IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SPEECH AND EMOTION RECOGNITION USING DEEP LEARNING

[1]Dr.Senthil Kumar.M, [2]Surendar P, [3]Subhash S
[1]Associate professor, [23]UG Student,
[1]Department of Computer Science and Engineering
[1]*SRM Valliammai Engineering College, Kattankulathur, Chengalpattu district, Tamil Nadu, India*

Abstract - In recent years, emotion recognition has become a rapidly increasing research subject. Machines, unlike humans, lack the ability to perceive and express emotions. However, automatic emotion recognition can improve human-computer connection, minimising the need for human intervention. The manufacturing properties of these signals are extracted using signal processing methods. The studies use the source's instantaneous fundamental frequency (F0), the system's formants and dominant frequencies, the zero-crossing rate (ZCR), and the combined features signal energy. Zero-frequency filtering (ZFF) is used to produce F0, while LP spectrum is used to obtain formants and dominating frequencies. A rectangular window of 200 samples is used to derive short-time signal energy (STE) and ZCR in the voiced and unvoiced regions. The main goal of SER is to improve the human-machine interface. It can also be utilised in lie detectors to check a person's psychophysiological condition. Speech emotion recognition has recently found applications in health and forensics. In this work, pitch and prosody traits are used to distinguish seven emotions. The majority of the speech features used in this study are time-domain features. The emotions were classified using a Long Short Time Memory (LSTM) classifier.

Index Terms **– Speech Emotion Recognition,Long Short Time Memory , Mel-Frequency Cepstrum.**

I. INTRODUCTION

In today's digital era, speech signals become a one of the mode of communication between humans and machines which is possible by various types of technological advancements. Speech recognition techniques with methodologies signal processing techniques made leads to Speech-to-Text (STT) technology which is used mobile phones as a mode of communication. Speech Recognition is the fastest growing research topic in the generation which attempts to recognize speech signals. This leads to Speech Emotion Recognition (SER) growing research topic in which lots of advancements can lead to advancements in various field like automatic translation systems, machine between to human interaction, used in synthesizing speech from text so on. In contrast the paper focus to survey and review various include such as speech extraction features, emotional speech databases, classifier algorithms and so on. Problems present in various topics were addressed. This paper is organized as follows. Section 2 describes background information about speech recognition, emotion recognition system, applications of emotion recognition. Section 3 explains the methods of feature extraction and optimization from speech signals. compares various speech emotional databases prepared for research. contains various classifier algorithms for classifying speech signals according to the emotion inferred. Finally, a conclusion is given

II. RELATED WORKS

Nithya Roopa S et.al [1] has suggested to recognize speech and emotion using deep learning even though there are many technique available in the Machine learning concepts. Various database has been studied to identify the emotion using speech signals. The used technology is Inception Net for Emotion recognition. The datasets used were IEMOCAP but the accuracy the emotion recognition using speech signals is 35%. Ting – Wei Sun has proposed that even though there are many techniques used for speech and emotion recognition, the accuracy of the output depends on the speech acoustic features that are being selected. It incorporates speaker gender information and does not rely on any speech acoustic factors. The project want to take advantage of the abundant information in speech raw data without using any artificial means. In general, classical acoustic parameters must be manually selected as classifier input for emotion classification in speech emotion recognition systems. To achieve emotion recognition, the network uses deep learning algorithms to automatically select significant information from raw voice signals for the classification layer. It can prevent the omission of emotional data that can't be directly quantitatively described as a speech acoustic characteristic. To boost recognition accuracy even more, we include speaker gender information to the suggested system. Mao Li et al has suggested that the absence of publicly available large-scale labelled datasets has long plagued SER. We investigate how unsupervised representation learning on unlabeled datasets can help SER to overcome this challenge. We show that the contrastive predictive coding (CPC) approach may learn salient representations from unlabeled datasets, resulting in better emotion recognition. This technique obtained state-of-the-art concordance correlation coefficient (CCC) performance on IEMOCAP for all emotion

primitives (activation, valence, and dominance) in our studies. Furthermore, when compared to baselines, our technique significantly improved performance on the MSPPodcast dataset. The author in [5] has proposed that A neural network (NN), also known as an artificial neural network (ANN) or simulated neural network (SNN) in the case of artificial neurons, is a connected group of natural or artificial neurons that processes information using a mathematical or computational model based on a connectionistic approach to computation.

III. MATERIALS AND METHODS

This section explains for the proposed methodology, emotion database used for research, Inception model.

A. EMOTION FOR DATABASE

The Speech Analysis and Interpretation Laboratory (SAIL) has created a new crop called "interactive emotional dyadic motion capture database" (IEMOCAP), which is used in this paper. Because this data is rarely used, this initiative digs deeper into it. Corpus Data is made up of ten actors in dyadic sessions who have markers on their faces, heads, and hands that offer extensive information about their facial expression and hand movements during planned and spontaneous spoken communication scenarios. There are twelve hours of audio-visual data in the database. This module selects audio snippets from a variety of sessions. These 10-second audio samples were categorised into one of the emotions classes by various annotators. The audio-visual data is separated into five sessions, with audio in.wav format and video in.mp4 format. Various annotators classify the actor's emotions into seven categories during the data collection sessions. The database comes with all of the data.

TRANSFORMATION LEARNING

Transfer learning is a machine learning model that applies speech emotion knowledge learned from one problem to another. Transfer learning clearly solves a large number of issues in a short amount of time. When it comes to reducing computing costs and achieving accuracy with less training, Transfer Learning is used.

B. INCEPTION NET V3 MODEL

Inception Net v3 To create an emotion recognition model, a model is used. Inception is a modified version of the GoogLeNet Architecture. The model is used to automatically classify and identify images based on their content. Google Picture Search uses Inception-v3 for image classification. In the ILSVRC 2012 classification challenge validation, Inception-v3 had the best error rate of 5.6 percent. Depicts the entire architecture of the Inception Net v3 Model. The Inception Net model consists of an Inception module that concatenates all of the output from the 1x1, 3x3, 5x5 filters.

IV. IMPLEMENTATION

In the section which contains many explanations about experimental setup, libraries using for the Deep learning which helps in emotion recognition

A. SYSTEM SETUP

For performing the pratical experimentally I"ve used system setup consist of Corei7 6th Generation 3.7 GHz Processor,Samsung SSB of 512 GB memory space,NVIDIA GeForce GT 730 2GB GPU Card with Ubuntu 16.04 installed. For deep learning I"ve used Tensor Flow 1.5 for implementing the Inception net model and Tensor Board for visualizing the learning, graphs, histograms and so on.

B.TRAINING METHOD

All images labeled with respective emotions are prepared for training the model. The proposed CNN model was implemented using TensorFlow. The spectrogram images were generated from the IEMOCAP are resized to 500 x 300. More than 400 spectrograms were generated from all the audio files in the dataset. For each emotion, Image range of about 500 for each class of emotion is collected from the corpus database. The training process was run for 20 epochs with a batch size set to 100. Initial learning rate was set to 0.01 with a decay of 0.1 after each 10 epochs. Training data model was performed on a single NVidia GeForce GT 730 with 2 GB onboard memory. The training took around 35 minutes and the best accuracy was achieved after 28 epochs. On the training set, a loss of 0.71 was achieved, whereas 0.95 loss was recorded on the test set. An accuracy of 35.95 % was achieved per spectrogram. It is important to notice here that the overall accuracy is very low. These may be due to transfer learning used and less dataset for each class of emotion

V RESULTS AND DISCUSSION

An accuracy rate of about 35.6% is achieved from the data model for predicting. The emotions. It is evident that 0.8 is the highest accuracy rate achieved during validation of data. Some of the reason for less accuracy rate are, Transfer Learning is used to train the model, there could've been less spectrograms used for training, which leads to the less accuracy. There are also less data set used for the training process which also leads to the case.

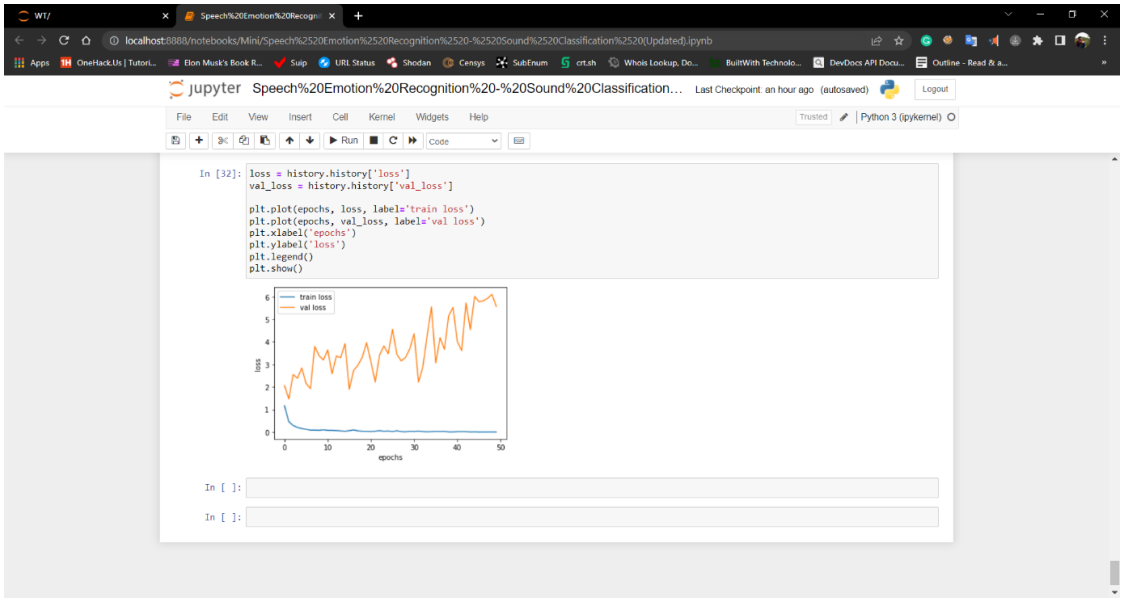

Fig.1 Loss in prediction

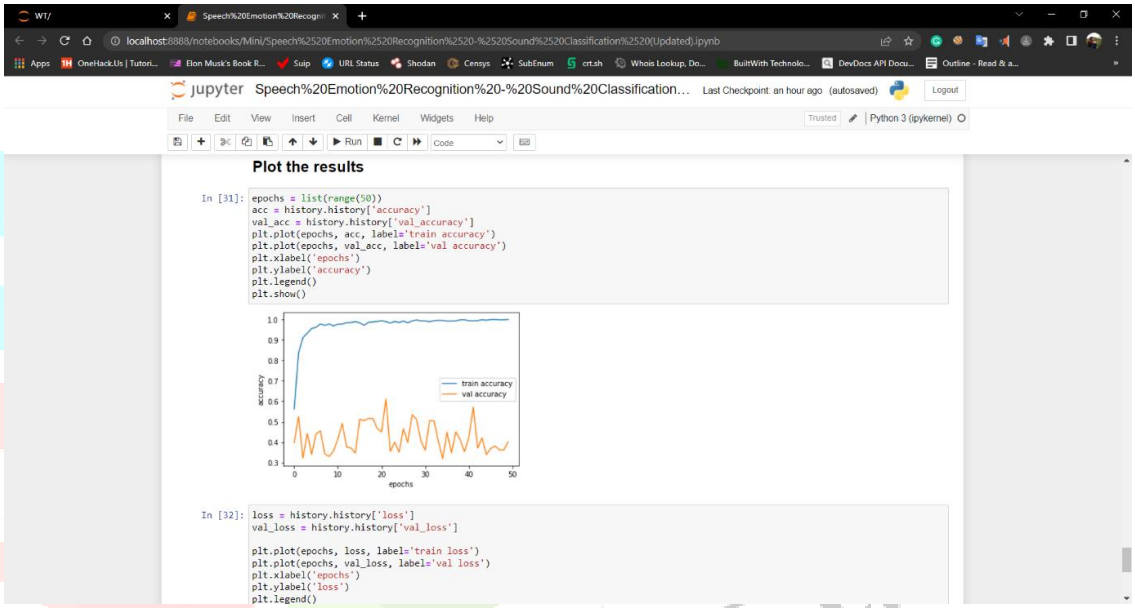

Fig 2 Accuracy of the prediction

VI CONCLUSION

Various investigations and surveys about Emotion Recognition, Deep learning techniques used for recognizing the emotions are performed. It is necessary in future to have a system like this with much more reliable, which has endless possibilities in all fields. This project attempted to use inception net for solving emotion recognition problem, various databases have been explored, IEMOCAP database is used as dataset for carrying out my experiment. Trained my model using TensorFlow. Accuracy rate of about 38% is achieved. In future, real time emotion recognition can be developed using the same architecture.

REFERENCES

[1] TING-WEI SUN "End-to-End Speech Emotion Recognition With Gender Information", Digital Object Identifier 10.1109/ACCESS.2020.301746 – 2020

[2] Nithya Roopa S., Prabhakaran M, Betty.P "Speech Emotion Recognition using Deep Learnin", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4S, November 2018

[3] Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke , Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, Chao Wang "CONTRASTIVE UNSUPERVISED LEARNING FOR SPEECH EMOTION RECOGNITION"

[4] Dr. M.Senthil Kumar, "An automated Neuro model for software effort estimation and RMMI using competitive Learning" in International Journal of Computer Technology and application, Published in: Vol 3, No 6. Publication year: 2012 Page No: 2060 - 2065, ISSN: 2229-6093.

[5] Suraj Tripathi, Abhay Kumar , Abhiram Ramesh, Chirag Singh , Promod Yenigalla "Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions"

[6] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-Text and Speech-to-Speech Summarization," vol. 12, no. 4, pp. 401–408, 2004.

[7] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011.
[8] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[9] J. Schmidhuber, "Deep Learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85–117, 2015. Speech Emotion Recognition using Deep Learning 250 Published By: Blue Eyes Intelligence Engineering & Sciences Publication Retrieval Number: E1917017519 5. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," Proc. 28th Int. Conf. Mach. Learn., pp. 689–696, 2011.
[10] F. Dipl and T. Vogt, "Real-time automatic emotion recognition from speech," 2010. 7. S. Lugovic, I. Dunder, and M. Horvat, "Techniques and applications of emotion recognition in speech," 2016 39th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2016 - Proc., no. November 2017, pp. 1278–1283, 2016.
[11] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture," Acoust. Speech, Signal Process., vol. 1, pp. 577–580, 2004.
[12] S. G. Koolagudi and S. R. Krothapalli, "Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features," Int. J. Speech Technol., vol. 15, no. 4, pp. 495–511, 2012.
[13] J. Rong, G. Li, and Y. P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, 2009.
[14] F. Noroozi, N. Akrami, and G. Anbarjafari, "Speech-based emotion recognition and next reaction prediction," 2017 25th Signal Process. Commun. Appl. Conf. SIU 2017, no. 1, 2017.
[15] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649