Human Voice Recognition System with Backpropagation Neural Network Method

Mohammad Bagus Dimas Prayugo(✉), Nanda Azzahrotun Nafisa, Azis Yulianas, and Hisyam Fahmi

Mathematic Department, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang, Indonesia
19610001@student.uin-malang.ac.id

Abstract. The system on the computer can make everything run quickly and efficiently, so that it becomes a tool in information processing. One of the computer systems is an Artificial Neural Network (ANN). Along with technological advances, events that require computational models to perform speech recognition can be useful for science, as well as for making practical applications such as voice-based security systems. Artificial neural network is a method of grouping and separating data that has a working system like a neural network in humans. Artificial neural networks can pick up patterns that have been perfectly studied and well received. Backpropagation is a systematic method for training multiple layers of artificial neural networks. The backpropagation network model is composed of an input layer, at least one hidden layer and an output layer. Voice data in the form of signals is converted into discrete data by LPC and FFT methods. The activation function used is the sigmoid function, 2 hidden layers and the number of neurons 15. Optimal training was obtained in the 4th experiment with an MSE error of 0.19413 with a time of 11 s with 678 iterations. System accuracy to training data is 90%, and accuracy to test data is 40% . This means that the level of system accuracy can run well.

Keywords: Artificial Neural Network · Backpropagation · Voice · Mean Square Error (MSE)

1 Introduction

The rapid development of technology cannot be separated from the influence of computers. The system on the computer can make everything run quickly and efficiently, so the computer becomes a very important tool in information processing. One of these computer systems is an Artificial Neural Network (ANN). Artificial Neural Network (ANN) is a network consisting of a group of small processing units modeled on the basis of a human neural network [1]. An artificial neural network is a model that mimics the biological workings of neural networks in humans.

Voice is a gift from Allah SWT, where every human being is given a variety of voices, so that they have their own characteristics in humans [2]. One of the means to identify a

E. Susanti et al. (Eds.): ICGT 2022, AER 221, pp. 432–442, 2023.
https://doi.org/10.2991/978-94-6463-148-7_43

person's personality is to use voice. As time goes by and technological advances bring up an event that requires a calculation model to perform speech recognition that can be useful for science, as well as for making practical applications such as voice-based security systems.

One of the application in the field of technology is pattern recognition. Because the voice pattern is so complete and unique, the voice signal identification process is assisted by calculations that can extract the special characteristics of the voice signal. In this case, an artificial neural network method is used to assist in identifying pattern images, one of which is a signal image that is converted in graphic form.

One of the most commonly used types of artificial neural networks is an artificial neural network using the backpropagation method. This method is a method that is carried out based on the output value approach to the comparison value [3]. Various applications that can be used using an artificial neural network using the backpropagation method include speech recognition.

2 Basic Theory

2.1 Neural Network

A neural network is an artificial representation of the human brain which always tries to simulate the learning process in the human brain [4]. Artificial understanding in this case is used because the neural network is applied using a computer program that can complete the activity process.

Humans have a brain with a complex arrangement. About 10,000,000,000 neuron in the human brain are interconnected [5]. In the human brain consists of neurons in which neurons work according to signals received from other neurons and transmit them to other neurons. One neuron consists of dendrites, axon, and cell body [7].

2.2 Sound

Sound is a mechanical compression or longitudinal wave propagating through a medium. The medium in question is an intermediate substance which can be solid, gas or liquid. Normally sound travels through air, and sound cannot travel in a vacuum. The vibration of an object produces sound, when it vibrates there is a different pressure so that it arises in the surrounding air and becomes sound waves [8].

Voice recognition is a way to recognize or identify voice so that it can be used. Voice recognition can be divided into three approaches, namely artificial intelligence approach, pattern recognition approach, and forensic acoustic approach [9].

2.3 Artificial Neural Network

An artificial neural network is a method of grouping and separating data that has a working system similar to that of a human neural network. Artificial neural networks are created to solve problems, such as identifying or classifying patterns due to the learning process. The artificial neural network model mimics the shape of the neural network

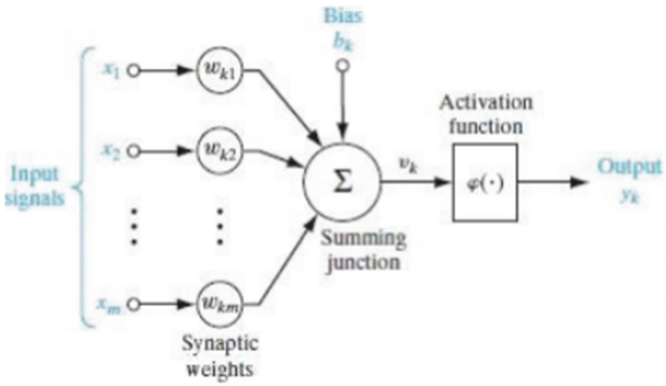

Fig. 1. The flow of the artificial neural network system

found in humans. The creation of an artificial neural network system is intended to allow computers to identify certain shapes, patterns, or structures because the computer system does not have intelligence, even though a computer system can actually perform calculations, such as in a short time identifying human faces rather than humans themselves [4].

Artificial neural networks have the ability to take perfectly learned patterns and produce similar and well-accepted responses to previously learned patterns. According to Drs. Jong Jek Siang, M. Sc, artificial neural networks are determined by 3 things [9]:

a. Network Architecture

   The architecture of a network plays an important role in determining the success of the goals obtained because the same architecture cannot solve all the problems. In an artificial neural network, it consists of simple computational units called "artificial neurons", where each unit is connected to other units via a weight connector and then these units calculate the number of weights that have been inputted and look for the output using activation function (Fig. 1).

b. Training/Learning/Algorithm method (the method used to determine the connection weight).

c. Activation Function

   The activation function of the artificial neural network is used to determine the output of a neuron, the net input (linear combination of input and weight) is the argument of the activation function. There are 3 activation functions that are often used, namely the threshold function, the sigmoid function and the identity function.

2.4 Backpropagation

Backpropagation or reverse error propagation is an artificial neural network method that is commonly used to minimize by adjusting the weights/values based on the difference from the desired output. Backpropagation is one of the systematic methods in training many layers of artificial neural networks. The backpropagation network model is composed of multilayer or many layers, namely the input layer, at least one hidden layer and

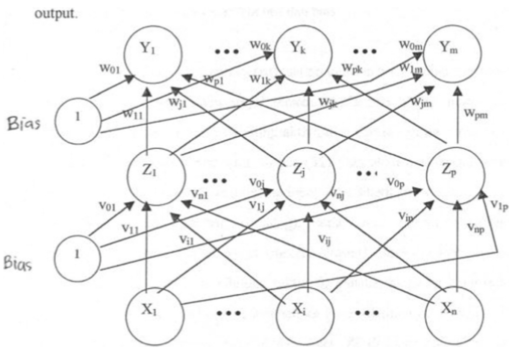

Fig. 2. Backpropagation neural network model

the output layer. Each layer contains a number of neurons or data processing units or which perform their respective functions. As for examples are as in Fig. 2.

2.5 LPC (Linear Predictive Coding)

LPC is a speech signal analysis technique that provides good quality and efficient feature extraction used in calculations [8]. The LPC model is derived from the initial idea that the sample of the voice signal at the n-th time, $s(n)$, can be thought of as a linear composite of several p. It can be denoted as follows:

$$\hat{S}_n \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_i s(n-i) \quad (1)$$

$$\hat{S}_n = \sum\nolimits_{i=1}^{p} a_i S_{n-1} \quad (2)$$

The voice signal in the nth example n-th, $s(n)$ can be approximated with some of the previous examples by using the prediction coefficient ai, the constants in the voice signal analysis block are assumed to be coefficients $a1, a2, \ldots, ai$. If the value of the square of error between the actual signal sn and the predicted signal sn is small, then the estimated value of ai is the best result.

2.5.1 Preemphasis

Signal data is taken to minimize the area of signal change. The digital form of the voice sample is filtered to even out the spectral signal using a first order Finite Impulse Response (FIR) filter. The purpose of this filter process is to obtain a finer frequency spectral shape of the sound signal.

$$\tilde{s}(n) = s(n) - 0.9375 \times s(n-1) \quad (3)$$

where:
$\tilde{s}(n)$ = Preemphasis function
$s(n)$ = Signal data n-th
$s(n-1)$ = Signal data $n-1$-th

2.5.2 Frame Blocking

The compressed voice signal is divided into frames in which each frame contains N voice clips and several adjacent frames are separated by M voice clips.

2.5.3 Windowing

In the windowing step, the weighting function is carried out on each frame that was created in the previous step. The type of window that is often used is the Hamming window which has the following general form:

$$w(n) = 0.54 - 0.46 \times \cos\frac{2\pi n}{N-1}, 0 \le n \le N-1 \tag{4}$$

2.5.4 Autocorrelation Analysis

Each frame that has been input into the windowing is autocorrelated with the highest autocorrelation value which is the order of the LPC analysis.

$$r_{\varepsilon}(u) = \sum\nolimits_{v=0}^{V-1-u} \overline{x}_{\varepsilon}(v)\overline{x}_{\varepsilon}(v+u),\ u = 0, 1, \ldots, p. \tag{5}$$

2.6 FFT (Fast Fourier Transform)

FFT or Fast Fourier Transform is a mathematical computational technique which is used to transform digital signals from analog signals based on frequency. FFT is an algorithm to perform calculations on discrete Fourier transforms efficiently and quickly. The data obtained by FFT is the input data for the artificial neural network.

$$S(f) = \smallint_{-\infty}^{\infty} s(t)e^{j2\pi ft}dt \tag{6}$$

where:

$S(f)$ = signal at frequency
$S(t)$ = signal at time
$s(t)e^{j2\pi ft}dt$ = signal value constant
f = frequency
t = time

Based on the integral equation above, it can be seen that the Fast Fourier Transform can be used in calculating the frequency value, signal wave phase and amplitude. As for the calculation of the frequency spectrum of a signal contained in the computer, a Disrete Fourier Transform (DFT) algorithm is needed to replace the signal in the time domain so that it becomes a signal in the frequency domain.

$$F(u) = \frac{1}{N}\sum\nolimits_{X=0}^{X=N-1} f(x)\exp\left[-\frac{2j\pi ux}{N}\right] \tag{7}$$

$$F(u) = \frac{1}{N}\sum\nolimits_{X=0}^{X=N-1} f(X)\left(\cos\left(\frac{2\pi ux}{N}\right) - j\sin\left(\frac{2\pi ux}{N}\right)\right) \tag{8}$$

2.7 MSE (Mean Square Error)

Mean Square Error (MSE) is one of the calculation methods used to evaluate the prediction results. Each error or remainder is squared [10]. This method manages large forecasting errors because the errors are squared. The calculation for MSE is as follows:

$$MSE = \frac{1}{N}\sum\nolimits_{i=1}^{N}(t_k - y_k)^2 \tag{9}$$

MSE = mean square error calculation.
t_k = target
y_k = training output results
N = amount of data.

3 Research Methodology

3.1 Designing System

The system to be designed is a system that can be used to identify various kinds and many types of voice models that are appropriate to the problems that have been formulated. This system will identify the voice of the subject that has been previously entered into a folder. This system will be designed to recognize a person's voice after going through several processes as shown in Fig. 3.

3.2 Input Voice

The voice recording process used in this research was obtained by taking samples of men, women and small children. As for the words that are spoken is the word "Hi" for each person. The recorded sound is a WAV format sound. Voice recordings are collected in one folder for easy searching.

3.3 Voice Acquisition

Acquisition is the process of taking something or a certain object which will later be added to something or a certain object that is already owned. In this case, voice acquisition means the process of extracting the maximum voice from the data obtained by cutting or removing the parts that are not needed. This serves to equalize the format of the desired voice.

The voice signal data obtained is filtered using BPF (Band Pass Filter) to reduce the noise in the voice. Then the filtered voice will be sampled to get a discrete form of the voice signal. As for converting the voice signal to discrete, a condition is required, namely, nyquist. Nyquist is the minimum sampling rate (number of sensors per inch) that results in a signal that still contains all of the original signal information that is twice the maximum frequency in the original signal as shown in Fig. 4.

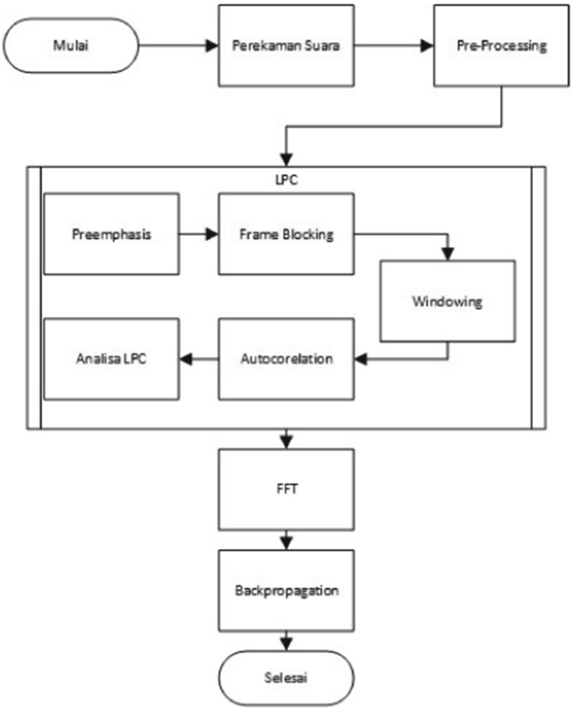

Fig. 3. Voice recognition system flowchart

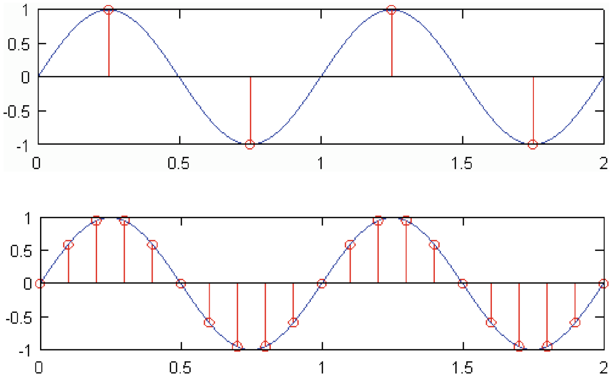

Fig. 4. Voice signal nyquist process

3.4 LPC (Linear Predictive Coding) Process

Signal data is taken to minimize the signal change area. This results in a finer spectral form of the voice signal. The compressed voice signal is divided into frames where each frame contains N voice clips and several adjacent frames are separated by M voice clips. After the voice signal is completed in the frame blocking step, in the windowing step the weighting function is performed on each frame that was created in the previous step. In the autocorrelation analysis step, each frame that has been inputted into the windowing is autocorrelated with the highest autocorrelation value which is the order of the LPC analysis. The results of the autocorrelation analysis are then carried out by LPC analysis to obtain quality and efficient feature extraction which is used in calculations.

3.5 FFT (Fast Fourier Transform) Process

FFT or Fast Fourier Transform is a mathematical computational technique used to transform digital signal data from an analog signal based on frequency. The voice data that has been processed on the LPC is transformed with FFT to facilitate data processing on the artificial neural network.

3.6 Backpropagation Neural Network

The backpropagation neural network trains the network to get a balance of each network's ability to recognize the patterns used during training. The ability of the network to respond correctly to input patterns is similar to the pattern used during training. The backpropagation artificial neural network used in this research consists of input, hidden layer, hidden neuron, activation function, and the number of epochs.

4 Analysis and Results

4.1 Collecting Voice Data

The voice data obtained were 40 votes taken from 10 participants with 4 votes for each participant. Voice recording is done by spreading Google Forms, sending via WhatsApp, and direct recording via Smartphone. The voice data that has been obtained is collected in one folder to make data processing easier.

4.2 Voice Data Processing

Voice data in the form of signals is converted into discrete data. The sound signal is filtered by using a Band Pass Filter (BPF) to reduce noise in the voice. The filtered voice will then be normalized and trimmed to the voice. Normalization and voice cutting aims to eliminate the silent voice that exists in the voice. The voice is then converted into a discrete voice signal so that it can be calculated. As for the process of converting into discrete signals using Linear Predictive Coding (LPC) and Fast Fourier Transform (FFT) with the help of the Matlab program.

Based on the FFT processing that has been done, the vector of each sound with a length of 64 frequency signal waves is obtained. As for the results of the FFT as shown in Fig. 5.

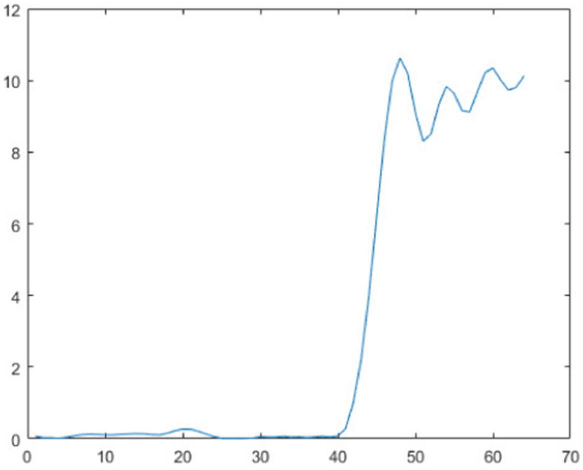

Fig. 5. FFT on voice signal

4.3 Data Type

The data consists of two types, training data and test data. Training data is data that is used to train the algorithm in finding suitable data. The training data consisted of 30 voices consisting of 3 voices each participant. Test data is data used to test and determine system performance obtained at the testing phase. Test data consists of 10 voices from each participant.

4.4 Training Data

The voice data used as training data are 30 voices and 2 layers are used. The training process is carried out using the backpropagation artificial neural network method. The activation function used is the sigmoid function. The error value calculation used is the mean square error (MSE). Signal training experiments were carried out 4 times with different iterations. The parameters used are; error = 0, learning rate (a) = 0.01, hidden layer = 2, and number of neurons = 15.

Based on the training experiment in Table 1, the optimal training was obtained in the 4th experiment. The training time obtained to approach the error target 0 is 0.19413. The training time obtained is 11 s with 678 iterations.

From the training data it is found that there are 3 votes that are predicted to be incorrect. So that the accuracy value of the training data can be calculated sing Eq. (10).

$$accuracy = \frac{total\,voice\; - failure}{total\,voice} \cdot 100\% \tag{10}$$

It can be concluded that the accuracy of training data is 90%.

4.5 Testing Data

The voice data used in testing the test data there are 10 voice data. Testing the test data by using the Matlab application produces the following results in Table 2. From the data

Table 1. Network training results for training data

Trial	Iteration	Time (Second)	*Mean Square Error* (MSE)
1	2000	13	6.6751
2	45	13	0.27858
3	8	28	1.4493
4	678	11	0.19413

Table 2. Test results of test data

Name	Target	Training Result	*Error*	Classification	Conclusion
Azis	1	1.3027	−0.3027	1	True
Dara	2	2.6862	−0.68618	3	False
Dimas	3	6.708	−3.708	7	False
Donny	4	2.9472	1.0528	3	False
Ike	5	4.8657	0.13425	5	True
Nada	6	3.2998	2.7002	3	False
Nanda	7	9.3898	−2.3898	9	False
Thalia	8	7.876	0.12404	8	True
Tirta	9	1.3194	7.6806	1	False
Wahyu	10	9.8003	0.19969	10	True

there are 10 voices data of which 4 out of 10 voices data are correct. So that the accuracy of the test data in recognizing voice is obtained by 40%.

5 Conclusion

Based on the results of the research that has been carried out, it is found that the backpropagation artificial neural network system can be used in the identification of human voices. Before training and testing with Backpropagation there are several steps that need to be done, namely converting voice signal data into discrete data with Linear Predictive Coding (LPC) and Fast Fourier Transform (FFT). Parameters used; error = 0, learning rate (a) = 0.01, hidden layer = 2, number of neurons = 15. Optimal training was obtained in the 4th experiment with MSE 0.19413 with a travel time of 11 s. The accuracy of the training data is 90% and the accuracy of the test data is 40%, so it can be concluded that the accuracy of the system can run quite well.

References

1. Solikhun, S., Safii, M., & Trisno, A. (2017). Jaringan Saraf Tiruan Untuk Memprediksi Tingkat Pemahaman Sisiwa Terhadap Matapelajaran Dengan Menggunakan Algoritma Backpropagation. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *1*(1), 24–36. https://doi.org/10.30645/j-sakti.v1i1.26
2. Fatman, Y., & Islamiyati. (2020). Pengenalan Suku Kata Bahasa Indonesia Menggunakan Metode LPC dan Backpropagation Neural Network. *JOINTECS (Journal of Information Technology and Computer Science)*, *5*(3), 155–166.
3. Zaitun, Warsito, & Pauzi, G. A. (2015). Sistem Identifikasi dan Pengenalan Pola Citra Tanda-Tangan Menggunakan Sistem Jaringan Saraf Tiruan (Artificial Neural Networks) Dengan Metode Backpropagation. *JURNAL Teori Dan Aplikasi Fisika FMIPA Universitas Lampung*, *03*(02), 93–101.
4. Adinugraha, T. A. C. (2016). *Prediksi Jumlah Pendapatan Asli Daerah Kabupaten Boyolali dengan Metode Jaringan Syaraf Tiruan Backpropagation*. Universitas Islam Negeri Maulana Malik Ibrahim Malang.
5. Loppies, S. H. D. (2018). Implementasi Jaringan Saraf Tiruan Backpropagation Untuk Deteksi Wajah Dalam Citra Digital. *Musamus Journal of Technology & Information*, *1*(1), 1–7. https://doi.org/10.35724/mjti.v1i1.991
6. Setiawan, P. (2021). *Pengertian, Jenis Dan Struktur Neuron (Sel Saraf)*. Gurupendidikan.Com. https://www.gurupendidikan.co.id/neuron-sel-saraf/#Struktur_dan_Bagian_Neuron_Sel_Saraf
7. Ferdinando, H. (2010). Dasar-dasar Sinyal dan Sistem. *ANDI*.
8. Dinuriyati, I. S. (2019). *Klasifikasi Pengenal Suara Kicau Burung Menggunakan Metode Linear Predictive Coding (LPC) dan Nearest Neighbor*. Universitas Islam Negeri Maulana Malik Ibrahim Malang.
9. Siang, J. J. (2005). Jaringan Syaraf Tiruan dan Pemograman Menggunakan Matlab. In ANDI (Ed.), *ANDI Yogyakarta*. ANDI OFFSET.
10. Ramadhanty, A., Cholissodin, I., & Dewi, C. (2017). Prediksi Jumlah Permintaan Koran Menggunakan Metode Extreme Learning Machine. *Repository.Ub.Ac.Id*, 1–7.