IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006329 |

Speech Emotion Detection Using State of the Art CNN and LSTM

Anchit Banga[1], Bhavik Baheti[2], Dipesh Sachdev[3], Yash Jajoo[4]

Department of Computer Engineering, Sinhgad Institute of Technology, Lonavala, India[1,2,3,4]

ABSTRACT: We humans find speech as the best and essential way to show our feelings or emotions. The world we live in has given birth to so many ways through which we can communicate with each other. In professional life we see, mails carry a heavy importance and play a vital role for each individual. Whereas, in day-to-day life platforms like WhatsApp, Facebook and Instagram have provided a very comfortable way to express ourselves. Every day we witness those emotions play a fundamental role when we communicate so to detect and analyse each emotion it has become chief constituent for this age of technology in distant messages. In the modern world emotion recognition and detection of the same has always been a highly complicated task, as we all know that the emotions and feelings are very personal. No software or individual has general agreement or a detailed guide to define, evaluate and classify them. We've prescribed a Speech emotion recognition system as an accumulation of procedures in which we operate and categorically classify the information and signals we get from the speech. The mechanism seeks use in various demands such as virtual assistants or a conversation between customer care and the client. In our analysis we've attempted to determine the basic human emotions by the speech we record and then examine the audible characteristics. In our analysis and examination, we've shown a speech emotions detection system in which we have improved the way we recognise and classify the emotions from the speech through detailed selection and our approach focuses on categorising audio perceptions on the basis of sentiments.

KEYWORDS: Speech Emotion, CNN, LSTM, Deep Learning, MFCC, Signal Preprocessing.

I. INTRODUCTION

Recently, deep learning algorithms have successfully addressed problems in various fields, such as image classification, machine translation, and speech recognition. Lots of improvements have been made to the research for processing different kinds of speech tasks. One of the most challenging tasks is understanding the user's emotion using the speech that is recorded. Machine Learning classifiers are not emotionally aware. So the first thing that we have to build is a robust classifier that displays good performance predicting emotions regardless of application use. In particular, the speech emotion recognition task is one of the most important problems in the field of paralinguistics. Deep learning has accelerated the progress of recognizing human emotions from speech, but there are still deficiencies in the research of SER, such as data shortage and insufficient model accuracy. The goal of the system is to predict emotional content in speech and classify it into different labels such as happy, sad, neutral, angry, disgust, fear and surprise. There are two important challenges in this area. Firstly, there is insufficient data available for the neural networks to train on and the second is that emotion must be learned from low level speech signals. In this paper we are using the CNN (Convolutional Neural Network) along with LSTM (Long Short-Term Memory) to increase the accuracy of the current available models. Our proposed model outperforms previous state-of-the-art methods by 72% to 75% when applied to the RAVDESS dataset.

II. LITERATURE SURVEY

Over the last decade many researchers have investigated statistical methods and algorithms to characterize emotion from speech. Cao et al. [1] proposed a ranking SVM method for synthesizing information about emotion recognition to solve the problem of binary classification. The SVM algorithms used every utterance of data as a query and instructed to mix all predictions for multi class classification. Thus ranking-based SVM achieved higher accuracy in recognizing emotional utterances than conventional SVM methods.

Another research was done by Chen et al. [2]. It was basically to improve speech detection using three level emotion speech recognition. It classified different emotions using the rate of Fisher. This was the input for the multi SVM

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006329 |

classifier. Four comparative experiments including Fisher + SVM, PCA + SVM, Fisher + ANN and PCA + ANN were done. Consequence indicates in dimension that Fisher is better than PCA.
Nwe et al. [3] proposed a system for emotion classification of speech signals. It used LFPC (log frequency power coefficients) and not MFCC(mel-frequency Cepstral coefficients). Results were compared between both of them and it concluded that LPFC is a better option for feature extraction.

III. METHODOLOGY

The model we have chosen is a Time Distributed Convolutional Neural Network. The Fundamental idea of the respective network is to implement a mechanism throughout the log-mel spectrogram. Every window in this mechanism as defined will be the access of convolutional neural network which is composed by the four of local feature learning blocks and then whatever we yield from these convolutional networks, it will then be catered in a repeated neural network which is made by the two cells LSTM (Long Short-Term Memory) to learn the long-term contextual dependencies. Then at the final stage we predict the emotions from the recorded speech with the help of a complete connected layer and with the use of SoftMax activation.

Fig.1 shows the knowledge about how the system of speech emotion recognition system was built:

- At the initial stage there's a voice transcription
- Secondary stage consists of discretizing the signals from the audio.
- At this stage the extraction of Log-mel-spectrogram takes place.
- Using the rolling window in order to split the spectrogram
- Use of the model that is pre-trained so that we could make a prediction.·

To limit overfitting, we have trained our model with:

- The enhancement of the data from the audio.
- Early stopping
- Creating the best model and then continuing with it.

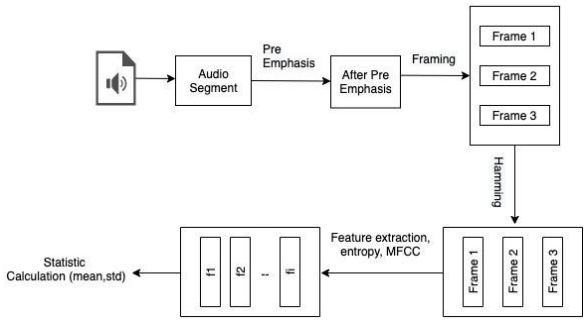

Fig. 1 Feature extraction from a audio signal

3.1 SIGNAL PREPROCESSING

Following points describe audio signal preprocessing before audio features extraction.

3.1.1 Framing:

After preprocessing the signal using a filter, the audio needs to be broken down into smaller windows called frames. We do this because frequencies in a signal change over time because a fourier transform would not be efficient enough over the entire signal. Window size is usually ranging from 20ms to 40ms where 25ms is the standard size. Our audio speech lies in the frame length for 16kHz which is then splitted into 400 samples/windows.In order to prevent frequency contours we use windowing and also the framing of signals so to prevent the presumption made by the Fast Fourier Transform and to avoid spectral leakage we use windowing. Therefore, we separate the clip into smaller time frames using this technique.

3.1.2 Hamming:

It is also known as the extension of the Hann window. It helps in reducing the ripple on the signal peaks and gives a clear idea of the actual frequency spectrum. It improves the quality of the harmonics of sound. Doing this at the

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006329 |

beginning and end of the frame can be a subtle example of this. If we don't match the frames then it will look like discontinuity in the signal and will show up as nonsense in the Discrete Fourier Transform. Applying Hamming function makes sure that beginning and end match up while smoothing the signal.

3.1.3 Log Mel Spectrogram:
Spectrogram is computed using overlapping signals of the windowed segment. When we convert frequencies into the Mel scale it is known as Mel Spectrogram. The frequency which is on the y-axis is converted into a log scale. It is a logarithmic transformation of the received signal's frequency.

3.1.4 MFCCs:
Mel-frequency cepstral coefficients (MFCCs) are parametric representation of the speech signal, that is commonly used in automatic speech recognition, but they have proved to be successful for other purposes as well,among them speaker identification and emotion recognition.They are claimed to be robust of all the features for any speech tasks. They enable a representation of a signal that is very close to human perception.

3.2 MODEL SELECTION
3.2.1 Dataset:
The RAVDESS database was used in order to evaluate our methodology in this paper. It contains the emotions speech of male and female actors (gender balanced) that were asked to pretend six different emotions (happy, sad, angry, disgust, fear, surprise and neutral) at two levels of emotional intensity.

3.2.2 Feature Extraction:
After the extraction of the required features in speech is done we proceed to the next step of matrix features per audio data. We also calculate global statistics of the extracted features such as mean, standard deviation,skewness , 1% percentile, 99% percentile, min, max and range between min and max. We obtain a vector for every audio signal.l

3.2.3 Classifier:
Time Distributed Convolutional Neural Network:
To post-process the audio signal,we reduced the dimensionality of the overall vector using Principal Component Analysis and also normalized the signal using normalization techniques. This is to reduce the overall computation and memory consumption. Feature selection and feature extraction are used to reduce the dimensionality of the vector. We used CNN (Convolutional Neural Network) as our classifier. RNN (Recurrent Neural Networks) are mostly used in sequential tasks but combining this with CNN architecture of feature extraction we can train to classify speech signals.
The only input required by the CNN is the log mel spectrogram (described above). The main reason we used CNN is to use it along with LSTM networks to apply fixed size and time-step in the rolling window.
Each of these windows will be the entry of a convolutional neural network, composed by four Local Feature Learning Blocks (LFLBs) and the output of each of these convolutional networks will be fed into a recurrent neural network composed by 2 cells LSTM (Long Short Term Memory) to learn the long-term contextual dependencies. Finally, a fully connected layer with softmax activation is used to predict the emotion detected in the voice.

IV. EMPIRICAL RESULTS

In the next section we will try to get as close as possible to the state of the art performances.
In this part, we present the results obtained with the deep learning model whose architecture has been presented in the previous sections.
To limit overfitting during the training phase, we split our data set into train (80%), validation (15%) and test set (5%). We also added early stopping to stop the training when the validation accuracy starts to decrease while the training accuracy steadily increases. We chose Stochastic Gradient Descent with decay and momentum as optimizer and a batch size of 64.
Fig. 2 and Fig. 3 represent loss (categorical cross-entropy) and accuracy for both train and validation set:
Our model allows us to obtain a maximum score of 74% on the validation set. Thanks to a Keras functionality called ModelCheckpoint we have been able to save the weights associated with the best model, allowing us to obtain a score of 72% on the test set.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006329 |

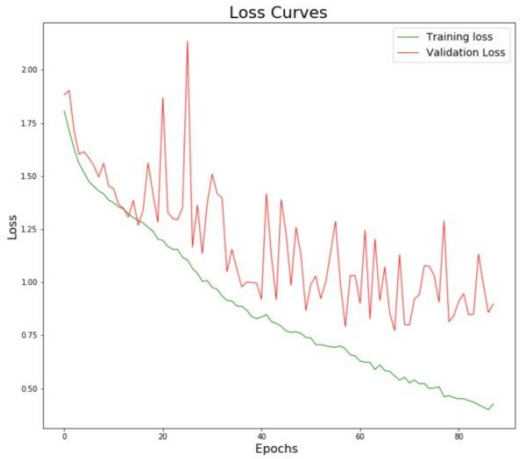

Fig. 2 Loss Curves

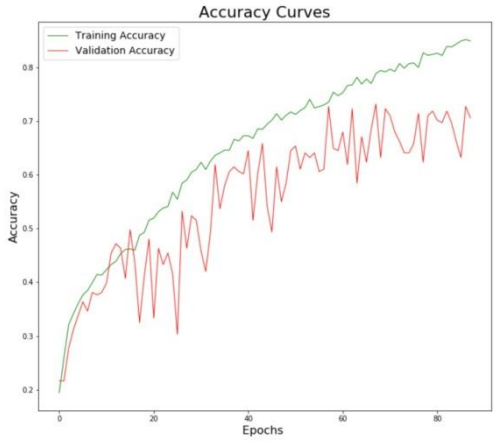

Fig. 3 Accuracy Curves

The use of deep learning and time distributed convolutional neural networks allow us to achieve a 10% higher performance compared to the traditional approach using SVM.

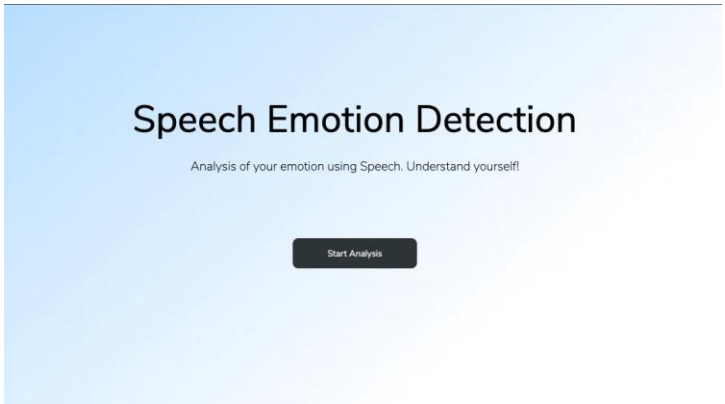

Fig. 4 Speech Emotion Detection Webapp

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006329 |

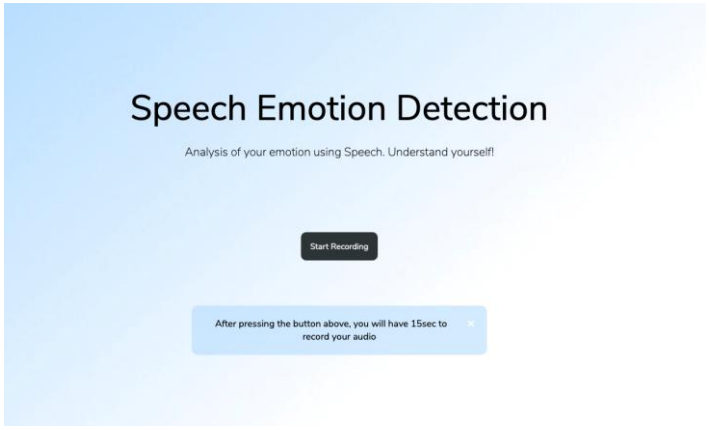

Fig. 5 Audio Recording page

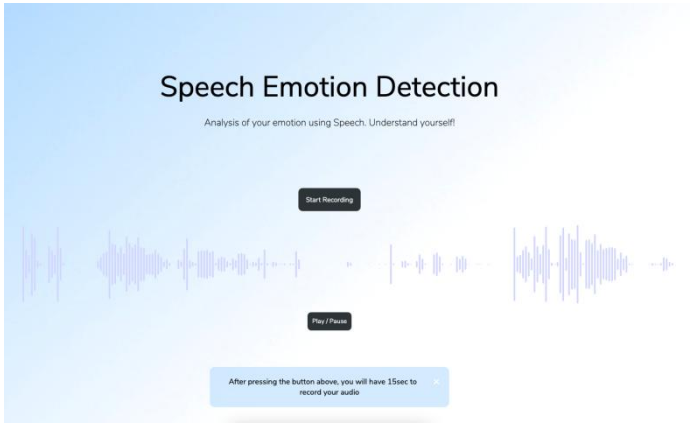

Fig. 6 Audio recording capture

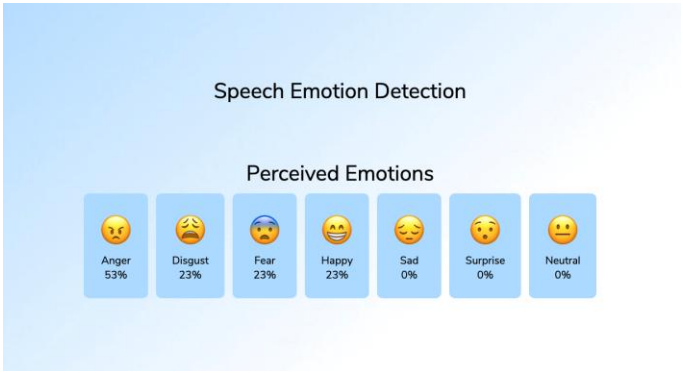

Fig. 7 Final analysis

V. CONCLUSIONS

The aforementioned speech emotion recognition system delivers fairly substantial results. Our rate of prediction for the recognition is somewhere around 65% for 7 various emotions (Happy, Sad, Angry, Scared, Disgust, Surprised, Neutral) and 75% for 6-way emotions (Happy, Sad, Angry, Scared, Disgust, Neutral). In order to achieve the better results and also to get closer to the current state of knowledge, nevertheless we'll focus on improving the model and carry out even more refined classification in the further phase of this project, for example: Hidden Markov Model (HMM) Convolutional Neural Networks (CNN) which seem to be capably better candidates for SER. These classifiers which we have shown or presented in our project are trained for only short-term and not on/for global statistics features, unlike SVM classifiers. HMM and CNN have been considered as more beneficial in order to better capture the

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006329 |

temporal dynamic incorporated in the speech. Also in order to operate a multimodal model for the recognition of emotions, we will need to set up the withdrawal of silence and then probably build the speaker who identifies and not bias the emotions predictions within the speech domain.

REFERENCES

1. H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,"
2. T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models,"
3. L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models,"
4. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).
5. B.Basharirad, and M.Moradhaseli. Speech emotion recognition methods: A literature review. AIP Conference Proceedings 2017.
6. L.Chen, M.Mao, Y.Xue and L.L.Cheng. Speech emotion recognition: Features and classification models. Digit. Signal Process, vol 22 Dec. 2012.
7. T.Vogt, E.Andre´ and J.Wagner. Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation. Affect and Emotion in Human-Computer Interaction, 2008.
8. T.Vogt and E.Andre. Improving Automatic Emotion Recognition from Speech via Gender Differentiation. Language Resources and Evaluation Conference, 2006.
9. T.Giannakopoulos. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details