HUMAN EMOTION RECOGNITION BASED ON SPEECH USING BACK PROPAGATION ALGORITHM

Muhammad Iqbal G Putra
School of Electrical Engineering
Telkom University
Bandung, Indonesia
iqbalgputra@gmail.com

Andrew Brian Osmond
School of Electrical Engineering
Telkom University
Bandung, Indonesia
abosmond@telkomuniversity.ac.id

Casi Setianingsih
School of Electrical Engineering
Telkom University
Bandung, Indonesia
setiacasie@telkomuniversity.ac.id

Abstract

The current technological developments are growing very fast and have enormous benefits on our lives. To examine more about how a computer recognizes emotions in humans through the sound media that is processed first and adjust the computer language or in other words Emotion Recognition. This final project uses Neural Network method. Neural Network or also called in bahasa is Jaringan Syaraf Tiruan (JST) is an effective system that can change its structure based on flowing information. Neural Network in this final project is used to conduct training and data testing. While Fast Fourier transform (FFT) is used to process sound data and convert from analog to digital, which is used as input of Neural Network and processed asone of the materials to get the decision whether the voice data used contain emotion or the system cannot detect the emotion.

So after the authors do the test system that has been made obtained 100% accuracy on the test of cross validation, this result shows that the system is very good for use in detecting emotion in a voice. While the test using Indonesian language test data system get 79.1667% accuracy and the test using the English language test data accuracy obtained for 89.1892%, it shows that the system is made better use of English test data because the data train on the system using data practicing english speaking.

Keywords : *Emotion Recognition, Neural Network, Jaringan Saraf Tiruan, FFT.*

I. INTRODUCTION

Technology development is currently developing very fast and has enormous benefits in our lives. Several studies have been conducted in the field of artificial intelligence including and one of them is emotion recognition or an emotional recognition system on human faces with many methods, one of which is the SVM (Support Vector Machine) method. From the results of research, emotional recognition on the human face has an accuracy of 74% [1]. This type of emotion like anger, sadness, pleasure, fear, disgust, and neutral has been agreed upon by experts as a basic emotion. Like the results of a journal from the University of La Laguna by Manuel G. Calvo [2]. A good human and computer interaction system must be able to recognize interpret, and process human emotions. In addition to the SVM method there is also the Naive Bayes method, Couchy Naïve Bayes, VSM (Vector Space Model). All of these methods can be done to process human emotions in the form of images. From the results of the study it can be concluded that smart technology can recognize human faces from photographs. In addition to input using photo / picture media, the human face is also input using sound media, from the resulting sound it can be seen that the object is sad, angry, happy, or neutral. This final project will process the sound with two methods, namely the Fast Fourier Transform and Neural Network methods or in other words, Artificial Neural Networks (ANN) Fast Fourier transform is here to transform the sound signal into a frequency signal, meaning the sound recording process is stored in digital form frequency spectrum so that further data can be processed by the Neural Network method. The Neural Network method functions to obtain training data for data verification and system accuracy.

.

II. SYSTEM OVERVIEW

The simulated system in this experiment has a flow of research process to get better result. The flow of the research process is presented in the form of flowchart on speech recognition system based on speech as follows:

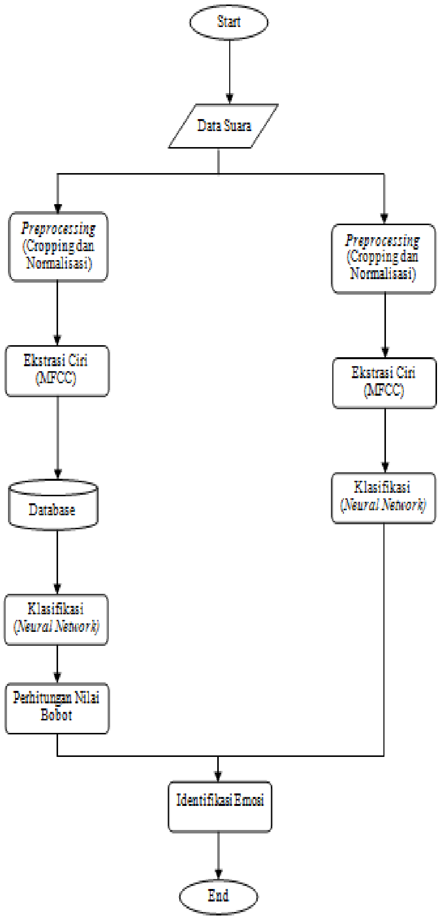
Start
Data Suara
Preprocessing (Cropping dan Normalisasi)
Preprocessing (Cropping dan Normalisasi)
Ekstrasi Ciri (MFCC)
Ekstrasi Ciri (MFCC)
Database
Klasifikasi (Neural Network)
Klasifikasi (Neural Network)
Perhitungan Nilai Bobot
Identifikasi Emosi
End

Figure 2.1 Process of emotion detection using speech

. Test data consists of 8 types of voice data that contain emotions as training data. Details of the types of training data can be seen in the table below :

Table 2.1 Tabel of emotion

No	Emotion data	Amount Of data	Persentase
1	Angry	95	10.41%
2	Sad	125	13.69%
3	Netral	84	9.20%
4	Calm	164	17.96%
5	Disgust	139	15.22%
6	Fearful	68	7.45%
7	Happy	111	12.16%
8	Surprised	127	13.91%
Jumlah		913	100%

2. Pre-processing

Preprocessing on the system is the process of quantization sampling and encoding.

1. Sampling is the process of taking a sample from a sound signal with a frequency band between 300-3400 Hz, where this process is carried out by an amplitude modulator.

2. Quantization is the process of providing fixed prices for PAM signals; the small size is adjusted to the nearest comparator voltage price.

3. Coding is the process of converting PAM signals into digital signals.

3. Feature Extraction

Fast Fourier Transform is an optimal computational algorithm that implements Discreet Fourier Transform (DFT) with fast calculation techniques and utilizes the periodical properties of fourier transforms. FFT is a mathematical operation that aims to decompose a time domain signal to a frequency domain signal. FFT is DFT with a fast calculation technique that utilizes the periodic nature of fourier transformation.

a. Pre-Emphasis
Pre-emphasis is a filter that aims to strengthen high frequency. The filter uses first order with equation 2.1, the typical value is 0.97 in the test.

b. Frame Blocking
This stage aims to divide the signal into frames so that it can do fourier transform which will affect the mel value.

c. Windowing
Aiming to analyze unlimited repetitive times so that the data created by Fast Fourier Transform can be processed by the next step.

d. Fourier Transform
This stage aims to change the domain from the time domain to the signal domain so that it can be processed by the next step. The equation used is 2.3 for the value of N in each frame.

e. Filter Bank and Mel-frequency wrapping
This stage aims to calculate the value of scale. To calculate it using the triangular filter equation using 40 filter banks, then get the filter coefficient value using DCT calculation.

f. Cepstrum
This stage aims to get the mel value which later becomes a pattern for the classification method. Data is saved into a file format .mfc. The following is an example of a mel value:

```
[0.748859  0.887164  0.784837  0.691527  0.887620
0.776858  0.769570  0.846908  0.765208  0.783258
0.882245  0.799754  -0.003987  1.155037  0.524826
0.673646  0.363743  0.446378  0.468476  0.454827
0.479051  0.472398  0.440639  0.501011  -0.004041
1.158567  0.478539  0.617268  0.545264  0.429493
0.546261  0.567528  0.390837  0.465143  0.592720
0.493404  -0.004034  1.160890  0.508957  0.591929
0.474563  0.432195  0.510009  0.512385  0.439554
0.476909  0.473159  0.444976  -0.003861  1.154781
0.576729  0.750883  0.504735  0.488693  0.479062
0.494489  0.439959  0.472717  0.295540  0.411154
0.000000]
```

4. Classifier

It is a stage to determine the output in accordance with the method chosen. In the system created using the back propagation method on the Neural Network. The value of the feature in the .mfc file is then labeled and will be processed by the neural network to be compared. Back Propagation works on the Neural Network is to minimize errors in output generated by the network. Backpropagation is used to look for gradient errors from the network to the weight of a modified network. This gradient error will be used to find the weight value that will minimize errors.

5. Identify Emotions

Finally is the stage to match the sound of each emotion from the existing data obtained after training then the system will compare the value of the data to be tested with the value of the data obtained after the training, the range of comparison values with the data value to be tested is 0.125, if the value the data is different from the test data of 0.125, the system considers it an error or it can be said that the emotion tested is different from the desired emotion.

III. NEURAL NETWORK (NN)

Neural Networks or Neural Networks can be called information processing techniques that are inspired by biological nerve cell systems. Artificial Neural Networks resemble the human brain in two ways, namely knowledge acquired by the network through the learning process and the strength of the relationship between nerve cells (neurons) used to store knowledge. The Artificial Neural Network adopts the basis of the biological nervous system, receives input from a data or from the output of nerve cells in nerve tissue. Each input comes through an existing relationship, and each input and output information pattern given to the Artificial Neural Network is processed in neurons. These neurons accumulate in layers called neuron layers.

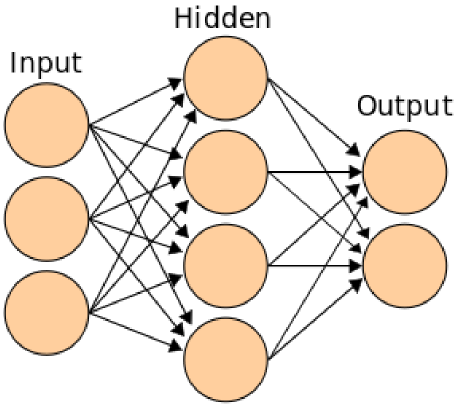

Figure 3.1 neural network structure

1. **Network Device**

Artificial Neural Networks consist of a number of different layers and vertices per layer.

1. Input Layer: consists of unit node units that act as input data processing processes on the neural network.

2. Hidden Layer: consists of unit node units which are analogous to hidden layers and act as layers that pass the response from the input.

3. Output Layer: consists of node units that play a role in providing solutions from input data.

2. **Classification of Artificial Neural Networks**

Artificial Neural Networks can be distinguished based on the level of activation of the output:

1. Singel Layer Perceptrons(SPL)

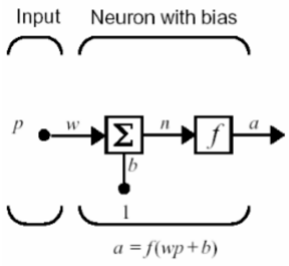

Figure 3.2 Single Layer Perceptrons

SPL or Single Layer Perceptrons are artificial neural network groups that use one input layer and one output. The function used is hard limitting ie the output unit will be worth one, if the amount of input weight is greater than the bias value.

2. Multi Layer Perceptrons(MLP)

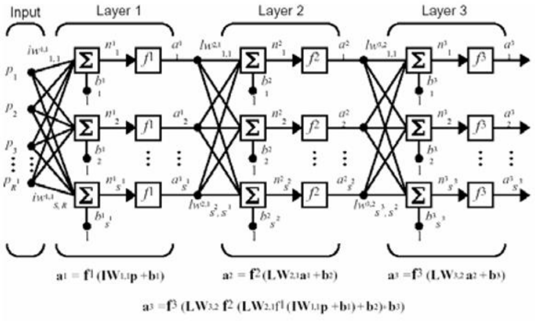

Figure 3.3 Multi Layer Perceptrons

This type is a type of artificial neural network that uses an advanced path with at least one layer hidden or hidden layer. The problem with using this type is how many hidden layers are used to get optimal results or in other words to get the smallest error.

3. **Analysis of Word Recognition**

In the word recognition system using Artificial Neural Networks (ANN) or can also be called Neural Network (NN). NN here is used as a determinant of the results of the system that is to recognize the voice with input from the MFCC. The first step of NN to choose stimuli randomly after that is to calculate the response to the next stimuli to calculate the error obtained and then renew the weight, after which it is repeated until it gets the smallest possible error.

1. Calculating the response to the stimulus
   - ○ Calculating the hidden layer

     h=sigmoid(W*s+bias)

   Calculating Response
2. o=sigmoid(v*h+bias)
3. Sigmoid Transfer Function

$$\text{sigmoid}(x)=1/(1+e^{-x})$$

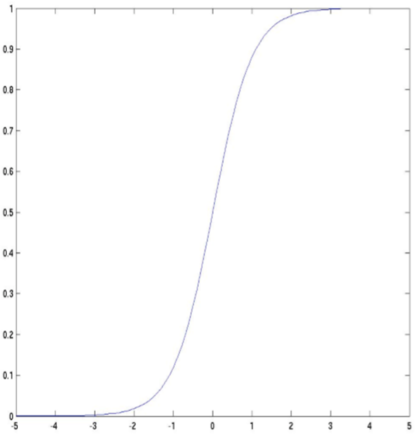

Figure 3.4 Sigmoid transfer function

4. Calculating error

   a.To get the stimuli, the error between the target and the response.

   b. t-o

   c.the value is at 0 or 1

   d. the value of o is between 0 and +1
5. Renew weight

   a. $v = v_{previous} + y(t-o)h^T$

   b. v is weight between hidden-layer and *output*

   c.γ(gamma) is learning rate

IV. EXPERIMENTAL RESULT

There are several tests of the system with several scenarios that are designed by changing the parameters in the method of feature extraction and the classification method are as follows:

1. **Validation**

In the validation of this system, the author uses a cross validation method and uses voice data that is partially separated in the sound data collection of exercises with k = 10, so that in this validation the author gets 100% accuracy because it uses the same sound data as the data used for training. The words used for this validation are "Children talking at the door" and also "The dog is sitting by the door". Using 24 actors consisting of 12 women and 12 men.

Table 4.1 9 result using cross validation method

No	Data Aktual	Net Result
862	1	1.002
863	1	0.9923
864	1	0.99785
865	1	0.99957
866	1	1.0058
867	1	0.99987
868	1	0.99956
869	1	0.99998
871	1	0.99885
Akurasi		**100%**

2. Testing using different epoch

In this test the results of different epochs were tested for 50, 150, 300. The aim was to find the best accuracy with not too much epoch.

- Testing using 50 epoch

Table 4.2 result of testing using 50 epoch

No	Data Aktual	Net Result
862	*1*	0.89101
863	1	1.06341
864	1	0.978917
865	1	0.978917
866	1	1.01963
867	1	0.515592
868	*1*	0.515592
869	1	0.515592
871	1	0.515592
Akurasi		**82.7231%**

- Testing using 150 epoch

Table 4.3 result of testing using 150 epoch

No	Data Aktual	Net Result
862	*1*	0.903249
863	1	0.903545
864	1	0.910448
865	1	0.756746
866	1	0.756746
867	1	0.756854
868	*1*	0.756842
869	1	0.758984
871	1	0.758485
Akurasi		**100%**

Testing using 300 epoch

Table 4.4 result of testing using 300 epoch

No	Data Aktual	Net Result
862	*1*	0.98529
863	1	0.98529
864	1	0.944632
865	1	0.843502
866	1	0.84535
867	1	0.845302
868	*1*	0.84567
869	1	0.88458
871	1	0.845886
Akurasi		**98.627%**

Table 4.5 result of changing epoch

Percobaan ke-	Nilai Epoch	Akurasi
1	50	82.7211%
2	100	89.7025%
3	150	98.627%

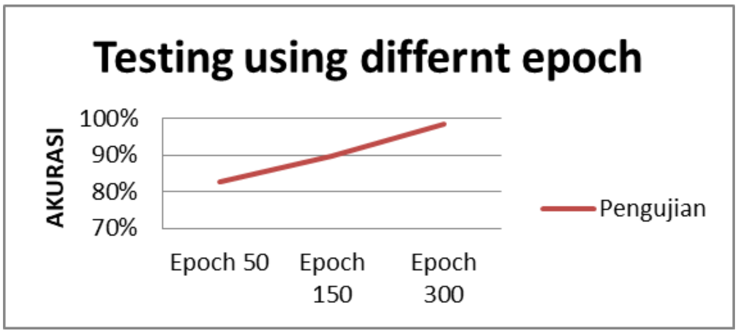

Figure 4.1 Chart of testing using different epoch

3. Manual Testing

In this manual testing the author uses the sound recorded using a mobile phone and will be tested on the system, the final result of this manual testing is to get the desired accuracy. Manual testing is done several times and carried out by 1 man and 1 woman, the language used is Indonesian and English, in the first test the word actor uses the word "baik " for Indonesian and "later" for English, on Testing 3 words- said the actor using the word "kamu lagi apa" for Indonesian and "talking about it" for English, in testing 5 words the actor used the word "orang itu sedang jatuh cinta" for Indonesian and "why do you like it" for England.

4. Testing using indonesian language

In this test the data to be tested using Indonesian language with one female speaker and one male speaker, each male or female emotion actor produces 3 test data which in total all 48 test data are divided into 8 emotions. The results of this test are as follows.

Table 4.6 result of testing using indonesian language

No	Data Aktual	Net Result
1	0.71486	0.292776
2	0.71486	0.318522
3	0.71486	0.293415
4	0.857143	0.352269
5	0.857143	0.312626
6	0.857143	0.32398
7	1	0.32297
8	1	0.330398
9	1	0.327565
10	1	0.35458
Akurasi		**5.40541%**

5. Testing usning english language

In this test data will be tested using English with one female speaker and one male speaker, each male or female emotion produces 3 test data which, if total, 48 test data are divided into 8 emotions. The results of this test are as follows.

Table 4.7 result of testing using english language

No	Data Aktual	Net Result
1	0.71486	0.492776
2	0.71486	0.718522
3	0.71486	0.793415
4	0.857143	0.452269
5	0.857143	0.312626
6	0.857143	0.42398
7	1	0.442297
8	1	0.34444
9	1	0.45445
10	1	0.552465
Akurasi		**59.40541%**

6. Analysis on system testing

After the system is run, there is a difference in accuracy from the first test with one second test word with three words and a third test with five words as follows:

Table 4.8 Table of diffrent accuration on diffrent language

Bahasa Yang Digunakan	Akurasi
Bahasa Indonesia	5.40541 %
Bahasa Inggris	59.4595 %

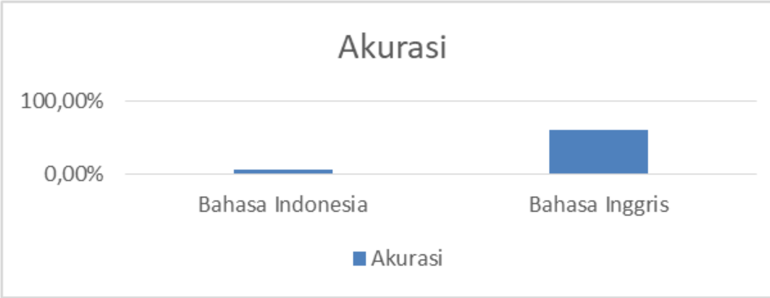

Figure 4.2 Diffrent accuration between language

VI. CONCLUSION

The following is an analysis of the 1 word accuracy comparison test with 3 and 5 words. and a comparative analysis of the language used is english and indonesian :

1. In the first test, the system is tested, whether the system proves to be feasible to use or not to recognize the emotions contained in the test data, when testing is done the test data used uses some training data used for training, this method is called the "cross validation" method when the results are obtained the accuracy shows 100% which means the system is ready for use. Practice the data used in English and say 6 words. This training data is used in each subsequent test.
2. In the first test starting from the comparison of the test data used and the sample data used combined with the times, this can be seen when times are added, accuracy increases, this shows that epoch can affect the system in order to recognize emotions correctly.
3. In the next test, manual testing was carried out using Indonesian language test data, the accuracy obtained was 5.41% compared to testing using English test data which had an accuracy of 59.4595%. In this test it appears that the English test data is superior to the Indonesian language test data, because it uses English language test data which has characteristics more similar to English language training data as well.
4. In the picture of the test results shown there are various words and numbers, can be explained as follows:
   - Data Aktual
     Is the value that is interpreted as an emotional label, the value obtained when taking features and attached to each test data that will be tested.
   - Net Result
     This is the value obtained after calculating the new weight after the training process is carried

out. If stated in mathematical notation, the equation is as follows:

$$Y=f(x1*w1+x2*w2+ ... +XmWm)$$

Or in vector notation :

$$y = f(x + w)$$

x = row vector consisting of m elements,
w = vector column consisting of m elements,
y = scalar amount,
f = nonlinear function.

Simply stated, x is the characteristic value obtained from the MFCC method and w is the weight value obtained when the training is carried out.

- Galat
  Is the diffrent between Net result and Aktual result
- Accuration result
  Accuracy value is the value obtained after calculating the correct data divided by the total data tested multiplied by 100%. Correct data is obtained when the error value is 0> = and <0.3.

REFERENCES

[1] Wibowo, Mochammad Arief W.,2012, Perancangan dan Analisi Deteksi Ekspresi Emosi Pada Model Citra Wajah Dua Dimensi Dengan Menggunakan Metode SVM (Support Vector Machine) dan Logika Fuzzy, Bandung, Tugas Akhir, Program Sarjana Telkom University.

[2] G. Calvo, Manuel., 2008, *Detection of Emotional Faces: Salient Physical Features GuideEffective Visual Search.* IEEE.

[3] Rawat, Arti., 2015, *Emotion Recognition through Speech Using Neural Network*, Volume 5, Issue 5.

[4] Yahya, Arganka, Suyanto., 2012, Deteksi Emosi Melalui Pengenalan Suara Menggunakan *Linear Predictive Coding* (LPC) dan *Hidden Markov Model* (HMM). Bandung, Program Studi S1 Teknik Informatika, Universitas Telkom.

[5] Nurlaily., 2009, Pencocokan Pola Suara Dengan Algoritma FFT dan FC.Medan.

[6] Rao, K.S,; Koolagudi, S.G. 2013. *Robust Emotion Recognition using Spectral and Prosodic Features*. India, Indian Institute Technology Kharagpur. Kharagpur, West Bengal. India.

[7] Singh, Dilbag. 2012. *Human Emotion Recognition System*. Guru Nanak Dev University Amristar (Punjab), India.