First International Conference on Computer Engineering
International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
Volume 5 Issue 8, November-December-2020

Patient Feedback using Speech Emotion Recognition

Rutuja Patil[1], Siddhi Salunke[1], Pournima Ubale[1], Mayur Talole[1], Prof. Ajita Mahapadi[2]

[1]Department of Computer Engineering, Dr. D. Y. Patil School of Engineering, Lohegaon, Maharashtra India

[2]Assistant Professor, Department of Computer Engineering, Dr. D. Y. Patil School of Engineering, Lohegaon, Maharashtra India

ABSTRACT

Hospitals attempt to collect feedback from the patients to study their sentiment towards their services and facilitates provided by the hospitals to improve their environment. In present scenario feedbacks are taken in written form and are not truly maintained by hospital staff , and this technique does not reveal the true sentiments of the patients ,but this SER feedback provides a chance to highlight certain aspects. In this paper, a method has been proposed for emotion recognition by speech based on speech features and speech transcriptions, such as Spectrogram, Mel-Frequency Cepstral Coefficients (MFCCs) ,helps to retain emotion related low level characteristics in speech where as text helps capture semantic meaning both of which help in different aspects of emotion detection.

Keywords: Speech Emotion Recognition, SER, Speech Transcriptions, Speech Features .

I. INTRODUCTION

Feedback is an event that occurs when the output of a system is used as input back into the system as part of a chain of cause and effect. Feedback plays an important role in hospital by helping to adopt new knowledge and prevent repetitive mistakes. Feedback is a process which helps the organization to monitor, evaluate, and regulate the overall working environment. Good feedback practice provides useful information to the organization in improving the teaching and learning experience.

Traditionally feedback in hospital is filled out manually through forms due to which patients pay no attention about filling the form seriously. Also this process is time-consuming and very tedious job. Their might be also duplications of data and the information entered maybe false or misleading . There are many chances to lose data. Hospital might try to maintain only the positive feedback for its reputation , and this might be misleading to the people. This paper focuses on generating Patient feedback that takes voice as an input, analyze it using Speech Emotion Recognition(SER) & conveys us the feedback of the patients through their emotions. Also it generates ratings through feedback such as Excellent ,satisfactory , non- satisfactory, need improvement etc.

© 2020 IJSRST | Print ISSN: 2395-6011 | Online ISSN: 2395-602X | Published : 18/12/2020 - 216-220

Emotion Recognition using speech(SER)

- Speech is the most important feature communication tool .
- It is a biometric feature like fingerprint and carries the emotional state of the speaker.
- Therefore, speech data extracted from real talks may result more realistic emotional features than textual data.

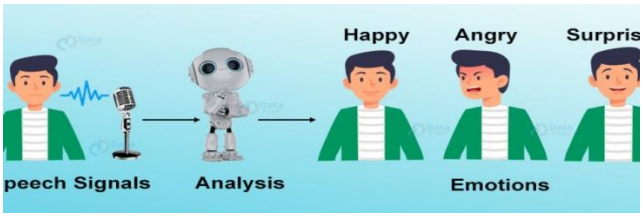

Fig1. Speech Emotion Recognition

- Speech Emotion Recognition (SER) systems is a collection of methodologies in which speech signals are classify and process to detect the emotions.

- <u>Applications of SER</u>
  - Human communication with machine
  - Psychological consultation
  - Patient care
  - Call centers

- <u>Objective:</u>
  - To automate the work of taking the feedbacks of patients.
  - To simplify the task of Patients.
  - To efficiently organize the feedback record.
  - To reduce errors and duplication of records.
  - To maintain True data

II. LITERATURE SURVEY

Sr. No	*Paper*	*Remarks*
1.	Review on Emotion Recognition using Speech	features of speech can be used to identify the difference between several emotional statements
2.	Speech Emotion Recognition using deep learning.	Some of the classifications algorithms like K-NN, Random Forest are used to classify Emotions
3.	Deep Learning Based Emotion Recognition system using Speech features and Transcriptions	Transcription and Speech Features such as Spectrogram and MFCC help to retain the Emotion.

III. PROPOSED METHOD

This paper focuses on speech features and speech transcriptions such as Spectrogram and MFCC, which together provide low-level features required and the necessary semantic relationships to classify among different emotions correctly.

We perform experiments on both speech features and speech transcriptions individually and also together, which gives us greater accuracies than the previous methods. Various combinations of data has been used as inputs in different Deep Neural Network(DNN) architectures, the details of architectures is show below.

1. CNN model based on transcriptions (Model 1)

Transcription is a printed or written version of something. Speech transcriptions is transcribes a spoken audio into text and returns block of text with its semantic meaning for each portion of the transcribed audio.

Speech transcriptions plays an important role in recognising emotions using speech. It is necessary because because it is hard to know the context in which this word has been used.

For e.g.

"good"-can have positive meaning.

-can be used in sarcastic way.

The CNN model described in this section takes speech transcriptions, in the form of word embeddings, as input to detect emotion. CNNs can directly be applied to word embeddings without prior information on their semantic contexts. This strongly suggests that language can be interpreted as a signal no different from other signals.

A word embedding is trained representation for text where words that have similar meaning have a same representation.They are distributed representation of text that is one of the key breakthroughs for the impressive performance of deep learning methods on challenging natural language processing(NLP) problems.

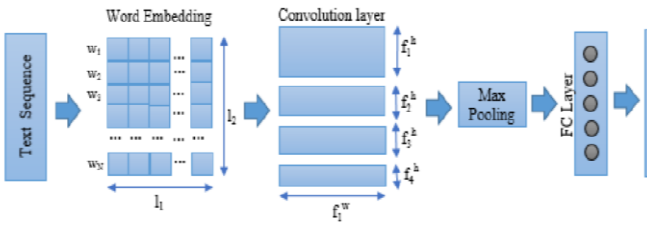

Fig 2. Transcriptions–based on CNN model

The architecture of the Transcription-based CNN model as shown in the fig 2 is used in this paper for our purpose.

Transcription sequences (embedded vectors), which given as an input to this model, are convolved with kernels of different sizes. The utrmost number of the words in any given utterance has been set to 128, which covers entirety of the IEMOCAP dataset. One feature from each of the different convolutional layers is picked by the max-pool layer. These features are fed to the single FC layer of our model. Finally, a softmax layer is used to perform classification. We experimented with batch-normalization, a technique which helps prevent the model from over-fitting and also from being too sensitive to the initial weight, and also varied dropout rates from 0.25 to 0.75. An improvement in convergence rate is also observed with the use of batch-normalization.

2. CNN model based on speech features (Model 2)

Spectrograms and MFCCs are commonly used to represent speech features.

Spectrogram:

A spectrogram displays signal strength over time at the various frequencies present in a waveform. Spectrograms can be two-dimensional graphs with a third variable represented by color, or three-dimensional graphs with a fourth color variable.

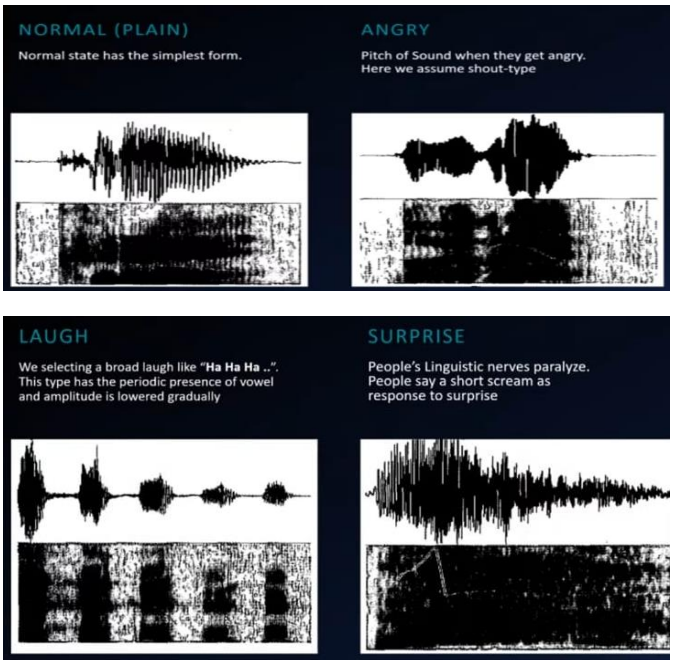

Fig 3. Spectrograms

Fig 3 represents different spectrogram for the different pitch or for different Emotions.

The obtained Spectrogram magnitudes are then mapped to the Mel-scale to get Mel-spectrograms. 128 Spectrogram coefficients per window are used in this model. The Mel-frequency scale puts emphasis on the lower end of the frequency spectrum over the ones which are higher , which imitates the perceptual hearing capabilities of the humans. Along with the parameters mentioned above , we have used "librosa" python package to compute the mel-spectrograms.

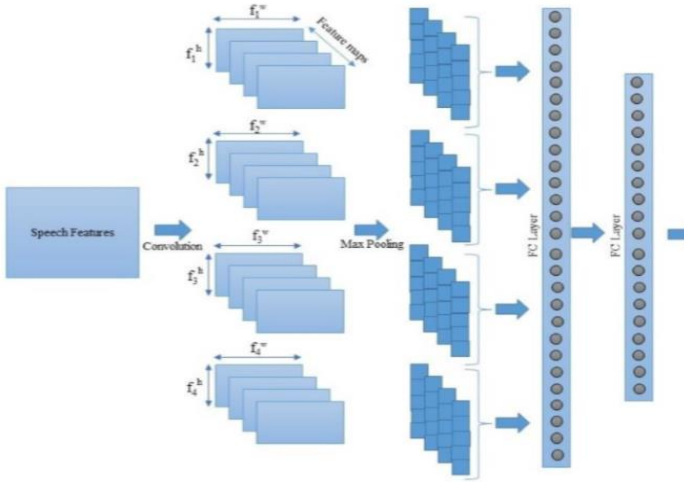

Fig 4 Spectrogram/MFCC based CNN model

Fig. 4 explains the 2D CNN architecture which is used to detect emotion using Spectrograms . A set of 4 parallel-2D convolutions are applied on the Spectrogram to extract speech features. The input shape of the Spectrogram image is 128 x 256 (number of Mel's x number of windows). 200 2D-kernels are used for each of the parallel convolution steps. Figuring out the optimal kernel size is a difficult and time taking task, which may depend on several factors all which cannot be clearly defined. To prevent choosing one single kernel size that could possibly be sub-optimal we decided to use kernels of different sizes, each of which is fixed for a single parallel path, so that we can take advantage of the different patterns that are chosen by each of the kernel. The sizes of each of the kernels in their respective parallel CNN paths are 12 x 16, 18 x 24, 24 x 32, and 30 x 40. The features generated in the said convolution layers are then fed to their respective max-pool layers, which extracts 4 features from each filter as the pool size is exactly half along the width and height of the convolution output. The extracted features are fed to the Fully Connected (FC) layer. This model makes use of two FC layers of sizes 400 and 200. Batch normalization is applied to both the FC layers. We experimented with dropout rates varying between 25% and 75% for the first FC layer but excluded it completely from the second FC layer.

3. Combined CNN model based on both speech transcriptions and speech features(model3)

A text-based CNN model fails to capture the low-level features of speech signals and due to which it cannot achieve a very high emotion detection accuracy. As mentioned above the combined Spectrogram and MFCC based CNN model achieves an improvement of 4% over existing state-of-the-art methods. The combined Text-MFCC model performs even better and beats the benchmark class.

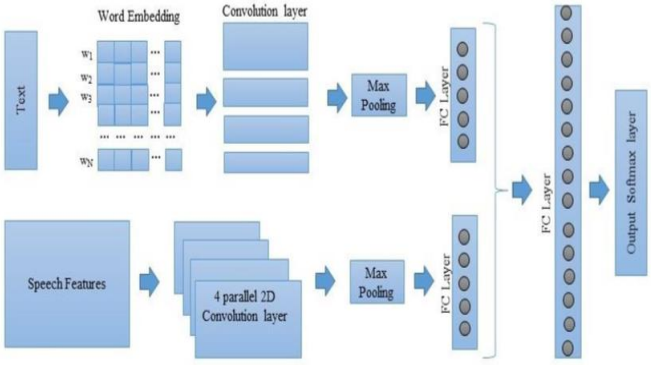

Fig5. Combining Speech features and spectrogram

In Model 3 the Spectrogram channel consists of 4 parallel 2D-CNN layers with kernels of different sizes. MFCC channel also consists of 4 parallel 2D-CNN layers as compared to the Spectrogram channel. Outputs from Spectrogram and MFCC channels are fed to one FC layer each. The outputs of both the FC

layers, after normalization, are concatenated and fed to the 2nd FC layer. The final step is to feed the outputs of the last FC layer to a softmax layer.

IV. ARCHITETURE DIAGRAM

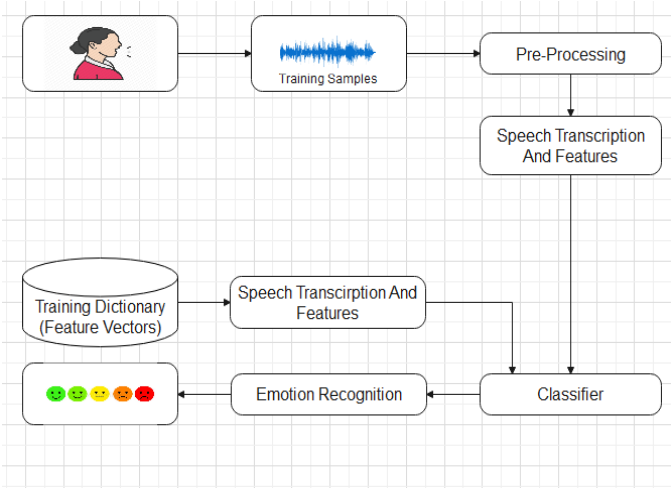

Fig 6. Architectural Diagram

V. DATASET

We are using IEMOCAP dataset [14] for proposed method. The IEMOCAP dataset consists of five sessions from both scripted and spontaneous act. From total 9 categories of emotion, only 3 emotion categories are used i.e. anger, excitement, neutral and sadness. 4936 out of 10039 turns is the total number of utterance used. The distribution of utterances for each class is almost identical to make the dataset balance. From many modalities, only speech and text are used on this research. Speech signals which are present in the dataset are processed at sampling rate of 16kHz with average length 4.5s. and for the text, average words per turn is 11.4 words while longest utterance has 554 words. Text and speech not aligned when it is processed for feature extraction. The processing for both modalities is performed independently and simultaneously through its networks.

VI. CONCLUSION

Therefore, in this paper multiple CNN based architecture have been proposed which includes speech features and speech transcriptions . model 1 provides greater accuracy , which then later improves greater accuracy when it is combined with model 2 .The model 3 which include both speech transcriptions(model 1) and speech features(model 2) results in an overall emotion detection accuracy greater than previous model by 7% , that is 76.10%. The proposed model can be used to detect emotions of the patients based on their feedback using speech.

VII. REFERENCES

[1]. Suraj Tripathi1, Abhay Kumar1*, Abhiram Ramesh1*, Chirag Singh1*, Promod Yenigalla1, “Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions” , Samsung R&D Institute India – Bangalore , arXiv.org, 2019.

[2]. Nithya Roopa S., Prabhakaran M, Betty.P, Nov 2018. “Speech Emotion Recognition using Deep Learning”. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4S, November 2018

[3]. k Ashok Kumar, J L Mazher Iqbal . “Machine Learning Based Emotion Recognition using Speech Signal” . International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1S5, December, 2019