Ann PUBLISHER
Journal Engineering Sciences (Improsci)
e-ISSN 3031-7088
p-ISSN 3032-3452

Implementation Of Speech Recognition For Voice Command Use

Ade Johar Maturidi[1], Didan Osman[2], Muhamad Sulaeman[3].
[1]Universitas Sindang Kasih Majalengka, Majalengka, Indonesia, ade.johar@gmail.com
[2]Universitas Muhammadyah Cirebon, Cirebon, Indonesia, didan.secret@gmail.com
[3] Universitas Sindang Kasih Majalengka, Majalengka, Indonesia, ade.johar@gmail.com
Corresponding Author : ade.johar@gmail.com

Abstract:
Background. Physical limitations of a person sometimes make it impossible to operate a computer with only a keyboard and mouse,
Aims. One tool that can be used is a voice command, which is part of speech recognition technology.
Methods. The voice signal will be normalized first, and then the coefficient values will be calculated using the Linear Predictive Coding (LPC) and Fast Fourier Transform (FFT) methods. After the coefficient value is obtained, recognition is performed using the backpropagation method of the Artificial Neural Network.
Conclusion. The artificial neural network backpropagation method is used because it can adjust its own weights and produce error values that we can determine, thereby improving accuracy.
Implementation. This study implements a voice command system using MARF as its speech engine and Java as its programming language.

Keywords: LPC, FFT, Speech recognition, Artificial Neural Network, Backpropagation, Voice Command

INTRODUCTION

The use of computers as data-processing tools has become commonplace. However, most data processing involves character-based data (alphabets, numbers, symbols, etc.) input from a keyboard. However, many other types of data can be used as input, such as voice, fingerprints, and retinal scans. Systems that utilize such input data are called biometric recognition.

Biometric recognition is a system for recognizing or identifying individuals based on their unique biological characteristics. Its function is not only to support security systems by identifying individuals, but also to support military operations, among other purposes.

Biometric recognition applications include retinal scans (identification based on blood vessel patterns in the retina), fingerprint recognition (identification of unique fingerprint patterns for each individual), face recognition (identification based on facial features and expressions), and voice recognition (identification based on voice patterns).

Voice recognition itself is divided into two types: speech recognition and speaker recognition. Speech recognition is the process by which a computer recognizes spoken words without regard to their identity. An example of speech recognition is voice commands. Speaker recognition, on the other hand, is the recognition of a person's claimed identity based on their voice (specific characteristics can include intonation, depth of voice, and so on).

Sometimes users cannot fully execute commands using only a keyboard and mouse. For example, users with visual impairments (blindness) require assistive devices to use a computer. One such tool is voice command technology, which allows users to operate a computer solely by voice. Voice recognition technology itself is not new. However, research and development of this technology in Indonesia is still minimal.

Elements of novelty that can be identified in this study:

1. The implementation of voice commands uses a combination of LPC + FFT + ANN backpropagation methods based on the MARF engine and Java, not just an algorithmic simulation, but in the form of a running system.
2. The system was developed for user accessibility (a computer for users with physical limitations), which provides a social applicative context.
3. Offers specific ANN configuration parameters (2 hidden layers, 18–13 neurons, lr 0.3, momentum 0.6) for performance optimization.
4. The study emphasizes the relationship between:
    a. amount of input data
    b. Sound Variations
    c. error rate as a practical contribution to an ANN-based system.

So that the novelty lies not solely in the algorithm but in the implementation of the ANN-based voice command system using MARF, with a focus on accessibility and ANN parameter experimentation.

LITERATURE REVIEW

Definition of Speech

Speech is defined as "the ability or expression of actions that convey thoughts, feelings, or observations (perceptions) through spoken words."

The McGraw-Hill Encyclopedia of Science and Technology, published by McGraw-Hill Companies, Inc., translates the word "speech" as:

"A collection of audible sounds produced by vibrating air through the coordinated movement of a specific group of anatomical structures. Humans attach symbolic value to these sounds for communication."

"A collection of sounds produced through the vibration of air by properly moving a specific group of anatomical structures. Humans typically attribute symbolic values to these (different) sounds for communication."

Oxford University Press, in “The Oxford Companion to the Body,” World of the Body, defines the word “speech” as:

“Speech involves the voluntary initiation and engagement of a complex set of muscles around the larynx, throat, and mouth, along with rhythmic breathing and the use of expiratory muscles. Like other voluntary movement patterns, speech originates in the cerebral cortex. Other parts of the brain (especially the cerebellum), along with sensory feedback, modify and regulate the nerve impulses sent to motor neurons whose axons activate the relevant muscles. In this case, the motor neurons in question are located in the brainstem, and their axons travel along the lower cranial nerves to the muscles of the vocal apparatus. Effective speech also depends on motor neurons in the cervical and thoracic regions of the spinal cord that serve the respiratory muscles.”

Speech involves a response based on the mind’s own needs (its own desires). Also, it involves a combination of muscles around the larynx, throat, and mouth, along with pauses in the rhythm of inhalation and exhalation. As with any form of action caused by a thought impulse, speech is formed in the cerebral cortex. Appropriate. Immediately afterward, the motor nerves are recognized by the brain, and the nerve branches of the neurons (axons) move along the lower cranial nerves until they reach the muscles in the voice-producing organs. The effectiveness of pronunciation (speech) depends on the motor nerves found in the cervical and thoracic nerves, which are part of the spinal cord, to serve the respiratory muscles.

Understanding Speech Recognition

Speech recognition is a technology that enables computers with input peripherals to translate what humans say. Essentially, SR (speech recognition) is a sequence of algorithms that processes analog signals into digital signals. From these digital signals, the computer extracts and interprets certain parameters as words. The speech recognition phase is diagrammatically illustrated in the diagram below:

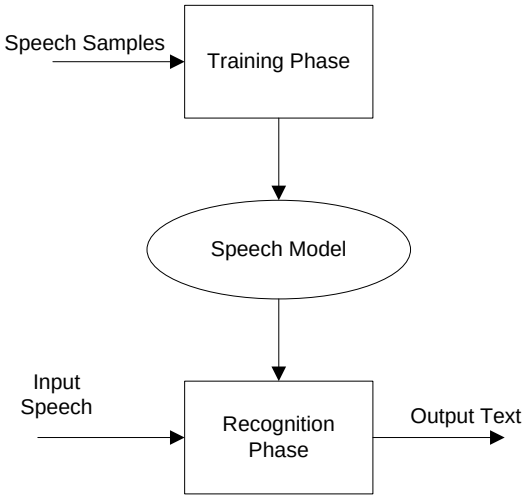

Figure 1 Speech Recognition System Outline

All SR systems operate in two phases: training and recognition. In the training phase, the system is taught a reference form for different speech signals, such as words, phrases, or phonemes, which will later serve as vocabulary. Each reference is taught to the system by speaking it, and the system averages the spoken word parameters, which become the statistical characteristics of the speech signal. The second phase, recognition, compares the unknown input with the reference obtained during training.

Linear Predictive Coding

The steps of LPC analysis for speech recognition, according to Rabiner, L.R., Juang, B.H., in their book "Fundamentals of Speech Recognition," are: Preemphasis.

In this step, digital word samples are filtered using a first-order FIR filter to smooth the spectral signature of the sampled word signal.

$\tilde{s}(n) = s(n) - \tilde{a}s(n-1)$ (2.1)

Frame Blocking

In this step, the emphasized word signal is divided into frames, each containing N word samples and adjacent frames separated by a distance of M.

To extract short-time features from a speech signal, the signal is divided into short chunks, called frames. The duration of each frame varies between 20 and 30 ms, and the speech signal within each frame is assumed to be stationary. To reduce the boundary effects of each segment, windowing (e.g., Hamming Windowing) is necessary for each frame. Overlapping frames can also achieve smoother results each time.

Windowing

In this step, a weighting function is applied to each frame created in the previous step.

$w(n)=0.54-0.46 \cos (2\pi n/(N-1)), 0 \leq n \leq N-1$

There are several types of windowing, including Hamming, Hanning, Bartlett, Rectangular, and Blackman. The windowing equation is as follows:

Window Hamming

$$W_{ham}(n) = \begin{cases} 0{,}54 - 0{,}46\cos[2\pi n/(N-1)] & 0 \leq n \leq N-1 \\ 0 & \text{elsewhere} \end{cases}$$

Window Hanning

$$W_{han}(n) = \begin{cases} (1 - \cos[2\pi n/(N-1)])/2 & 0 \leq n \leq N-1 \\ 0 & \text{elsewhere} \end{cases}$$

Bartlett Window

$$W_B(n) = \begin{cases} 2n/(N-1) & 0 \leq n \leq (N-1)/2 \\ 2 - 2n/(N-1) & (N-1)/2 \leq n \leq N-1 \\ 0 & \text{elsewhere} \end{cases}$$

Blackman Window

$$W_{Bl}(n) = \begin{cases} 0{,}42 - 0{,}5\cos\left[\frac{2\pi n}{(N-1)}\right] + 0{,}08\cos\left[\frac{4\pi n}{N-1}\right] & 0 \leq n \leq N-1 \\ 0 & \text{elsewhere} \end{cases}$$

Where:

N = Number of data points in one window

n = nth data point

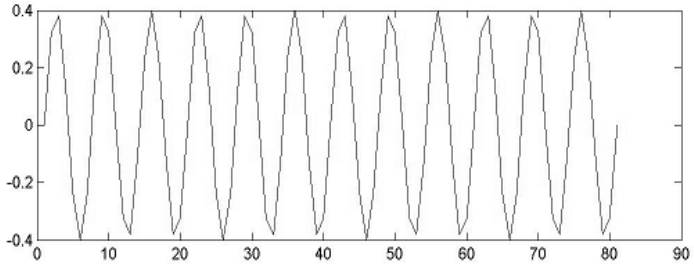

Figure 2. Sine wave signal without windowing

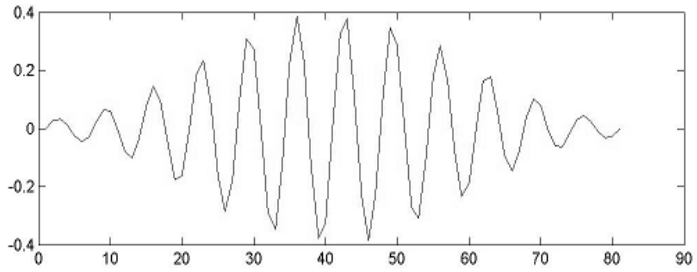

Figure 3. Sine wave signal with Hamming windowing

Autocorrelation Analysis

At this stage, each windowed frame is autocorrelated with the highest autocorrelation value being the LPC order, typically an LPC order of 8 to 16.

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{X}_l(n)\tilde{X}_l(n+m)$$

LPC Analysis

The next step is LPC analysis, where the autocorrelation values in each frame are converted into a set of LPC parameters: the LPC coefficient, the reflection coefficient, and the log area ratio coefficient.

Converting LPC Parameters to Cepstral Coefficients

The cepstral coefficients are Fourier transform coefficients that represent the log magnitude spectrum.

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, 1 \le m \le p$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, m > p$$

Fast Fourier Transform

The FFT (Fast Fourier Transform) algorithm is a method for transforming audio signals into frequency signals. This means the recorded audio is stored digitally as a frequency-domain waveform.

The Fourier Transform is a highly efficient method for solving discrete Fourier transforms, widely used for signal analysis purposes such as filtering, correlation analysis, and spectrum analysis. The Discrete Fourier Transform (DFT) is a sequence defined in the discrete-frequency domain that represents the Fourier transform of a finite-duration sequence. The DFT plays a crucial role in implementing various signal processing algorithms due to its computational efficiency across applications.

The Fast Fourier Transform (FFT) is a continuation of the Discrete Fourier Transform (DFT). This Fourier transform converts a signal from the time domain to the frequency domain. This allows the signal to be processed using spectral subtraction.

The FFT is a special form of the Fourier integral equation:

$$H = \int h(t)e^{-jwt}\,dt$$

By changing the variables, time(t), frequency(w) into discrete form, the discrete Fourier transform (DFT) equation is obtained, namely:

$$H(kwo) = \sum_{N=0}^{N-1} h(nT)e^{-jkwonT}$$

Simplified with T=1 sample time N=sample time N=sample frequency k so that it becomes:

$$H(k) = \sum h(n)e$$

With: k : 0,1,2,…,N-1

The FFT is used for faster computation and can reduce the number of multiplications from N2 to NlogN multiplications. The FFT uses 512 points, and because the FFT results are symmetric, the output is only 256 data points. The results of the FFT process will yield symmetric signal points, so the data taken is only half the total, from which the maximum value is then taken.

From reading the article, several important research gaps have not been answered:

1. The system is still speaker dependent, proving that the accuracy drops drastically when tested on other users. There is no exploration towards the independent speaker model.

2. The system accuracy rate is still low (71.25%) and has not been compared to:
    a. Other ANN methods
    b. Modern machine learning methods (SVM, HMM)
    c. deep learning (CNN, RNN/LSTM)
3. There is no discussion about:
    a. More advanced noise filtering
    b. Large datasets or benchmarking
    c. real-time processing constraints
4. There were no trials on diverse voice conditions (sick, accent, fast/slow, different languages).
5. The system has not integrated modern approaches such as:
    a. signal enhancement
    b. MFCC (Mel Frequency Cepstral Coefficients)
    c. deep learning speech recognition
6. Lack of implementation exploration for direct accessibility of users with disabilities (e.g., user testing).

In conclusion, the main gaps are the need to improve accuracy and generalization across multiple speakers, as well as to integrate modern speech recognition technology.

METHOD

The research method used is an experimental method with a Research and Development (R&D) approach. This research develops a system (Artificial Neural Network-based speech recognition), thus falling within the characteristics of R&D. The system design, training, and testing processes include parameter settings (hidden layers, neurons, learning rate, momentum), which are characteristic of experimental research. System accuracy is measured through direct testing with different input data and users to evaluate system performance empirically. Therefore, this research uses experimental research to develop and test the system (experiment-based R&D).

DISCUSSION

The state of the art in this study can be summarized as follows:

1. The use of the LPC & FFT method as a sound feature extraction technique is a method that has long been needed in speech recognition systems due to its stability and ability to represent frequency signals accurately.
2. Backpropagation ANN was chosen for the classification process because it can adjust weights adaptively and minimize errors during the training process.
3. The system is designed to implement voice commands as a medium of human-computer interaction, especially for users who have physical limitations in using the keyboard and mouse.
4. The implementation platform uses the MARF speech engine and Java as the programming languages.
5. The focus of the research is on speaker-dependent recognition, which means that the system learns and reacts based on the voice of a particular user.

Thus, this study follows the trend of existing ANN-based speech recognition but is implemented in the context of voice commands intended for accessibility applications.

System Analysis and Design

The following are the phases used in speech recognition design:

1. Recording Phase

   In this phase, the user's spoken voice is recorded using a microphone or similar device. There are two methods for this voice recording process: through this program or using another voice recording device. The only audio format recognized by this program is .wav.

2. Extraction Phase

   In the extraction phase, the recorded voice is processed using several methods, such as normalization and FFT. After analysis, unique coefficients are generated that can be used to differentiate voices.

   The normalization method is used to separate the primary voice from other sounds/audio interference, such as noise. Therefore, normalization is expected to increase the accuracy of the calculations.

   The FFT method is used to identify the characteristic coefficients of a voice, which will then become important data for subsequent processing.

3. Training Phase

This training phase is used to train new voices that have not been previously recognized. In this phase, an Artificial Neural Network method using the backpropagation algorithm is used.

The coefficient values generated from the previous phase are used as training data. The more data used for training, the smaller the error value is expected to be, thus increasing the accuracy. Furthermore, this phase also performs classification, which involves grouping speech signals based on the output they will produce. After training, weight values are generated, which will be used in the recognition phase.

4. Recognition Phase

   This recognition phase is used to recognize spoken voices and determine whether they match the previously drilled data. If they match, the program will proceed to the next phase. If they do not, an error message will be displayed.

   The weight values generated in the training phase will be used in conjunction with the input voice coefficient values.

5. Execution Phase

   This execution phase will be executed if there were no errors in the previous phase. In this phase, the program will open the application program that was drilled.

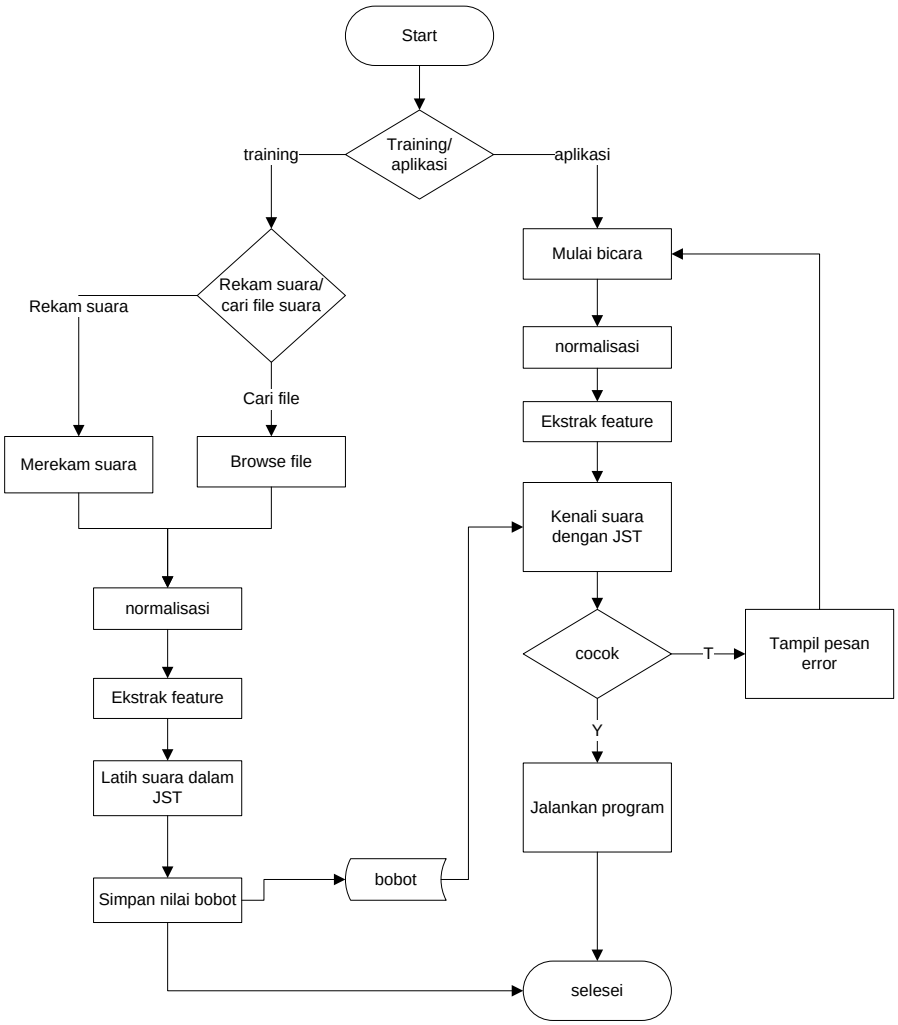

Figure 4. Program Flowchart

Use Case Diagram

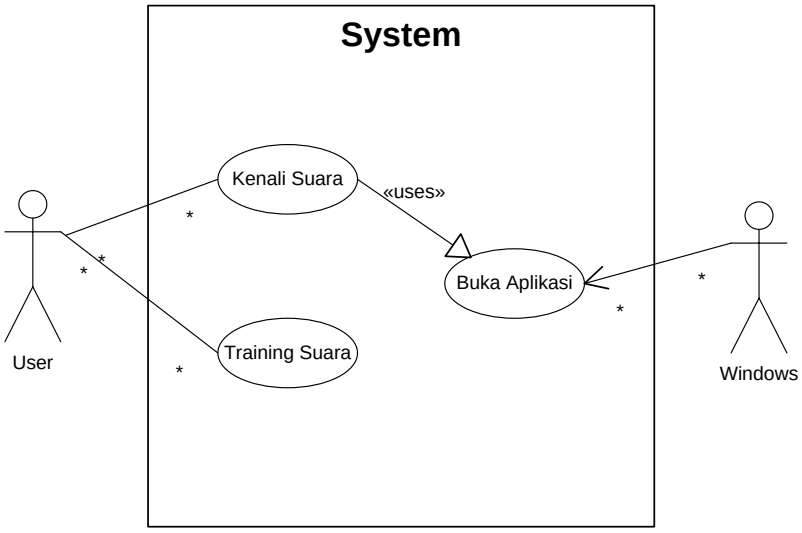

Figure 5. Use Case Diagram

Program Structure

To make it easier to understand and use the program, the program structure can be seen as follows:

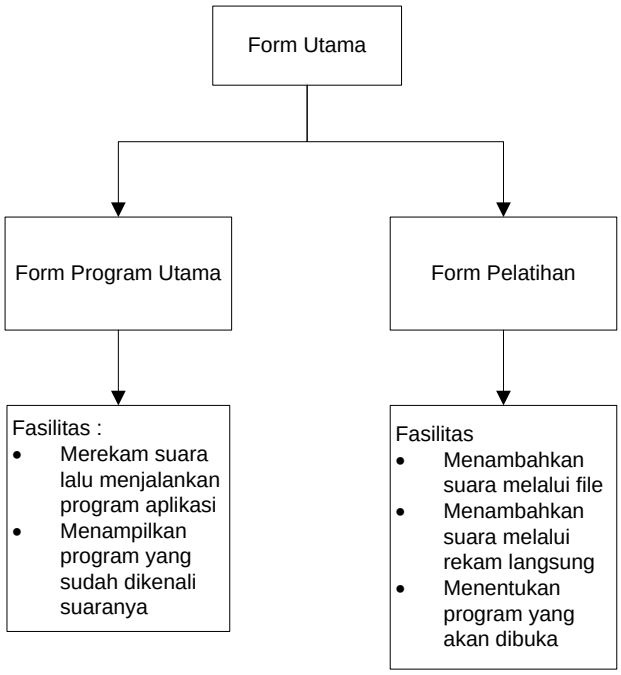

Figure 6. Program Structure

Voice Input Training Form

In this form, users can add data to be trained by recording their voice directly.

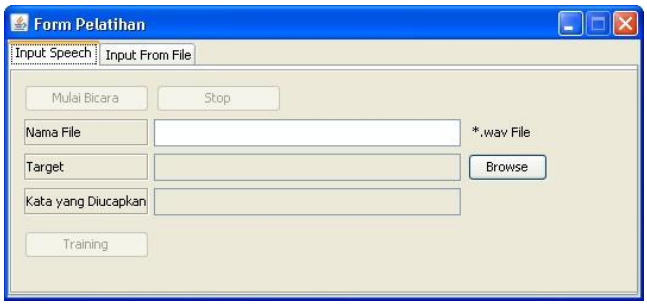

Figure 7. Voice Training Form Display

CONCLUSION

To recognize a command, at least two pieces of input data are required. The more input data used for a command, the more the error accuracy can be minimized. Different voice patterns can affect data accuracy; even a single person's voice with different conditions

(flu, cough, pronunciation, etc.) can affect the resulting voice pattern. This system uses an Artificial Neural Network with parameters , using two hidden layers with 18 and 13 neurons in each layer. The learning rate used is 0.3 and the momentum used is 0.6. After training, the overall system accuracy was 71.25%. Testing the system with other users showed low accuracy because the system never recognized the speaker's voice. With increasing variety in the input data used (in this case the number of votes and the number of speakers), the system will become more global, or in other words, the system will not care whose voice is entered.

Implication

Based on the research and testing results of the voice command recognition system, several important implications arise that require consideration, both for system development and for its real-world implementation.

First, the requirement for at least two input data sets to recognize a single command indicates that the quality and quantity of training data significantly impact system performance. The greater the variety of input data used, the lower the resulting error rate. The implication of this finding is that speech recognition system development should prioritize collecting sufficient and diverse data to ensure the system's performance is more accurate and stable.

Second, differences in voice patterns, both between individuals and within a single individual under different conditions, significantly impact accuracy. This implies that speech recognition systems are susceptible to changes in physical conditions and pronunciation. Therefore, systems need to be designed to accommodate these voice variations, for example, by adding training data from various voice conditions or applying voice normalization techniques.

Third, the use of an Artificial Neural Network with a two-hidden-layer configuration and specific parameter settings yielded moderate accuracy. These results imply that the choice of network architecture and training parameters plays a crucial role in determining system performance. Furthermore, the low accuracy in testing with other users indicates that the system is still speaker-dependent, requiring further development to enable it to recognize new users' voices more effectively.

Fourth, the increasing diversity of input data, both in terms of the number of voices and the number of speakers, makes the system more global or speaker-independent. An important implication of this finding is that by adding representative training data, speech recognition

systems can be developed to be more flexible and less dependent on specific speaker identities, making them more readily applicable to broad and diverse usage environments. Overall, the results of this study imply that developing speech recognition systems requires a well-thought-out data collection strategy, appropriate model design, and a continuous training process to achieve higher levels of accuracy and reliability in real-world applications.

BIBLIOGRAPHY

Dima Batenkov, "Fast Fourier Transform", key paper in computer science seminar 2024.

Djon Irwanto, S.Kom, MM "Membangun Object Oriented Software dengan Java dan Object Database". Jakarta : Elex Media Computindo,2019

Edward R. Jones, Ph.D, "*An Introduction to Neural Networks*", white paper, Visual Numerics, Inc., 2024

Fausett, Laurene . "Fundamentals Of Neural Network". Englewood Cliffs, New Jersey : Prentice-Hall.Inc, 2008

"Fast Fourier Transform", Modul Perkuliahan Numerical Analysis E3, I3, FMN050, Centre for Mathematical Sciences Lund University, Sweden, 2011.

Ganesh K. Venayagamoorthy, "Teaching neural networks concepts and their learning Techniques", presented at American Society for Engineering Education Midwest Section Conference, 2012.

Gressia Melissa, "Pencocokan Pola Suara (Speech Recognition) Dengan Algoritma FFT Dan Divide And Conquer", Makalah IF strategi algoritmik, Institut Teknologi Bandung, 2012.

Jont B. Allen, " Articulation and Intelligibility", USA : Library of Congress Cataloging-in-Publication Data, 2005.

Jochen Fröhlic, "Neural Net Component in a object Oriented Class Structure*"*.

Mark Watson, "Practical Artificial Intelligence Programming With Java", USA : Creative Commons Attribution-Noncommercial-No Derivative Works, 2008.

Mimi Tantomo, "Perbandingan Keakuratan antara Jaringan Syaraf Tiruan Back Propagation dan Self Organizing Maps untuk Speech Recognition"*,* paper in www.scribd.com.

Richard G. Baldwin, "Fun with Java, Understanding the Fast Fourier Transform (FFT) Algorithm", in www.developer.com/java/other/article.php diakses pada tanggal 16 Maret 2025 Pk.08.27.

Rudi Adi Pranata, Resmana, "Pengenalan suara manusia dengan metode LPC dan Jaringan Syaraf Tiruan Propagasi Balik", dipresentasikan pada Seminar Nasional I Kecerdasan Komputasional Universitas Indonesia, Jakarta, Indonesia, 2019.

Thiang, Hadi Saputra, "*Sistem Pengenalan Kata dengan Menggunakan Linear Predictive Coding dan Nearest Neighbor Classifier*", Jurnal, Universitas Kristen Petra, Jakarta, 2005

"Topik lanjutan pengolah wicara", modul kuliah Praktikum Pengolahan Informasi Wicara

http://www.developer.com/java/other/article.php/3374611 diakses pada tanggal 23 Maret 2025 Pk. 08.23.

http:/www.developer.com/java/other/article.php/3457251 diakses pada tanggal 23 Maret 2025 Pk. 08.31.

http://mathworld.wolfram.com/DiscreteFourierTransform.html diakses pada tanggal 23 Maret 2025 Pk. 09.01
http://marf.sourceforge.net diakses pada tanggal 15 Juni 2025 Pk 08.47