International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Speech Emotion Recognition using CNN

Abdul Ajij Ansari[1], Ayush Kumar Singh[2], Ashutosh Singh[3]

[1, 2, 3]*Student, Department of Computer Science & Engineering, Galgotias College of Engineering & Technology, Greater Noida, Uttar Pradesh, India*

Abstract - *In order to obtain emotional-related response from computers and other intelligent machines, the first and decisive step is accurate emotion recognition. This paper presents the implementation of this function with the deep learning model of Convolutional Neural Networks (CNN). The architecture was an adaptation of an image processing CNN, programmed in Python using Keras model-level library and TensorFlow backend. The theoretical background that lays the foundation of the classification of emotions based on voice parameters is briefly presented. According to the obtained results, the model achieves the mean accuracy of 79.33% for five emotions (happiness, fear, sadness, neutral, anger), which is comparable with performances reported in scientific literature. The original contributions of the paper are: the adaptation of the deep learning model for processing the audio files, the training of the CNN with a set of recordings in English language and an experimental software environment for generating test files.*

***Key Words*: Speech Emotion Recognition, Convolutional Neural Network.**

1. INTRODUCTION

In today's digital era, speech signals become a mode of communication between humans and machines which is possible by various technological advancements. Speech recognition techniques with methodologies signal processing techniques made leads to Speech-to-Text (STT) technology which is used mobile phones as a mode of communication. Speech Recognition is the fastest growing research topic in which attempts to recognize speech signals. This leads to Speech Emotion Recognition (SER) growing research topic in which lots of advancements can lead to advancements in various field like automatic translation systems, machine to human interaction, used in synthesizing speech from text so on. In contrast the paper focus to survey and review various speech extraction features, emotional speech databases, classifier algorithms and so on. Problems present in various topics were addressed.

Speech Recognition is the technology that deals with techniques and methodologies to recognize the speech from the speech signals. Various technological advancements in the field of the artificial intelligence and signal processing techniques, recognition of emotion made easier and possible. It is also known as Automatic Speech Recognition". It is found that voice can be next medium for communicating with machines especially when computer-based systems. Since there is an enormous development in the field of Voice Recognition. There are many voice products has been developed like Amazon Alex, Google Home, Apple Home Pod which functions mainly on voice-based commands. It is evident that Voice will be the better medium for communicating to the machines

1.1 CONVOLUTION NEURAL NETWORK

Convolutional neural networks (CNNs) are one of the most popular deep learning models that have manifested remarkable success in the research areas such as 14 object recognition, face recognition, handwriting recognition, speech recognition, and natural language processing. The term convolution comes from the fact that convolution—the mathematical operation—is employed in these networks. Generally, CNNs have three fundamental building blocks: the convolutional layer, the pooling layer, and the fully connected layer. Following, we describe these building blocks along with some basic concepts such as SoftMax unit, rectified linear unit, and dropout.

1.2 CONVOLUTIONAL LAYER

Convolutional layers in CNNs use convolution instead of multiplication to compute the output. As a result, the neurons in the convolutional layers are not connected to all the neurons in their preceding layers. This architecture is inspired by the fact that neurons of the visual cortex have local receptive field. That is, the neurons are specialized to respond to the stimuli limited to a specific location and structure. As a result, using convolution introduces sparse connectivity and parameter sharing to CNNs, which decreases the

number of parameters in deep neural networks drastically. Figure demonstrates the convolution of a kernel, which is a 2 × 2 matrix, with a one-channel 3 × 3 image. The output is a volume of 2 × 2 × 1. Generally, the size of output is (nh −f + 1) × (nw −f + 1) ×nf, where nh is the height of the input, nw is the width of the input, and nf is the number of kernels. The depth of the kernel is determined by the depth of the input.

***Figure 1**: Human Speech Emotion Recognition*

2. LITERATURE REVIEW

Complete review on the speech emotion recognition is explained in which reviews properties of dataset, speech emotion recognition study classifier choice. Various acoustic features of speech are investigated and some of the classifier methods are analyzed in which is helpful in the further investigation of modern methods of emotion recognition. This paper investigated the prediction of the next reactions from emotional vocal signals based on the recognition of emotions, using different categories of classifiers. Some of the classification algorithms like K-NN, Random Forest are used in to classify emotion accordingly. Recurrent Neural network arises enormously which tries to solve many problems in the field of data science. Deep RNN like LSTM, Bi-directional LSTM trained for acoustic features are used. Various range of CNN are being implemented and trained for speech emotion recognition are evaluated. Emotion is inferred from speech signals using filter banks and Deep CNN which shows high accuracy rate which gives an inference that deep learning can also be used for emotion detection. Speech emotion recognition can be also performed using image spectrograms with deep convolutional networks which is implemented.

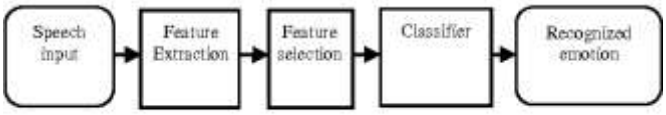


***Figure 2** structure of the speech emotion recognition*

3. PROBLEM FORMULATION

Here the focus lies on a hybrid recommender approach, which combines content-based and collaborative based approaches. It shows that many of the disadvantages of existing systems become obsolete by combining known concepts with new ones. A hybrid recommender approach contains the use of multiple algorithms at the same time. Algorithm to be used are listed below.

3.1 Matrix Factorization

When a user gives feed back to a certain movie they saw (say they can rate from one to five), this collection of feedback can be represented in a form of a matrix. Where each row represents each user, while each column represents different movies. Obviously, the matrix will be sparse since not everyone is going to watch every movie, (we all have different taste when it comes to movies).

One strength of matrix factorization is the fact that it can incorporate implicit feedback, information that are not directly given but can be derived by analyzing user behavior. Using this strength, we can estimate if a user is going to like a movie that (he/she) never saw. And if that estimated rating is high, we can recommend that movie to the user.

The concept of matrix factorization can be written mathematically to look something like below.

$$\hat{r}_{ui} = q_i^T p_u.$$

3.2 K-nearest neighbor

KNN belongs to supervised learning domain and is majorly used in pattern recognition, and data mining. The K-nearest neighbor scheme requires training set and desired classification for each item.

When we need to make a classification for new data item, its distance to each data in the training set is to be made. Only the k closest entries in the training set ae considered. The new data item is then placed in the class that holds the most number of items for this set of k closest data items.

3.3 Decision Tree

Decision Trees are used for processing a large amount of data and thus it is used in data mining. It is most useful in classification problem and is easy to understand by humans. Structure of decision tree includes root, nodes, branches and leaf nodes. Every internal node represents a test on attribute, each branch denotes the outcome of the test and leaf nodes holds a class name. When the tree is built completely, it is then applied to each data in the data set and results in classification for the tuple. It can be used to propose conditions like fog, wind, rain, thunder, pressure and humidity.

3.4 Naive Bayes Classifier

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

4. PROPOSED WORK

We initiate our approach by first transforming all the recommending papers (in our dataset) into a paper-citation relations matrix in which, the rows and the columns respectively represent the recommending papers and their citations. Our approach aimed to deal with scenarios in which: (a) A researcher who finds an interesting paper after some initial searches, wants to get more other related papers similar to it. (b) A student received a paper by his supervisor to start a research in the topic area covered by it. (c) A reviewer wants to explore more based on a received paper that addresses a subject matter which he is not a specialist in. (d) A researcher who wants to explore more from his previous publication(s). In all these cases, we consider a situation where the references and citations of the possessed paper that indicate the user's preferences are publicly available (which is usually the case in almost all the major academic databases).

Algorithm 1. Algorithm representing proposed approach.

Input: Target Paper

Output: Top-N Recommendation

Given a target paper p_i as a query,

1. Retrieve all the set of references Rf_j of the target paper p_i from the paper-citation relation matrix C.
    a. For each of the references Rf_j, extract all other papers p_{ci} that also cited Rf_j other than the target paper p_i.
2. Retrieve all the set of citations Cf_j of the target paper p_i from the paper-citation relation matrix C.
    a. For each of the citations Cf_j, extract all other papers p_{ri} that Cf_j referenced other than the target paper p_i.
3. Qualify all the candidate papers p_c from p_{ci} that has been referenced by at least any of the p_{ri}
4. Measure the extent of similarity $W^{p_{i \to p_c}}$ between the target paper p_i and the qualified candidate papers p_c
5. Recommend the top-N most similar papers to the user.

We accept the user's query in order to identify the target-paper. Once the target paper is identified, we

apply algorithm 1. The algorithm retrieves all the target paper's references and citations. For each of the references, it extracts all other papers from the web (google scholar to be precise) that also cited any of those target paper's references. In addition, for each of the target paper's citations, it extracts all other papers from the web that referenced any of those target paper's citations (in other words, all the references to the target paper's citations) and we refer to these extracted papers as the target papers nearest neighbors. For each of the neighboring papers, we qualify candidate papers that are co-cited with the target paper and which has been referenced by at least any of the target papers references. We then measure the degree of similitude between these qualified candidate papers and the target paper by measuring their collaborative similarity using Jaccard similarity measure given by Eq (1). We then recommend the top-N most comparable papers to the researcher.

Jaccard similarity does not only measure the extent of similarity between our target paper and any of the qualified candidate papers but also measures their deviations. Given two papers *X* and *Y*, each with *n* binary attributes, the Jaccard coefficient *J*, is a useful measure of the overlap that *X* and *Y* share with their attributes. Each attributes of *X* and *Y* can be either

$$J = W^{Pi \to Pc} = \frac{Z_{11}}{Z_{01} + Z_{10} + Z_{11}}$$

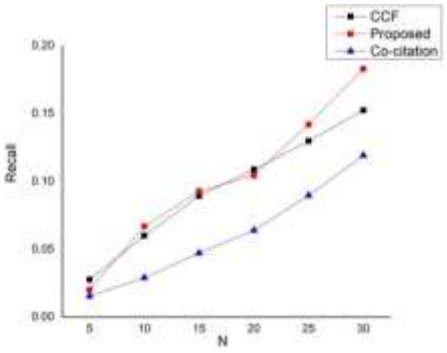
0 or 1. The Jaccard similarity coefficient *J*, is given as

where,

Z_{11} Represents the total number of attributes where *X* and *Y* both having a value of 1.

Z_{01} Represents the total number of attributes where the attribute of *X* is 0 and the attribute of *Y* is 1.

Z_{10} Represents the total number of attributes where the attribute of *X* is 1 and the attribute of *Y* is 0.



5. IMPLEMENTATION

In the current study, we implemented convolutional neural networks (CNNs) to classify speech utterances based on their emotional contents. In addition to three widely used benchmarks for recognition of emotion from speech utterances, we used a private database to train and evaluate our models. We used TensorFlow (an open-source library written in Python and C++) as the programming framework to implement our CNN models. This chapter describes the experimental setup of the current work. The first section introduces the databases administered in our study. The second section explains the preprocessing procedure. The third section describes the training and test setup used to train and evaluate our models. Finally, this chapter ends by introducing the baseline architecture of the CNNs implemented in the current work.

6. DATABASES

6.1 RAVDESS: BRITISH ENGLISH DATABASE

The RAVDESS is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats. The set of

7356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were

characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity and test-retest intrarater reliability were reported. Corrected accuracy and composite "goodness" measures are presented to assist researchers in the selection of stimuli. All recordings are made freely available under a Creative Commons license and can be downloaded at https://doi.org/10.5281/zenodo.1188976.

6.2 TRAINING AND TEST SETS

Our models were trained and evaluated using 5-fold cross-validation. That is, the data were partitioned into 5 folds. The first fold was used as a test set whereas the other folds were used to train our models. Then, the second fold was used to test our models while the remainder of the folds were used for training, and so on. To reduce overfitting and the adverse effect of small size of databases, the data sets were augmented by adding white Gaussian noise with +15 signal to noise ratio (SNR) to each audio signal either 10 times or 20 times. The SNR is defined as 10 log10(Pspeech/Pnoise), where P is the average power of the signal. The data augmentation resulted in two types of data sets used to train our models: data sets with 10 times augmentation (10x) and data sets with 20 times augmentation (20x). We used the original data without noise to test our models. The augmented data were used only for training. Finally, the labels of the training and the test data were encoded as one-hot vectors. Table shows the class labels of each database. The number of training epochs was varied between 100 to 4000. The favorable training epoch was set to 100 dues to the computation and time expenses.

	Stimulus Type			
Emotion	EMODB	SAVEE	EMOVO	BTNRH
happy	1	1	1	1
sad	2	2	2	2
angry	3	3	3	3
scared	4	4	4	4
bored	5	-	-	-
surprisd	-	5	5	-
disgustd	6	6	6	-
neutral	7	7	7	5

6.3 ARCHITECTURE

The baseline architecture of the deep neural network that was implemented in the current study was a convolutional neural network with two convolutional layers and one fully connected layer with 1024 hidden neurons. Depending on the number of classes, either a 5-way or a 7-way SoftMax unit was used to estimate the probability distribution of the classes. Every convolutional layer was followed by either a max-pooling or average-pooling layer. Recti ed Linear Units (ReLU) were used in convolutional and fully connected layers as activation functions to introduce nonlinearity to the model. The initial kernel size of convolutional layers was set to 5 5 with stride of 1. The initial number of kernels was set to 8 and 16 for the first and the second convolutional layers, respectively. The kernel size of pooling layers was set to 2 2 with stride of 2. Cross-entropy was used as the loss function and the Adam optimizer was employed to minimize the loss function over the mini batches of the training data. The size of the mini batches was set to 512. The number of training iteration was 100. The networks developed in this study took between 30 minutes to 2 days to be trained using Graphics Processing Units (GPUs). Broadly speaking, GPUs are used instead of CPUs to accelerate the speed of computation since GPUs have several cores and can handle a large number of concurrent threads. We used the Crane cluster of the Holland Computing Center at University of Nebraska-Lincoln to run our experiments. Further, we incorporated the dropout algorithm into the fully connected layer to improve the performance of our networks whenever the symptoms of overfitting were diagnosed. Figure illustrates the fundamental building blocks of the CNN model in the current work.

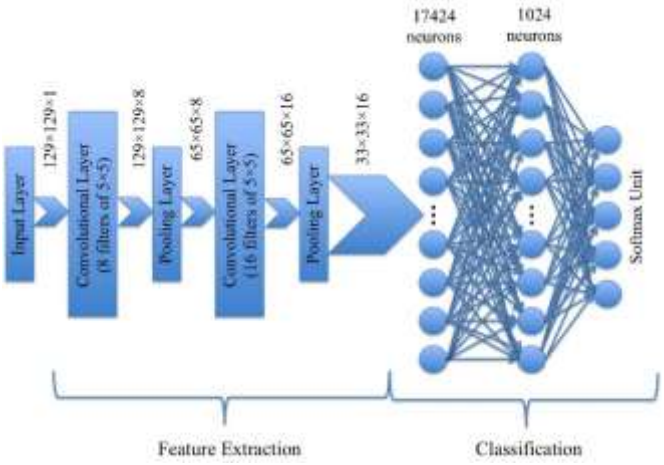


Figure 3*: The baseline architecture of the CNN used in the current study to classify speech utterances based on their emotional states.*

7. RESULT ANALYSIS

An accuracy rate of about 35.6% is achieved from the data model for predicting the emotions. It is evident from the below figure that 0.8 is the highest accuracy rate achieved during validation of data. We ran several language-dependent gender-independent experiments on each database. We embarked on our study by implementing the baseline CNN architecture introduced in Chapter 4. Subsequently, we modified the hyperparameters such as the size of convolution kernels and the deletion probability of the dropout algorithm hing on the performance of the models. This chapter aims to present the results of these experiments and to discuss the outcomes. Some of the reason for less accuracy rate are, Transfer Learning is used to train the model, there could've been less spectrograms used for training, which leads to the less accuracy. There are also less data set used for the training process which also leads to the case.

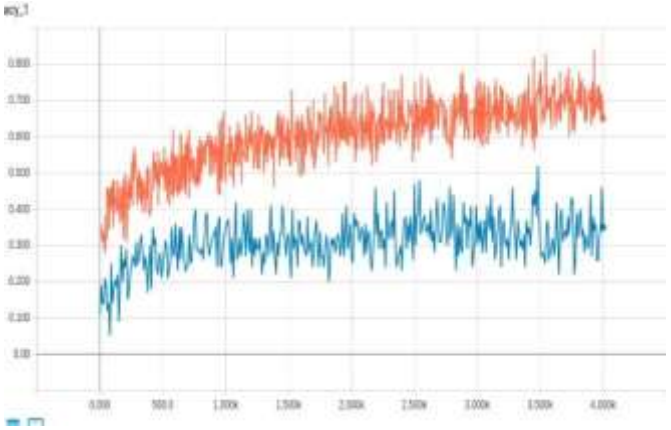


Figure 4: Accuracy rate of the Data Model

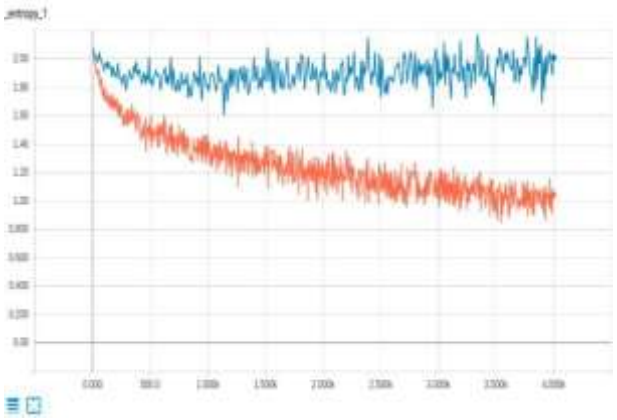


Figure 5: Cross Entropy

8. CONCLUSION

Various investigations and surveys about Emotion Recognition, Deep learning techniques used for recognizing the emotions are performed. It is necessary in future to have a system like this with much more reliable, which has endless possibilities in all fields. This project attempted to use inception net for solving emotion recognition problem, various databases have been explored. Trained my model using TensorFlow. Accuracy rate of about 39% is achieved. In future, real time emotion recognition can be developed using the same architecture.

9. REFERENCES

[1] Dong Yu and Li Deng. AUTOMATIC SPEECH RECOGNITION. Springer, 2016.

[2] Samira Ebrahimi, Vincent Michalski, Kishore Konda, Goethe Roland Memisevic, Christopher Pal— Recurrent Neural Networks for Emotion Recognition in Video||, Kahou École Polytechnique de Montréal, Canada ; Universität Frankfurt, Germany; Université de Montréal, Montréal, Canada; 2015.

[3] Ray Kurzweil. The singularity is near. Gerald Duckworth & Co, 2010.

[4] Demis Hassabis, Dharshan Kumaran, Christopher Summer eld, and Matthew Botvinick. Neuroscience-inspired artificial intelligence.

[5] Marvin Minsky. The emotion machine: Commonsense thinking.

[6] artificial intelligence, and the future of the human mind. Simon and Schuster, 2007.

[7] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. Artificial intelligence: a modern approach, volume 2. Prentice hall Upper Saddle River, 2003.

[8] Lawrence R Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition, volume 14. PTR Prentice Hall Englewood Clis, 1993.

[9] Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. A maximum likelihood approach to continuous speech recognition. In Readings in speech recognition, pages 308{319. Elsevier, 1990.

[10] Stephen E Levinson, Lawrence R Rabiner, and Man Mohan Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. The Bell System Technical Journal, 62(4):1035{1074, 1983.