SPEAKER RECOGNITION SYSTEM WITH MFCC FEATURE EXTRACTION AND NEURAL NETWORK BACKPROPAGATION

Eko Riyanto [1)]
Informatics Engineering STMIK HIMSYA
Semarang, Indoneesia
email: ekoriyanto89@gmail.com

Suryono [2)]
Master Information System Diponegoro University
Semarang, Indonesia
email: suryonosur@gmail.com

***Abstract*—Sound is the identity of human beings who are unique and inherent in the human body. Human speech recognition system using human voice that was extracted by the MFCC method, generating matrix and stored in a database. The process of identifying the human voice with voice to match tested and matched to the matrix that exists in the database using artificial neural network algorithm. A number of sound 30 files of 300 people. A sound file in a matrix waveread 2 x 176520 then extracted with MFCC be 10 x 501 (10 Chanal) further compression by PCA method to be 10 x 4 per voice and direshape to 1 x 40 (change column so rows) the end result of the extraction of sound before entering the neural network 40 x 300. Sound pattern recognition using artificial neural networks have a 80% accuracy rate**.

Keywords-component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Sound is the identity of human beings than fingerprint, iris, face and DNA attached to the human body that are unique. Characteristic frequency of the human voice is on the resulting articulator. ASR pronunciation technology (Automatic Speech Recoginition) allow computers to human speech recognize, although it is still limited to a certain language [1].

Speech recognition by proposing a fully-connected network in the hidden layer between input and output. Extraction of features used in speech recognition is LPCC and MFCC. Artificial neural networks have the ability to solve the problem of non-linear wave-like sound [2].

Artificial neural networks can be used to recognize isolated sound. Stage noise speech recognition consists of 2 stages, the pre-processing DSP (Digital Signal Processing) and post processing with artificial neural network with three different models of multi layer back propagation, Elman and probabilistic neural networks [3]

II. THEORY AND METHODOLOGY

2.1. Pattern Recognition

Artificial neural networks (ANN) are composed of simple elements operating in parallel. These elements are inspired by biological neural system so that it can mimic the human brain works. ANN has been trained to perform complex functions in various areas, such as identification, pattern recognition, classification, image restoration, as well as for system control [4]

Simple introduction to the classification using back propagation neural network successful almost 97% of the pattern [5].

Pattern recognition (pattern recognition) can be defined as the process classification of objects or patterns into several kateori or class, and aims to decision making[6].There are three pattern recognition approach, which is the syntax, statistics, as well as through artificial neural networks. Approach with neural network pattern is the approach by combining statistical approach and syntax.

Authentication action to determine or confirm something or someone as authentic, namely claims by or about the object or individual is correct.

2.2. Authentication

Authentication of a person in general include the verification of the identity of the person[7]. Authentication methods used by humans are divided into three:

1. Users themselves / biometrics (fingerprints, retina, DNA, iris, voice)
2. Something owned (identification card)
3. Something to keep in mind (password, PIN)

Pattern recognition with Neural Networks will make intelligent systems by providing training and some rules and statistical data that is used by the system to make decisions[8].

2.3. Speech Signal

Speech signal is a the signal produced from the human voice during a conversation. Speech signal is a complex combination of variations in air pressure that passes through vocal cords and vocal tract, the mouth, tongue, teeth, lips, and palate. Speech resulting from a collaboration between the lungs (pulmonary), glottis (the vocal cords) and the articulation tract (mouth and nasal cavities). Voice signal consists of a series of sounds, each of which stores a piece of information. Based on how it is developed, the sound can be divided into voiced and unvoiced. Voiced speech sounds or noise resulting from vibration of the vocal cords, while unvoiced sounds generated from the friction between the air in the vocal tract. Speech signal has some characteristics,

such as the pitch and intensity of sound that is useful in analyzing the voice signal. Pitch is the frequency of the signal is often called intonation. The intensity of sound is the sound power level.

2.4. Feature Extraction

Specific characteristics to be extracted from the input signal will sound speaker recognition. Extraction is the best parametric representation of the acoustic signal to produce better performance recognition.

cSpecific characteristics to be extracted from the input signal will sound speaker recognition. Extraction is the best parametric representation of the acoustic signal to produce better recognition performance. Mel Frequency cepstral coefficients (MFCC) is one of the most successful feature extraction in speech recognition, and coefficients obtained through the analysis filter bank. The steps are performed in the pre-extraction is emphais, frame blocking, windowing, filter bank analysis, logarithmic compression and discrete cosine transform. Overall process can be seen in Figure 1.

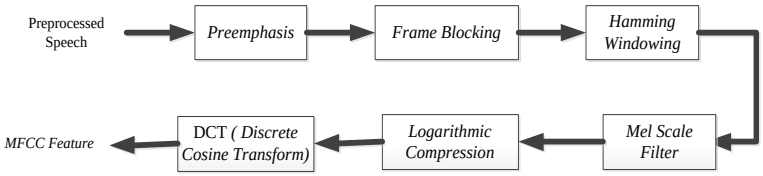

Figure 1. The Process feature extraction with MFCC

1) Pre-Emphasis

Pre-processing voice signals from both databases are provided to pre MFCC feature extraction stage pressure

Pre-Emphasis is a process designed to increase the frequency of some magnitude with respect to the magnitude of other frequencies. The first order of the FIR filter with the transfer function in the z domain.

$$F(z) = 1 - \delta z^{-1} \quad (1)$$

Pre-emphasis coefficient δ located between $0 \leq \delta \leq 1$

$$e(w_i^{r\prime}) = \vartheta e(w_i^{r\prime}) - \delta\vartheta(w_i^{r\prime} - 2) \quad (2)$$

$$e(w_j^{r\prime}) = \vartheta e(w_j^{r\prime}) - \delta\vartheta(w_j^{r\prime} - 2) \quad (3)$$

2) Frame Blocking

Statistical characteristics of the speech signal is invariant only in the short time that interval. Here the signal is blocked into sample frame $f_{N,}$ with adjacent frames being separated by the sample FM (*Frame Shift*). If k^{th} speakers frame is $x_k e(w_i^{r\prime}), x_k e(w_j^{r\prime})$, and K *frame* the overall voice signal, then

$$x_k(w_i^{r\prime}) - \vartheta\left(f_{M^r} + w_i^{r\ \prime}\right), 0 \leq w_i^{r\prime} \leq f_{M^r} - 1 \quad (4)$$

$$(0 \leq k \leq f_{M^r} - 1$$

$$x_k(w_j^{r\prime}) - \vartheta\left(f_{M^t} + w_j^{t\ \prime}\right), 0 \leq w_j^{t\prime} \leq f_{M^t} - 1 \quad (5)$$

$$0 \leq k \leq f_{M^t} - 1$$

3) Windowing

The next process is windowing, where each frame windowed to reduce the signal discontinuities at the beginning and end of the frame. Windowing is chosen to record the edge sinya in every frame. If windowing is defined as bellow :

$$w(w_i^{r\prime}), 0 \leq w_i^{r\prime} \leq f_{M^r} - 1 \quad (6)$$

$$w(w_j^{r\prime}), 0 \leq w_j^{t\prime} \leq f_{M^t} - 1 \quad (7)$$

Then,

$$x_k(w_i^{r\prime}) = x_k(w_i^{r\prime})w(w_i^{r\prime}), 0 \leq w_i^{r\prime} \leq f_{M^r} - 1 \quad (8)$$

$$(w_j^{t\prime}) = x_k(w_j^{t\prime})w(w_j^{t\prime}), 0 \leq w_j^{t\prime} \leq f_{M^t} - 1 \quad (9)$$

Hamming Window is the best option in the voice recognition, which integrates all of the frequency of the nearest line. Hamming Window equation as below

$$w(w_i^{r\prime}) = 0.54 - 0.46cos\left(\frac{2\pi(w_i^{r\prime})}{f_{M^r - 1}}\right) \quad (10)$$

$$w(w_j^{t\prime}) = 0.54 - 0.46cos\left(\frac{2\pi(w_j^{t\prime})}{f_{M^t - 1}}\right) \quad (11)$$

4. Filter Bank Analysis

A filter bank analysis conducted to convert any instance the time of f_N to the frequency. Transformation of fourier applied to convert convolution of signals glottal and response in the domain of the vocal tract signal to domains frequency.

Magnitude of frequency response any filter triangular is equal to the center and the frequency and reduce to zero on two linear frequency nave near tapis. Furthermore, each output tapis is the sum of component filter spectral. Mel scale is defined as follows :

$$M^f = 2595 log_{10}\left(1 + \frac{f}{700}\right) \quad (12)$$

A filter that same called as bank mel scale filter and frequency response of filter bank simulate process perception done in the ear.

5. Logarithmic Compression

Function compression algorithms output filter obtained from analysis filter bank. Output compression filter f_m th expressed as follows:

$$X_{f_{M^r}(\ln\ )} = \ln\left(X_{f_{M^r}}\right), 1 \leq f_{M^r} \leq f_{M^r} \quad (13)$$

$$X_{f_{M^t}(\ln\ )} = \ln\left(X_{f_{M^t}}\right), 1 \leq f_{M^t} \leq f_{M^t} \quad (14)$$

6. Discrete Cosine Trnasformation

Discrets Cosine Transform (DCT) applied to output filter and first coefficients were brought together as vekctor feature of the framework speaker specific. A coefficient k K^{th} MFCC between $1 \leq K \leq C$ as follows :

$$MF_k^r(w_i^{r\prime}) = \sqrt{\frac{2}{f_{M^r}} \sum X_{f_{m^r(\ln}} \cos(\pi k(f_{m^r} - 0.5) f_{m^r})} \quad (15)$$

$$MF_k^t(w_j^{t\prime}) = \sqrt{\frac{2}{f_{M^t}} \sum X_{f_{m^t(\ln}} \cos(\pi k(f_{m^t} - 0.5) f_{m^t})} \quad (16)$$

Where C is sequence Mel Scale Cepstrum.

2.5. Neural Network Backpropagation

Backpropagation is a supervised learning algorithm, and usually used by a perceptron with a lot of layers to change the weights that connect with existing neurons in the hidden layer. Backpropagation algorithm using the error output to modify the weights-weights in the backward direction (backward). To get this error, advanced propagation phase (feedfoward) should be done first. The architecture neural network backpropagation as follow

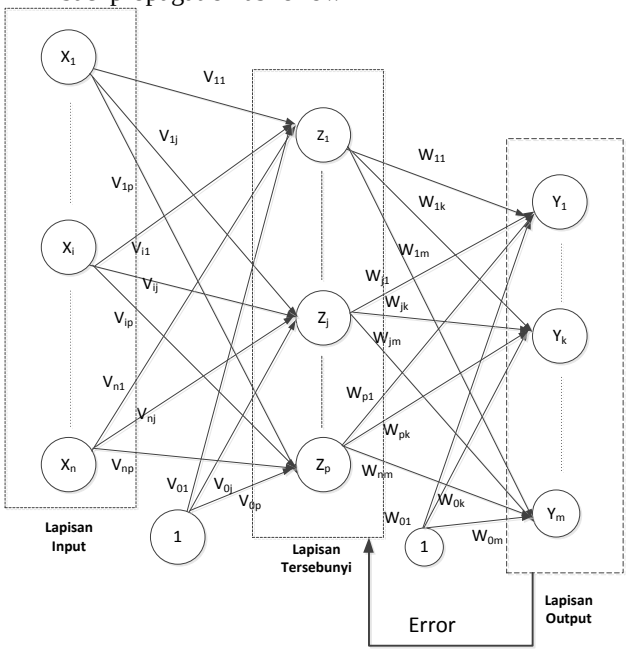

Figure 2. The architecture neural network backpropagation

III. Software Development Methods

Software development method used is the System Development Life Cycle (SDLC) including analysis, design, implementation, testing, and evaluation as shown Figure 3.

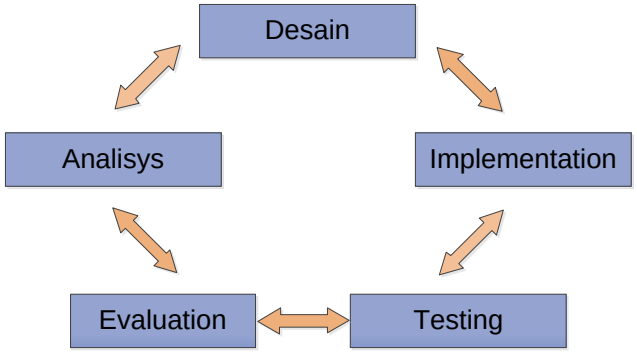

Figure 3. The SDLC method

Software development model for building human speech recognition system as shown in Figure 4.

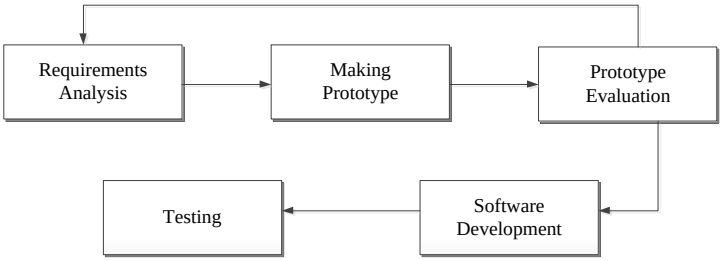

Figure 4. Software development

3.1. Desaign Display

1. The Main Desaign

The principal design program, for displaying the title, buttons training and testing can be seen from figure 5.

Figure 5. The main design of the speaker recognition system.

2. Learning Desaign

Design training data is a step to perform training for some of the data with a recorded voice format. *wav*. There are two processes of training in the pre-processing and feature extraction, as shown in Figure 6.

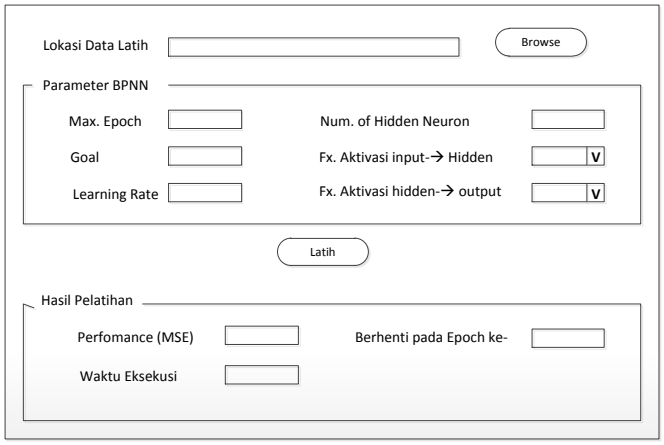

Figure 6. Learning Desaign

3. Testing Desain

Design of testing speaker recognition is step process of matching data, displaying location of the sound file to be tested or it can be record directly the sound using a microphone.

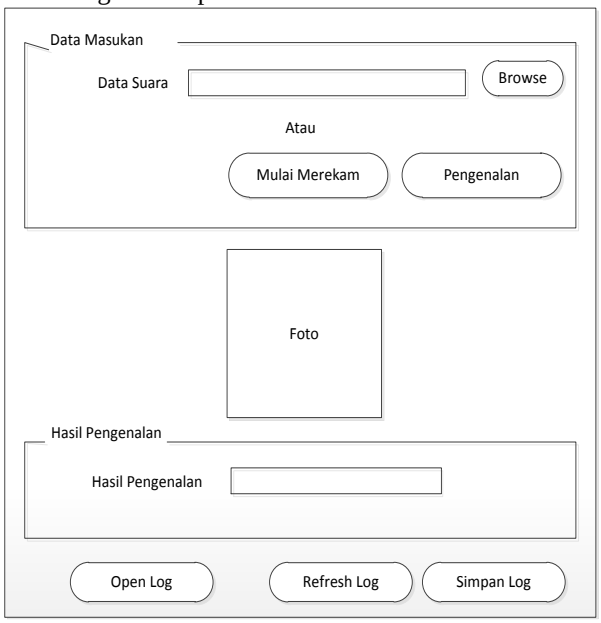

Figure 7. Desaign testing

Can be seen in Figure 7, the data input to display data voice file or that can be tested directly by recording his voice. By clicking recognition results can be seen in the results column recognition.

4. Desaign System

The design steps the speaker recognition system is divided into two steps of training and testing steps voice sound to match the data voice.

Artificial neural networks training stage with sound and then place the extracted feature vector of characteristics and generate each and are stored in the database, and next up is the process of matching data as seen in Figure 8.

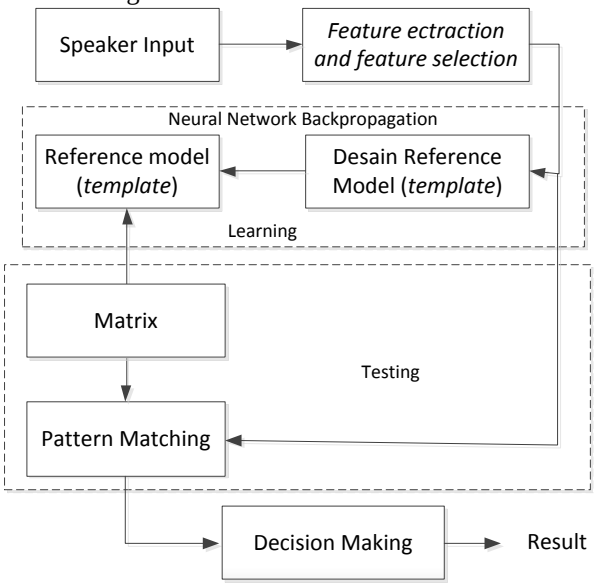

Figure 8. Desain Recognition System

IV. RESULT

4.1. Training artificial neural networks

Artificial neural networks have been trained up a few times to get a high performance and short time in training then testing was done to get a good performance value. The following Table 1.

Table 1.Ttraining results.

max epoch	Goal	Learning	PCA	Function activation input to hidden	Function activation hidden to output	hidden neuron	MSE	Stop at epoch	Times (s)
5000	0,00001	0,01	3	Tansig	purelin	5	0.027781	21	22.973
5000	0,00001	0,01	3	Tansig	purelin	10	0.02917	10	25.095
5000	0,00001	0,01	3	Tansig	purelin	20	0.022276	13	71.931
5000	**0,00001**	**0,01**	**4**	**Tansig**	**purelin**	**5**	**0.028766**	**11**	**13.938**
5000	0,00001	0,01	4	Tansig	purelin	10	0.029114	9	21.672
5000	0,00001	0,01	4	Tansig	Purelin	20	0.039284	10	66.821
5000	0,00001	0,01	5	Tansig	Purelin	5	0.031845	9	11.259
5000	0,00001	0,01	5	Tansig	Purelin	10	0.030467	10	30.284
5000	0,00001	0,01	5	Tansig	Purelin	20	0.025711	11	92.908

4.2. Test Result

Table 2. Is a test results of the test data as many 10 times from each testers name.The above table will show the level of accuracy of each personnel in a test of the recorded sound can be directly calculated by the formula

$$\text{Accuracy} = \frac{\text{Total correct identification}}{\text{Total voice test}} \times 100\% = \frac{245}{300} \times 100\% = 81{,}67\%$$

V. CONCLUSION

Spaker recognition system can research some conclusions as follows.

Speaker recognition system design using characteristic extraction method MFCC (Mel Frequency cepstral Coefficients) and matching the data with the voice behind the propagation neural network has a high degree of accuracy with the MSE (Mean Square Error) which includes small as 0.028766.

In the speech recognition system is tested with accuracy 81,67%.

VI. FURTHER WORK

Some research will be developed next is as follows:

Recognition systems necessary data sound of which various on condition in surroundings diverse so, when tested in environmental conditions diverse has the accuracy is high.

Speaker recognition systems with additional device high-resolution accuracy have a high as well.

VII. REFERENCES

[1] Voll, K, 2007, A Hybrid Approach to Improving Automatic Speech Recognition Via NLP, *Springer*, 514-525.

[2] Ahmad, A.M., Ismail, S, dan Samaon, D.F., 2004, Recurent Neural Network with Bacpropagation through Time for Speech Recognition, *ISCIT* 26, 98-102.

[3] Dede, G, dan Sazli, M.H., 2009, Speech recognition with artificial neural networks.*Digital Signal Processing 20*, 763-768.

[4] Demuth, H, dan Beale, M, 1994, *Neural Network Toolbox for Use with Matlab*, Mass Natick, The Math Work, Inc.

[5] Kamruzzaman, J, dan Aziz, S.M., 1998, A Neural Network Based Character Recognition System Using Double backpropagation, *Malaysian Journal of Computer Science 11(1)*, 58-64

[6] Theodoridis, S. dan Koutroumbas, K. (2006), Pattern Recognition, 3rd edition, Academic Press, San Diego.

[7] Kandim, K. And Yuwanly.2006.Sistem Autentifikasi Dengan Pengenalan Iris.*BINUS, 7-15*.

[8] Kusumadewi, Sri.2003.*Artificial Intelegent (Teknik dan Aplikasinya).*Graha Ilmu, Yogyakarta.