International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Speech Emotion Recognition Using Convolutional Neural Network (CNN)

[1]Apoorv Singh, [2]Kshitij Kumar Srivastava, [3]Harini Murugan

Abstract--*The Automated Speech Emotion Recognition is a tough process because of the gap among acoustic characteristics and human emotions, which depends strongly on the discriminative acoustic characteristics extracted for a provided recognition task. Different persons have different emotions and altogether a different way to express it. Speech emotion do have different energies, pitch variations are emphasized if considering different subjects. Therefore, the speech emotion detection is a demanding task in computing vision. Here, the speech emotion recognition is based on the Convolutional Neural Network (CNN) algorithm which uses different modules for the emotion recognition and the classifiers are used to differentiate emotions such as happiness, surprise, anger, neutral state, sadness, etc. The dataset for the speech emotion recognition system is the speech samples and the characteristics are extracted from these speech samples using LIBROSA package. The classification performance is based on extracted characteristics. Finally we can determine the emotion of speech signal.*

Key words*---Speech emotion, Deep learning, Tensor flow, CNN*

I. INTRODUCTION

Deep Learning

Deep Learning in a single term we can understand as Human Nervous System. Machine Vision Deep learning sets are made to learn over a collection of audio/image also known as training data, in order to rectify a problem. The various deep learning models trains a computer to visualize like a human.

Deep learning models based on the inputs to the nodes can visualize. Hence network type is like that of a Human Nervous System, with every node performing under a larger network as a neuron. So, deep learning models are basically a part of Artificial Neural Networks. Algorithms of Deep learning learns in depth about the input audio/image as it passes over every Neural Network Layer. Low-level Characteristics like edges are detected by learning given to the initial layers, and successive layers collaborate characteristics from prior layers in a more philosophical representation.

Images, sounds, censor data and other data are those digital forms patterns which Deep Learning recognizes. For prediction we are pre-training the data and constructing a training set and testing set (results are known). As our prediction obtains an optimum node such that the predicted node provides the satisfactory output.

[1]*Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India, apoorvsingh52@gmail.com*
[2]*Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India, kshitijsd.7@gmail.com*
[3]*Asst Prof., Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India, harini.officialmail@gmail.com*

DOI: 10.37200/IJPR/V24I8/PR280260
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

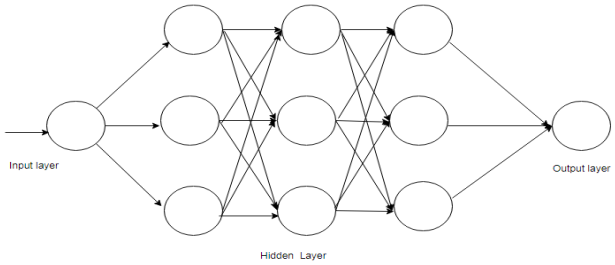
Basis of the neurons are in different levels and created to predict at every level and the most-optimum predictions, and thereafter for the best-fit outcome we use the data. It is treated as true machine intelligence.

A Convolutional Neural Network (CNN)[2] is a sort of feed-ahead artificial network in which the joining sequence among its nodes is motivated by presenting an animal visual-cortex.

Single cortical neurons give response to the stimuli at a prohibited area of region known as the receptive areas. The receptive areas of various nodes semi-overlap so that they can match the visual area. The reply of a single node for stimuli among its receptive area could be mathematically through the convolution operations. Convolutional network was motivated by natural procedures and are varieties of multi-layer perceptron formulated to use least quantity of pre-processing. They have broad use in image and video recognition, recommendation systems and NLP.

The dimensions of the Characteristics Map (Convolved Features) is regulated by following parameters:

- Depth: Representing the filter count we used in the convolution operation.
- Stride refers to size of the filter, if the size of the filter is 5x5 then stride is equal to 5.
- Zero-padding: Padding the input matrix with 0swas often convenient around the border, in order to apply filter to 'Input Audio' matrix's bordering elements. Using zero padding size of the characteristics map can be governed.



II. LITERATURE SURVEY

Here, Xinzhou Xu [3] et al generalized the Spectral Regression model exploiting the joins of Extreme Leaning Machines (ELMs) and Subspace Learning (SL) was expected for overlooking the disadvantages of spectral regression based Graph Embedding (GE) and ELM. Using the GSR model, in the execution of Speech Emotion Recognition (SER) we had to precisely represent theses relations among data. These multiple embedded graphs were constructed for the same.Demonstrationover4Speech Emotional Corpora determined that the impact and feasibility of the techniques compared to prior methods that includes ELM and Subspace Learning (SL)techniques. The system output can be improved by exploring embedded graphs at more precise levels. Only Least-Square Regression along with l2-norm minimization was considered in the regression stage.

Zhaocheng Huang[4] et al uses a heterogeneous token-used system to detect the speech depression. Abrupt changes and acoustic areas are solely and collectively figured out in joins among different embedding methods. Contributions towards the detection of depression were used and probably various health problems that would affects vocal generation. Landmarks are used to pull out the information particular to individual type of articulation

DOI: 10.37200/IJPR/V24I8/PR280260
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

at a time. This is a hybrid system. LWs and AWs hold various information. AW holds section of acoustic area into single token per frame, and on the contemporary the abrupt changes in speech articulation are shown by LWs. The hybrid join of the LWs and AWs permits exploitation of various details, more specifically, articulatory dysfunction into conventional acoustic characteristics are also incorporated.

Peng Song [5] offers Transfer Linear Subspace Learning (TLSL) framework for cross corpus recognition of speech. TLSL approaches, TULSL and TSLSL were taken in count. TLSL aims to extract robust characteristics representations over corpora into the trained estimated subspace. TLSL enhances the currently used transfer learning techniques which only focuses on searching the most portable components of characteristics TLSL can reach even better results compared to the 6 baseline techniques with stats significance, and TSLSL gives better outcomes compared to TULSL, in fact all the transfer learning is more accurate than usual learning techniques. TLSL significantly excels TLDA,TPCA, TNMF and TCA, the excellent transfer learning techniques based on characteristics transformation. A bigset back that these early transfer learning methods possess was that they concentrate on searching the portable components of characteristics that tend to ignore less informative section. The less informative parts are also significant when it comes to transfer learning results experimented that TLSL is implemented for cross-corpus recognition of speech emotion.

With this paper Jun Deng [6] et al focused on unsupervised learning with automatic encoders of speech emotion recognition. Significantly work was on joining generative and discriminative training, by partially supervised learning algorithms designed to settings where non-labeled data was available. The process had been sequentially evaluated with 5 databases in different settings. The proposed technique enhances recognition performance by learning the prior knowledge from non-labeled data in conditions with a smaller number of libeled examples. These techniques can solve the problems in mismatched settings and incorporate the learnings from different domains into the classifiers, eventually resulting in outstanding performance. This shows that the model is having the capacity to make good use of the combination of labeled and non-labeled data for speech emotion recognition. The residual neural network displayed that intense architectures make the classifier beneficial to pull out complicated structure in image processing.

Ying Qin[7] et al presented Cantonese-speaking PWA narrative speech which is a base of completely automated assessment system. Experiments on the text characteristics driven by the proposed data could detect out the impairment of language in the aphasic speech. The AQ scores were significantly correlated with the text characteristics learned by the Siamese network. The improvised representation of ASR output was leveraged as the confusion network and the robustness of text characteristics were felicitated to it. There was an immediate requirement of improving the performance of ASR on aphasic speech for generation speech that has more robust characteristics. It was necessary that the databases of pathological speech and other languages to apply this proposed methodology. As seen clinically the most desirable one is automatic classification of aphasia variant along with this large-scale accumulation of data is needed substantially.

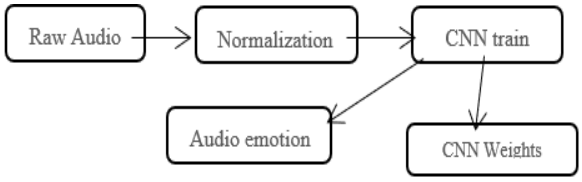
DOI: 10.37200/IJPR/V24I8/PR280260
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

A. Inference

The various Automatic speech recognition (ASR) in noise surrounding the person requires a multichannel improvement of speech with a mic array. Using the beam formation, the multichannel speech improvement can be approached. We can lay focus on speech that comes from one direction and noise is cancelled from the other direction which are basically the spatial information. This approach improved the results of improvised ASR in Chime Challenge with the help of this approach. There are many varieties of beam forming for e.g. Minimum Variance Distortion-less Response (MVDR),Multi-Channel Wiener Filtering (MWF), Generalised Side Lobe Cancelling (GSC) and Generalized Eigen-Value (GEV),performed at the time-frequency (TF) domain. Although DNN-based beam forming performs good in handful and regulated in demonstration environments, it possesses two big issues in world. Firstly, due to over fitting to the training data having many pairs of noisy and disturbed speech spectrograms and Ideal Binary Mask (IBM) have resulted in low performance of ASR under unknown environments. Secondly the physical meanings and generative processes of characteristics such as Inter-Channel Level and Phase Variants (ILDs and IPDs) are not taken into consideration and they are kept simply as an in input to DNNs.

III. METHODOLOGY

The speech emotion recognition application is executed using convolutional neural network. Following is the architecture of the system:



Training Model and Testing Model

A training data is fetched to the system which consists the expression label and Weight training is also provided for that network. An audio is taken as an input. Thereafter, intensity normalisation is applied over the audio. A normalised audio is used to train the Convolutional Network, this is done to ensure that the impact of presentation sequence of the examples doesn't affect the training performance. The collections of weights come out as an outcome to this training process and it acquires the best results with this learning data. While testing, the dataset fetches the system with pitch and energy, and based on final network weights trained it gives the determined emotion. The output is represented in a numerical value each corresponds to either of five expressions.

There are 3 emotions that are being detected based on the person's bpm value, those are Relaxed/Calm, Joy/Amusement, Fear/Anger. The produced art's colors and shapes are parallel to the detected emotion based on the principles of "color psychology" and "shape psychology".

DOI: 10.37200/IJPR/V24I8/PR280260
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

A. Algorithm

//Anaconda with Jupyter Notebook Tool in Python language.

Step 1: The sample audio is provided as input.

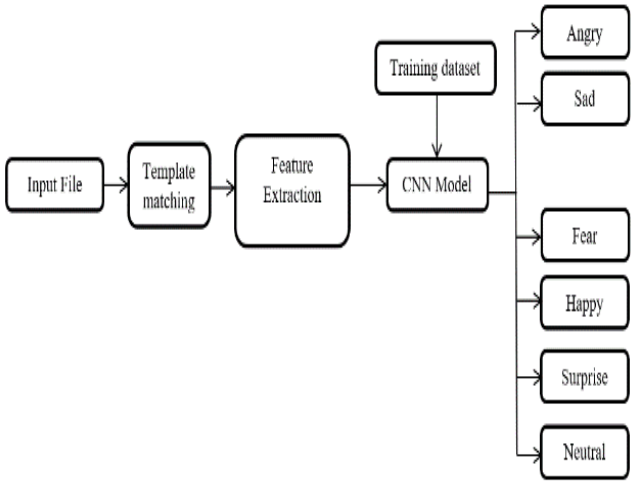
Step 2: The Spectrogram and Waveform is plotted from the audio file.

Step 3: Using the LIBROSA, a python library we extract the MFCC (Mel Frequency Cepstral Coefficient) usually about 10–20.

//Processing software

Step 4: Remixing the data, dividing it in train and test and there after constructing a CNN model and its following layers to train the dataset.

Step 5: Predicting the human voice emotion from that trained data (sample no. - predicted value - actual value)



B. Dataset

We are making use of RAVDESS dataset. It is downloaded from kaggle.com. It holds "1440 files: 60 trials/actor multiplied with 24 actors = 1440 trials". The RAVDESS consists of 24 professional voices (12 feminine, 12 masculine), speaking2 lexically-matched sentences in the even North-American accent. Happy, sad, angry, fearful, calm, disgust and surprise are the various speech emotion expressions used. Every expression is generatedin2 levels of emotional intensity (light, bold), with a neutral expression. Every file out of 1440 files has an unique filename. The filename holds a 7-part numerical identifier (e.g., 03-02-05-01-02-02-11.wav). They constitute the evoking features.

"Modality=>01 = full-AV, 02 = video-only, 03 = audio-only.

Vocal channel =>01 = speech, 02 = song

Emotion=>01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised.

Emotional intensity=>01 = normal, 02 = strong."

NOTE: For the 'neutral' expressions there are no bold intensity available.

Statement=>

DOI: 10.37200/IJPR/V24I8/PR280260
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

01 = "Kids are talking by the door",

02 = "Dogs are sitting by the door".

Repetition=>

01 = 1st repetition,

02 = 2nd repetition

Actors => 01 to 24.

01,03,05,07……23 are Male actors.

02,04,06,08……24 are female actors.

C. Modules

In our CNN model we have four important layers:

1. Convolutional layer: Identifies salient regions at intervals, length utterances that are variable and depicts the feature map sequence.
2. Activation layer: A non-linear Activation layer function is used as customary to the convolutional layer outputs. In this we have used corrected linear unit (ReLU) during our work.
3. Max Pooling layer: This layer enables options with the maximum value to the Dense layers. It helps to keep the variable length inputs to a fixed sized feature array.
4. Dense layer

- Audio Feature Extraction and Visualizations. (module01)

Characteristics extraction is required for classification and depiction. The audio signal is a 3D signal in which 3 axes indicate time, amplitude and frequency. We will use librosa to analyze and extract characteristics of any audio signal. (.load) function pulls an audio file and decrypts it into a 1D array which is of time series x, and SR is actually sampling rate of x. By default SR is 22 kHz. Here I will show one audio file display with the use of (IPython.display) function. Librosa.display is important to represent the audio files in various forms i.e. wave plot, spectrogram and colormap.

Wave plots use loudness of the audio at a particular time. Spectrogram displays various frequencies for a particular time with its amplitude.

- To train the model for accuracy calculation. (module02)

Within this module we train the model for accuracy estimations. 1st, import necessary modules. Then pull the dataset. We will receive the sampling rate value with librosa packages and mfcc function. Thereafter this value holds other variables. Now audio files and mfcc value hold a variable consequently it will add a list. Then zip the list and hold two variables x & y. Then we have represented (x, y) shape values with the use of numpy package.

DOI: 10.37200/IJPR/V24I8/PR280260
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

- Implementation process of CNN model. (module03)

Speech represented in the form of image with 3 layers. While using CNN, do consider, 1st and 2nd derivatives of speech image with time and frequency. CNN can predict, analyze the speech data, CNN can learn from speeches and identify words or utterances.

- Classification of speech emotions. (module04)

When testing we provide the audio input. Next, we run the audio in order to hear with ipython.disply packages. Thereafter plot the audio features with librosa.display.waveplot packages. Extract the Characteristics using librosa.load. It converts one data frame and display structured form. Further it compares loaded model by predict function batch size 32. Ultimately it displays the output from the audio file what sort of expression/emotion that audio file has.

IV. RESULTS AND DISCUSSIONS

We experimented an audio file to get its characteristics by plotting the waveform &spectrogram. (Fig.1 and Fig.2)

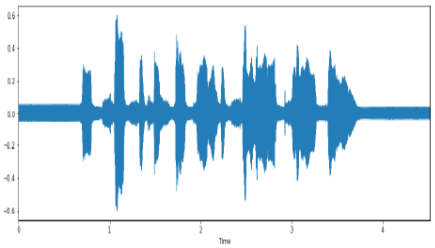


Figure 1: Time Domain Plot of the Speech signal

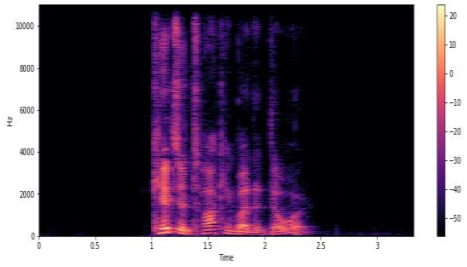


Figure 2: Frequency Domain Plot of the Speech signal

After training various models it came out with the most optimum accuracy of 82% with SoftMax activation layer, “rmsprop” activation layer,18 layers, Batch-Size = 32 and with 1000 epochs.

DOI: 10.37200/IJPR/V24I8/PR280260
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 216, 128)	768
activation_1 (Activation)	(None, 216, 128)	0
conv1d_2 (Conv1D)	(None, 216, 128)	82048
activation_2 (Activation)	(None, 216, 128)	0
dropout_1 (Dropout)	(None, 216, 128)	0
max_pooling1d_1 (MaxPooling1	(None, 27, 128)	0
conv1d_3 (Conv1D)	(None, 27, 128)	82048
activation_3 (Activation)	(None, 27, 128)	0
conv1d_4 (Conv1D)	(None, 27, 128)	82048
activation_4 (Activation)	(None, 27, 128)	0
conv1d_5 (Conv1D)	(None, 27, 128)	82048
activation_5 (Activation)	(None, 27, 128)	0
dropout_2 (Dropout)	(None, 27, 128)	0
conv1d_6 (Conv1D)	(None, 27, 128)	82048
activation_6 (Activation)	(None, 27, 128)	0
flatten_1 (Flatten)	(None, 3456)	0
dense_1 (Dense)	(None, 10)	34570
activation_7 (Activation)	(None, 10)	0

Total params: 445,578
Trainable params: 445,578
Non-trainable params: 0

Figure 1: Model Summary

The below figure shows the training and testing loss on our dataset. As the graph says that both “training and testing” errors reduces as number of epochs to the training model increases.

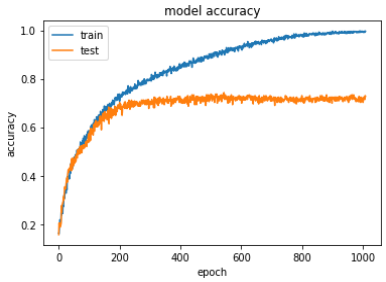


Figure 2: Modal loss plot

From above plot we can also infer that the number of suitable epochs is around 200 as the accuracy of test data remains constant after 200 epochs. Post model training we must depict out test data emotions with 75% avg. accuracy and 82.08% at most accuracy. The following table displays our depiction with the actual values and the predicted values.

	actualvalues	predictedvalues
58	male_fearful	male_happy
59	male_fearful	male_fearful
60	male_fearful	male_fearful
61	male_fearful	male_fearful
62	male_sad	male_sad
63	male_fearful	male_fearful
64	male_happy	male_happy
65	female_angry	female_angry
66	female_angry	female_fearful
67	male_angry	male_angry

Figure 3: Actual values and Predicted values.

DOI: 10.37200/IJPR/V24I8/PR280260
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

V. CONCLUSION

After constructing various models, we got the better CNN model for the emotion distinction task. We reached 71%accuracy from the previously available model. Our model would've performed better with more data. Also our model performed very well when distinguishing among a masculine and feminine voice.

Our project can be extended to integrate with the robot to help it to have a better understanding of the mood the corresponding human is in, which will help it to have a better conversation as well as it can be integrated with various music applications to recommend songs to its users according to his/her emotions, it can also be used in various online shopping applications such as Amazon to improve the product recommendation for its users. Moreover, in the upcoming years we can construct a sequence to sequence model to create voice having different emotions. E.g. asad voice, an excited one etc.

REFERENCES

1. Y. Chen, Z. Lin, X. Zhao, S. Member, G. Wang, and Y. Gu, "Deep Learning-Based Classi fi cation of Hyperspectral Data," pp. 1–14, 2014.
2. L. Chua and T. Roska, "The CNN Paradigm," vol. 4, no. 9208, pp. 147–156, 1993.
3. X. Xu, J. Deng, E. Coutinho, C. Wu, and L. Zhao, "Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition," *IEEE*, vol. XX, no. XX, pp. 1–13, 2018.
4. Z. Huang, J. Epps, D. Joachim, and V. Sethu, "Natural Language Processing Methods for Acoustic and Landmark Event-based Features in Speech-based Depression Detection," *IEEE J. Sel. Top. Signal Process.*, vol. PP, no. c, p. 1, 2019.
5. P. S. Member, "Transfer Linear Subspace Learning for Cross-corpus Speech Emotion Recognition," vol. X, no. X, pp. 1–12, 2017.
6. J. Deng, X. Xu, Z. Zhang, and S. Member, "Semi-Supervised Autoencoders for Speech Emotion Recognition," vol. XX, no. XX, pp. 1–13, 2017.
7. Y. Qin, S. Member, T. Lee, A. Pak, and H. Kong, "Automatic Assessment of Speech Impairment in Cantonese-speaking People with Aphasia," *IEEE J. Sel. Top. Signal Process.*, vol. PP, no. c, p. 1, 2019.
8. M. D. Zeiler *et al.*, "ON RECTIFIED LINEAR UNITS FOR SPEECH PROCESSING New York University , USA Google Inc ., USA University of Toronto , Canada," pp. 3–7.

DOI: 10.37200/IJPR/V24I8/PR280260
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020