Indian Journal of Natural Sciences *www.tnsroindia.org.in ©IJONS*

Vol.14 / Issue 80 / Oct / 2023 *International Bimonthly (Print) – Open Access* *ISSN: 0976 – 0997*

RESEARCH ARTICLE

Real Time Emotion Recognition Using Convolutional Neural Networks

G.S.K.Gayatri Devi[1]*, L.Sailaja[2], P.Kiran[3], B.Priya[3] and G.Govardhan Raju[3]

[1]Professor, ECE Department, Malla Reddy Engineering College, Secunderabad-500100, Hyderabad, India
[2]Assistant Professor, ECE Department, Malla Reddy Engineering College, Secunderabad-500100, Hyderabad, India
[3]B.Tech, Malla Reddy Engineering College, Secunderabad-500100, Hyderabad, India

Received: 20 June 2023 Revised: 26 Aug 2023 Accepted: 03 Oct 2023

***Address for Correspondence**
G.S.K.Gayatri Devi
Professor, ECE Department,
Malla Reddy Engineering College,
Secunderabad-500100, Hyderabad, India

ABSTRACT

Human emotion detection plays very important role in like Artificial Intelligence particularly in the field of image processing, machine learning etc. The human voice presents vaarious emotions which is an indication of what is going on in mind. Real time emotion recognition is to train the machine to identify and process human emotions. The different modulations which occur in our voice reflects various states of the person. In this paper, a voice input is used to identify the state of person into one of seven predefined emotions by building a multi class classifier. Convolutional neural networks (CNNs) are used for training over voice inputs. Various experiments are conducted with different depths and layers to improve accuracy. The authors present the real-time implementation of emotion recognition which provides accurate results for multiple voice inputs. The results obtained from the research are rather appealing.

Keywords: CNN, Deep learning, Human Emotions, Neural network

INTRODUCTION

Emotion detection from speech [1] or audio is a challenging problem in audio signal processing. A person's characteristics like age, gender, his mental state can be found out from his voice. Recognition of a person's emotion is one among the above. It is completely known fact that the human's speech encloses linguistic content, identity as well as the emotion of speaker. Human emotion plays a major role in daily human interactions. It helps in understanding the feelings of others and also helps others in understanding our feelings. Speech is one such important human emotion. Human- Technology interface is significant in both the quantitative and qualitative

terms. The emotional state and body language of a person can be obtained from Speech communication. Although an enormous effort is invested in recognizing the emotions of a person from speech but still much research is needed. Emotions are universal but their understanding, interpretation and sections are particular and partly culturally specific. Based on the art survey of results in emotion detection, we decided to implement the emotion detection from voice, as most appropriate in the context of application intended. There are many applications to detect the emotion of the persons like in audio surveillance, web-based E learning, and commercial applications, clinical studies, and entertainment etc. Emotion identification can be used as a voice tag in different database access systems. This voice tag is used in telephony shopping, and ATM machine as a password for accessing that particular account. Human emotion recognition is widely gaining more popularity as a research topic. Emotion recognition systems have numerous potential applications, including mental health diagnosis, human- robot interaction, and marketing research. One popular approach to emotion recognition is to use facial expression analysis, where CNNs are trained to detect and classify facial expressions associated with different emotions. Researchers have explored different CNN architectures and training strategies to optimize the performance of these systems.

Zadeh, Milad *et. al.[*2] presented a deep learning based framework for human emotion detection. The framework proposed makes use of the Gabor filters for feature extraction and then the deep CNN. The results depict that the speed and accuracy of training the neural network cabn be increased by the proposed features. Rashid, Munaf *et. al.* [3] presented the design of human emotion recognition system based on sound and spatio-temporal characteristics. The system proposed conducted tests for both genders on audio visual emotion data set.The simulations and results showed that only 74.15 % accuracy can be obtained from visual features alone. An accuracy of 67.39% can be obtained from audio features alone. A significant accuracy of 80.27% can be obtained by mixing both audio and video features. Katsigiannis *et. al.* [4] presented a multimodal database containing EEG and ECG signals which are recorded during affect elicitation using audio-visual stimuli. Anurag Jain *et al*. [5] introduced an algorithm for conversion of emotions. This algorithm needs a database of neural utterances while these can be used for deriving other expressive style utterances. The algorithm proposed relies on linear modification model. Kumar et. al. [6] presented a new approach to envisage human emotions using CNN and to indicate the emotion intensity on a human face.

Salunkeet *et.al.* [7] proposed an artificially intelligent system which can use facial expressions to recognize human emotions. The proposed network has three Convolutional layers which is followed by max pooling and ReLU. Ruiz-Garcia *et. al* [8] provided a deeply-trained model which uses facial expression images for human emotion recognition. Investigational results were presented. Kartali *et. al*.[9] gave an evaluation of five different approaches for four basic human emotion recognition. A comparison of three deep-learning approaches based on CNN and two conventional approaches were given. In [10], Verma *et. al*. proposed a new architecture named Venturi Architecture and evaluated its performance in terms of training accuracy, testing accuracy, training loss and testing loss. Subramanian *et.al* in [11] discussed about the application of emotion recognition where seven different emotions such as happy, sad, neutral, angry, surprise, fear and disgust are obtained. A new hybrid based method was introduced by Rahul *et. al* [12] based on RNN and CNN. These achieved good results by retrieving some parts of the database. The available literature shows that very few works on human emotion recognition based on voice input have been reported earlier. The present work aims to develop a real-time emotion recognition system based on convolution neural networks (CNN's).

MATERIALS AND METHODS

The system is divided broadly into data set formation; pre-processing, feature extraction and classification. MatLab R2014a version is used for programming the system. The Roberson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [1], which consisted of both male and female speech audio samples was used to test for 3 emotions-angry, happy and sad. As a first step, the audio database is divided into two sets namely training and testing sets. Each signal from both the sets is pre processed to make it suitable for data gathering and analysis. In succession, the

features are extracted from the preprocessed signal. The feature vectors are the input to the multi class support vector machine (SVM) classifier which forms a model corresponding to every emotion. Then the test signal is subjected to testing with every model so as to categorize and find out its emotion. Figure 1 shows the recognition model.

Proposed System

Implementation of the DWT

There are several ways to implement DWT. The most straightforward approach is to implement the filters in a Laguerre network (considering first order all-pass filters, A(z),which are reset every N samples). In the second approach, we can implement the filtering by a matrix-vector multiplication in two steps: first we divide the all-pass IIR transfer functions into N terms, and then sample the frequency responses of the warped filter bank to obtain the DWT matrix through an inverse discrete Fourier transform (IDFT).

The second approach is used here which acts as filter bank for an N-tap Finite impulse response filter. Here the inner multiplication of input vector and filter coefficient vector represents the filtering and decimation by N. From Parseval's relation, this is again equal to the inner product of the conjugate DFT of the input and the DFT of the filter coefficients, which is equal to the sampled value of Fk(ejw) for w = (2pk=N) where k = 0;1; :::;N -1: Similarly, we can approximate the result of the filtering with Fk(A(ejw)) as the inner product of the input vector and the IDFT of the sampled sequence of More detailed description about the DWT and its implementations can be found below.

Applying for Speech enhancement:

$$Y_k(t) = X_k(t) + N_k(t), \qquad k = 0, 1, \cdots, M-1$$

A noise signal is assumed to be added to speech signal x. The resulting signal is represented with y. Taking the DWT gives us,

where k denotes the kth the frequency bin, M is the total number of frequency components, and t is the frame index in the time domain, respectively. Given a frame of noisy speech signal, the basic assumption adopted in a speech enhancement approach could be described by the following hypotheses:

The frequencies of these formants are controlled by modifying the shape of the tract. This, for example, could be done by changing the tongue position. An important part of many speech coders and decoders, is the modeling of the vocal tract as a short term filter. The transfer function of vocal tract modeling filter requires to be updated only if it's shape changes. The excitation to the vocal tract filter can be given by forcing air through the vocal cords. Based on their excitation mode, speech sounds can be divided into three types. Sounds are produced when flow of air from lungs to vocal tract is interrupted and this produces air pulses as input. Pitch of the sound can be given as the rate at which opening and closing takes place. By varying the shape of cords, tension in the cords and air pressure behind the cords, pitch can be varied. A high amount of periodicity can be observed which varies between 2ms and 20ms.

This section describes the methodology proposed in this study. This study starts with inputting the real-time voice, followed by the implementation of Convolutional Neural Network (CNN) for recognizing the emotion. Succeeding, the recognized emotion will be displayed. The proposed flowchart in this study is as shown in fig.2. CNN is deep learning based approach which can get high accuracy in recognition. Table 1 show the analysis conducted during training process. The purpose of this analysis is to find the best ratio for dividing the images. The images used for testing the accuracy of application are 80 audio signals. A total number of 80 audio signals have been used during the accuracy testing conducted. From Table 2, there exists a FALSE result to indicate the application is wrongly recognize the emotion. The reason of obtain false result is because the application failed to recognize the correct emotion based on the input voice. For example, the expected result should be sad, but the application recognizes it as happy. Hence, the similarities of the voice cause the application a false result. To conclude overall accuracy performance, the average accuracy is calculated. Equation 1 shows the formula for accuracy calculation.

A=(N/T)*100% (1)

Where A= Accuracy

N=Number of correct predictions

T=Total number of all cases

Here, the numerator indicates prediction predicted by the application, the denominator indicates number of images that were tested. The quotient is multiplied with 100% to get the percentage of accuracy. In this work, the total accuracy is reported to be 85%.

RESULTS AND DISCUSSION

The results are simulated using Matlab software. Figure 3 shows the GUI window displayed showing options for inputting voice signal. To use the program, the user would select the "Browse" option to select an input voice file. This is depicted in figure 4. Once the file is selected, the program displays it in signal form, which may be a graphical representation of the sound wave. The “preprocess” allows many preprocessing functions like filtering, noise reduction etc. to be performed to improve the accuracy of the voice recognition. The "DWT Features" option likely extracts features from the signal using Discrete Wavelet Transform (DWT), which can be used as inputs for the neural network. The "Database" option likely allows the user to access a database of previously trained data to compare against the input signal. The "NN Training" option likely allows the user to train the neural network on a dataset. Finally, the "Classifier" option likely applies the trained neural network to classify the input signal and provide a voice recognition output.

“Preprocess" option in the user interface window allows applying a filtering algorithm to the input signal. This filtering step likely helped to remove unwanted noise or interference from the signal, making it easier to extract useful features for voice recognition. The step “filtering” is important because it helps in improving accuracy and reliability of further stages. By eliminating noise and other unwanted signals, the filtering step can help to reveal the underlying patterns and features in the input signal that are relevant for the task at hand. Figure 5 shows about DWT. It is a powerful signal processing tool that can be used to analyze and extract features from signals. DWT decomposes a signal into a series of wavelet coefficients at different scales and positions, which can be used to identify important features or patterns in the signal. After applying the DWT to the preprocessed signal, the program likely displayed a wavelet transform signal, which is a graphical representation of the wavelet coefficients at different scales and positions. This signal can provide valuable information about the underlying features and patterns in the signal that are relevant for voice recognition.

Features

Input Voice Signal Features:3.3711e+06
Input Signal Features:
DWT:
Maximum Signal Level:0.6744
Minimum Signal Level: -1.0471
Avg Signal Level: 3.6577e-04
Peak Level:0.6744
Median Filter Signal Level:0.0209
Standard Deviation: 0.0687
Histogram : -3.3712e+06
Entropy Level:3.4420
Zero Crossing Rate: 0.0151
Fundamental Energy Level:19.3565

Loading option allows signal to be loaded into database. This helps in storing and organizing data for speech recognition applications. The screenshot is shown in figure 6.The database can be used to store information about the signal, such as its features, labels, or other metadata, which can be used to train and test voice recognition models.

Figure 7 shows NN training. NN training is a critical step in building an effective voice recognition system. During training, the NN algorithm learns to recognize patterns and features in the input signal that are associated with specific words or phrases. The weights and biases of the network can be adjusted to minimize errors between actual predicted values. The training process is time consuming mainly depending on qmount of data and complexity of architecture employed. NN-based systems can achieve high accuracy. The classifier is also a trained ML algorithm which predicts the emotion based on given voice input. The classifier uses the features extracted from the DWT transform and the preprocessed signal to make its prediction. The classifier program runs and displays the predicted emotions like **"happy", "sad", "angry", or "neutral"**. The screenshot is shown in figure 8.

CONCLUSION

A human emotion recognition system based on voice input using CNN is presented. The application is able to recognize four types of emotions which are happy, sad, normal, and disgusting. The Convolutional Neural Network (CNN) used the Mobile Net algorithm with a custom dataset and evaluated using a confusion expression is a valuable expression that portrays human. The developed application achieved an average accuracy of 92.50% in term of the sensitivity and specificity, it able to achieve 85.00% and 95.00% respectively. Hence, the implementation of CNN in recognizing emotion successfully achieved promising results and could be able to contribute to the succession work in CNN. In future the addition of CNN with any Artificial Intelligence method can be carried for further improvement in performance.

REFERENCES

1. Koolagudi, Shashidhar G., and K. SreenivasaRao. "Emotion recognition from speech: a review." *International journal of speech technology, vol.* 15, pp. 99-117,2012.
2. Zadeh, Milad Mohammad Taghi, Maryam Imani, and BabakMajidi. "Fast facial emotion recognition using convolutional neural networks and Gabor filters." *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*. IEEE, 2019.
3. Rashid, Munaf, S. A. R. Abu-Bakar, and Musa Mokji. "Human emotion recognition from videos using spatio-temporal and audio features." *The Visual Computer,* vol.29,pp. 1269-1275,2013.
4. Katsigiannis, Stamos, and NaeemRamzan. "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices." *IEEE journal of biomedical and health informatics,* vol.22, No.1,pp. 98-107,2017.
5. Jain, Anurag, S. S. Agrawal, and NupurPrakash. "Transformation of emotion based on acoustic features of intonation patterns for Hindi speech and their perception." *IETE Journal of Research,* vol. 57, No.4, pp. 318-324,2011.
6. G. A. R. Kumar, R. K. Kumar and G. Sanyal, "Facial emotion analysis using deep convolution neural network," 2017 International Conference on Signal Processing and Communication (ICSPC), Coimbatore, India, 2017, pp. 369-374, doi: 10.1109/CSPC.2017.8305872.
7. Salunke, Vibha V., and C. G. Patil. "A new approach for automatic face emotion recognition and classification based on deep networks." *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. IEEE, 2017.
8. Ruiz-Garcia, Ariel, *et al*. "Deep learning for emotion recognition in faces." *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*. Springer International Publishing, 2016.
9. Kartali, Aneta, *et al*. "Real-time algorithms for facial emotion recognition: A comparison of different approaches." *2018 14th Symposium on Neural Networks and Applications (NEUREL)*. IEEE, 2018.
10. Verma, Abhishek, Piyush Singh, and John Sahaya Rani Alex. "Modified convolutional neural network architecture analysis for facial emotion recognition." *2019 international conference on systems, signals and image processing (IWSSIP)*. IEEE, 2019.

11. Subramanian, R. Raja, *et al*. "Design and evaluation of a deep learning algorithm for emotion recognition." *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2021.
12. Rahul, Mayur, *et al.* "A New Hybrid Approach for Efficient Emotion Recognition using Deep Learning", *IJEER* ,vol.10, No.1, pp. 18-22,2022.

Training the CNN model:

Table 1:Analysis of Training Process

	Splitting images (%)	Accuracy (%)	Loss (%)	Error Rate	Error Rate (%)
Train	**90**	**94**	**53**		
Test	**5**	**80**	**94**	**9/46**	**19.56**
Validation	**5**	**84**	**66**		
Train	80	94	54		
Test	10	78	79	19/91	20.87
Validation	10	89	66		
Train	70	93	53		
Test	15	78	79	28/100	28
Validation	15	79	79		

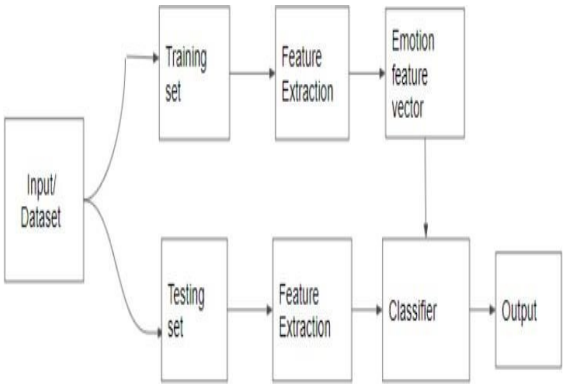

Figure 1: Recognition Model

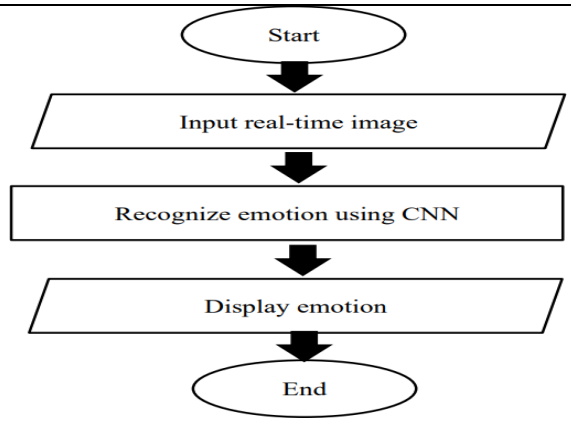

Figure 2.Proposed flow chart

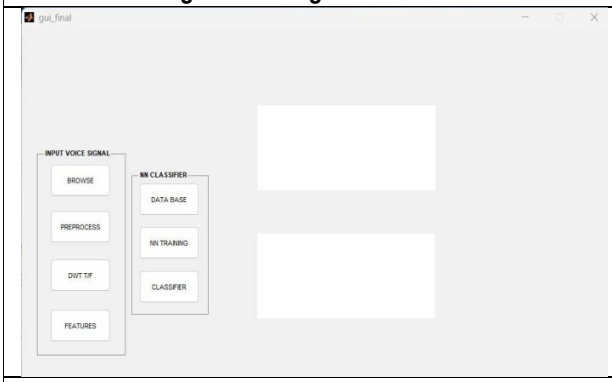

Figure 3:GUI Window

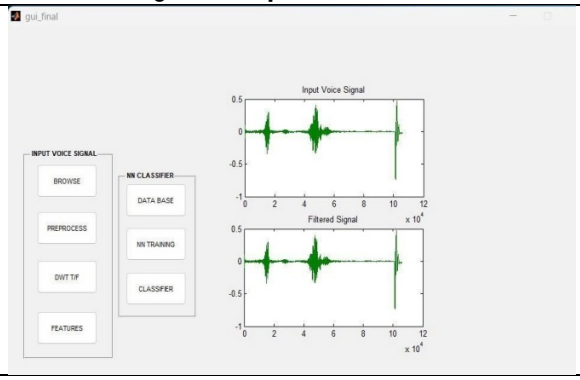

Figure 4:Browse the Input signal

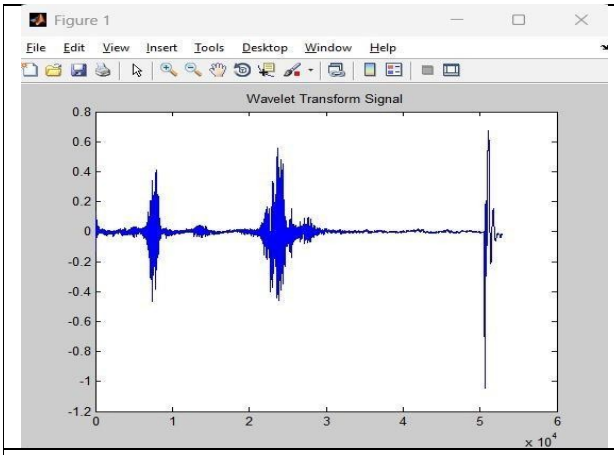

Figure 5: Wavelet Transform Signal

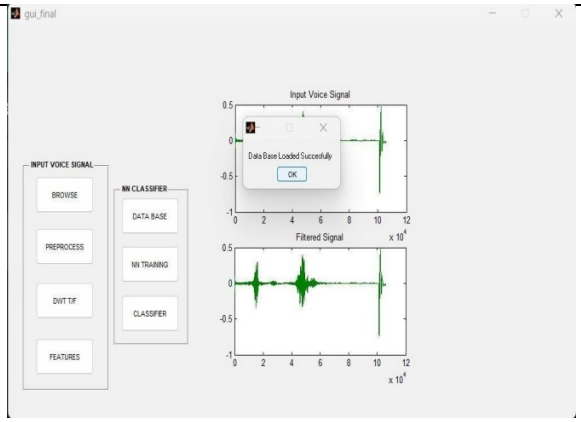

Figure 6: Database

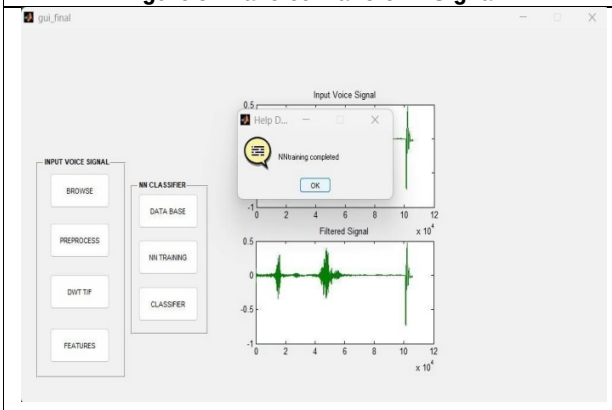

Figure 7: NN Training

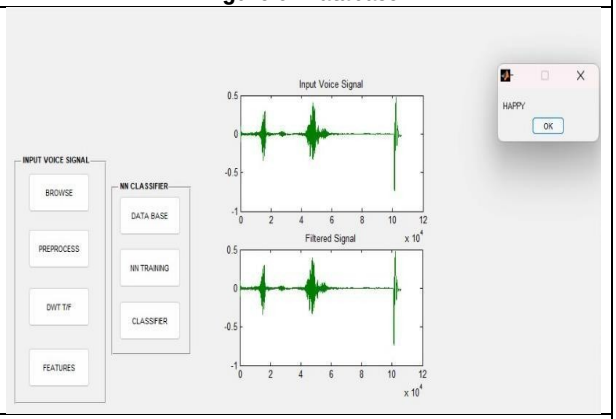

Figure 8: Classifier gives result