Baby Emotional Voice Recognition Using Feed Forward Neural Network

1. Introduction:

Machine learning is the science of getting computers to act without being explicitly programmed. Many researchers think it is the best way to make progress on this through learning algorithm called neural network which mimics how to human brain work. An Artificial Neural Network (ANN) is designed for information processing like biological nervous systems. There are so many Neural Networks applications, like Speech Recognition, Character Recognition, Modeling human behavior, Classification of patterns, Image analysis, Stock market prediction etc. This project is concerned with developing a model for recognizing baby emotional voice using feed forward neural network. The features of baby emotional voice were extracted by the method of Fast Fourier Transformation. In this project, single layer feed forward neural network has been used; this network is generated and trained by Matlab Neural Network Toolbox.

2. Feed Forward Neural Network:

The simplest type of feed-forward (FF) network consist with one input layer, one hidden layer, and one output layer shown in figure -1 that use supervised learning and Figure-2 describes the internal processing at each node. The weights between the connections of nodes from one layer to next layer are initialized randomly and then in each iteration it is updated according to the formula given below.

There is a connection from one layer to next layer and have a weight of each connection denoted by W_{ji}. The input data X_i is presented to input layer and passed to the next layer. The weights are initialized randomly and then they are updated until convergence the network. Output of each node Y is calculated and weight is updated by the following equations.

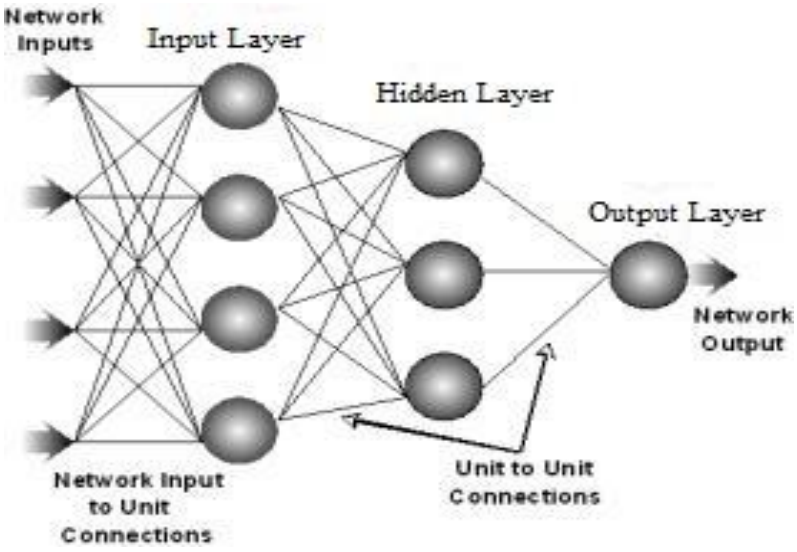

Figure. 1. Feed Forward Neural Network

Input of a node : $X= \sum_{i=1}^{n} WiXi$

Output of a node: $Y = 1/(1+ \exp^{-x})$

Weight update: $W_{New} = W_{Old} + \eta\delta Xi$

Output Layer: $\delta=(T_0 - Y_o)*Y_o*(1 - Y_o)$

T_o = Target output

Y_o = Actual output

η = Learning rate

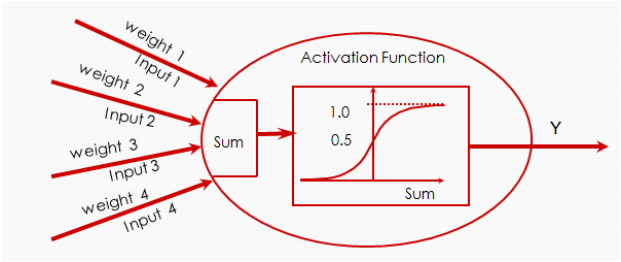

Figure. 2. Processing at each Node

3. Implementation:

In this project, we used five different categories (angry, cooing, laughing, hungry, and sadness) of baby emotional voice for training and testing the network. To make a sample database, these five categories of voice were recorded (two channel) during two second from baby's and save it as *wav* file. Two sample recorded voice of each type are used for training network and another recorded voice used for testing the network. At first step for sampling data, we have converted these wav file to analog signal using wavread matlab function. It returns only first 22050 samples from each channel.

The second step is the most important part of all recognition systems is the feature extraction has been performed by fast Fourier transformation (FFT) that translates the speech signal to some digital form having meaningful features. Then conjugate and transpose are performed on this data set. This same process is applied on two sample recorded wav file of each type emotional voice. Finally, we take the average of those sampled data which is used as input for training the network. The training algorithm has been shown in figure-3 and system flowchart in figure-3.

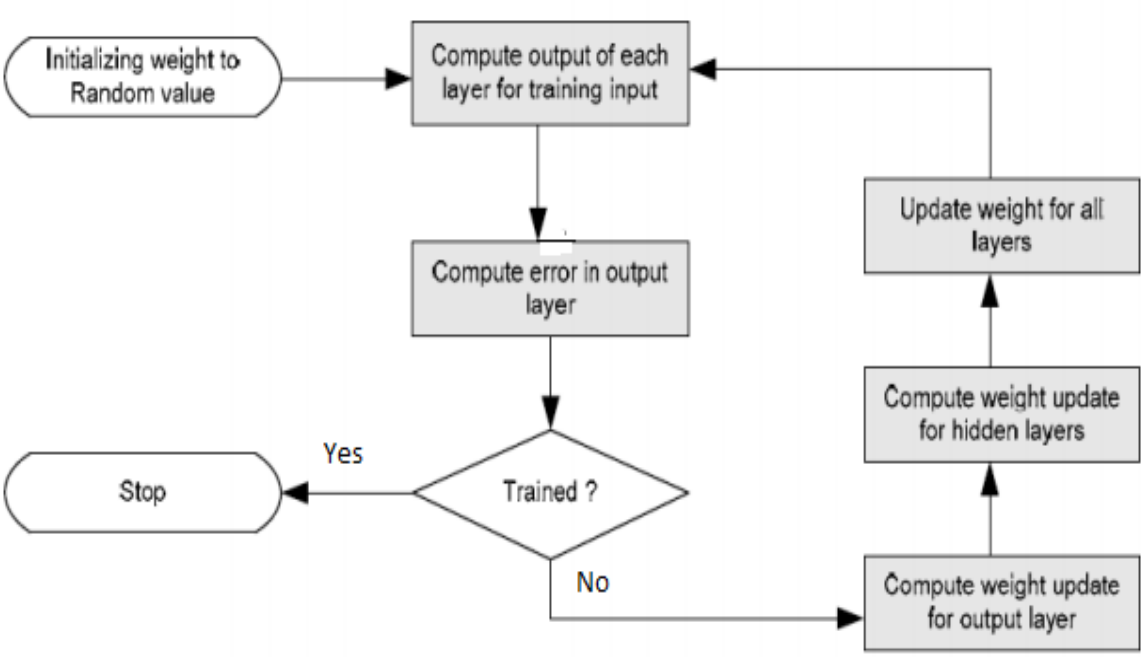

Figure. 3 Back-propagation training algorithm

As it is supervised learning and we have already get our target output t, now we need input of a node to start training the network. For determining input we used frq2mel(frq) function which converts a vector of frequencies (in Hz) to the corresponding values p on the Mel scale which corresponds to the perceived pitch of a tone. After that have to know the range (maximum and minimum value) of Mel scale value so that we can give this range as input during creating network.

For training and testing network, we will create two networks. Net1 is used as training network and Net2 is used as testing purpose. In Net1, input layer have 2206 neurons, hidden layer have 80 neurons and output layer have 5 neurons. Similarly, in Net2 contain 2206 neurons in input layer, 80 neurons in hidden layer and 1 neuron in output layer. Net1 returns [5x2206] matrix where the rows hold the value for angry, cooing, laughing, hungry and sadness. Net2 returns [1x2206] matrix which hold the value for unknown voice. Then compare the output of Net2 with Net1 by determining the root means square difference between them called error calculation. Which one gives the minimum error that means the tested unknown voice is more close to that type of voice compared with others types. Finally, we can recognize the unknown voice.

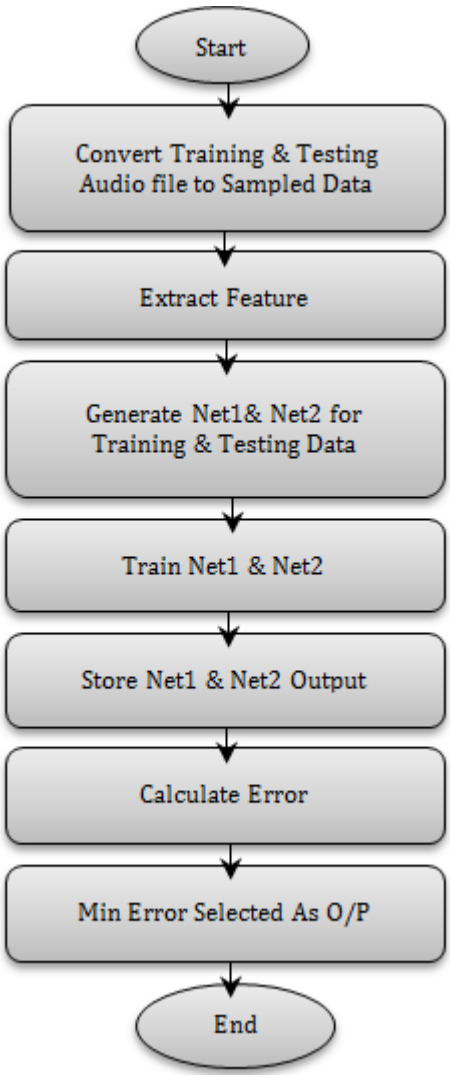

Figure.4. Baby Emotional Voice Recognition System flowchart

To create a Feedforward backpropagation network and training the network we used newff() and train() function respectively. Now we want to create feed forward neural network net1 with one hidden layer, with tangent sigmoid as transfer function in hidden layer and linear function for output layer, and with gradient descent with momentum backpropagation training function, just simply use the following commands:

```
Net1 = newff(minmax(p),[80,5],{'tansig' 'purelin'},'trainlm','learnp');

Net1 = train(Net1,p,t);
```

There are 80 neurons (it can vary) in hidden layer and 5 neurons (fixed) in output layer because we will train the network with different five categories of voice. Here, p is input vector and t is target and tansig, trainlm, purelin, and, learnp these take default value. Now, we are ready to training the network that is done according to following flowchart.

4. Working Procedure:

For running the program we have to maintain some steps. The training (named babyEmoVoice) and testing file have to save in matlab bin directory. Then for recognizing the voice, copy a wav file from testing folder and paste it to babyEmoVoice folder and rename it as unknown.wav Now for recognizing another voice, first delete the unknown.wav file from babyEmoVoice and do the same procedure which I described above. The graphical user interface has been shown in figure-5.

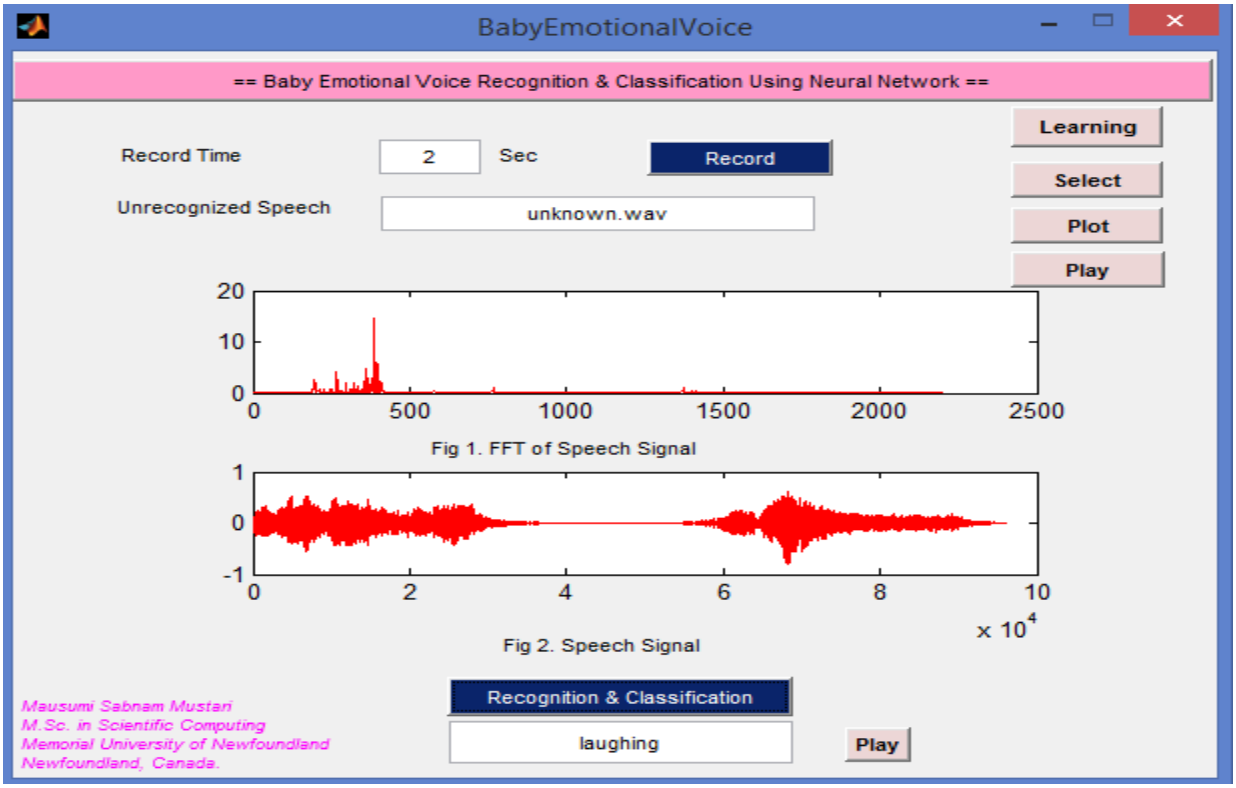

Figure.5. Graphical User Interface

Now after running the program or “write” guide in command prompt, then do the following steps sequentially.

Step 1: hit learning button and wait until the others button actived.
Step 2: hit the select button and select the unknown.wav file from babyEmoVoice folder.
Step 3: hit the plot button and it will show the original speech signal and FFT speech signal.
Step 4: hit the play button and it will play our selected unknown emotional voice.
Step 5: hit the Recognize & Classification button and it will display bellow the name of category which is most similar to our selected unknown voice.
Step 6: hit the play button which play the most similar voice from training wav file.

5. Experimental Result Analysis:

The network has been trained by two sample emotional voices for each category and the recorded time duration for training and testing sample emotional voices are only two sec. The system performance can be improved by training the network with more number of training emotional voice of each category and also need longer recorded time. The detailed results are shown in Table-1. The following experimental result has been produced from the training and testing file stored same age group baby emotional voice.

Emotional voice	Number of Sample for Testing	Unknown Emotional voice for testing	
		No. of Properly Recognized Voice	Recognition Rate (%)
Angry	**3**	**2**	**66%**
Cooing	**3**	**3**	**100%**
Laughing	**3**	**2**	**66%**
Hungry	**3**	**1**	**33%**
Sadness	**3**	**3**	**100%**
Total	**15**	**11**	**73%**

Table1: Details Recognition results