Speech Emotion Recognition using Deep Learning

Nithya Roopa S., Prabhakaran M, Betty.P

Abstract: Emotion recognition is the part of speech recognition which is gaining more popularity and 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 model are explained in this paper. Some of the issues on database, existing methodologies are addressed in the paper. Inception Net is used for emotion recognition with the paper. Inception Net is used for emotion recognition with IEMOCAP datasets. Final accuracy of this emotion recognition model using Inception Net v3 Model is 35%(~).

Index Terms: speech recognition; emotion recognition; automatic speech recognition; deep learning; image recognition; speechtechnology; signal processing; image classification

I. INTRODUCTION

IN today’s digital era, speech signals become a mode of communication between humans and machines which is possible by various technological advancements. Speech recognition techniques with methodologies signal processing techniques made leads to Speech-to-Text (STT) technology[1] which is used mobile phones as a mode of communication. Speech Recognition is the fastest growing research topic in 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 to human interaction, used in synthesizing speech from text so on. In contrast the paper focus to survey and review various 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. Section 4 compares various speech emotional databases prepared for research. Section 5contains various classifier algorithms for classifying speech signals according to the emotion inferred. Finally, a conclusion is given in section 6.

II. BACKGROUND INFORMATION

A. SPEECH RECOGNITION

Speech Recognition is the technology that deals with techniques and methodologies to recognize the speech from the speech signals. Various technological advancements in the field of the artificial intelligence and signal processing techniques, recognition of emotion made easier and possible. It is also known as ‘Automatic Speech Recognition’. It is found that voice can be next medium for communicating with machines especially when computer-based systems. A Need for inferring emotion from spoken utterances increases exponentially. Since there is an enormous development in the field of Voice Recognition. There are many voice products has been developed like Amazon Alex, Google Home, Apple HomePod which functions mainly on voice-based commands. It is evident that Voice will be the better medium for communicating to the machines.

B. EMOTION RECOGNITION

The Basically, Emotion Recognition deals with the study of inferring emotions, methods used for inferring. Emotion can be recognized from facial expressions, speech signals. Various techniques have been developed to find the emotions such as signal processing, machine learning, neural networks, computer vision. Emotion analysis, Emotion Recognition are being studied and developed all over the world. Emotion Recognition is gaining its popularity in research which is the key to solve many problems also makes life easier. The main need of Emotion Recognition from Speech is challenging tasks in Artificial Intelligence where speech signals is alone an input for the computer systems. Speech Emotion Recognition (SER) is also used in various fields like BPO Centre and Call Centre to detect the emotion useful for identifying the happiness of the customer about the product, IVR Systems to enhance the speech interaction, to solve various language ambiguities and adaption of computer systems according to the mood and emotion of an individual.

C. SPEECH EMOTION RECOGNITION

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

Revised Version Manuscript Received on 25 November, 2018.

Nithya Roopa S, Assistant Professor, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India.

Prabhakaran M, PG Scholar, Kumaraguru College of Technology, Coimbatore. Tamilnadu, India.

Betty. P, Assistant Professor, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India.

Survey on speech emotion recognition [2] which helps a lot in exploring speech emotion recognition

D. DEEP LEARNING

Deep Learning [3] is machine learning techniques which data models are designed bound to a specific task. Deep learning in neural networks is used for various tasks such image recognition, classification tasks, decision making, pattern recognition etc. [4] Various other Deep Learning techniques such as multimodal deep learning used for feature extraction, image recognition made at ease[5].

E. APPLICATIONS OF EMOTION RECOGNITION

Emotion Recognition is used in call center for classifying calls according to emotions[6]. Emotion Recognition serves as the performance parameter for conversational analysis[7] thus identifying the unsatisfied customer, customer satisfaction so on. SER is used in-car board system based on information of the mental state of the driver can be provided to the system to initiate his/her safety preventing accidents to happen[8].

III. LITERATURE SURVEY

Complete review on the speech emotion recognition is explained in [9] which reviews properties of dataset, speech emotion recognition study classifier choice. Various acoustic features of speech are investigated and some of the classifier methods are analyzed in [10] which is helpful in the further investigation of modern methods of emotion recognition. This paper [11] investigated the prediction of the next reactions from emotional vocal signals based on the recognition of emotions, using different categories of classifiers. Some of the classification algorithms like K-NN, Random Forest are used in [11] to classify emotion accordingly. Recurrent Neural network arises enormously which tries to solve many problems in the filed of data science. Deep RNN like LSTM, Bi-directional LSTM trained for acoustic features are used in [12]. Various range of CNN are being implemented and trained for speech emotion recognition are evaluated in [13]. Emotion is inferred from speech signals using filter banks and Deep CNN[14] which shows high accuracy rate which gives an inference that deep learning can also be used for emotion detection. Speech emotion recognition can be also performed using image spectrograms with deep convolutional networks which is implemented in [15].

IV. PROPOSED METHODOLOGY

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

A. EMOTION DATABASE

IEMOCAP corpus Database [16] is prepared by the Speech Analysis and Interpretation Laboratory (SAIL), at the University of Southern California (USC) a new corpus named “interactive emotional dyadic motion capture database” (IEMOCAP) is used in this paper. Since this data is rarely used, so this project explores more on this dataset. Corpus Data consists of ten actors in dyadic sessions with markers on the face, head, and hands, which provide detailed information about their facial expression and hand movements during scripted and spontaneous spoken communication scenarios. The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions (happiness, anger, sadness, frustration and neutral state). Database consist of twelve hours of audio-visual data. I’ve choose audio clips of various sessions. Based on certain annotators, these audio clips of 10 seconds(approx.) are classified into one of the emotions classes. All the audio-visual data is divided into five sessions audio data in .wav format and video data in .mp4 format. During the sessions of capturing the data, actor’s emotions is evaluated by various annotators into seven range of emotions. All the data are given along with the database.

B. TRANSFER LEARNING

Transfer learning is one of the machine learning models which uses the knowledge gained from solving one problem is incorporated to solve another problem. It is evident that Transfer learning solves many problems within short interval of time. Transfer Learning is incorporated whenever there is any need to reduce computation cost, achieve accuracy with less training

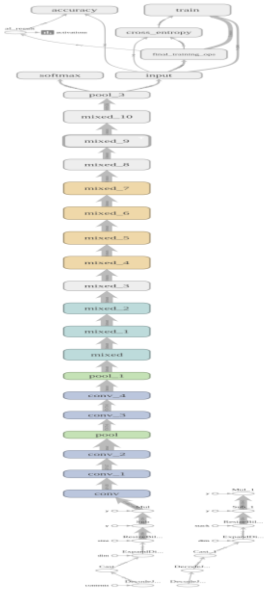

Fig 1. Inception Net Architecture

C. INCEPTION NET V3 MODEL

Inception Net v3 [17] Model is used to build an emotion recognition model. Inception is evolved from GoogLeNet Architecture with some enhancements. Inception model is used for automatic image classification and image labelling according to the image. Inception-v3 is used for image classification in Google Image Search.

Inception-v3 achieved top 5.6% error rate in ILSVRC 2012 classification challenge validation.

Figure 1 which illustrates the complete architecture of Inception Net v3 Model. Inception Net model which consist of Inception module which concatenates all the output of 1x1,3x3,5x5 filters.

Inception net consist of network in a network in a network which consist of three inception modules that are embed inside the inception architecture which helps in reduction of numerical array.

D. PREPARATION OF TRAINING DATASET

All the audio clips from the IEMOCAP databases are pulled out from various sessions. Using the emotion evaluation report which is given along with the database, various wav files are labeled and categorized into seven range of emotions as mentioned earlier. Speech signals in .wav format are converted into spectrogram images within the emotion class.

V. EXPERIEMENTAL SETUP

This section which contains explanations about experimental setup, libraries used for Deep learning which helps in emotion recognition.

A. SYSTEM SETUP

For performing the experiment I've used system setup consist of Core i7 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.

VI. RESULT & ANALYSIS

An accuracy rate of about 35.6% is achieved from the data model for predicting the emotions. It is evident from the below figure that 0.8 is the highest accuracy rate achieved during validation of data.

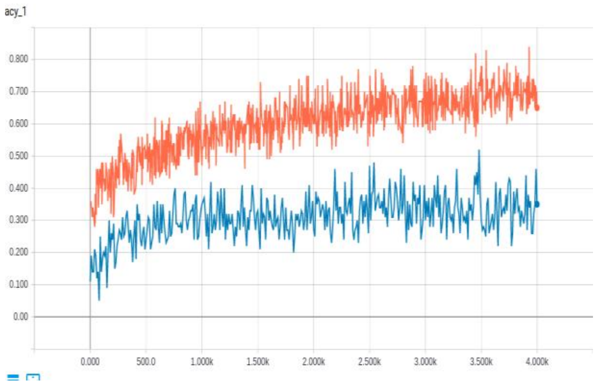

Fig.2 Accuracy rate of the Data Model

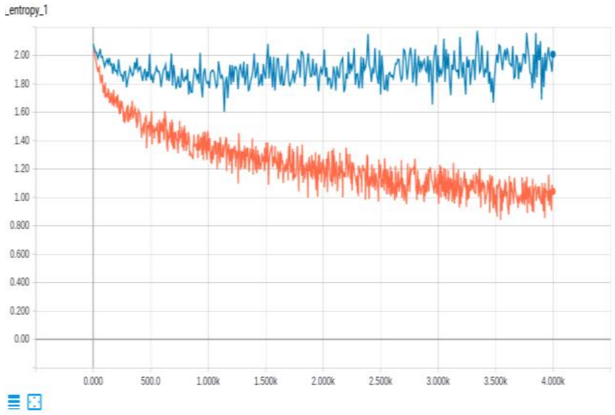

Fig.3 Cross Entropy

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.

VII. 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. 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.
2. 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.
3. Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
4. J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

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.
6. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. C.-W. Huang and S. S. Narayanan, “Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition,” pp. 1–19, 2017.
14. H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition,” *Neural Networks*, vol. 92, pp. 60–68, 2017.
15. A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, “Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network,” *2017 Int. Conf. Platf. Technol. Serv.*, pp. 1–5, 2017.
16. C. Busso *et al.*, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
17. C. Szegedy, V. Vanhoucke, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” 2014.