Fatchul Arifin *), Tri Arief Sardjono **), Mauridhi Hery **)
*) Electronic Engineering Department, Yogyakarta State University
Kampus UNY karangmalang Yogyakarta 55281
fatchul@uny.ac.id, fatchul_ar@yahoo.com
**) Electrical Engineering Department, Sepuluh Nopember Institute of Technology,
Kampus ITS, Keputih, Surabaya, 60111
t.a.sardjono@ee.its.ac.id, hery@ee.its.ac.id

Abstract— The laryngectomiced patient has no ability to speak normally because their vocal chords have been removed. This drastic change in the human body causes a loss of the ability of speech for the patient. The option for the patient to speech again is esophageal speech or using an electro-larynx. However, the sound has a poor quality and it is often not understandable. Meanwhile, the voice recognition technology has been increased rapidly. The voice recognition technology also can be used to identify a normal, esophageal, and electrolarynx speech correctly.

This paper describes a system for speech identification for normal, esophageal, and electrolarynx voice. Two main parts of the system, feature extraction and pattern recognition were used in this system. The Linear Predictive Coding - LPC is used to extract the feature and characteristic of the human voice. The pattern recognition will recognize the sound patterns correctly. The ***Gradient Discent, Gradient discent with momentum and learning rate, and Levenberg-Marquardt (LM) are used to recognize the pattern of those voices***. All three methods will be compared, and analyze to know which ones can provide the fastest and highest validity.

From the test results, it is known that the ***LM training*** methods give the fastest time, with a validity reached 88.2%. With Intel atom processor N270 1.60 GHz CPU, its learning process takes 0.54 seconds. Meanwhile ***Gradient descent*** training method gives the longest time (660.64 seconds), but has a higher validity, and even reached 100%.

Keywords-voice recognition; Electrolaryn; esophagus; Gradient Discen; Gradient discent with momentum and learning rate; Levenberg-Marquardt (LM)

ElectroLarynx, Esopahgus, and Normal Speech Classification using Gradient Discent, Gradient discent with momentum and learning rate, and Levenberg-Marquardt Algorithm

I. INTRODUCTION

Malignant cancer of the larynx in RSCM hospital is the third ranking after disease of ear and Nose. The average number of larynx cancer patients in RSCM is 25 people per year. More than 8900 persons in the United States are diagnosed with laryngeal cancer every year. The exact cause of cancer of the larynx until now is unknown, but it is found some things that are closely related to the occurrence of laryngeal malignancy: cigarettes, alcohol, and radioactive rays.

Ostomy is a type of surgery needed to make a hole (stoma) on a particular part of body. Laryngectomy is an example of Ostomy. It is an operations performed on patients with cancer of the larynx (throat) which has reached an advanced stage. The impact of this operation will make the patients can no longer breathe with their nose, but through a stoma (a hole in the patient's neck).

Human voice is produced by the combination of the lungs, the valve throat (epiglottis) with the vocal cords, and articulation caused by the existence of the oral cavity (mouth cavity) and the nasal cavity (nose cavity) [3]. Removal of the larynx will automatically remove the human voice. So that post-surgery of the larynx, the patient can no longer speak as before.

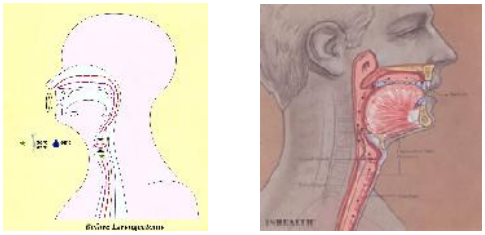

Fig. 1. Before the larynx removed

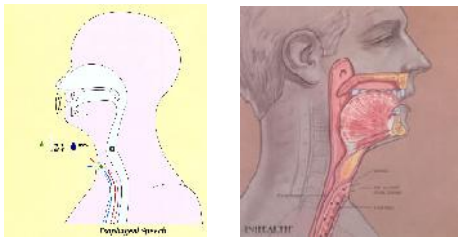

Fig. 2. After the larynx was removed

Several ways to make Laryngectomes can talk again has been developed., for example:

- *Esophageal Speech,*
- *Tracheoesophageal*
- *Electrolarynx Speech.*

Esophageal speech is a way to talk with throat as high as the original vocal cords as a source of sound. The vibration comes from swallowed air, before entering into the stomach[1]. The steps in practice of esophageal speech are blowing, winding, forming a syllable, and speaking. [2].

Tracheoesophageal is a device which implanted between the esophagus and throat. The voice source of this method is esophagus [4]. It can happen, when laryngectomies speaking, the flow of air into the stoma have been closed. So the air will lead to the esophagus through the vocal cords replacement has been planted. This method produces a satisfactory sound, but it has high risk infection risk.

Fig. 3. An implanted voice device

Another device for helping laryngectomies to speak is Electrolarynx. This tool is placed on the lower chin and make the neck vibrates to produce a sound. The sound that produced by electrolarynx is monotone and no intonation at all. So it likes robots and not attractive.

(a) (b)

Fig. 4. a. Electrolarynx,
b. Laryngectomies who spoke with electrolarynx

Meanwhile research in the Speech recognition and its application is now going rapidly. A lot of application of speech recognition was introduced. Some of them are: dialing the phone using voice (eg "call home"), entering simple data into a data base using voice, providing a simple command to a particular machine, etc [5]. Expectation that this technology also can be used by electrolarynx and esophageal speech.

This paper describes how to recognize the voice of electrolarynx, esophageal and normal speech accurately by using gradient discent, gradient discent with momentum and learning rate, and levenberg-marquardt (LM).

II. METHODOLOGY

There are two main parts of the human voice recognition systems, the voice extraction and the pattern recognition. Voice extraction will take important parts of the human voice characteristics, while pattern recognition is used to identify patterns of human voices accurately. In order the sound that will identify have the same duration, it will be done splitting process.

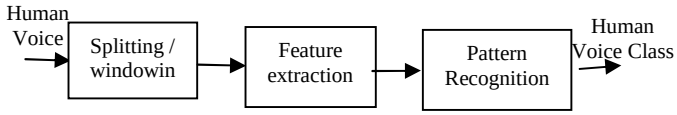

Fig. 5. The general model of voice recognition

From the various existing journal it was known that there are some method for extracting characteristic of human voice, some of them are: Linier Predictive Coding (LPC), wavelet transform(WT), Fast Fourier Transform (FFT), dan Mel-frequency cepstral coefficients (MFCC). On the other hand to perform pattern recognition is also available many choices of method, among others: Artificial Neural Network (ANN), Vector Quantization (VQ), and vector Gaussian models (GVM). [6].

In this research, the LPC method was used for feature extraction step, and ANN method was used for pattern recognition step. However, to increase the sensitivity of the feature extraction, the output of the LPC will be sent to FFT first before it would be processed by ANN. Fig. 6. shows the block diagram of Electrolarynx, Esophagus, and normal voice classification.

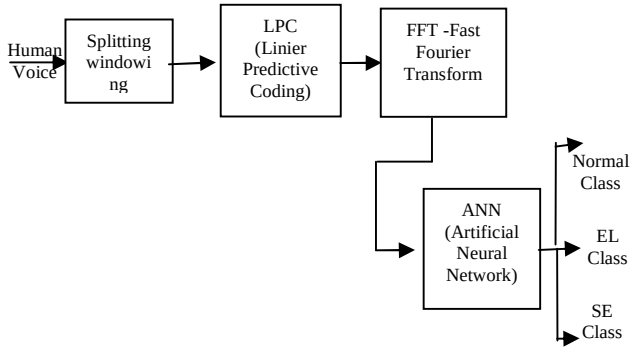

Fig. 6. Block Diagram of Electrolarynx, Esophagus, and normal voice classification

A. *LINIER PREDICTIVE CODING (LPC)*

Linear Predictive Coding (LPC) is one of the techniques of signal analysts. This method is able to identify important information contained in the voice signal. It will produce simple data, but full of information (it minimizes loss of information). Cepstral coefficients, gain, and pitch are parameters generated by the LPC.

The steps in doing LPC can be explained as follows:

a. Preemphassis

At this stage the voice signal is filtered by a single-order FIR filter to refine the spectral signal which has been sampled. It also will reduce noise ratio, so it will increase the signal quality and balance the voice spectrum.

Mathematical model of this stage can be written as follows:

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1) \qquad (1)$$

$$0{,}9 \le a \le 1.$$

b. Frame Blocking

As it was mentioned before LPC is a linier method of human voice analyzing. In order the human voice can be considered linear the analysis must be performed on the short time. At this stage, the voice signal is divided into some frames with each frame containing N sample. The frames adjacent are separated by M. The length of the frame used in LPC that is still considered to be linear is 10-30 milliseconds.

In the mathematic formula, the frame blocking can be written as:

$$xl = \tilde{s}(Ml + n) \qquad (2)$$

where $n = 0,1,2,\ldots,N-1$

$l = 0,1,\ldots,L-1$

xl is the first frame of voice signal

N is the length of data in one frame,

and L is the number of frames.

c. Windowing

Frame blocking process can lead to discontinuation of the voice signal. It can cause spectral leakage or aliasing in the signal. To overcome it, the result of the frame blocking must be processed by windowing. Output signal of windowing process can be written as follows.

$$\tilde{x}_l(n) = x_l(n).w(n) \qquad (3)$$

where $0 \le n \le N-1$

$\tilde{x}_l(n)$ = Output of windowing process

$x_l(n)$ = output of *frame blocking*

$w(n)$ = windowing function

N = The length of data in one frame

There are a lot of window function w(n) that can be used. A good window function should be narrowed in the main lobe and should

be widened on the side lobe. Here are few types of window functions that can be used:

Table 1. Various kinds of windowing [8]

Window type	w_n
Rectangular	1
Hanning	$0{\cdot}5 + 0{\cdot}5 \cos [n/(m+1)]$
Hamming	$0{\cdot}54 + 0{\cdot}46 \cos (n/m)$
Blackman	$0{\cdot}42 + 0{\cdot}5\cos (n/m) + 0{\cdot}08 \cos (2n/m)$
Kaiser	$I_0(\ [1 - n^2 / m^2]^{1/2}) / I_0(\ )$

d. Autocorelation Analysis

At this stage each frame that has been done by windowing process will be done by autocorrelation process. Mathematic model of the autocorrelation process can be written as follows:

$$r(m) = \sum_{n=0}^{N-1-m} \tilde{x}(n).\tilde{x}(n+m) \qquad (4)$$

m = 0,1,2,…,p

e. LPC Analysis

At this stage autocorrelation value of each frame was converted into a set of LPC parameters, reflection coefficient, and logarithmic area ratio coefficient)

f. Conversion of LPC Parameters to Cepstral Parameters

At this stage LPC parameters which are obtained will be converted into cepstral coefficients. Cepstral coefficient is a Fourier transforms coefficients which represent the log magnitude spectrum.

B. FAST FOURIER TRANSFORM (FFT)

Fourier transform is a method to change the time domain signal into frequency domain. This characteristic is important in signal processing because frequency domain provides a clearer picture to be observed and manipulated. In the frequency domain signal is represented as a series of values that indicate the number of signal units in a particular frequency.

Fast Fourier transform (FFT) is an efficient algorithm to compute the discrete Fourier transform (DFT) and its inverse. FFT was first developed by Cooley and Tukey in 1965. The using of FFT was popular, because the FFT can perform calculations faster and able to simplify the number of DFT multiplication of N2 into N logN multiplication.

C. ARTIFICIAL NEURAL NETWORK

Artificial neural network is an algorithm system adopting the ways of human brain work. It has thousands of nerve cells, called neurons. This system has a lot of processors in parallel and distributed. Each processor (neuron) can store knowledge as a result of learning, which will be used to make decisions in next time. Comparison of the human brain nerve cells with artificial neural network architecture can be seen in Fig. 7, and 8.

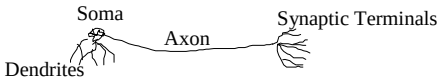

Fig. 7. Neurons of human nerve cells

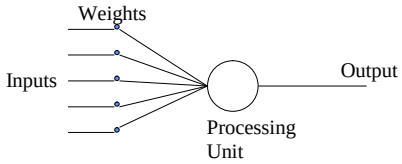

Fig. 8. Neuron of Artificial Neural Network

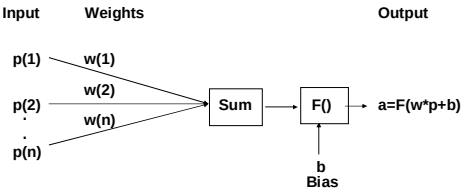

Fig. 9. Details one neuron in ANN

More detail, the description one neuron in the design of artificial neural networks.

$$A = F [W*p + b] \qquad (5)$$

Where:

P = Input Pattern
W = Weight Pattern
F = Activation function

The output of the neuron is obtained by multiplying the input with the weight added by the bias, then inserted into the function activation. Weights and biases obtained from the learning process.

In the human brain nerve cells, there are millions of neurons. Similarly, in the design of artificial neural networks, It can consist of many neurons. Neurons can be in one layer or multiple layers. The relationship of neurons with other neurons can be connected feed forward all or backward.

As mentioned above, this system has the ability to learn and adapt to the environment. The learning process in ANN is the process of finding the the best value of weight (W) and bias (b) for the system.

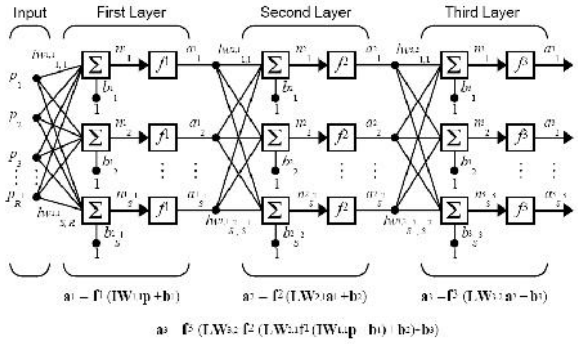

Fig.10. Multi layer feed forward

In general, the updating weight is done by:

$$\mathbf{w}_{kj}(n+1) = \mathbf{w}_{kj}(n) + \Delta\mathbf{w}_{kj}(n)$$
$$\mathbf{b}_{kj}(n+1) = \mathbf{b}_{kj}(n) + \Delta\mathbf{b}_{kj}(n) \quad (6)$$

Weights and biases after learning were obtained by the weights and biases before learning to be added to the delta learning outcomes.

In the design of ANN, in general there are four types of learning algorithms. They are hebbian learning, error correction, competitive learning, and Boltzman learning. One of the error correction algorithms which are very popular is Back Propagation, i.e. to calculate weights and biases through error correction. This correction started from calculating output in forward, then calculate the error backward to the previous layers until the input layer. [9]

In the back propagation, how to update weights and biases can also be done in a variety of ways. Some of its well-known are the gradient descent, gradient descent with momentum and adaptive learning rate, and Levenberg Markov (LM).

III. EXPERIMENTAL

In this paper, human speech will be classified into three classes: normal, electrolarynx and esophagus speech. The speech was recorded from Laryngectomies joined in Societies Esophagus Speech east of java (Dr_Soetomo hospital), who is familiar with proficient of Electrolarynx and esophagus speech.

A. *The Sampling and Normalization process of Voice Signal*

In this research voice signal was recorded and sampled at 8000 Hz with a resolution of 8 bits (1 byte). The speed of sampling is based on the assumption that the human voice signals is in the region of 300-3400 Hz frequency. In order to meet the Nyquist criterion, the sampling frequency must be:

$$f_s \geq 2xf_h \quad (7)$$

$$f_h = f_{in}tertinggi$$

Furthermore, the signal will be normalized in order to get signal with the same size. The normalization process is done by adding some additional data (with a zero value) if the data sample results do not meet the required amount or by cutting the amount of data if the sample results exceed the amount of input required. In this research, the voice signal is limited by 0.5 seconds (It is assumed that pronouncing one word is less than 0.5 seconds). In this research, the voice signal is limited by the duration of 0.5 seconds (assumed to be pronounced one word less than 0.5 seconds). With 8000 Hz sampling, the amount of data obtained as many as 4000 pieces.

B. *Finding LPC Coefficient Process*

As explained earlier, that the LPC is a method of voice analysis linearly. In order it is linier, the voice analysis must be conducted in short frame. In this research, Voice signal will be divided into some short frame. Each frame has 30 milliseconds duration time, and between adjacent frames was separated 10 milliseconds duration time.

So with 8000 Hz sampling, each frame will contain 240 bytes of data with the distance between frames 80 bytes of data. In other words, there is 160 bytes of data overlap, while the data do not overlap in each frame is 80 bytes. Thus the number of frames is

$$\frac{4000-160}{80} = 48.$$

In this paper, for the calculation of cepstral coefficients, the highest LPC order was set to 10, so the output of LPC will get 48x10 pieces of data.

C. *FFT Process*

The process of Fast Fourier Transform (FFT) was performed after obtaining cepstral coefficients as many as 480 data. FFT use 512 points. Because the FFT is symmetric, the FFT output is taken only half of it. It is 256 data. The 256 data are grouped into 32 blocks, so that each block contains 8 data. Furthermore each block is calculated its average. From here It will get the output of the FFT is 32 data. Then this data was sent to the Artificial Neural Network.

D. *Voice Recognition Process Using ANN*

The feed forward architecture with 3 layers is used in this research. In the input layer there are 32 nodes. This is in accordance with the amount of data output from the FFT. While the number of neurons in the output layer is 3, each will show electrolarynx, esophagus, and normal speech. Network training is done by providing: 3 voice of electrolarynx, 3 voice of esophagus, and 3 voice of normal to the network. The allowed limits maximum error is 0.04. This means that if the error had reached that value, the training is stopped.

In this paper, three kinds of learning methods gradient descent, gradient descent with momentum and adaptive learning rate, and Levenberg-Marquardt (LM) were compared.

IV. RESULTS AND DISCUSSION

A. *Gradient Descent Learning Algorithm*

Performance training of gradient descent can be seen at Fig. 11. System error reaches 0.005 on 20645-th iterations. Using Intel atom processor N270 1.60 GHz CPU, its learning process takes 660.4 seconds. Then the system will be tested its validity. Tests conducted on 17 data (4 electrolarynx voices, 4 normal voices and 9 esophagus voices).

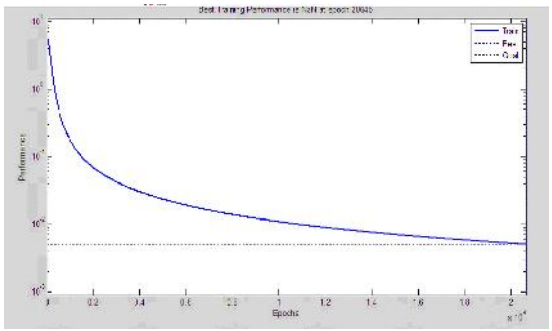

Fig. 11. Performance of Gradient Descent Algorithm

Test results shows that all the data can be recognized correctly, so the system can be said to have 100% accuracy.

Table 2. The results of Gradient Descent Algorithm

No	Sample		Result
1.	Electro Larynx	Data_EL_1.wav	OK
2.		Data_EL_2.wav	OK
3.		Data_EL_3.wav	OK
4.		Data_EL_4.wav	OK
5.	Esophagus (SE)	Data_SE_1.wav	OK
6.		Data_SE_2.wav	OK
7.		Data_SE_3.wav	OK
8.		Data_SE_4.wav	OK
9.	Normal	Data_normal_1.wav	OK
10.		Data_normal_2.wav	OK
11.		Data_normal_3.wav	OK
12.		Data_normal_4.wav	OK
13.		Data_normal_5.wav	OK
14.		Data_normal_6.wav	OK
15.		Data_normal_7.wav	OK
16.		Data_normal_8.wav	OK
17.		Data_normal_9.wav	OK

B. *Gradient descent with momentum and adaptive learning rate* Algorith

Performance training of *gradient descent with momentum and adaptive learning rate* can be seen at Fig. 12. System error reaches 0.0082 on 120-th iterations. Using Intel atom processor N270 1.60 GHz CPU, its learning process takes 3.84 seconds. Then the system will be tested its validity. Tests conducted on 17 data (4 electrolarynx voices, 4 normal voices and 9 esophagus voices).

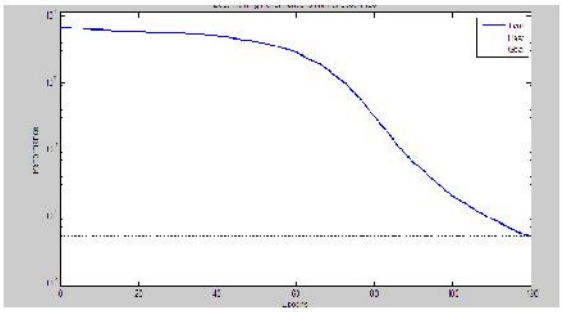

Fig. 12. Performance training of G***radient Descent with Momentum and Adaptive Learning Rate Algorithm***

Test results show that only one the data that can not be recognized correctly. It can be said that the system has accuracy:

$$\frac{16}{17} \text{x} 100\ \% = 94\ \% \cdot$$

Table 3. The result of ***gradient descent with momentum and adaptive learning rate***

No	Sample		Result
1.	Electro Larynx	Data_EL_1.wav	OK
2.		Data_EL_2.wav	OK
3.		Data_EL_3.wav	OK
4.		Data_EL_4.wav	OK
5.	Esophagus (SE)	Data_SE_1.wav	OK
6.		Data_SE_2.wav	Wrong
7.		Data_SE_3.wav	OK
8.		Data_SE_4.wav	OK
9.	Normal	Data_normal_1.wav	OK
10.		Data_normal_2.wav	OK
11.		Data_normal_3.wav	OK
12.		Data_normal_4.wav	OK
13.		Data_normal_5.wav	OK
14.		Data_normal_6.wav	OK
15.		Data_normal_7.wav	OK
16.		Data_normal_8.wav	OK
17.		Data_normal_9.wav	OK

C. *Levenberg-Marquardt (LM) Learning Algorithm*

Performance training of gradient descent can be seen at Fig. 13. System error reaches 0.0001 on 17-th iterations. Using Intel atom processor N270 1.60 GHz CPU, its learning process takes 0.054 seconds.

Then the system will be tested its validity. Tests conducted on 17 data (4 electrolarynx voices, 4 normal voices and 9 esophagus voices).

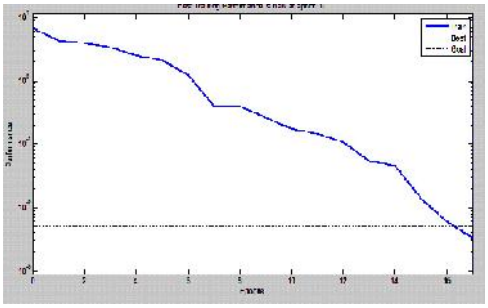

Fig. 13. Performance of *Levenberg-Marquardt Algorithm*

Test results shows that only two data that can be recognized correctly, so the system can be said to have accuracy:

$$\frac{15}{17} \text{x} 100\,\% = 88.2\%$$

Table 4. Result of *Levenberg-Marquardt (LM)*

No	Sample		Hasil
1.	Electro Larynx	Data_EL_1.wav	OK
2.		Data_EL_2.wav	OK
3.		Data_EL_3.wav	OK
4.		Data_EL_4.wav	OK
5.	Esophagus (SE)	Data_SE_1.wav	OK
6.		Data_SE_2.wav	Wrong
7.		Data_SE_3.wav	OK
8.		Data_SE_4.wav	OK
9.	Normal	Data_normal_1.wav	OK
10.		Data_normal_2.wav	OK
11.		Data_normal_3.wav	OK
12.		Data_normal_4.wav	OK
13.		Data_normal_5.wav	OK
14.		Data_normal_6.wav	OK
15.		Data_normal_7.wav	Wrong
16.		Data_normal_8.wav	OK
17.		Data_normal_9.wav	OK

If the three kinds of test results are compared, it will tell us that Levenberg-Marquardt learning provides the fastest solution. Only with 17 iterations, It has been able to find the expected results, while gradient descent with momentum and adaptive learning rate algorithm requires 120 iterations. Moreover, compared with gradient descent, it needs 20.645 iterations for reaching limit error. But in terms of accuracy, Levenberg-Marquardt algorithm has a lower accuracy than momentum and adaptive learning rate or gradient descent. Comparison of the three methods can be seen in Table 5.

However, for processes that do not require too high validity, the LM algorithm is the best choice

Table 5. Comparison of Three learning methods

No	Name of Training Algorithm	Epoch	Time consumption comparison	Accuracy
1	Gradient Descent	20645	660.64 s	100 %
2	Gradient descent momentum and adaptive learning rate	120	3.84 s	94 %
3	LM	17	0.05 s	88.2 %

V. CONCLUSSION

The classification of the human voice, normal, electrolarynx and Esophagus voice were analyzed. The system was built consists of two main parts, they are feature extraction (using the LPC method) and pattern recognition (using artificial neural network). The results show that the system has been able to classify those voices correctly.

The comparison of gradient descent, gradient descent with momentum and adaptive learning rate, and Levenberg-Marquardt (LM) shows that the Levenberg-Marquardt algorithm (LM) provides faster solutions than the other two algorithms. In the other hand accuracy of Levenberg-Marquardt algorithm is lower than the other two methods. For things that do not require a very high accuracy, LM method has become a very attractive option.

BIBLIOGRAPHY

1. Nury Nusdwinuringtyas, Tanpi pita suara: bicara kembali, Blog spot, Februari, 200
2. American Cancer Society. Cancer facts and figures-2002
3. *Fellbaum, K.:* Human-Human Communication and Human-Computer, Interaction by Voice. Lecture on the Seminar "Human Aspects of Telecommunications for Disabled and Older People". Donostia (Spain), 11 June 1999
4. www.webwhispers.org/news/oct2004, Nopember 2009
5. http://en.wikipedia.org/wiki/Speech_recognition , Nopember 2009
6. Mohammed Bahoura, Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes, Internationala journal: Computers in Biology and Medicine 39 (2009) 824 -- 843 , journal homepage: www.elsevier.com/locate/cbm
7. (*http://digilib.petra.ac.id/jiunkpe/s1/info/2005/jiunkpe-ns-s1-2005-26402028-7608-kenal_suara-chapter2.pdf)*
8. *(http://www.dsptutor.freeuk.com/WindowFunctionPlot/notes.html)*
9. Mauridhi hery, Agus Kurniawan, Supervised Neural Network, Graha ilmu Surabaya, 2006