

Speech Based Emotion Recognition

Aishwarya Prabha Kumar, Aiswarya Milton Lopez, Akhila Anjanan, Aneena Thereesa

Department of Computer Science and Engineering
Toc H Institute of Science & Technology, Ernakulam
Kerala – India

ABSTRACT

Voice or speaker recognition is the ability of a machine or program to receive and interpret dictation or to understand and carry out spoken commands. Voice recognition has gained prominence and use with the rise of AI and intelligent assistants, such as Amazon's Alexa, Apple's Siri and Microsoft's Cortana. Communication through voice is one of the main components of affective computing in human-computer interaction. In this type of interaction, properly comprehending the meanings of the words or the linguistic category and recognizing the emotion included in the speech is essential for enhancing the performance. In order to model the emotional state, the speech waves are utilized, which bear signals standing for emotions such as boredom, fear, joy and sadness. In the first step of the emotional reactions using neural networks. The sequence is extracted based on the speech signals being digitized at tenths of a second, after concatenating the different speech signals of each subject. The prediction problem is solved as a nonlinear auto-regression time-series neural network with the assumption that the variables are defined as data-feedback timeseries.

Keywords:- Voice Recognition, Artificial Intelligences, Linguistic, Neural Networks

I. INTRODUCTION (HEADING 1)

The ability of humans is vocal interaction is based on linguistic statements and emotional signal that are intrinsically produced. The main elements of vocal communication are the properties of the voice that both humans and animals use to interpret the meanings of the implement novel strategies for enhancing the performance of the interaction between human and computer, which is referred to as Human-Computer Interaction (HCI). HCI develops recognizing the properties of human abilities, in order to implement similar interaction systems in computers.

In fact, humans can transfer and receive the information explicitly, for example, according to the meanings of the words, or implicitly, by incorporating an emotional expression algorithm to be used by a machine while considering only the vocal be performed based on their linguistic, non-linguistic and paralinguistic properties. Nonlinguistic approaches perform the recognition task through signal processing. This research takes a non-linguistics approach for emotion recognition and prediction based on voice.

II. LITERATURE REVIEW

A. Emotion Recognition Using Wireless Systems

This system can be divided into three major components: data collection, feature extraction and model learning. We first collect heart rate data through an android application and store them into a database. After that, we extract suitable features and labels for the problem. Using the features and labels, we find an appropriate model for predict user's emotion. Our system is illustrated in Figure 2.1.

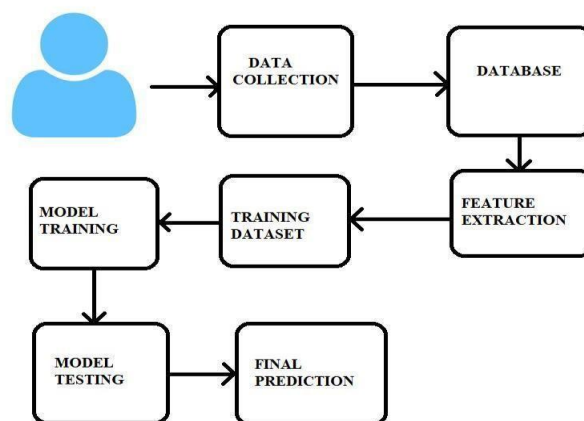


Figure 2.1- Basic working of emotion recognition using wireless systems.

In this system at first emotion and heart rate is collected. To collect data, the system has an Android application, Emotion and Heart Rate Collection that can link to our wearable device to collect heart rates from registered users. To prepare a dataset for choose a model in the system, several participants use the Android app and then, they will record their emotions when they have free times. This process can be easily done by several clicks in our app. In addition, we also collect all possible information from users: age, email, gender and career.

The collection of heart-rate signals is based on six common emotions, they are fear, anger, sadness, disgust, neutral and happiness. All (emotions are clustered into three groups: negative emotion (fear, anger, sadness, and disgust), neutral emotion (neutral), and positive emotion happiness).

After data collection, feature extraction is done. Extracting valuable features from heart-rate signals is one of the most important tasks to in our system. To capture enough signals for building a good emotion prediction system, we consider different time windows in the collected data.

After extract several kinds of feature vectors, we need to find a suitable learning model for our system. Since our current approach is supervised learning, we will analyze several types of supervised learners to obtain the best one. From the collected dataset, we will divide it into two groups: 70% for training and 30% for testing.

Our dataset includes 950 heart rate data with recorded emotions from participants. The detail of collected samples is given in the Table 2.1.

From the raw dataset (only including heart-rates and recorded emotions), three datasets are created with respect to three kinds of time windows. The model selected was SVM and an accuracy of 79% was obtained.

| Type of emotions | Number of samples |
|------------------|-------------------|
| Fear | 92 |
| Anger | 61 |
| Sadness | 59 |
| Happiness | 340 |
| Neutral | 300 |
| Disgust | 98 |
| Total | 950 |

Table 2.1- Data set for emotion recognition using wireless system

B. Emotion Recognition from Videos

In this system we use videos for predicting the emotion. When a videos is gives a input to the system it will first acquire frames from the video source. And from these frames a single frame is selected to analyze and it is submitted to the API. The analysis results are consumed and it is returned from the API call.

The acquired frame is analyzed in 3 different approaches Simple Approach, Parallelizing APIs, Producer Consumer Design.

In the simple approach the frame received is converted into bit patterns and the bit patterns are analyzed with the help of C and #c in this method the emotion is predicted using IfElse condition. In this method at a time only a single frame is analyzed so it takes a lot of time to predict the result. That is a time lag is happening in this method. To overcome this disadvantage parallelizing API approach was introduced. In this approach they have used a frame buffer is used in this method at a time only frame is analyzed from the buffer. So

the frames have to wait for a long time in the buffer. So a latency occurs in this approach.

So to overcome this disadvantage the Producer Consumer method was implemented. Instead of consuming analysis results as soon as they are available, the producer Simply puts the tasks into a queue to keep track of them. We also have a consumer thread, that is taking tasks off the queue, waiting for them to finish, and either displaying the result or raising the exception that was thrown. By using the queue, we can guarantee that results get consumed one at time, in the correct order, without limiting the maximum frame-rate of the system.

c. Emotional Recognition From Brain activity

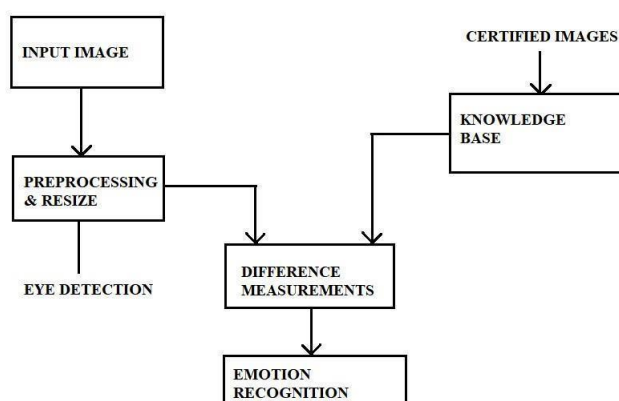
Emotion adjusts the state of the human brain, and directly or indirectly influences several processes. The step-by-step methodology to be followed for Human emotion recognition system using neural networks:

1. Study and analyze various old techniques for human emotion recognition.
2. Based upon above analysis a simulator is developed for Human emotion recognition by using MATLAB 7.5 version (I have only used the simulator it is developed by a team of Luigi Rosa 'ITALY').
3. Results achieved after the execution of program are compared with the earlier outputs.

It contains certified images which we will use for comparisons for the sake of emotion recognition. These images are highly qualified and these are stored in given database. Whenever any input is given to system, system will find the relevant picture from knowledge base by comparing input to certified images and gives a output (emotion).

The main goal of preprocessing and resize step is to enhance input image and also remove various type of noises. After applying these techniques, we need to resize the image that is consider only human face this is done by using eye selection method. It will find the difference between the input image and the certified images (stored in knowledge base) and give result to emotion recognition step.

Figure2.2- Working of emotion recognition system using brain



activity

III. PROPOSED SYSTEM

3.1 Convolution Neural Network Algorithm

A convolutional neural network consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, RELU layer i.e. activation function, pooling layers, fully connected layers and normalization layers. Description of the process as a convolution in neural networks is by convention. Mathematically it is a cross-correlation rather than a convolution (although cross-correlation is a related operation). This only has significance for the indices in the matrix, and thus which weights are placed at which index. There are mainly 5 layers in CNN: Convolution layer, pooling/sub-sampling layer, fully connected layers, fully connected, receptive field, weights.

1.Convolution Layer

When programming a convolutional layer, each convolutional layer within a neural network should have the following attributes:

Input is a tensor with shape (number of images) x (image width) x (image height) x (image depth).

Number of convolutional kernels.

- o Width and height of kernels are hyper-parameters.
- o Depth of kernels must be equal to the image depth. Convolutional layers apply a convolution operation to the input, passing the result to the next layer. The convolution emulates the response of an individual neuron to visual stimuli.

Each convolutional neuron processes data only for its receptive field. Although fully connected feedforward neural networks can be used to learn features as well as classify data, it is not practical to apply this architecture to images. A very high number of neurons would be necessary, even in a shallow (opposite of deep) architecture, due to the very large input sizes associated with images, where each pixel is a relevant variable. For instance, a fully connected layer for a (small) image of size 100 x 100 has 10000 weights for *each* neuron in the second layer. The convolution operation brings a solution to this problem as it reduces the number of free parameters, allowing the network to be deeper with fewer parameters.^[11] For instance, regardless of image size, tiling regions of size 5 x 5, each with the same shared weights, requires only 25 learnable parameters. In this way, it resolves the vanishing or exploding gradients problem in training traditional multi-layer neural networks with many layers by using backpropagation.

2. Pooling

Convolutional networks may include local or global pooling layers. Pooling layers reduce the dimensions of the data by combining the outputs of neuron clusters at one layer into a single neuron in the next layer. Local pooling combines small clusters, typically 2 x 2. Global pooling acts on all the

neurons of the convolutional layer. In addition, pooling may compute a

max or an average. *Max pooling* uses the maximum value from each of a cluster of neurons at the prior layer. *Average pooling* uses the average value from each of a cluster of neurons at the prior layer.

3.Fully connected

Fully connected layers connect every neuron in one layer to every neuron in another layer. It is in principle the same as the traditional multi-layer perceptron neural network (MLP). The flattened matrix goes through a fully connected layer to classify the images.

4.Receptive Field

In neural networks, each neuron receives input from some number of locations in the previous layer. In a fully connected layer, each neuron receives input from every element of the previous layer. In a convolutional layer, neurons receive input from only a restricted subarea of the previous layer. Typically the subarea is of a square shape (e.g., size 5 by 5). The input area of a neuron is called its receptive field. So, in a fully connected layer, the receptive field is the entire previous layer. In a convolutional layer, the receptive area is smaller than the entire previous layer.

5.Weights

Each neuron in a neural network computes an output value by applying some function to the input values coming from the receptive field in the previous layer. The function that is applied to the input values is specified by a vector of weights and a bias (typically real numbers). Learning in a neural network progresses by making incremental adjustments to the biases and weights. The vector of weights and the bias are called a *filter* and represents some feature of the input (e.g., a particular shape). A distinguishing feature of CNNs is that many neurons share the same filter. This reduces memory footprint because a single bias and a single vector of weights is used across all receptive fields sharing that filter, rather than each receptive field having its own bias and vector of weights.

IV. EXPERIMENT AND RESULT

In the proposed system firstly a Non Linguistic input (Voice) is recorded using a microphone. And the voice is passed through different layers of CNN for processing. First the layer is passed through the convolution layer in which the parameter and feature extraction is done. The features considered are Scope, Emotional Intensity and Two Baseline Emotions. The feature scope is important for classification because here we have used the dataset RAVDESS, which contains 7356 clips for training and testing. So the scope of correct prediction of emotion is really high.

The second feature considered is the Emotional Intensity. All emotions have been performed at two levels of emotional intensity, normal and strong. Intensity is one of the most salient aspects of emotion, and has a prominent role in several theories of emotion. Intensity often forms one of several orthogonal axes in a multidimensional emotional space.

Perceptually, intense facial and vocal expressions are identified more accurately than their less intense counterparts. Thus, intense displays may be useful when researchers seek clear, unambiguous emotional exemplars.

Another feature considered is the two baseline emotions. The two baseline emotions considered are neutral and calm. Many studies incorporate a neutral or “no emotion” control condition. However, neutral expressions have produced mixed perceptual results at times conveying a negative emotional valence. The calm expression is not contained in any other set of dynamic conversational expressions.

After passing through the convolutional layer, it is passed through the activation layer. In activation layer, if any of the parameters are having negative value it will be rounded to zero or positive value. And then it is passed through the dropout layer. This is done to avoid overfitting of the parameter values which is then passed through the max pooling layer. In this layer we reduce the number of parameters so that the classification is really simple.

After passing through the max pooling layer it is again passed through convolution layer and activation layer. And then it is passed through the flatten layer to reduce the size of the input and to get fully connected network. And then it is passed through the last layer dense layer which performs classification.

Classification done using comparing the class labels of different emotion. Each emotion is given different labels, such as 0 for neutral, 1 for calm, 2 for happy, 3 for sad, 4 for angry, 5 for fearful, 6 for disgust and 7 for surprised.

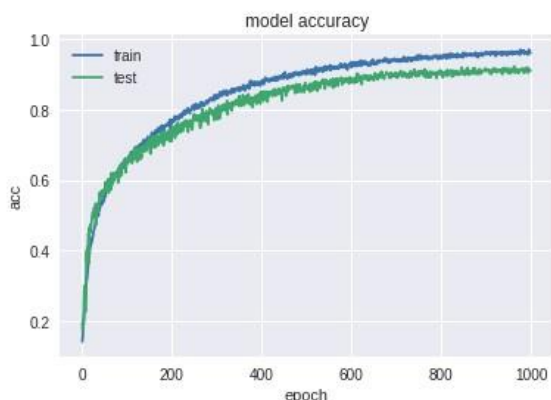


Figure 4.1- Receiver operating characteristics.

| Layer (type) | Output Shape | Param # |
|--------------------------------|-----------------|---------|
| conv1d_3 (Conv1D) | (None, 40, 128) | 768 |
| activation_4 (Activation) | (None, 40, 128) | 0 |
| dropout_3 (Dropout) | (None, 40, 128) | 0 |
| max_pooling1d_2 (MaxPooling1D) | (None, 5, 128) | 0 |
| conv1d_4 (Conv1D) | (None, 5, 128) | 82048 |
| activation_5 (Activation) | (None, 5, 128) | 0 |
| dropout_4 (Dropout) | (None, 5, 128) | 0 |
| flatten_2 (Flatten) | (None, 640) | 0 |
| dense_2 (Dense) | (None, 8) | 5128 |
| activation_6 (Activation) | (None, 8) | 0 |
| Total params: 87,944 | | |
| Trainable params: 87,944 | | |
| Non-trainable params: 0 | | |

Figure 4.2- Model summary.

V. CONCLUSION

Emotion recognition has wide possibilities in the current era where emotional support can be provided as per the mood of the person. Emotion recognition has a wide range of application in intelligent assistant systems and robotics. In the present scenario robot doesn't have the ability to recognize the emotion of the person communicating with it. In all the existing intelligent assistant system it can be used only as a search engine or to perform task it is asked to. So by implementing emotion recognition in intelligent assistant system it can be used to recognize the emotion the person and act accordingly.

All the existing papers on emotion recognition and prediction is based on facial images, videos and text. Text based emotion recognition is not that efficient because in text based recognition the system will be predicting the emotion of the person depending on the words used. But the disadvantage is that if a new word that is given as input to the system that is not being used to train the system, the system will not be predicting the corresponding emotion. Also the words used to represent an emotion can be used for other emotions also. For example, we usually express “hahaha” as a happy text but this can be used to tease a person with a second meaning.

To overcome the challenges of text based emotion prediction rather than using text, the human voice is given as an input. Then the input voice is passed through different layers of CNN computes the values of the parameters considered and compare it with the trained dataset and predicts the emotion of the input given. And finally music playlist will be shown and user can listen to the music.

REFERENCE

- [1] B. Jiang and X. Hu, "A Survey of Group Key Management," 2008IEEE, 2008 International Conference on Computer Science and software Engineering, Vol. 3, pp. 994-1002, Dec. 2008, doi: 10.1109/CSSE.2008.1282.
- [2] Michael E Whitman and Herbert J Mattord, "Principles of Information Security", Vikas Publishing House, New Delhi, 2004I.
- [3] http://en.wikipedia.org/wiki/Communication_in_small_groups
- [4] <https://smartlaboratory.org/ravdess/>
- [5] https://www.microsoft.com/en-us/research/wpcontent/uploads/2016/02/CNN_ASLPTrans2-14.pdf
- [6] http://www.csnow.in/xadm/data_entry_module/project/project_upload/57c53dbfcc7bd2.64636014.pdf
- [7] <https://machinelearningmastery.com/introduction-pythondeep-learning-library-keras/>