be made for better emotion recognition,

# REFERENCES

1. B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP’03)., vol. 2. IEEE, 2003
2. K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in Fifteenth annual conference of the international speech communication association, 2014
3. S. K. Bhakre and A. Bang, ”Emotion recognition on the basis of audio signal using Naive Bayes classifier,” 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016
4. Noroozi, F., Sapi´nski, T., Kami´nska, D. et al. Vocal-based emotion recognition using random forests and decision tree. Int J Speech Technol 20, 239–246 (2017).
5. A. M. Badshah, J. Ahmad, N. Rahim and S. W. Baik, ”Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network,” 2017 International Conference on Platform Technology and Service (PlatCon), 2017, pp. 1-5, doi:10.1109/PlatCon.2017.7883728.
6. Zhang Y YDu JWang Z Ret al. Attention based recognition 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Honolulu, USA: IEEE, 2018: 1771-1775.
7. H. S. Kumbhar and S. U. Bhandari, "Speech Emotion Recognition using MFCC features and LSTM network," *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 2019, pp. 1-3, doi:10.1109/ICCUBEA47591.2019.9129067
8. Mustaqeem, M. Sajjad and S. Kwon, ”Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM,” in IEEE Access, vol. 8, pp. 79861-79875, 2020