Third International Conference on "Materials, Computing and Communication Technologies"
International Journal of Scientific Research in Computer Science,
Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Emotion Recognition from Speech Using Deep Learning

Anusha V, Mary Little Flower T
Department of Electronics and Communication Engineering, St. Xavier's Catholic College of Engineering, Chunkankadai, Nagercoil, Tamil Nadu, India

Article Info
Volume 8, Issue 7
Page Number : 121-123

Publication Issue :
May-June-2022

Article History
Accepted: 01 June 2022
Published: 20 June 2022

ABSTRACT

Speech Emotion Recognition (SER) plays a major and challenging research in human computer inter- actions. Speech is the expression of ideas and thoughts by means of articulate vocal sounds and the fastest, most natural method of communication between humans. Emotion is a strong feeling deriving from one's circumstances, mood, or relationships with others. Speech Emotion Recognition (SER) is the task of recognizing the emotion from speech signal of the speaker. In recent years, deep neural networks have been used in determining emotional states. Inspired by this, Convolution Neural Network (CNN) with seven layer model is proposed. In this work, Berlin speech emotion database having seven emotional states is used. Mel spectrograms were extracted from the speech signal using Fast Fourier Transform (FFT) and the emotions are classified using Convolution Neural Network (CNN). The proposed model is trained on Mel spectrograms obtained from Berlin emotion dataset with 70% for training and tested with 30% from the same dataset. This method achieves the speech emotion recognition accuracy of 83.89%.

Index Terms—Emotions, Melspectrogram, Convolutional Neu- ral Networks, Berlin emotion dataset

I. INTRODUCTION

Speech signal is the most natural way to communicate among human beings. Researchers are constantly working to apply this mode in the domain of human-machine in- teraction. However, it require machines to interpret human spoken phrases intelligently and understand it semantically [1]. Emotional intelligence (EI) is the ability of a tool to rec- ognize a persons emotions. . Speech emotion recognition (SER) defines the process of recognizing human emotions from speech with its infuential afective states [2]. Speech emotion recognition plays an important role in the HMI. As well known, thanks to recent advances in accurate speech recognition and replenishing wide availability of speech recog- nition devices, the speech emotion recognition (SER) has

made great progress in the last score years since researchers are increasingly engaged on the SER experiments. In spite of that nevertheless, owing to taking account the speakers' sexuality, age even physical state in the process of the SER, it's very difficult to draw any affective information and specific emotion from voice of an individual. In order to conquer various situation, researchers have run amounts of trials in the each stage of SER, such as the stage of signal preprocessing, features extracted and classification [3]. The emotional state is an important element in the interactions of human beings. It influences many aspects of communication such as facial expressions, voice characteristics, and semantic contents [4]. Speech signal processing has been revolutionized by deep learning. More and more researcher achieved excellent results in certain applications using deep belief networks (DBNs), convolutional neural networks [5]. In this paper, we will utilize Berlin emotion dataset for the speech signal. From the speech signal melspectrogram is extracted and the extracted melspectrogram is given as input to the Convolution Neural Network. The experimental result shows that the CNN classi- fies the seven emotions(sad,happy,fear,angry,neutral,boredom and disgust) effectively.

The remainder of the paper is organized as follows. Section II presents the related work. The proposed method is described in Section III. The obtained results are reported in Section IV. Conclusions are drawn in Section VI

II. RELATED WORK

Abdul Malik et.al [1] proposed a method in which the emotion is recognized from speech using spectrograms and deep CNN. From spectrogram images a discriminative features are extracted by three convolutional layers and three fully con- nected layers. The outputs are predicted for seven emotions. The accuracy for this set is quite accurate. In second set to determine the suitability of transfer learning a fine-tuned pre trained AlexNet mode is used. The accuracy for second set is not quite accurate. Different aspects of the feature extraction, content representation and classification are analysed.

A. Christy et.al [2] developed a multimodal speech emotion recognition and classification method using CNN techniques. Emotion recognition is an emerging application in the field of Artificial intelligence. A person emotion can be identified with the tone and pitch of the voice. The algorithms like linear regression, decision tree, random forest, SVM and CNN are used for classification and prediction once relevant features. The Ryerson Audio-Visual Database of emotional speech and song is used for the dataset. Classification is performed with decision tree, random forest and support vector machine and the accuracy is calculated.

George Trigeorgis et.al [6] proposed a solution for the prob- lem of 'Context-aware' emotion relevant feature extraction by using Convolutional Neural Network with LSTM method. Convolutional model operates on the raw signal, to perform an end-to-end spontaneous emotion prediction task from the speech data. Time-continuous prediction of spontaneous and natural emotions is tested on speech data. RECOLA database is used for input speech sample. The method achieves better performance in comparision to traditional designed on the RECOLA database. The network learns an intermediate leads to improved performance.

Jianfeng Zhao et.al [7] developed a method to learn high- level emotional features from the raw audio clips and log-mel spectrogram to recognise speech emotion and three deep CNN architectures is designed. The merged CNN is evaluated on two different public emotional speech databases that is Berlin Emo DB and IEMOCAP. The emotional utterances of the selected databases are all performed by ten speakers, the data of eight subjects are chosen as the training set, the training data of other two subjects are chosen as the testing set. The results

show that the designed CNN architectures can learn hierarchical features and model high-level abstractions of the emotional information from the raw audio clips and log mel spectrograms.

Shiqing Zhang et.al [8] proposed a multiscale deep con- volutional LSTM framework for spontaneous speech emo- tion recognition. For the deep segment level features is ob- tained challenging spontaneous emotional speech datasets i.e AFEW5.0 and BAUM-1s databases are used. The effect of LSTM architectures and the performance of different lengths of Mel-Spectrograms are analysed. The result is compared with the state of the arts and the performance is measured. A different emotion recognition results are obtained by com- baining CNN with LSTM at spectrograms. Finally, different emotion recognition results, obtained by combining CNN with LSTM at multiple lengths of segment-level spectrograms, are integrated by using a score-level fusion strategy. Experimental results on two challenging spontaneous emotional datasets, i.e., the AFEW5.0 and BAUM-1s databases, demonstrate the promising performance of the proposed method, outperform- ing state-of-the-art methods.

III. METHODOLOGY

Berlin emotional database is used for the input speech sample. By using Fast Fourier Transform the Mel spectrogram is extracted and fed as input to Convolutional Neural Network for the classification process. Typical Speech Emotion Recog- nition is composed of two main portions. A processing unit that extracts the most suitable features from speech signals. Classifier to recognize the hidden emotions in speech using the extracted features. Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data. The Classification is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. The proposed block diagram for emotion recognition from speech is shown in Figure 1.