International Journal of Artificial Intelligence & Robotics (IJAIR) E-ISSN: 2686-6269
Vol.1 No.1, 2019, pp.33-38

Speech to Text Processing for Interactive Agent of Virtual Tour Navigation

Dian Ahkam Sani[1*], Muchammad Saifulloh[2]
[1,2]Universitas Merdeka Pasuruan, Pasuruan, 67129, Indonesia
[1]dianahkam@unmerpas.ac.id,[2]ipungyellow1@gmail.com

ABSTRACT

The development of science and technology is one way to replace the method of human interaction with computers, one of which is to provide voice input. Conversion of sound into text form with the Backpropagation method can be understood and realized through feature extraction, including the use of Linear Predictive Coding (LPC). Linear Predictive Coding is one way to represent the signal in obtaining the features of each sound pattern. In brief, the way this speech recognition system worked was by inputting human voice through a microphone (analog signal) which then sampled with a sampling speed of 8000 Hz so that it became a digital signal with the assistance of sound card on the computer. The digital signal from the sample then entered the initial process using LPC, so that several LPC coefficients were obtained. The LPC outputs were then trained using the Backpropagation learning method. The results of the learning were classified with a word and stored in a database afterwards. The results of the test were in the form of an introduction program that able display the voice plots. the results of speech recognition with voice recognition percentage of respondents in the database iss 80% of the 100 data in the test in Real Time

Keywords : Sound, Linear Predictive Coding (LPC), Backpropagation .

I. INTRODUCTION

At this time, the development of technology can be said to have been very advanced. The development itself is inseparable from three main aspects, namely sound, sight, and touch. These three aspects become essential in the development of technology, especially artificial intelligence [1].

Many conveniences are offered for the benefit of human and computer interaction. For example, the speech recognition and turning the results into text. Speech Recognition is the conversion of an acoustic signal that is captured by a microphone or telephone to string words [2]. For some people, knowing a word is easy because humans have very good pattern recognition, but what about computers? For this reason, technology is needed to convert sound into text, so that the computer is able to recognize voice input and translate it into text.

In a previous study conducted by [3] entitled "Recognition of Human Voice Using the Linear Predictive Coding (LPC) Method", there was a statement that it was difficult for users or computer users to use a number of tools contained in a computer application, which was human desire to facilitate the use a tool, so that the recognition of natural characteristics of humans can be used. Sound is a form of biometrics that can be used as person identification. Speech recognition does not require special equipment and because of the different characteristics of the human voice.

Furthermore, in a study conducted by [4] entitled "Introduction of Text-Free Speech with the Vector Quantization (VQ) Method Through Linear Predictive Coding (LPC) Extraction", there is a statement that signal processing plays an important role, especially in science and communication technology. Intensive research in the field of signal processing causes communication technology to develop rapidly. One of them is the introduction of the speaker. The introduction of the speaker is a method used to find out the identity of someone who utters the information signal. This can be done because each individual has specific speech signal characteristics that can be distinguished by extraction with a coding technique. The coding technique commonly used in speech signal extraction is LPC (Linear Predictive Coding). This analysis produces an estimation of basic speech parameters, including pitch, formant, vocal path area equation, and compression (compression) of the speech signal to obtain a low bit-rate for transmission or storage purposes.

In a study conducted by [5] entitled "Converting Analog Data Into Digital Data and Digital Data Into Analog Data Using the Ppi 8255 Interface With Borland Delphi 5.0 Programming Language", the researcher stated that the development of modern technology is very necessary in current conditions to perfect technology F. Control system based on PC (Personal Computer) is one of the applications of modern technology where many of these computer applications can help humans in completing their work. One example of the application of using a PC-based control system is to control the process of analog to digital conversion

DOI: 10.25139/ijair.v1i1.2030

and vice versa in which intended to replace the control manually, so as to improve time and energy efficiency. Therefore, the technology is made to convert sound into text, so that the computer is able to recognize voice input and translate it into text form.

II. METHOD

Speech recognition with backpropagation method can be understood through feature extraction with the aim to clarify the characteristics of each sound pattern, including Linear Predictive Coding (LPC) sourced from books and internet literature.

A. System Flowchart

The research flow contains input process in real-time so that it can produce output in the form of text as in Fig. 1. This sub-chapter contains an explanation of the flowchart in the form of training data and testing data from the research conducted. The flowchart used is shown below:

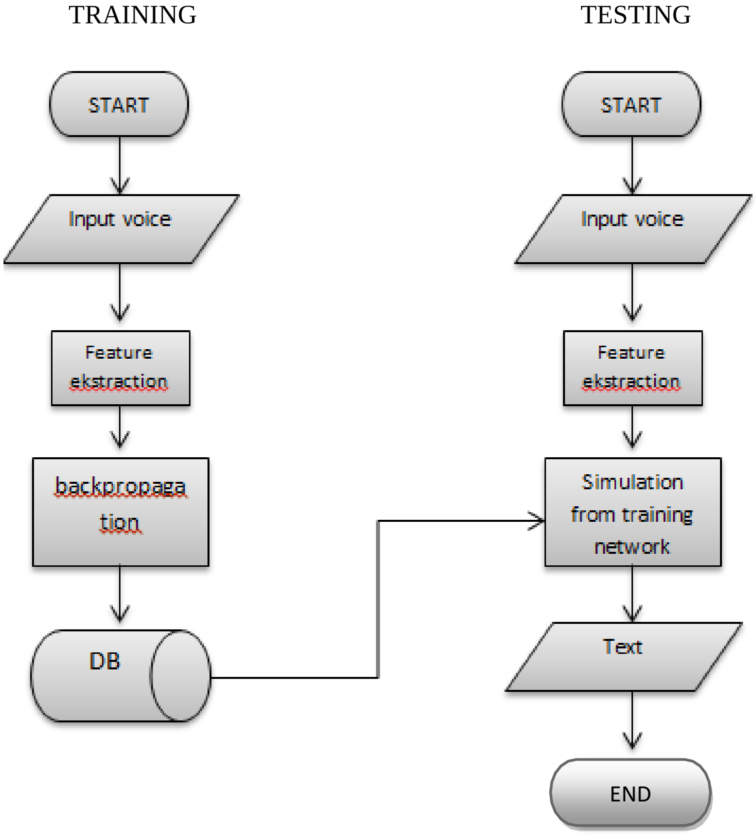


Fig. 1. Voice Recognition Flow Chart

The system starts from sampling analog signals into digital signals by the process of changing, filtering, and estimating audio signals into binary data by obtaining values of each audio signal characteristic required by digital signal processing. From the digital signal, the process of assigning sound parameters into a series of feature vectors in the form of summaries and relevant information from the sound can be done by the process of extracting Linear Predictive Coding (LPC) features. The output of the LPC results in the form of the LPC coefficient value is a feature of the spoken voice.

For the process of recognition and decision making, the Backpropagation method can be used. Backpropagation method will classify each sound with a word and will be stored in a database. In the search method, incoming sounds will be compared with those already in the database, then the output obtained will be displayed in text form.

B. Linier Prediktive Coding (LPC)

Feature Extraction is a brief description of the voice so that relevant and important information for the recognition subsystem is maintained. The final goal of the feature extraction process is to provide sound parameters into a row of feature vectors in the form of summaries and relevant information from the sound. The extracted feature is expected to be able to distinguish similar sounds, so that the model can be created without requiring large training data [2].

The sound that has been converted from an analog signal into a digital signal through a microphone and sound card is then extracted so that features are obtained. each sound represents a feature so that it has a clear difference [6].

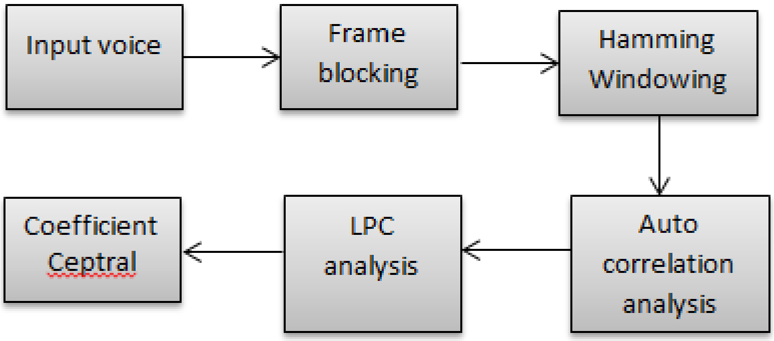


Fig. 2. Block Diagram of LPC Feature Extraction

Fig. 2 explains the LPC Feature Extraction Block Diagram with the following steps:

1. Voice Input

The human voice recorded with microphone and saved in WAV format will be processed to the next stage.

2. Frame Blocking

The recorded sound signal is divided into several frames, consisting of N sound samples with the distance between adjacent frames separated by M-samples. If M ≤ N, several adjacent frames will overlap and the LPC spectral estimation results will correlate from frame to frame. Conversely, if M> N, there will be no overlap between adjacent frames [7].

Given:
number of samples per frame (N) = (8000 x 0.04) 320 samples
overlapping amount of each frame (M) = N/2 (320/2) = 160 samples
number of frames (i-N/M)+1 = (8000-320/160)+1 = 49 Frames

3. Windowing

The function of windowing is to eliminate the effect of discontinuity caused by the previous process, so samples that have been divided into several frames need to be made continuous noise [6]. The type of window that is commonly used is Hamming window which has a general form as in formula (1):

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right),\quad 0 \le n \le N\text{-}1 \tag{1}$$

given:
w(n) : Hamming window
N : Number of sound samples

4. Autocorrelation Analysis

After going through the windowing process, each frame of the signal will be analyzed for autocorrelation. The highest autocorrelation value of p is the LPC order, which is 12, because the p value is usually between 8 to 16 using formula (2).

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m),\ m = 0,1,\ldots,p. \tag{2}$$

Given:
$r_1(m)$: autocorrelation coefficient
$\tilde{x}_l$: Input signal
P : LPC analysis order

5. LPC analysis

The next process is LPC analysis, which converts each p + 1 autocorrelation frame into the form of LPC parameters or commonly referred to as the LPC coefficient using formula (3).

$$a_m = \alpha_m^{(p)} \qquad 1 \le m \le p \tag{3}$$

The results of this data are used as input for Backpropagation neural networks.

C. BackPropagation.

The parameters generated by the above stages are inserted into the neural network using the BackPropagation learning method. The network training process flowchart can be seen in Fig.3.

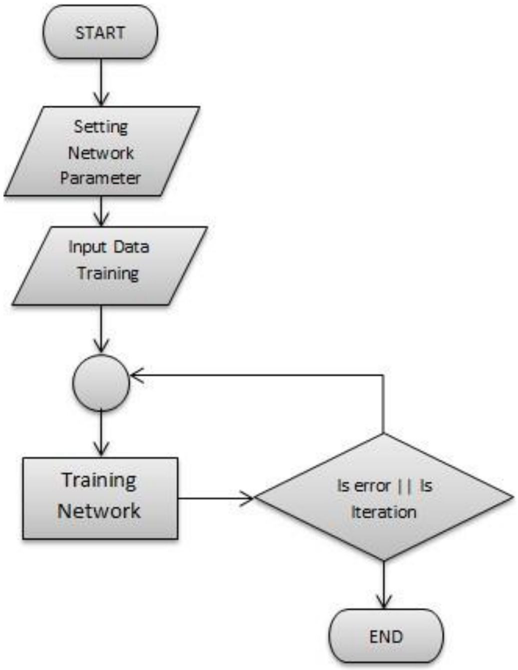


Fig. 3. BackPropagation Flow Chart

Referring to the process above, the formation of the network is to set targeted errors and set a maximum iteration with a value of bias = 1 for all input layers which are then hidden in the final data. Setting the output pattern target is to enter 100 predetermined outputs. In the initial weight initialization, a small randomization is performed with a random value between -0.5 and 0.5. For efficiency, the selection of Nguyen Widrow's initial weights can be used as a reference by adjusting the weights to the number of hidden neurons. The initial iteration (epoch) parameter setting is equal to 1. As long as the epoch is less than the maximum, the iteration will run. The next step is calculating the error between the Neural Network output pattern and the target pattern. When it reaches an error / maximum number of iterations, the learning process is complete.

The training data to be tested is 700 new voice data, consisting of 7 votes for one word sampled in the form of a matrix. The matrix is then simulated into a network that has been previously trained in order to get the results of the introduction of each word.

III. RESULT AND DISCUSSION

The backpropagation method used in the trial for the conversion of sound into text form is shown in Fig.4 below. In the picture, there is just one button that can be used to record to start a voice recognition program which is then followed by saying one voice.

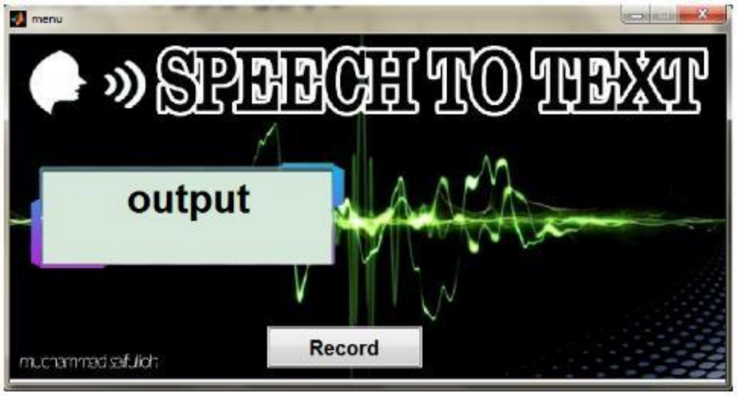
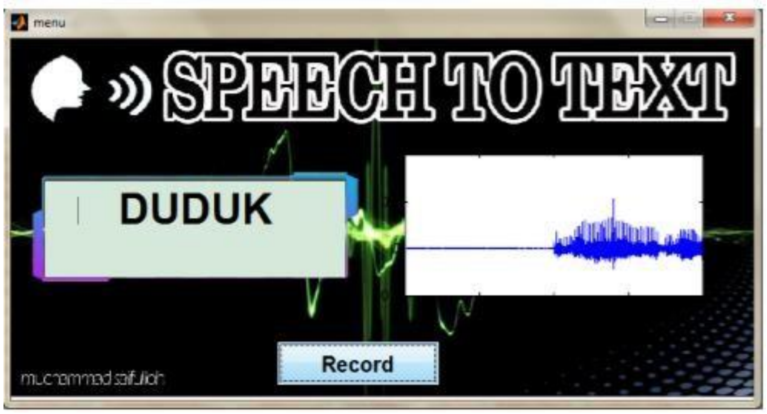


Fig. 4. *Backpropagation* Method test

DOI: 10.25139/ijair.v1i1.2030

The following were the results of voice recognition trial depicted in Fig. 5 with Static Text 1 box and a recording plot in the Axes1 box. Users can repeat the speech recognition process by pressing the "RECORD" button. The introduction program can be terminated by pressing the "X" button.



DUDUK

Fig. 5. Speech Recognition

The system of Sound Conversion into Text Forms Using the Backpropagation Method testing is done using a microphone. The Testing period used 100 data. The results of vehicle number plate recognition identification system testing were presented in Table I.

TABLE I
VOICE CONVERSION TEST

Input	Output
Walking	Correct
Running	Correct
Looking	Correct
Reading	Correct
Paying	Incorrect
Cleaning	Incorrect
Thinking	Correct
Note Taking	Correct
Register	Correct
Listening	Incorrect

From the results of testing 100 data, there were 80 successful identification data, while 20 data were incorrect.

$$A = \frac{Data\ benar}{jumlah\ data}\ x100\%$$

$$A = \frac{80}{100} \text{ x } 100\%$$

$$A = 80\%$$

The accuracy rate of speech recognition is 80%. Based on the sample data that has been tested, the recognition failure is caused by several things, namely crowded and noisy condition, etc, lots of noise, and the sound does not exist in the database.

IV. CONCLUSION

The speech to text recognition system has worked in accordance with the design and are able to convert sound into text and display the results in the MATLAB GUI. This system can recognize sounds correctly according to the Speeches inputted, where the level of recognition becomes better if the voice of the person is registered in the database. This study produces the best level of recognition, which is 80% out of 100 data. On the other hand, the level of speech recognition drops when words are spoken by people whose voices are not in the database, as well as in noisy and crowded rooms.

ACKNOWLEDGMENTS

This speech recognition system using the Matlab-based backpropagation method is certainly still far from perfection. Future research is expected to help the development of this speech recognition system to be better. Aspects of further research that can be done is to use different methods in the recognition system, so that the results of the voice recognition level can be compared and there is further development into the application as a medium for learning the pronunciation of various words in languages.

REFERENCES

[1] F. A. AHDA, “ANALISIS SUARA ALPHABET MENGGUNAKAN JARINGAN SYARAF TIRUAN PROPAGASI BALIK,” *Anal. SUARA Alph. MENGGUNAKAN Jar. SYARAF TIRUAN PROPAGASI BALIK*, vol. 60, no. 4, pp. 982–992, 2010.

[2] M. B. Gunawan, “KONVERSI SUARA KE TEKS MENGGUNAKAN METODE HIDDEN MARKOV MODEL,” *KONVERSI SUARA KE TEKS MENGGUNAKAN Metod. HIDDEN MARKOV Model*, p. 45, 2010.

[3] K. Anam, “Pengenalan suara manusia menggunakan metode,” 2013.

[4] F. AN, “Pengenalan Pengucap Tak Bergantung Teks dengan Metode Vector Quantization (VQ) Melalui Ektraksi Linear Predictive Coding (LPC),” pp. 1–8, 2004.

[5] R. A. SRI MELATI SAGITA, SITI KHOTIJAH, “PENGKONVERSIAN DATA ANALOG MENJADI DATA DIGITAL DAN DATA DIGITAL MENJADI DATA ANALOG MENGGUNAKAN INTERFACE PPI 8255 DENGAN BAHASA PEMROGRAMAN BORLAND DELPHI 5 . 0,” *ISSN 1979-276X*, vol. 6, no. 2, pp. 168–179, 2013.

[6] M. Irfandy, “Aplikasi Pengenalan Ucapan Dengan Jaringan Syaraf Tiruan Propagasi Balik Untuk Pengendalian Robot Bergerak,” *Apl. Pengenalan Ucapan Dengan Jar. Syaraf Tiruan Propagasi Balik Untuk Pengendali. Robot Berger.*, pp. 1–7, 2004.

[7] and A. A. Z. Sigit Nur Rohman, Achmad Hidayatno, “APLIKASI PENCIRIAN DENGAN LINEAR PREDICTIVE CODING UNTUK BALIK Landasan Teori,” *Apl. PENCIRIAN DENGAN LINEAR Predict. CODING UNTUK PEMBELAJARAN PENGUCAPAN NAMA HEWAN DALAM Bhs. Ingg. MENGGUNAKAN Jar. SARAF TIRUAN PROPAGASI BALIK*, pp. 151–158, 2012.

DOI: 10.25139/ijair.v1i1.2030