Voice Recognition System for User Authentication Using Gaussian Mixture Model

Novario J. Perdana [1], Dyah E. Herwindiati [2], Nor H. Sarmin [3]

[1] Departement of Information System
Faculty of Information Technology, Tarumanagara University
Jakarta, Indonesia
Email: novariojp@fti.untar.ac.id

[2] Departement of Informatics
Faculty of Information Technology, Tarumanagara University
Jakarta, Indonesia
Email: dyahh@fti.untar.ac.id [2]

[3] Department of Mathematical Sciences
Faculty of Science, Universiti Teknologi Malaysia
Johor Bahru, Malaysia
Email: nhs@utm.my [3]

2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET) | 978-1-6654-6837-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/IICAIET55139.2022.9936856

***Abstract*— The use of biometrics in the user authentication process is the leading choice today. One of the biometrics that can be used is the human voice. In this paper, a voice authentication system using the Gaussian Mixture Model (GMM) is proposed. GMM was chosen because of the ease and accuracy in classifying the data. Voice data features are extracted using Linear Predictive Coding (LPC) before being classified using GMM. Voice data was recorded directly from 30 respondents using laptops and smartphones. Additional devices in the form of earphones were added to get better results. The system's learning process has an accuracy of 84%, and the overall testing process has an accuracy of 82%. There are also differences in the accuracy of user authentication between data that use enhancements and those that do not. They are 87% and 72%, respectively.**

Keywords—biometrics, user authentication, voice

I. INTRODUCTION

The massive use of the internet today provides convenience for humans. Various human activities today are closely related to the internet. Activities such as chatting, socializing, and shopping are now online. This phenomenon has reached a specific stage for each user, where they need to enter their identity so that the application/device can recognize them and personalize their needs. However, behind the convenience provided, a real threat follows. The authorization and authentication of users who uses personal data make users vulnerable to cyber-attacks. It has become common, some of which have succeeded until they get the password to log in.

Cyber-attacks here had a disturbing sense of the physical and logical flow of the system, which is done intentionally to disrupt the three basic concepts of network security, namely confidentiality, integrity, and availability. The problem is realized, and solutions emerge in the form of authentication systems that aims to prevent and reduce the potential and impact of the attack. Authentication has several types ranging from the most common, namely Username and Password, to using certificates, Smart Cards, and Biometric authentication. User authentication is the first line to access different means of technology in which a set of services are tailored to users. Once authenticated, one can access their company intranet to consoles, databases, and applications. Many websites now have authentication methods to secure their systems. Users need to provide their information to log in to the system.

Based on statistical data, in most cyberattacks, around 80% of the root cause is passwords that are vulnerable to being infiltrated by hackers. Because of that, it is said to be less effective and requires a new alternative. Another alternative chosen is to use biometrics. Various alternatives have also been published. One of the most frequently used biometrics is fingerprint [1]–[3]. Chen et al., in their research, offered an authentication process for IoT devices using fingerprints. They recommend additional fingerprint authentication protocols for handshake communication between devices [1]. Meanwhile, Iancu and Constantinescu use a fuzzy logic control system for fingerprint recognition [3]. Others use face patterns [4], [5] and handwriting [6], [7].

This authentication process requires additional devices, such as fingerprint, retina, and writing scanners. It means that when the implementation requires a high cost. Therefore, the latest alternative that can be used is biometric authentication using voice. The use of voice as authentication has several advantages: it does not need direct contact with the device and does not require a particular device. A problem with using voice as authentication is the absence of a direct application circulating. Fluctuations in the consistency of voice data to be classified due to unexpected factors such as emotion in the voice, noise in raw data, availability of sound recording media that can function properly, and many other factors.

The process of voice recognition application begins with extracting features and classification. There are various methods for feature extraction, namely Linear Prediction Coefficients (LPC) [8], Discrete Wavelet Transform (DWT) [9], and Mel-Frequency Cepstrum Coefficients (MFCC) [10]. Among these methods, LPC is used here. LPC is one of the most powerful methods used in audio and speed signal processing. LPC extracts speech parameters such as formants and spectra. It provides a good model of the speech signal.

Classification techniques can be applied in classifying and recognizing the voice. The first that comes to mind is Gaussian Mixture Models (GMM). Gaussian Mixture Models (GMM) is a probabilistic model that assumes all data points are

978-1-6654-6837-4/22/$31.00 ©2022 IEEE

generated from a mixture of several Gaussian distributions with unknown parameters. One can think of Mixture Models as generalizations of K-Means clustering to combine information about the covariance structure of the data as well as latent Gaussian centers [11], [12].

In this article, we propose an authentication system with voice recognition. The system will use GMM for speech recognition and LPC for feature extraction. GMM has been known as one of the classification methods because it is simple and has good accuracy. This paper is organized as follows. Section 2 presents data input used in this research. Section 3 presents the steps of Linear Predictive Coding and Gaussian Mixture Model. Learning process is presented in section 4. Section 5 discusses the testing process of the system. Section 6 presents the conclusion.

The novelty of this research is to use GMM for voice recognition so that the system will use the system to authenticate the right user.

Table 1. Recorded Speakers, Including Speaker Tag, Age, Gender, Location, and Device that is used

Speaker Number	Age	Gender	Location	Device
1	24	M	Home	Smartphone with earphone
2	21	F	Home	Smartphone with earphone
3	20	F	Home	Smartphone with earphone
4	21	M	Lab	Smartphone with earphone
5	21	F	Lab	Smartphone with earphone
6	34	M	Home	Smartphone with earphone
7	28	M	Home	Smartphone with earphone
8	23	M	Lab	Smartphone with earphone
9	23	M	Home	Smartphone with earphone
10	21	M	Classroom	Smartphone with earphone
11	29	M	Home	Laptop with earphone
12	21	F	Lab	Laptop with earphone
13	21	F	Lab	Laptop with earphone
14	21	M	Lab	Laptop with earphone
15	21	M	Lab	Laptop with earphone
16	27	M	Lab	Laptop with earphone
17	21	F	Home	Laptop with earphone
18	20	M	Classroom	Laptop with earphone
19	21	M	Classroom	Laptop with earphone
20	25	M	Home	Laptop with earphone
21	21	F	Lab	Smartphone without earphone
22	21	F	Lab	Smartphone without earphone
23	21	M	Lab	Smartphone without earphone
24	21	M	Lab	Smartphone without earphone
25	21	M	Lab	Smartphone without earphone
26	21	M	Lab	Smartphone without earphone
27	21	M	Lab	Smartphone without earphone
28	21	F	Home	Smartphone without earphone
29	21	F	Home	Smartphone without earphone
30	21	M	Classroom	Smartphone without earphone

II. Data Input

Data are obtained from the voice recording results by willing respondents. All respondents were asked to read a sentence in Indonesian that had been prepared, and the sentence was the same for all respondents. Thirty respondents consist of twenty men and ten women. Each respondent voice is recorded for one minute ten times. Data recorded in different environmental conditions, such as laboratory rooms, classrooms, and private homes. Each environment has a different noise level and is recorded at different times. The aim is to understand the factors that impact the verification results of the GMM model. Table 1 displays the data.

III. Methodology

This section discusses the method of feature extraction and data classification. The use of both methods can be seen in Figure 1. More detailed steps are described in the following sections.

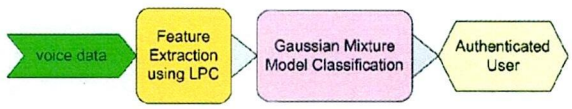

Fig. 1. Block diagram of the use of method for feature extraction and model classification

A. Linear Predictive Coding for Feature Extraction

The input data from voice recordings are processed into a numeric vector. It is then pre-processed, and the features are extracted using the Linear Predictive Coding (LPC) method. The extracted sound features are in the form of a vector with the parameter coefficient values of the LPC method is then used as input data in the Gaussian Mixture Model (GMM) method. LPC consists of several steps. The block diagram in Figure 2 illustrates the steps involved in feature extraction. The steps include pre-processing, pre-emphasis, framing, and windowing steps.

The raw speech data is distorted during the pre-processing step to reduce noise. It uses a High Pass Filter which will boost only the high-frequency components of the signal. This is done using (1).

$$Y(n) = s(n) - \alpha.s(n-1) \quad (1)$$

Where $Y(n)$ is after pre-emphasis signal, $s(n)$ is before pre-emphasis signal and α is the filter constant ranges from 0.9 to 1.0.

The result of the pre-emphasis signal is then sliced into frames. The number of frames is based on signal duration (T_s) multiplied by frame duration (M). Frames are taken as long as possible to get a better frequency resolution, while the shortest possible time is meant to get the best time domain. This process is called frame blocking. After that, we conduct windowing to the frames to minimize spectral distortions when blocking the speech signal. This is done in the form of (2). We use Hamming window for this purpose. Equation (3) represents the Hamming window $w(n)$.

$$X(n) = f_l(n)w(n) \quad (2)$$

$$w(n) = 0.54 - 0.46\cos(2n\pi / N-1),\ 0 < n < N-1 \quad (3)$$

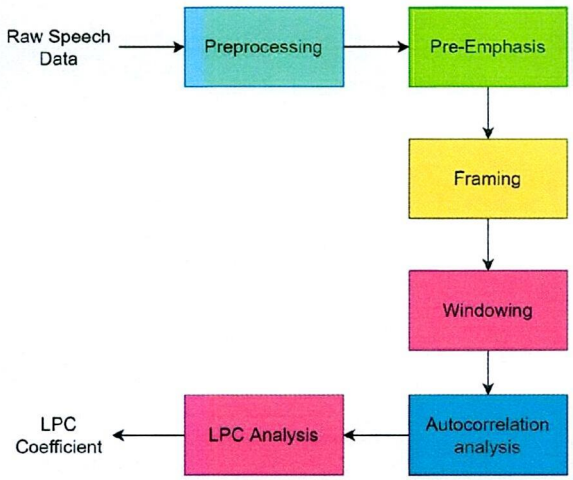

Fig. 2. Block Diagram of front-end Processing

Where $X(n)$ is the windowing result signal, $f_l(n)$ is the frame blocking result signal and N is the duration of the signal frame.

The window is applied to each frame. For each sample average C_k, and lagged value K, the autocorrelation step is be calculated using (4).

$$C_k = \frac{1}{t}\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y}),\ k=0,\ 1,\ 2,\ \ldots,\ K \quad (4)$$

$$r_k = {}^{C_k}/_{C_0},\ k = 0, 1, \ldots., K \quad (5)$$

The last step for LPC is to calculate the LPC coefficient. It is done by conversing the autocorrelation value to the LPC coefficient. The formulas are shown from (6) through (9).

$$E^{(0)} = r(0) \quad (6)$$

$$k_i = \frac{\{r(i) - \sum_{j=1}^{L-1} \alpha_j^{i-1} r(|1-j|)\}}{E^{i-1}}\ 1 \le i \le p \quad (7)$$

$$\alpha_j^{(i)} = \alpha_j^{i-1} - k_i \alpha_{i-j}^{i-1} \quad (8)$$

$$E^{(i)} = (1-k_i^2)E^{i-1} \quad (9)$$

The feature extraction step produces eight cepstral coefficients. These coefficients then become the input for GMM.

B. Gaussian Mixture Model for Data Classification

The classification process uses GMM. The working principle of the technique is briefly explained in this section. It is important to note that the classification is modelled during training, whereas, during testing, the model will classify/identify the speaker to whom the data belongs.

The Gaussian Mixture Model (GMM) is a probabilistic modelling technique that takes input data such as a sequence of feature vectors and creates one model per speaker. GMM models each source by a component probability density function (N component densities) and its mixture weights. Each component density is a product of the Gaussian component and a mixture weight. The formula for the probability density function of a one-dimensional gauss distribution is in (10).

$$p(x) = \sum_{i=1}^{K} \emptyset_i N(x|\mu_i, \sigma_i) \quad (10)$$

$$N(x|\mu_i, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right), \quad (11)$$

where K is the number of components, μ_i is the mean of i^{th} component, σ_i is the variance of i^{th} component, and $\emptyset_i$ is the weight of the i^{th} component.

We use the EM algorithm to estimate the mixture model's parameter. EM algorithm involves two steps: Expectation step and Maximization step. For GMM and a feature vector $x = \{x_1, x_2, \ldots, x_N\}$, the first step is calculating the expected sample data log-likelihood function as shown below.

$$L(x_i|\mu^k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (12)$$

The next step is the Expectation step. This step calculates the probability of data being in a cluster group, $P(b|x_i)$. It denotes in the formula below.

$$P(b|x_i) = \frac{P(x_i|b)P(b)}{P(x_i|b)P(b)+P(x_i|a)P(a)} \quad (13)$$

$$P(x_i|b) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x_i-\mu_b)^2}{2\sigma^2}\right) \quad (14)$$

The maximization step is the last step of GMM. In this step, we update the parameters of each iteration. We need to determine the new weight using (15) and (16).

$$\mu_b = \sum_{n=1}^{N} \frac{b_k G(x_n|\mu_k, \sigma_k^2)}{\sum_{k=1}^{k} b_k G(x_n|\mu_k, \sigma_k^2)} x_n \quad (15)$$

$$\sigma_b^2 = \sum_{n=1}^{N} \frac{b_k G(x_n|\mu_k, \sigma_k^2)}{\sum_{k=1}^{k} b_k G(x_n|\mu_k, \sigma_k^2)} (x_n - \mu_i)^2 \quad (16)$$

The method is based on the maximization of the likelihood of GMM in finding the model parameters. For a sequence of R training vectors, $X = [\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_n}]$, the GMM likelihood criterion is then calculated using (17). The decision rule is to select the model with the most significant score.

$$p\left(\frac{x}{y}\right) = \prod_{r=1}^{R} p\left(\frac{x_r}{y}\right) \quad (17)$$

IV. Learning Model Process and Evaluation

The process of learning is illustrated in Figure 3. Vectors obtained from the feature extraction process are labeled based on their respective classes. The labeled class is used for GMM classification. The output of the classification process is identification of the speaker.

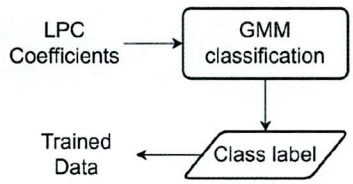

Fig. 3. Illustration of the learning process

We use 550 data for training the model. It is classified into 30 classes (C1, C2, …, C30), and each class holds 10 data. During the training process, GMM generates its class labels and calculates the error between the generated labels and the provided desired labels. The calculated error is then fed back to the model. This process is repeated until the error reaches the desired minimum level. Hence, in this way, GMM gets trained on the recorded voices and classifies the input model into the different speakers. Table 2 shows the training accuracies achieved by GMM in each class.

Table 2 shows that most of the classes achieved an accuracy of more than 70%. This shows that the model can classify training data well. The exceptions are classes C28 and C29, which only managed to get under 60%. This is because there is much noise in the recording, making the respondent's voice too faint to recognize. If examined further, the class with reasonable accuracy (>= 80%) is the recording data using earphones. Examples are C1, C2, C3…C9 to C10, which is the respondent's voice data recorded using a smartphone with the help of earphones. All these classes get an accuracy of 84%.

With this good accuracy, the model can be accepted and proceed to the testing phase. Based on the testing phase, it is expected that the model will classify the sound recording well to the sound owner. This classification will be part of speech recognition.

Table 2. Accuracy of Authentication using GMM

Class	Accuracy
C28	50%
C29	60%
C11; C14; C16; C21; C24; C25	70%
C10; C13; C15; C18; C20; C22; C26; C27	80%
C1; C2; C17; C30	90%
C3; C4; C5; C6; C7; C8; C9; C12; C19; C23	100%
Average	**84%**

V. Voice Authentication

The results obtained from the GMM model training are vectors containing a label from each voice data. The model is then tested using 150 data testing. The data is classified into the same 30 classes used in the training process. Each class is classified using five-voice data, for which 3 are cross folded with data from another class. The accuracy calculation is done manually. Table 3 shows the accuracy of each test.

Each class gets an accuracy above 50%, with the lowest accuracy being 60%. Ten classes get 60% accuracy, seven get 80% accuracy, and 13 get 100% accuracy. The average accuracy for the entire class is 82%. Furthermore, we also investigate the difference in results for using an additional device in the form of earphones. Figure 3 shows the difference in accuracy for cases with or without earphones. The system provides better recognition accuracy for data input using an earphone device than the ones without the earphone. This is because the extended device can reduce the data input noises.

Table 3. Accuracy of Testing Phase for Each Class

Class	Accuracy
C10; C11; C14; C15; C20; C21; C24; C27; C28; C29	60%
C13; C16; C18; C22; C25; C26; C30	80%
C1; C2; C3; C4; C5; C6; C7; C8; C9; C12; C17; C19; C23	100%
Average	**82%**

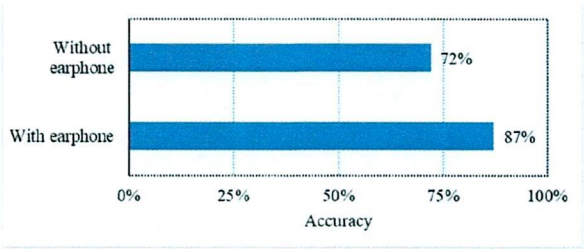

Fig. 4. Accuracy for each case

VI. Conclusion

Based on the experiment, GMM provides reasonably good accuracy, around 82%, despite GMM being a straightforward method. LPC method can be considered as a feature extraction method from voice data. Furthermore, they classify the data using GMM, although these two methods are not the most current methods. In the experiment, we use data input taken directly from the respondents by recording using a different platform, some using earphones and some using laptops and smartphones directly. The data input taken using earphones provides better accuracy than those without earphones.

This good result is the result of the performance of the GMM method. In the future, the experiment needs to be expanded by using more data. In addition, experiments using more diverse voice data also need to be carried out.

References

[1] D. Chen, N. Zhang, Z. Qin, X. Mao, Z. Qin, X. Shen and X. Li, "S2M: A Lightweight Acoustic Fingerprints-Based Wireless Device Authentication Protocol," IEEE Internet Things J., vol. 4, no. 1, pp. 88–100, Feb. 2017, doi: 10.1109/JIOT.2016.2619679.

[2] A. A. Darwish, W. M. Zaki, O. M. Saad, N. M. Nassar, and G. Schaefer, "Human Authentication Using Face and Fingerprint

Biometrics," in 2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks, Jul. 2010, pp. 274–278. doi: 10.1109/CICSyN.2010.40.

[3] I. Iancu and N. Constantinescu, "Intuitionistic fuzzy system for fingerprints authentication," Appl. Soft Comput., vol. 13, no. 4, pp. 2136–2142, Apr. 2013, doi: 10.1016/j.asoc.2012.11.001.

[4] Z. Akhtar, A. Buriro, B. Crispo, and T. H. Falk, "Multimodal smartphone user authentication using touchstroke, phone-movement and face patterns," in 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Nov. 2017, pp. 1368–1372. doi: 10.1109/GlobalSIP.2017.8309185.

[5] V. Soniya, R. S. Sri, K. S. Titty, R. Ramakrishnan, and S. Sivakumar, "Attendance automation using face recognition biometric authentication," in 2017 International Conference on Power and Embedded Drive Control (ICPEDC), Mar. 2017, pp. 122–127. doi: 10.1109/ICPEDC.2017.8081072.

[6] I. Griswold-Steiner, R. Matovu, and A. Serwadda, "Handwriting watcher: A mechanism for smartwatch-driven handwriting authentication," in 2017 IEEE International Joint Conference on Biometrics (IJCB), Oct. 2017, pp. 216–224. doi: 10.1109/BTAS.2017.8272701.

[7] D. Lu, D. Huang, Y. Deng, and A. Alshamrani, "Multifactor User Authentication with In-Air-Handwriting and Hand Geometry," in 2018 International Conference on Biometrics (ICB), Feb. 2018, pp. 255–262. doi: 10.1109/ICB2018.2018.00046.

[8] A. Bundy and L. Wallen, "Linear Predictive Coding," in Catalogue of Artificial Intelligence Tools, A. Bundy and L. Wallen, Eds. Berlin, Heidelberg: Springer, 1984, pp. 61–61. doi: 10.1007/978-3-642-96868-6_123.

[9] S. Hamzenejadi, S. A. Y. H. Goki and M. Ghazvini, "Extraction of Speech Pitch and Formant Frequencies using Discrete Wavelet Transform," in 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), 2019, pp. 1-5, doi: 10.1109/CFIS.2019.8692150.

[10] Sood, M., Jain, S. "Speech Recognition Employing MFCC and Dynamic Time Warping Algorithm," In: Singh, P.K., Polkowski, Z., Tanwar, S., Pandey, S.K., Matei, G., Pirvu, D. (eds) Innovations in Information and Communication Technologies (IICT-2020). Advances in Science, Technology & Innovation, 2021, Springer, Cham. https://doi.org/10.1007/978-3-030-66218-9_27.

[11] M. Raitoharju, Á. F. García-Fernández, R. Hostettler, R. Piché, and S. Särkkä, "Gaussian mixture models for signal mapping and positioning," Signal Process., vol. 168, p. 107330, Mar. 2020, doi: 10.1016/j.sigpro.2019.107330.

[12] A. N. Jadhav, and N. V. Dharwadkar, "A Speaker Recognition System Using Gaussian Mixture Model, EM Algorithm and K-Means Clustering," IJMECS, vol. 10, no. 11, pp. 19–28, Nov. 2018, doi: 10.5815/ijmecs.2018.11.03.