International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Speech Emotion Recognition (SER) through Machine Learning

Sana Fatema N. Ali[1], Prof. S. T. Khandare[2]
[1]ME Scholar , Babasaheb Naik College of Engineering, Pusad, India
[2]Associate Professor , Babasaheb Naik College of Engineering, Pusad, India

ARTICLEINFO

Article History:

Accepted: 01 March 2023
Published: 13 March 2023

Publication Issue
Volume 10, Issue 2
March-April-2023
Page Number
213-217

ABSTRACT

Emotion recognition is the part of speech recognition that is gaining more popularity and the need for it increases enormously. Although there are methods to recognize emotion using machine learning techniques, this project attempts to use deep learning and image classification method to recognize emotion and classify the emotion according to the speech signals. Various datasets are investigated and explored for training emotion recognition models are explained in this project some of the issues on the database and existing methodologies are addressed in the project. Inception Net is used for emotion recognition with the project. Inception Net is used for emotion recognition with Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets.

Keywords - Inception Net, IEMOCAP, Machine Learning.

I. INTRODUCTION

In today's computerized time, discourse signals become a method of correspondence among people and machines which is conceivable by different mechanical headways. Discourse acknowledgment strategies with philosophies signal handling procedures made prompt Discourse to-Message Speech To Text (STT) technology which is involved cell phones as a method of correspondence. Discourse Acknowledgment is the quickest developing exploration subject in which endeavors to perceive discourse signals. This prompts Discourse Feeling Recognition developing exploration points in which bunches of headways can prompt progressions in different fields like programmed interpretation frameworks, machine-to-human association, utilized in blending discourse from the text so on. Conversely, the paper concentration to study and audit different discourse extraction highlights, profound discourse data sets, classifier calculations, etc. Issues present in different subjects tended to.

A. BACKGROUND INFORMATION.

SPEECH EMOTION RECOGNITION

Speech Emotion recognition (SER) is the task of recognizing the emotional aspect of Speech irrespective of the semantic content. While humans can efficiently perform this task as a natural part of speech communication. Speech Emotional

Recognition is the way of extracting the emotions in the speech signal. Speech Emotion Recognition is a research area problem that tries to infer emotion from speech signals. Various survey state that advancement in emotion detection will make a lot of systems easier and hence make the world a better place to live. SER has its own application which is explained later. Emotion Recognition is a challenging problem in ways such as emotion may differ based on the environment, culture, and individual facial reaction leading to ambiguous findings; speech corpus is not enough to accurately infer the emotion; lack of speech database in many languages.

B. MEL FREQUENCY CEPSTRAL COEFFICIENT (MFCC)

Mel Frequency Cepstral Coefficient (MFCC) is the most popular feature extraction technique. Frequency bands are placed logarithmically here so it approximates the human system response more closely than any other system. Due to its advantage of less complexity in implementation of feature extraction algorithm, only sixteen coefficients of MFCC corresponding to the Mel scale frequencies of speech Cepstral are extracted from spoken word samples in database As shown in below figure the first step is pre-processing in which the signals are pre-processed before feature extraction. In framing the signal splits into a number of frames in time domain, then on each individual frame the hamming window is applied. Discrete Fourier Transform (DFT) is used to convert each frame from time domain to frequency domain. The filter bank is created by calculating the number of picks spaced on Mel-scale and again transforming back to the normal frequency scale. Discrete Cosine Transformation (DCT) is used to convert the Mel spectrum coefficient to the time domain.

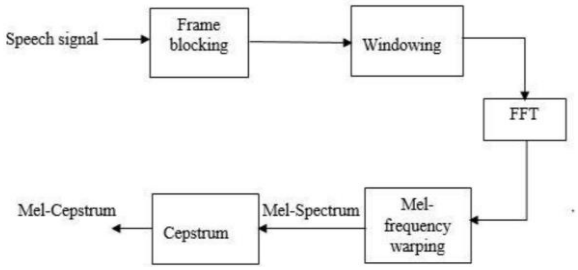

Fig 1 : Mel Frequency Cepstral Coefficient

II. LITERATURE REVIEW

Preeti Saini proposed Hindi automatic speech recognition using HTK. Isolated words are used to recognize the speech with 10 states in HMM topology which produced 96.61% [1]

Md. Akkas Ali presented automatic speech recognition technique for Bangla words. Feature extraction was done by Linear Predictive Coding (LPC) and Gaussian Mixture Model (GMM). Totally 100 words recorded 1000 times

which gave 84% accuracy [2].

Maya Money kumar, developed Malayalam word identification for speech recognition systems. The proposed work was done with syllable based segmentation using HMM on MFCC for feature extraction. [3]

Jitendra Singh Pokhariya and Dr. Sanjay Mathur introduced Sanskrit speech recognition using HTK. MFCC and two state of HMM were used for extraction which produces 95.2% to 97.2% accuracy respectively [4]

Geeta Nijhawan developed a real time speaker recognition system for Hindi words. Feature extraction done with MFCC using Quantization Linde, Buzo and Gray (VQLBG) algorithm. Voice Activity Detector (VAC) was proposed to remove the silence [5]

C. FEATURE EXTRACTION

In speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. The Feature extraction is usually performed in three stages. The first stage is called the speech analysis or the acoustic front end. It performs some kind of spectral temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals

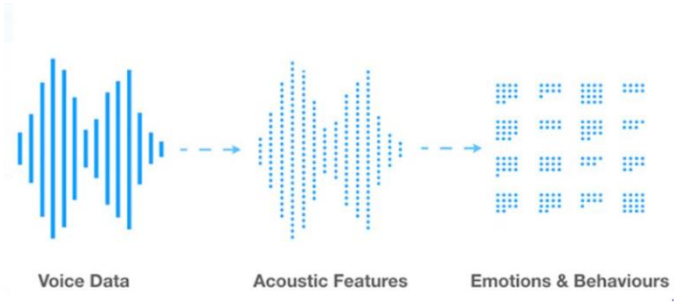

Fig 2 : Feature Extraction

The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage (which is not always present) transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer. Although there is no real consensus as to what the optimal feature sets should look like, one usually would like them to have the following properties: they should allow an automatic system to discriminate between different through similar sounding speech sounds, they should allow for the automatic creation of acoustic models for these sounds without the need for an excessive amount of training data, and they should exhibit statistics which are largely invariant across speakers and speaking environment. Automatic creation of acoustic models for these sounds without the need for an excessive amount of training data, and they should exhibit statistics which are largely invariant across speakers and speaking environments

D. RESEARCH METHODOLOGY

Inception Net v3 Model is used to build an emotion recognition model. Inception evolved from Google Net Architecture with some enhancements. The inception model is used for automatic image classification and image labeling according to the image. Inception-v3 is used for image classification in Google Image Search. Inception-v3 achieved the top 5.6% error rate in (ImageNet Large Scale Visual Recognition Challenge) ILSVRC 2012 classification challenge validation. Figure illustrates the complete architecture of the Inception Net v3 Model. Inception Net model which consists of an Inception module that concatenates all the output of 1x1, 3x3, and 5x5 filters. Inception net consists of the network in a network in a network which consists of three inception modules that are embedded inside the inception architecture which helps in the reduction of the numerical array.

Fig3: Inception Net v3 Model

E. APPLICATION

1. Emotion Recognition is used in a call center for classifying calls according to emotion.
2. Emotion Recognition serves as the performance parameter for conversational analysis thus identifying the unsatisfied customer, customer satisfaction so on.
3. SER is used in-car board systems based on information on the mental state of the driver can be provided to the system to initiate his/her safety preventing accidents to happen.

F. CONCLUSION

Different examinations and reviews about Feeling Acknowledgment, Profound learning procedures utilized for perceiving feelings are performed. It is essential for the future to have a framework like this with considerably more dependable and has vast potential outcomes in all fields. This undertaking endeavoured to utilize the initiation net for tackling feeling acknowledgment issues, different information bases have been investigated, and IEMOCAP data set is utilized as a dataset for doing my trial. Prepared my model utilizing Tensor Flow. An exactness pace of around 38% is accomplished. Later on, a constant feeling of acknowledgment can be created by utilizing similar engineering.

III. REFERENCES

[1]. Preeti Saini, Parneet Kaur and Mohit Dua et.al. "Hindi Automatic Speech Recognition Using HTK", International Journal of Engineering Trends and Technology (IJETT)", Vol.4, Issue 6, ISSN: 2231- 5381, June 2020, pp.2223-2229.

[2]. Md Akkas Ali, Manwar Hossain and Mohammad Nuruzzaman Bhuiyan et.al. "Automatic Speech Recognition Technique for Bangla Words", International Journal of Advanced Science and Technology, Vol. 50, January, 2020, pp.51-60.

[3]. Maya Money Kumar, Elizabeth Sherly and Win Sam Varghese, et.al. "Malayalam Word Identification for Speech Recognition System" An International Journal of Engineering Sciences, Special Issue iDravidian, Vol. 15 ISSN: 2229-6913(Print), December 2021, pp. 22-26.

[4]. Jitendra Singh Pokhariya and Dr. Sanjay Mathur, "Sanskrit Speech Recognition using Hidden Markov Model Toolkit", International Journal of Engineering Research & Technology (IJERT),Vol.3, Issue 10, ISSN: 2278-0181, October-2020, pp.93-98.

[5]. Geeta Nijhawan and Dr. M.K Soni, "Real Time Speaker Recognition System for Hindi Words", International Journal of Information Engineering and Electronic Business, Vol. 6, DOI: 10.5815/ijieeb.2019 .02.04, April 2020, pp. 35-4.

[6]. Vaibhav K. P., Speech Based Emotion Recognition Using Machine Learning, International Journal for Research in Applied Science and EngineeringTechnology,10.22214/ijraset.2021.3942,2021, Vol 9 (12), pp. 2093-2095.

[7]. Kogila Raghu and Manchala Sadanandam, A Perspective Study on Speech Emotion Recognition: Databases, Features and Classification Models, Treatment du signal,10.18280/ts.380631,2021, Vol 38 (6), pp. 1861-1873.

[8]. P Jothi Thilaga and S Kavipriya, Deep Learning based Speech Emotion Recognition System, Journal of University of Shanghai for Science and Technology 10.51201/just/21/121003,2021, Vol 23 (12), pp. 212-223.

[9]. Antonio Guerrieri, Eleonora Braccili and Federica Sgrò, Giulio Meldolesi et.al. Gender Identification in a Two-Level Hierarchical Speech Emotion Recognition System for an Italian Social Robot, 10.20944/preprints202112.0134.v1,2021.

[10]. Muhammad Zeeshan, Huma Qayoom and Farman Hassan et.al. Robust Speech Emotion

Recognition System through Novel ER-CNN and Spectral Features, 10.1109/isaect53699.2021.9668480,2021.

Cite this article as :

Sana Fatema N. Ali, Prof. S. T. Khandare, "Speech Emotion Recognition (SER) through Machine Learning", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 2, pp. 213-217, March-April 2023.
Journal URL : https://ijsrst.com/IJSRST231014