International Journal of Research in Engineering, Science and Management
Volume 4, Issue 3, March 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

End-to-End Speech Emotion Recognition Using Deep Learning

Ajai Jose Jacob[1*], Aswin A. Jacob[2], Ashly Mathew[3]
[1,2,3]*Department of Computer Applications, Saintgits College of Applied Sciences, Kottayam, India*

***Abstract*: The German Corpus database was used for testing the effectiveness of this technique. The database is used to test speech emotion recognition through deep neural networks. The technique uses convolutional pooling and fully connected layers. The database contains the audio recordings of ten actors of which five are male and other five are females, they contain seven emotional state of every actor of which only three emotional states are used. The audio recordings obtained from the database are divided into segments that are of 20 milliseconds, these segments may contain some that are empty which is identified by Voice Activity Detector(VAD) and removed. The remaining segments are divided into Training, Validation and Testing. Deep neural network is enhanced using stochastic gradient descent. After completing the experiment, the result obtained showed around 96% accuracy.**

***Keywords*: Deep neural networks, DNN architecture, Voice Activity Detector (VAD), Deep learning.**

1. Introduction

Significant amount of growth has been made in the field of Speech Recognition in the past few years, yet there is a need for a proper computer understanding of human emotions which may lead to further improve systems that support the interaction between humans and machines. Speech emotion recognition is primarily used to improve interactions between humans and machine.

The perfect way to reach this goal, as the modes are showing, might be to generate an end-to-end learning algorithm, that is able to process original input signals directly resulting in desired performance that can reduce the amount of work that needs to be done as well as reduce the knowledge required for the operation. Hence, in this article we Look into the possibilities in unison with this idea.

In the present time we are at a Deep learning revolution because during this short time it has improved the quality of many cutting edge domains which includes SR. An advantage of this method is that it lets us use complex multilayer models, these multilayer models represent data with multiple layers of abstraction. Both convolutional and recurring nets are used in this method hence Speech emotion recognition is not considered to be an exception.

2. Existing Methodology

Many of the normal image processing systems are deficient in detecting lung carcinomas at early stages of development. The primitive methods which are used in analysis of lung cancer are poor in providing precise or sufficient results which may result in more fatalities every year. The existing designs use simple Computed Tomography images in detection of lung cancers which are deficient as they are just two-dimensional images which makes it difficult for professionals to analyse, detect or calculate cancerous cells in the lungs. The extraction of features from a CT scan image is limited up to a certain extent. This system lacks in a number of features. It is strenuous to detect small malign nodules from a big picture of the whole lungs. So, we propose use of Convolutional Neural Network(CNN) using deep learning in providing a detailed picture of CT scan to improve the accuracy in detection of lung carcinomas at early stages.

3. Proposed Methodology

In this method a set of audio files are selected with different emotion state every audio file is split into chunks of 20 milliseconds with no overlapping. In the next step we separate the silent segments identified by VAD (Voice Activity Detector) and divide the segments into Training, Validation and Testing sets. Because there needs to be an even distribution of segments from each emotional state, segments are chosen for the experiment so that they constitute equal amount of segments from every class.

The data set used for this paper was German Corpus (Berlin Database of Emotional Speech). The data set contains about 800 sentences this includes 7 emotion classes which was recorded from 5 males and 5 female actors. Each actor was made to say 10 different sentences from each emotion class and some second versions. These recording of these sentences were done in an anechoic chamber which made use of high-quality machines and recorded the sentences in a sampling frequency of 48kHz which was later down-sampled to 16kHz.

In order to get comparable results and due to the fact that all classes are not equal we consider the sentences of all the actors but only 3 emotional classes are considered which are angry, sad and neutral. The removal of silent parts from the audio files are done using Google WebRTC Voice Activity Detector (VAD) which is incorporated into the preprocessing. All the audio files are the standardized so that they have zero mean and

*Corresponding author: ajaijose007@gmail.com

unit variance.

The audio files are split into segments of 20 milliseconds that have no overlap-vectors, it has a length of 320(16*20) which totals to a number of 39052 segments. The next step is to remove the silent segments which are identified by VAD, 098 segments are found to be silent and removed. The segments that remains after this process is divided into sets such as Training which is 79.56%, Validation which is 9.84% and Testing which is 10.60%. Since we have to acquire an even distribution of classes among segments we adjust the segments according to the class that has the smallest size, since we have done this only 73.86% of the non-silent segments are used in the method. This totals to a number of 21129 training segments in which 7043 segments in every class, 2613 validation segments 871 per class. The remaining segments were used as testing segments, it contained 2814 segments from 33 audio files (Eleven angry, Twelve neutral, Ten sad). The segments used were neither seen during testing or validation by the DNN.

DNN architecture:

The first two layers used is convolutional layer, it contains thirty-two kernels which are of the size 7x1 succeeded by an average pooling layer. In the third and the fourth layer also we use convolutional layer that has thirty-two kernels of size 13 which is similar to the previous layer succeeded with average pooling. The last two convolutional layers has 16 kernels, similar to the previous layer it has a size of 13x1. Next we divide the network into two sections, both these sections have only one pooling layer one of which has average pooling and the other has maximum pooling. which was later flattened and concatenated back to the main branch. Now after these processes the DNN contains only fully connected layers, the first of which has a size of 480, the second one has a size of 240 and the last one is an output Softmax layer which has three output neurons. The pooling layers used were having a pool size of 2. Zero padding was not performed since the border mode of each convolutional networks was set to ‘valid’.

4. Conclusion

This article has provided a detailed review of the deep learning techniques for SER. The Deep learning techniques such as DBM, RNN, DBN, CNN, and AE have been the content of much research in recent years. These deep learning techniques and their layer-wise architectures are briefly detailed based on the categorization of different natural emotion such as happiness, joy, sadness, neutral, surprise, boredom, disgust, fear, and anger. These methods give easy model training as well as the efficiency of shared weights.

The Limitations of Deep learning techniques include their large layer-wise internal architecture, less efficiency for temporally differing data to be inputted and over-learning during memorization of layer-wise information. These research works form a stand to evaluate the performances and limitations of currently used Deep learning techniques. Moreover, it highlights some promising directions for better future SER systems.

References

[1] Pavol Harar, Radim Burget and Malay Kishore Dutta, “Speech Emotion Recognition with Deep Learning,” 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), 2017.

[2] El Ayadi, M., Kamel, M.S. and Karray, F., Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), pp. 572-587, 2011.

[3] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 5200-5204.

[4] LeCun, Y., Bengio, Y. and Hinton, G., “Deep learning,” in Nature, 521(7553), pp. 436-444, 2015.

[5] R. Chakraborty and S. K. Kopparapu, "Improved speech emotion recognition using error correcting codes," 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Seattle, WA, USA, 2016, pp. 1-6.

[6] W. Fei, X. Ye, Z. Sun, Y. Huang, X. Zhang and S. Shang, "Research on speech emotion recognition based on deep auto-encoder," 2016 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), Chengdu, China, 2016, pp. 308-312.

[7] Chandni, G. Vyas, M. K. Dutta, K. Riha and J. Prinosil, "An automatic emotion recognizer using MFCCs and Hidden Markov Models," 2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, Czech Republic, 2015, pp. 320-324.