Proceedings of the 2nd International Conference on Communication and Electronics Systems (ICCES 2017)
IEEE Xplore Compliant - Part Number:CFP17AWO-ART, ISBN:978-1-5090-5013-0

Emotion Recognition from Speech using Convolutional Neural Network with Recurrent Neural Network Architecture

Saikat Basu
Computer Science and Engineering
MAKAUT, WB
Member IEEE
School of Medical Science and Technology
Indian Institute of Technology Kharagpur
Kolkata, India
Email: saikat.basu@smst.iitkgp.ernet.in

Jaybrata Chakraborty
Computer Science
Brainware University
Kolkata, India
Email: jaybrata1411@gmail.com

Md. Aftabuddin
MAKAUT, WB
Kolkata, India
Email: md_aftabuddin@yahoo.com.au

***Abstract*—Recognition of emotion is always a difficult problem, particularly if the recognition of emotion is done by using speech signal. Many significant research works have been done on emotion recognition using speech signal. The primary challenges of emotion recognition are choosing the emotion recognition corpora (speech database), identification of different features related to speech and an appropriate choice of a classification model. In this article we use 13 MFCC (Mel Frequency Cepstral Coefficient) with 13 velocity and 13 acceleration component as features and a CNN (Convolution Neural Network) and LSTM (Long Short Term Memory) based approach for classification. We chose Berlin Emotional Speech dataset (EmoDB) for classification purpose. We have approximately 80 percent of accuracy on test data.**

Keywords:- MFCC, CNN, LSTM, Emotion recognition from speech.

I. Introduction

Human machine interaction are widely used nowadays in many applications. One of the medium of interaction is speech. The main challenge in human machine interaction is detection of emotion from speech. Emotion can be recognised from different biological signals also [1], [2]. In this work our main concern was detecting emotion from speech. When two persons interact with each other they can easily recognize the underlying emotion in the speech, spoken by the other person. The objective of emotion recognition system is to mimic the human perception mechanisms [3]. There are several application in speech emotion recognition. Emotion can play an important role in decision making. If emotion can be recognized properly from speech then a system can act accordingly. An efficient emotion recognition system can be useful in the field of medical science [4] robotics engineering, call center application etc [5]. Human can easily recognize emotion of speaker. This can be achieved by many years of practice and observation. Human first analyzes different characteristics of particular speech and then using previous experience or observation he recognizes the emotion of the speaker. There is a need to build a human like system that can detect emotions effectively and efficiently. In this field several systems are proposed for recognizing emotional state of human from speakers voice or speech signal. Some universal emotions includes anger, happiness, sadness, surprise, neutral, disgust, fearful, stressed etc. For the last two decades several intelligent systems are proposed by researchers. These different systems also differs by the nature of features used for classification of speech signals. Some of the widely used spectral features are Mel-frequency cepstrum coefficients (MFCC) and Linear predictive cepstral coefficients (LPCC). Gausian Mixture Model (GMM), Supprot Vector Machine (SVM) and Hidden Markov Model (HMM) are used by researchers for classification using a supervised learning method. Xianglin Cheng et al. performed emotion classification using GMM and obtained the recognition rate of 81%. Only pitch and MFCC features are used for recognition of emotion [6].

Selection of features and size of the database plays important role for recognition scheme. The steps towards building of an emotion recognition system are, an emotional speech corpora is selected or implemented then emotion specific features are extracted from those speeches and finally a classification model is used to recognize the emotions. The main challenge of emotion recognition from speech is that each speech is of different length, now MFCC feature extraction method works in a sliding window method that means it set a 25ms frame over the speech signal and compute 13 cepstral coefficient from each frame those are used as features [7]. Now depending on various length MFCC return different number of frames. As a result from each speech signal we have different number of features which is not acceptable. Therefore we have done some preprocessing to make each speech signal of equal length. We have used CNN-LSTM architecture as our classification purpose, basically CNN is used for 2-dimensional input space there are some work where a spectrogram image generated from audio signal is used as input for CNN [8], [9] in our work we have used one dimensional input space containing 39 features per frame as a input for CNN. We also use the output from CNN as a input of LSTM network.

II. Speech Corpus

We have used Berlin Database of Emotional Speech (EmoDB) [10] as our speech corpus it consist of 535 utterances spoken by 10 different actors. All the recordings took place in the anechoic chamber of the Technical University Berlin, department of Technical Acoustics. This speech corpus contain seven emotional class namely happy, angry, anxious, fearful, bored, disgusted and neutral [10].

III. Feature Extraction

We have used mel frequency cepstral coefficient (MFCC) method for feature extraction. Proper choice of features play a very important part for emotion recognition.

A. Preprocessing

Before using MFCC we make some preprocessing on the data set. All the speech files are with .wav extension, first we compute amplitude values of each file with a sample rate of 16000 sample per second. Then we take a weighted average according to the length of speech files and make them equal by adding zeros to the smaller file to make them equal to the average length file and crop all the larger file for the same purpose. After this process all files became of equal size.

B. Mel Frequency Cepstral Coefficient

Mel frequency cepstral coefficients are computed on the basis of human hearing ability. In Mel frequency cepstral coefficients (MFCC) method, two types of filter are used. Some filter are spaced linearly at low frequency below 1 kHz and other are spaced logarithmically at high frequency above 1 kHz [7], [11]. The block diagram of MFCC is shown in Fig. 1.

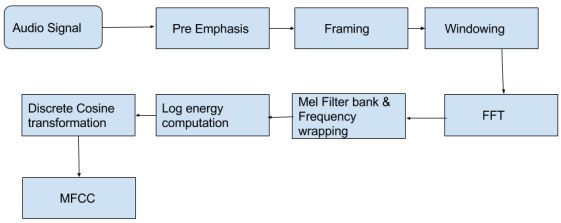

Fig. 1: Block diagram of MFCC

MFCC feature extraction process consists of a few steps as discussed below,

Pre-emphasis: Pre-emphasis is required to increase signal energy. In this process, speech signal is passed through a filter which increase the energy of signal. This increment of energy level gives more information.

Framing: In this process, speech sample is segmented into 20-40 ms frames. The length of human voice may vary, so for fixing the size of speech this processes is necessary. Although the speech signal is non-stationary in nature (i.e. frequency can be changed over the time period), but for a short duration of time, signal behave like a stationary signal.

Windowing: After framing process, the windowing process is performed. Windowing function reduce the signal discontinuities at the start and end of each frame. In this process, frame is shifted with a 10ms span. That means each frame contains some overlapping portion of previous frame.

Fast Fourier Transform (FFT): FFT is used to generate the frequency spectrum of each frame. Each sample of each frame converted from time domain to frequency domain by the FFT. FFT is used to find all frequencies present in the particular frame.

Mel scale filter bank: This is a set of 20-30 triangular filters applied to each frame. The mel scale filter bank identify how much energy exists in a particular frame. The mathematical equation to convert the normal frequency f to the Mel scale m is as follows,

$$m = 2595 \log \left(1 + \frac{f}{700}\right) \tag{1}$$

Log energy computation: After getting the filter bank energy of each frame, log function is applied to them. It is also inspired by human hearing perception. A human does not listen loud volume on a linear scale. If the volume of the sound is high, human ear can not recognize large variations in energy. Log energy computation gives those features for which human can listen clearly.

Discrete Cosine Transformation (DCT): In the final step DCT is calculated of the log filter bank energies.

We have used 25ms frames with 10ms of sliding. We have also used 26 band pass filters. From each frame we computed 13 MFCC features. We have also calculated energy within a frame. After getting 13 MFCC features, we also computed 13 velocity components and 13 acceleration components by calculating time derivatives of energy and MFCC [11].

$$\Delta C_m(t) = \sum_{\tau=-M}^{M} \tau C_m(t+\tau) / \sum_{\tau=-M}^{M} \tau^2 \tag{2}$$

Where $C_m(t)$ denotes static coefficient of t^{th} frame, calculate delta features based on preceding and following M frames

IV. Classifier

A classification system is an approach to set each speech to a proper emotion class according to the extracted features from speech. There are different classifiers available for emotion recognition. We have used one dimension CNN with LSTM for classification.

A. CNN

Convolutional neural network or CNN consist of several layer of convolution [12]. In CNN, nonlinear activation function for example rectified linear unit (ReLU) or Sigmoid function are used to the result. In neural networks nodes of input layer are connected with the nodes of hidden layer and

those hidden layer nodes are fully connected with nodes of output layer. In CNN, convolution are applied on input layer to generate the output. Each part of input are convoluted by different filters and combining them we get the final result. In CNN, there is a pooling layer and the purpose of these layer is sub sampling the input from a specified filter. A common pooling technique is max pooling. In max pooling a maximum value is selected from each filter. Pooling layer reduce the size of input . Max pooling layer can also perform over a window instead of performing the whole matrix [13]. After getting the output from CNN a neural network like Multilayer Perceptron (MLP) or Recurrent Neural Network (RNN) or Long Short Term Memory (LSTM) network can be used for training. CNN architecture is shown in Fig. 2.

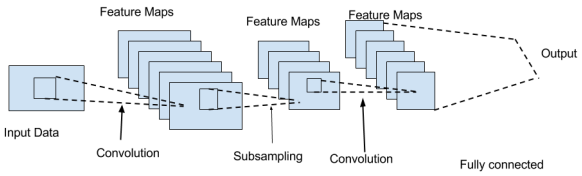

Fig. 2: CNN architecture

We use the output of CNN as input in LSTM.

B. LSTM

While classifying a set of temporal data, Recurrent Neural Network (RNN) architectures have outperformed other classifiers in temporal data classification. A simple RNN architecture is displayed in Fig. 3.

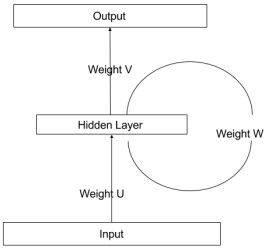

Fig. 3: A simple RNN with recurrent connections

RNNs exploit the temporal relations present in a sequence of data, thus will be very effective for speech signals. Unfortunately though the working principle of RNN is very promising it has a major drawback. It is referred as problem of Vanishing Gradient . It happens because RNN can unfold itself quite deep in time (depending on the length of input vector). If the input sequence is very long (which is a common scenario for emotion recognition using speech) RNN does not produce satisfactory result. A solution to this problem is to keep the error term always 1. This is known as Constant Error Carousel (CEC). This can be implemented by a variation of simple RNN node known as Long Short Term Memory (LSTM). The aim of using this new node structure is to use CEC by the help of different gate units as shown in Fig. 4.

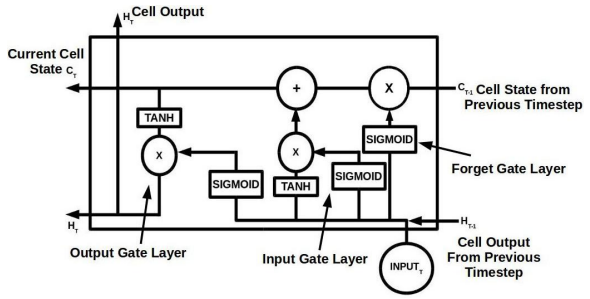

Fig. 4: LSTM Node as a memory cell

Unlike RNN one Node of the LSTM network is known as Memory Cell. Using the gates enables one to remove or add information to the cell state. This incorporates the ability to remember or forget information about states in a time frame that appeared long back in past [14]. The cell states and cell output at T^{th} time step are represented by C_T and H_T. The forget gate layer resets the states of the cell. Input gate layer decides how much of the input will affect the current cell state. The output gate layer decides how much output will affect the rest of the network. In this way LSTM eliminates the problem of Vanishing Error Gradient. This improvisation enables us to use Recurrent Neural Network (with LSTM) to train even longer sequences.

V. Experiments and Results

At first we divided the whole data set with 80% and 20% data. 80% data were used for training purpose and 20% data were used for validation purpose. After that we have computed the MFCC features with velocity and acceleration for each files of training dataset and test data set also. We provided those extracted features as initial input for convolution neural network. We use CNN with three convolution layer having 32, 16, 8 filter respectively. We have set 500 epochs for our network. We have used “adadelta” function as optimizer and “ReLU” as activation function. In LSTM network we have provided two hidden layer with 50 nodes in first layer and 20 nodes in second layer. We have used “softmax” as activation function for the final output nodes. We also used categorical cross entropy for computation of loss. After 500 epochs training accuracy reached at 96% and test accuracy reached at 80%. Confusion matrix for training data is shown in Table I. Fig. 5 and Fig. 6 are showing a graphical representation of error decay and accuracy with epochs.

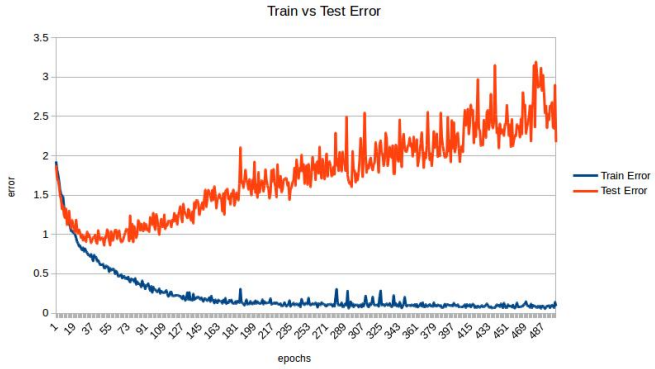

Fig. 5: Train and Test error with epochs

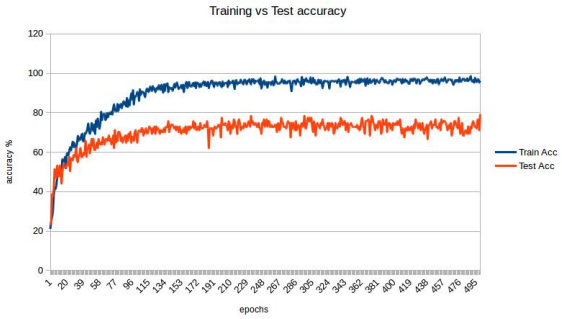

Fig. 6: Train and Test accuracy with epochs

TABLE I: Confusion matrix for validation data

	Fear	Disgust	Happiness	Boredom	Neutral	Sadness	Anger
Fear	11	0	1	0	2	0	0
Disgust	3	7	0	0	0	0	0
Happiness	5	1	9	0	0	0	0
Boredom	1	0	0	14	2	0	0
Neutral	0	0	0	6	10	0	0
Sadness	0	0	0	1	0	12	0
Anger	0	0	1	0	0	0	25

The diagonal elements of confusion matrix represents the actual recognition of emotion. For most of the emotions our network is able to recognize the proper emotion with a high level of accuracy.

VI. Conclusion

Use of CNN-LSTM model for recognition of emotion from speech is a effective step towards designing a generic emotion recognition system. Although the size of data set is not so large the performance of our proposed model is promising enough. Some normalized input data or use of Bidirectional LSTM instead of LSTM can lead to us more better solution. Also the training will produce more convenient outcome if we can feed a larger set of data to the system.

References

[1] S. Basu, N. Jana, A. Bag, M. Mahadevappa, J. Mukherjee, S. Kumar, and R. Guha. Emotion recognition based on physiological signals using valence-arousal model. In *Image Information Processing (ICIIP), 2015 Third International Conference on*, pages 50–55. IEEE, 2015.

[2] S. Basu, A. Bag, M. Mahadevappa, J. Mukherjee, and R. Guha. Affect detection in normal groups with the help of biological markers. In *India Conference (INDICON), 2015 Annual IEEE*, pages 1–6. IEEE, 2015.

[3] R. Elbarougy and M. Akagi. Cross-lingual speech emotion recognition system based on a three-layer model for human perception. pages 1–10, 2013.

[4] S. Basu, A. Bag, M. Aftabuddin, Md. Mahadevappa, J. Mukherjee, and R. Guha. Effects of emotion on physiological signals. In *2016 IEEE Annual India Conference (INDICON)*, pages 1–6, Dec 2016.

[5] D. J. France and R. G. Shiavi. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[6] Y. Wang and L. Guan. An investigation of speech-based human emotion recognition. pages 15–18, 2004.

[7] W. Han, C. F. Chan, C. S Choy, and K. P. Pun. An efficient MFCC extraction method in speech recognition. pages 145–148, 2006.

[8] Z. Huang, M. Dong, Q. Mao, and Y. Zhan. Speech Emotion Recognition Using CNN. pages 801–804, 2014.

[9] N. Anand. Convoluted Feelings Convolutional and recurrent nets for detecting emotion from audio data. pages 2–7.

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. pages 1517–1520, 2005.

[11] R. Hibare. Feature Extraction Techniques in Speech Processing : A Survey. *International Journal of Computer Applications*, 107(5):1–8, 2014.

[12] O. Abdel-hamid, L. Deng, and D. Yu. Exploring Convolutional Neural Network Structures and Optimization Techniques for Speech Recognition. Number August, pages 3366–3370, 2013.

[13] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech*, pages 223–227, 2014.

[14] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller. Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks. *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH), to be published*, 2016.