Speech Emotion Recognition using MFCC features and LSTM network

Harshawardhan S. Kumbhar
Dept. of Electronics and Telecommunication Engineering,
PCCOE, Pune
harsh.s.kumbhar@gmail.com

Sheetal U. Bhandari
Dept. of Electronics and Telecommunication Engineering
PCCOE, Pune
sheetalubhandari@gmail.com

Abstract— **Speech is a commonly used signal for interaction between humans, this leads to the usage of speech for human and machine interactions as well. Improvements in this interactive system reach toward speech emotion recognition (SER) system. SER gives sufficient intelligence for efficient natural communication between humans and machines. SER system classifies emotional states such as sadness, angry, neutral, and happiness from the speaker's utterances. This paper describes speech features and machine learning models that can be used for SER. For effective classification and to learn multidimensional complex data, a deep learning algorithm is used in this system. This paper also presents the preliminary results of a system with an MFCC feature and an LSTM algorithm.**

Keywords— SER; Machine learning; MFCC; LSTM

I. INTRODUCTION

Human commonly uses a vocal language for communication. This vocal language motivates researchers to think for speech communication with a machine. Multiple machines are developed based on this topic like assistance applications in a smartphone, speech to text converter and voice command operated machines. But this system lags in natural communication with humans; this activity can be improved by giving some intelligence to a machine. A machine can understand humans more efficiently when it can recognize human perception. Speech emotion recognition (SER) helps a machine to identify human emotions and react accordingly.

Safety, entertainment and biomedical these are the application field where SER is useful. Automatic call assistance system is the best example where customer emotions are necessary to be known for proper service so that call from an angry or disappointed customer is transfer to human instead of attending by an automatic system. Web-movies and computer tutorial applications can be controlled or responded according to the viewer's emotions for the better experience [2]. In the car, as a safety feature SER can be implemented [2]. If the SER system detects the mental status of the driver is disturbed so it will take controlling action in the car. SER system can be used as a diagnostic tool by a therapist.

The variations in speech due to speakers, speaking styles, speaking rate, and sentences make speech emotion recognition system more challenging [2]. Another challenge is that how certain emotion is expressed depends on speakers, culture, and environment. There are many techniques introduced by researchers using various speech features and deep learning algorithms. Speech features that can be used for analysis are pitch, energy, formants, linear predictor coefficients (LPC), linear frequency cepstral coefficients (LFCC), MFCC's and TEO, etc.

The speech feature values are very much dependent on the speaker. These values change person to person; this variation introduces challenges in classification. Normal classification algorithm cannot manipulate multidimensional complex data effectively. This reason leads to make use of deep learning algorithms in the SER system. Many researchers introduce learning algorithms for SER systems like CNN (Convolutional neural network), ANN (Artificial neural network), LSTM (Long short term memory) and SVM (support vector machine).

In this paper section II represents a Literature survey related to the speech emotion recognition system, section III describes implemented system methodology, section IV describes preliminary results observed on the implemented system.

II. LITERATURE SURVEY

In [1], Pavol Harar presented a method that achieved 96.97% accuracy on testing and 69.55% on file prediction. In this method, Deep Neural Network (DNN) architecture with convolutional, pooling and fully connected layers was used for emotion recognition.

Supriya B. Jagtap [3], Presented a system to detect seven emotions that are happiness, Anger, Boredom, Sadness, Surprise, Fear, and Neutral emotions. The study presents frequency information contained in speech signal are reduced into small numbers of coefficients using MFCC. This study also presents that the accuracy of the system depends on the database used for training.

In [4], Dhruvi Desai presented a review of emotion recognition through speech signals. This literature shows that MFCC provides a high level of recognition accuracy. Study shows that the classification accuracy of ANN is low compared to other classifiers. K-NN classifier gives faster computation makes it an optional technique.

In [5], M. S. Likitha presented a method of speech emotion recognition that has proven to be 80% efficient even in a noisy environment. This system uses the MFCC feature and standard deviation values to detect emotions.

In [7], Shumin presented methods based on LSTM-RNN models. This method has achieved 96.67% accuracy in case of angry emotion, 100% accuracy in case of sad emotion and for natural 86.67% accuracy is achieved. A literature survey shows that the MFCC feature is the most popular choice to identify emotions.

978-1-7281-4042-1/19/$31.00 ©2019 IEEE

III. Proposed Methodology

A. Feature Extraction

MFCC provides a high level of perception of the human voice and achieving high recognition accuracy [4]. Mel-Frequency Cepstral Coefficient (MFCC) is a popular and powerful analytical tool in the field of speech recognition [4]. MFCC reduces the computational complexity of the approach, gives better ability to extract the features and can find the different parameters like pitch and energy [3]. Mel Frequency Cepstral Coefficient (MFCC) reduces the frequency information of speech signal into the small number of coefficients which is easy to compute and extract the features [3]. It represents the short-term power spectrum of sound, based on linear cosine transform of a log power spectrum on a non-linear Mel scale of frequency. A survey shows that the MFCC gives good results as compared to other features for a speech-based emotion recognition system [4].

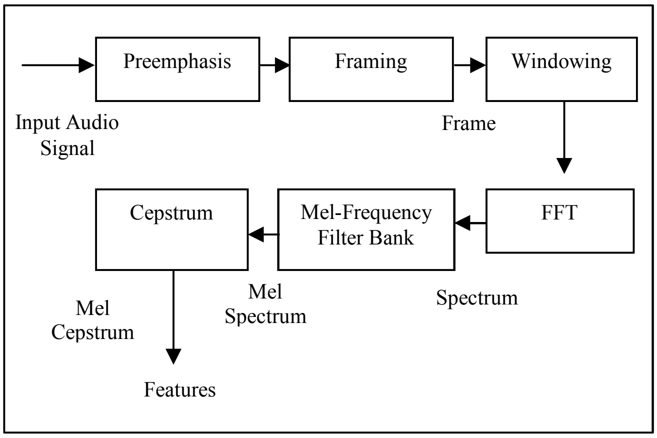

Fig. 1. Block diagram of MFCC process

Figure 1 shows the block diagram of the MFCC process. Pre-emphasis is the process of the speech signal in which pre-emphasis filter is used to achieve a smoother spectral. In the frame blocking process, the sound signal is segmented into multiple small overlapped frames in the presented methodology frame size of 20ms and the step between successive frames is also 20ms. Windowing is a process required for analyzing a section of long signals. This process removes the aliasing. Fast Fourier Transform (FFT) is used to convert a time-domain signal into a frequency spectrum. Mel-Frequency filter bank used for converting a linear frequency scale to the Mel-frequency scale. Mel- frequency scale is designed according to the perception of the human ear against the sound frequency. The scale of Mel-Frequency is a logarithmic scale, so it is sensitive to a lower frequency than a higher frequency. In the cepstrum process, Mel- spectrum will be converted into the time domain by using a Discrete Cosine Transform (DCT) to get the Mel-frequency Cepstrum coefficient (MFCC).

B. Dataset

The dataset used in the presented system is the Ryerson Audio-Visual Database[8] of Emotional Speech and Song (RAVDESS). This dataset contains 7356 files containing speech data of 12 female and 12 male actors. Actors are vocalizing two statements in a neutral North American accent. Speech includes seven emotions calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at normal and strong emotional intensity, with an additional neutral expression. The audio file format is 16bit, 48 kHz and file format is wave (.wav).

C. Machine Learning model

ML model used in the presented methodology is based on LSTM architecture. Long short-term memory (LSTM) is a modified version of artificial recurrent neural network (RNN) architecture. LSTM works better with a huge amount of data and enough training data. The main advantage of RNN over ANN is in the case of a sequence of data it gives better performance. In the case of speech processing signal is framed in small pieces this small section, for emotion detection the dependency of each section with the previous one should be considered. So in this case LSTM gives better performance.

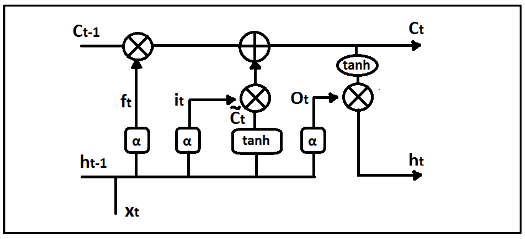

Fig. 2. LSTM structure

$$i_t = \sigma(x_t U^i + h_{t-1} W^i) \quad (1)$$

$$f_t = \sigma(x_t U^f + h_{t-1} W^f) \quad (2)$$

$$O_t = \sigma(x_t U^o + h_{t-1} W^o) \quad (3)$$

$$\tilde{C}_t = \tanh(x_t U^g + h_{t-1} W^g) \quad (4)$$

$$C_t = \sigma(f_t * C_{t-1} + i_t * \tilde{C}_t) \quad (5)$$

$$h_t = \tanh(C_t) * O_t \quad (6)$$

The above figure shows the structure of the LSTM cell and equations describing the LSTM model. U is the weight matrix contains the inputs to the hidden layer, W is the connection between current and previous layer. C is the internal memory of the unit, which is a combination of the previous memory, multiplied by the for-get gate, and the newly computed hidden state, multiplied by the input gate. $\tilde{C}$ is a candidate hidden state that is computed based on the current input and the previous hidden state [10].

TABLE I. Summery of Machine Learning Model

Layers	Activation	Output units	Parameters
LSTM	–	128	86016
Dropout	–	128	0
Dense	Relu	32	4128
Dense	Tanh	16	528
Dense	Relu	8	136
Dense	tanh	8	72
Dense (Output)	–	4	36

The above table shows a summary of the implemented machine learning model. 1st column shows layers and layer types, 2nd column shows activation functions, The

3rd column shows the number of units or neurons for a particular layer and the 4th column shows the number of parameters per layer. The dropout rate is 0.5 was set for the dropout layer. The total instances used for training were 90,916.

IV. Preliminary Results

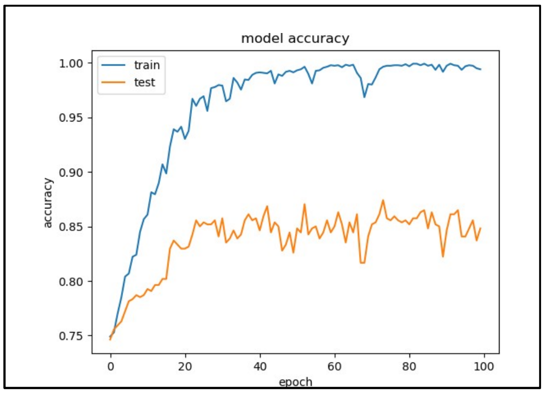

Fig. 3. LSTM model accuracy with respect to epoch

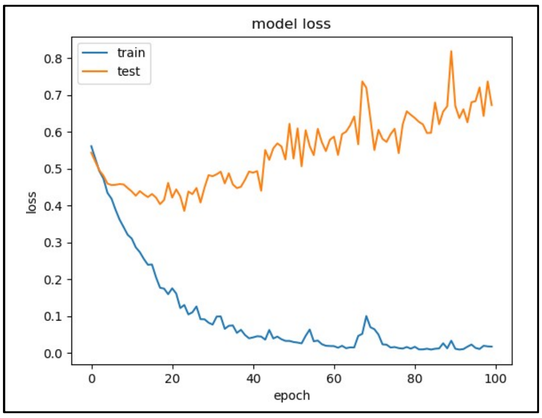

Fig. 4. LSTM model loss with respect to epoch

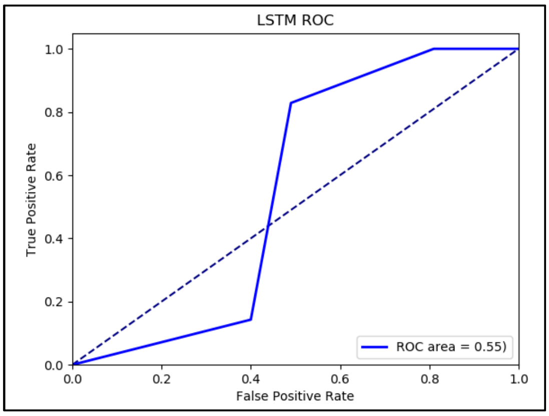

Fig. 5. ROC curve of implemented model

Above MFCC feature extraction and Machine learning model is implemented using python language. In MFCC feature extraction 39 number of coefficients are extracted. RAVDESS dataset was used as input to implement the machine learning model as summarized in table 1. The preliminary results are shown below. Figure 3 represents the accuracy curve for the implemented model. The model is trained for 100 epochs. Accuracy on test data is lesser than train data. The curve shows that after the 30th epoch the accuracy starts stabilizing. The average accuracy observed on test data is 0.8481 i.e. 84.81%.

Figure 4 shows the loss with respect to epoch count for implemented the model. On train data, it is observed that the number of epoch reduces loss. But in case of loss increases after 20 epochs, observation states that near 20 epochs model can achieve local minima of loss it will reduce the false-positive rate. The average loss observed in the model is 0.6721. So there is a chance to reduce loss to improve performance.

Figure 5 shows the receiver operating characteristic (ROC) curve. The figure shows that the ROC area is 0.55. There is a need to reduce the false-positive rate to achieve better performance.

V. Conclusions

In this work, a speech emotion recognition system with the LSTM model and MFCC feature is presented. It is observed that MFCC is a popularly used feature and gives better results for emotion detection in SER. There is a future scope in ROC curve improvement. The area observed under the ROC curve is 0.55. 67.21% loss is observed in this model, which needs to improve. The positive point observed in the implemented model is that the system achieves 84.81% accuracy. Still, there is a scope for improvement using a combination of different feature and optimizing ML model for a better true positive rate.

References

[1] P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, 2017, pp. 137-140.

[2] Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition, Vol. 44, Issue 3, 2011.Pages 572-587.

[3] Supriya B.Jagtap, Dr.K.R.Desai, Ms. J. K. Patil ” A Survey on Speech Emotion Recognition Using MFCC and Different classifier”, 8th national conference on emerging trends in engg and technology, 10th march 2018.

[4] Dhruvi desai “Emotion recognition using Speech Signal: A Review”, International Research Journal of Engineering and Technology (IRJET),Volume: 05 Issue: 04 , Apr-2018

[5] M.S. Likitha,Sri Raksha R. Gupta, K. Hasitha and A. Upendra Raju,” Speech Based Human Emotion Recognition Using MFCC” IEEE WiSP- NET 2017 conference, 2017.

[6] N. Sugan, N. S. Sai Srinivas, N. Kar, L. S. Kumar, M. K. Nath and A. Kanhe, ”Performance Comparison of Different Cepstral Features for Speech Emotion Recognition,” 2018 International CET Conference on Control, Communication, and Computing (IC4), Thiruvananthapuram, 2018, pp. 266-271.

[7] S. An, Z. Ling and L. Dai, ”Emotional statistical parametric speech synthesis using LSTM-RNNs,” 2017 Asia-Pacific Signal and Informa- tion Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017, pp. 1613-1616.

[8] S. R. Livingstone, K. Peck, and F. A. Russo, “Ravdess: The ryerson audio-visual database of emotional speech and song,” in Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science, 2012.

[9] Varsamopoulos, S & Bertels, Koen & G. Almudever, Carmen. (2018). Designing neural network based decoders for surface codes.

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

[11] Fei, W., Ye, X., Sun, Z., Huang, Y., Zhang, X. and Shang, S., 2016, June Research on speech emotion recognition based on deep auto encoder. (CYBER), 2016 IEEE International Conference on (pp. 308-312). IEEE.

[12] Pan, Y., Shen, P. and Shen, L., 2012. Speech emotion recognition using support vector machine. International Journal of Smart Home, 6(2), pp.101-108.