ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 11 March 2025; Accepted: 25 November 2025

https://doi.org/10.17559/TV-20250311002457
Original scientific paper

Voice Recognition-Based Room Access Security System Using a Modified VGG16 for Bahasa Indonesia

Suci DWIJAYANTI*, Bhakti Yudho SUPRAPTO, Adji SULTHONI, Hera HIKMARIKA

Abstract: Conventional security systems are vulnerable to break-ins, leading to a growing demand for access control systems that leverage biometric technologies, such as voice identification. However, speech recognition has not been widely adopted in practical security systems, particularly in the context of Bahasa Indonesia. To address this gap, this study developed a voice recognition-based access control system that utilizes deep learning algorithms to identify utterances in Bahasa, Indonesia. Three conventional convolutional neural network (CNN) architectures, VGG16, AlexNet, and a modified VGG16, were evaluated using both offline and online testing methods. Offline testing involved the use of test data, while online testing was conducted with real-time microphones. A short-time Fourier transform was employed to extract features in the form of spectrograms, which were subsequently processed by the CNNs. The modified VGG16 architecture achieved the highest accuracy, with 100 epochs and a training loss of 0.0014. The offline test results demonstrated that the modified VGG16 achieved a voice recognition accuracy of 95.09%. Additionally, online testing in our in-house control systems and robotics laboratory yielded an average real-time voice recognition accuracy of 80%. This model, based on the modified VGG16 architecture, exhibited superior performance and is well-suited for implementation in indoor access security systems.

Keywords: access control, convolutional neural networks, deep learning, security system, short-time Fourier transform, voice recognition

1 INTRODUCTION

Rapid technological advancements have brought both benefits and challenges to human existence, with security system-related issues being particularly prominent among the latter. A commonly used security system is the conventional access control system, which operates based on mechanical principles. These systems provide a low level of security and are vulnerable to crimes such as vandalism and theft. Consequently, various efforts have been made to incorporate technological advancements into security systems that can protect valuable assets [1] and offer a sense of security and comfort to their owners [2].

Security systems that identify individuals using unique data offer a potential solution to this problem. Examples of person-identification-based security systems include PIN- and password-based locks, as well as radio-frequency identification (RFID). However, unauthorized individuals can easily replicate or steal critical information, which limits the effectiveness of these methods. The use of natural human characteristics as input for a security system, known as biometric recognition, addresses this challenge. Biometric recognition involves identifying a person based on unique physical traits such as facial features, fingerprints, voice, retina patterns, and signatures [3]. Biometric recognition serves two essential functions: identification and verification. This technology uses pattern-recognition techniques and can recognize patterns in fingerprints, palm lines, faces, irises, and voices.

Human face recognition is the most commonly used biometric in security systems [4-7]. However, the information provided by facial recognition lacks reliability, stability, and distinctiveness in terms of recognizability. Changes in an individual's facial features or external environmental factors can affect the consistency of collected facial data. Human voice is another biometric that can be used for identification. Speech Recognition offers several advantages over other biometric methods, as it does not require expensive hardware and only requires a small file size [8]. Voice recognition has been utilized to facilitate automation in smart homes [9-12].

Several methods have been explored to enhance the accuracy of speech recognition. In speech recognition research, one critical factor is the selection of effective feature extraction techniques, as they directly impact the recognition accuracy. A system that translates English to Indonesian, based on speech recognition using mel-frequency cepstral coefficients (MFCC) for feature extraction and the hidden Markov model (HMM) method for classification, was proposed in [13]. The utterance of sentences comprising two to three words is not ideal for processing, as in HMM classification, the larger the number of words processed, the greater the time required for the probability calculation model. Thus, this method is computationally inefficient. Additionally, it is vulnerable to sound distortion caused by noise. Sound quality can deteriorate if the environment surrounding the sound source contains significant background noise. Noise-distorted voice recordings can adversely affect the identification process. Meanwhile, [14] proposed a combination of MFCC and principal component analysis (PCA) to improve the accuracy of Indonesian speech recognition. Fan and Liu compared several input features and proposed using the full-mel spectrogram (FBANK) for feature extraction. FBANK was selected because the spectral structure it provides is clearer at low frequencies and sparser at high frequencies, thus matching the nonlinear perceptual characteristics of the human ear [15]. Another study by Ismail et al. [16] combined support vector machines with dynamic time warping to improve speech recognition. In [17], mel-frequency was utilized in conjunction with a convolutional neural network (CNN) for speech recognition. Spectral entropy of speech signals was employed in [18] to identify the speaker. Another study by Chauhan et al. compared various feature extraction methods, such as linear predictive coding (LPC), MFCC, and zero crossing rate (ZCR), with two different classifiers: artificial neural networks and support vector machines [19].

Thus, previous studies have been limited to discussing biometric sounds in security systems. Furthermore, in the aforementioned studies, the recognition methods were still

traditional (e.g., HMM, support vector machines, and artificial neural networks) and relied on feature extraction. Although [17] used CNN for recognition, its computational process for evaluating MFCC was relatively complex compared to that for spectrograms. Additionally, only a few studies have demonstrated voice recognition in Bahasa Indonesia. To address this gap, this study applied spectrograms, visual representations of short-time Fourier transform (STFT), to identify the speaker. STFT offers an advantage over the fast Fourier transform (FFT); while the latter can accurately obtain frequency information from time data and vice versa, STFT can bridge the time and frequency domains, even in the absence of time data from a signal. Thus, STFT can simultaneously provide both time-specific and frequency-specific information. The use of spectrograms was discussed in [20]; however, it has not yet been implemented for security systems. Therefore, in this study, spectrograms were used for feature extraction in speech recognition, and a prototype security system was developed for real-time door access control using voice. This prototype was implemented in a control system laboratory as a security system. Unlike [21, 22], which utilized Google Assistant, this prototype employs a CNN algorithm to enhance the quality of voice recognition and open an entrance door in real time. CNNs are commonly used to recognize objects or images; however, they have not been widely employed for voice recognition. In this study, VGG16 was chosen as the architecture for voice recognition, as it has demonstrated robustness in [23]. Meanwhile, [24] also utilized VGG16 for speech recognition; however, it relied on lip images instead of speech signals. Therefore, VGG16 provides a straightforward and well-understood convolutional design whose layer-wise structure facilitates interpretation and modification for spectrogram-based audio inputs.

The contributions of this study are as follows:

- Speaker identification was developed for Bahasa Indonesia and applied to real-time room access control.
- The performance of two well-known architectures, namely VGG16 and AlexNet, as well as the proposed modified VGG16 architecture, was analyzed. The performance of the modified VGG16 architecture was compared with that of the first two architectures.

The remainder of this paper is organized as follows: Section 2 outlines the methodology employed in this study, Section 3 presents and discusses the results, and Section 4 summarizes the conclusions drawn from this study.

2 METHODS

2.1 Data Collection

Voice data used in this study were obtained from 102 respondents (students from our university). Each respondent uttered a 5-digit code in Indonesian 20 times; the first two digits of the 5-digit code represent the class year, and the last three digits correspond to the student identification number. Thus, a total of 2040 voice data points were collected. The voices were recorded in a closed room using a microphone (Fifine K669b, FIFINE MICROPHONE, China). The sampling frequency was set to 16 kHz, and each sample was recorded for 1-5 s.

In this study, the dataset was designed for user identification to control room access. To increase data diversity and improve model generalization, each respondent's speech was recorded in five variations: slow, moderate, and fast speaking speeds, as well as loud and soft voice levels. Additionally, the dataset was balanced across participants to ensure fair class representation. The speech samples were recorded under controlled conditions with added noise variations to simulate realistic environments.

2.2 Voice Identification Process

The first step in this process involved reducing noise that occurred during recording using pre-emphasis and bandpass filters. The pre-emphasis aimed to isolate sound signals from noise interference; however, it retained the high frequencies of the sound recording, which contained the sentences spoken by the speaker during the recording process. The pre-emphasis was calculated using the following equation:

$$[i] = [i] - \alpha S[i],\ 0.9 \le \alpha \le 1 \tag{1}$$

where $x[i]$ represents the i-th data point in the signal after pre-emphasis, and $S[i]$ represents the i-th data point in the sound sample before the pre-emphasis process. This was followed by the bandpass filter, which was part of the audio signal processing. This filter isolates or removes unwanted frequency components from the audio signals.

Following pre-processing, the sound signal was extracted using the STFT algorithm. In this process, a Hamming window was used as the input for the Fourier transform. STFT was used to transform the speech signals from the time domain to the frequency domain. The STFT process is expressed by Eq. (2).

$$X[k] = \sum_{i=1}^{N} x(i)\, w(i-k)\, e^{-\frac{j2\pi n}{K}} \tag{2}$$

where $X[k]$ is the result of the FFT process and $x(i)$ denotes the i-th element in the signal sequence within the time domain. The spectrogram serves as a two-dimensional (2D) image representation of a speech signal, as illustrated in Fig. 1.

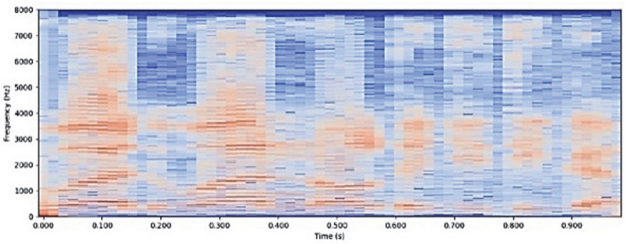


Figure 1 Sample obtained from a spectrogram with a size of 1400 × 500 pixels for training

The spectrogram image was input into the CNN in JPG format during the training phase. To ensure consistency in quality and preprocessing, all audio signals were preprocessed, normalized, and filtered. In addition, the same window size, overlap, and frequency sampling parameters were applied so that the resulting spectrogram images maintained consistent visual and statistical properties, thereby enhancing the model's generalization

capability. The model developed during the training stage was then employed to identify voices in both offline and online contexts. The dataset was divided into 70%, 20%, and 10% for training, validation, and testing, respectively.

2.3 CNN

CNN is a type of artificial neural network commonly used for image classification and recognition (Fig. 2). It consists of several layers, including convolution, rectified linear unit (ReLU), and pooling layers. In this study, CNN was employed for feature learning. Classification was performed on the subsequent layer after averaging the input, which was passed through a fully connected layer followed by a softmax layer. Fig. 3 illustrates the CNN architecture used for speech processing in this study.

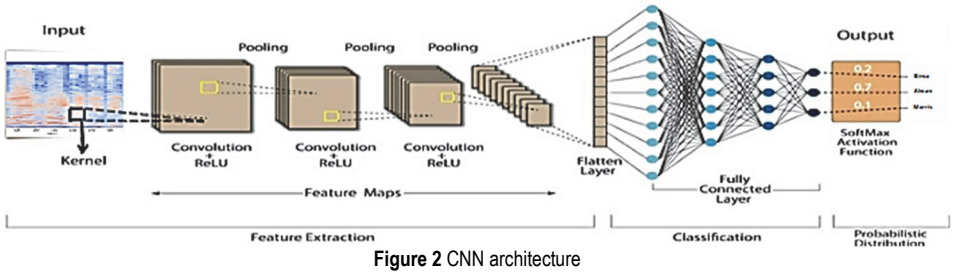


Figure 2 CNN architecture

Table 1 VGG16 architecture (Model A)

Layer (type)	Input shape	Parameter
input_1 (InputLayer)	224, 224, 3	0
block1_conv1 (Conv2D)	224, 224, 64	1729
block1_conv2 (Conv2D)	224, 224, 64	36928
block1_pool (MaxPooling2D)	112, 112, 64	0
block2_conv1 (Conv2D)	112, 112, 128	73856
block2_conv2 (Conv2D)	112, 112, 128	147584
block2_pool (MaxPooling2D)	56, 56, 128	0
block3_conv1 (Conv2D)	56, 56, 256	295168
block3_conv2 (Conv2D)	56, 56, 256	590080
block3_conv3 (Conv2D)	56, 56, 256	590080
block3_pool (MaxPooling2D)	28, 28, 256	0
block4_conv1 (Conv2D)	28, 28, 512	1180160
block4_conv2 (Conv2D)	28, 28, 512	2359808
block4_conv3 (Conv2D)	28, 28, 512	2359808
block4_pool (MaxPooling2D)	14, 14, 512	0
block5_conv1 (Conv2D)	14, 14, 512	2359808
block5_conv2 (Conv2D)	14, 14, 512	2359808
block5_conv3 (Conv2D)	14, 14, 512	2359808
block5_pool (MaxPooling2D)	7, 7, 512	0
Flatten	25088	0
dense (Dense)	4096	102764544
dense_1 (Dense)	4096	16781312
dense_2 (Dense)	103	421991

Table 2 AlexNet architecture (Model B)

Layer (type)	Input shape	Parameter
conv2d (Conv2D)	54, 54, 96	34944
max_pooling2d (MaxPooling2)	26, 26, 96	0
conv2d_1 (Conv2D)	26, 26, 256	614656
max_pooling2d_1 (MaxPooling2D)	12, 12, 256	0
conv2d_2 (Conv2D)	12, 12, 384	885120
conv2d_3 (Conv2D)	12, 12, 384	1327488
conv2d_4 (Conv2D	12, 12, 256	884992
max_pooling2d_2 (MaxPooling2D)	5, 5, 256	0
flatten (Flatten)	6400	0
dense (Dense)	4096	26218496
dropout (Dropout)	4096	0
dense_1 (Dense)	4096	16781312
dropout_1 (Dropout)	4096	0
dense_2 (Dense)	102	417894

In the convolutional layer, a convolution operation was performed to extract image features (spectrograms). The feature map was generated by applying the input spectrogram to an $n \times n$ matrix known as the feature detector. The output was then rectified using the ReLU layer to introduce nonlinearity. Following this, the pooling layers reduced the dimensions of the feature map. In this study, we applied maximum pooling to select the maximum value from the feature map. Since the feature map is in matrix form, it was converted into column vectors during the smoothing process.

Table 3 Modified VGG16 architecture (Model C)

Layer (type)	Input shape	Parameter
input_1 (InputLayer)	224, 224, 3	0
block1_conv1 (Conv2D)	224, 224, 64	1729
block1_conv2 (Conv2D)	224, 224, 64	36928
block1_pool (MaxPooling2D)	112, 112, 64	0
block2_conv1 (Conv2D)	112, 112, 128	73856
block2_conv2 (Conv2D)	112, 112, 128	147584
block2_pool (MaxPooling2D)	56, 56, 128	0
block3_conv1 (Conv2D)	56, 56, 256	295168
block3_conv2 (Conv2D)	56, 56, 256	590080
block3_conv3 (Conv2D)	56, 56, 256	590080
block3_pool (MaxPooling2D)	28, 28, 256	0
block4_conv1 (Conv2D)	28, 28, 512	1180160
block4_conv2 (Conv2D)	28, 28, 512	2359808
block4_conv3 (Conv2D)	28, 28, 512	2359808
block4_pool (MaxPooling2D)	14, 14, 512	0
Flatten (Flatten)	25088	0
dense (Dense)	1024	102761472
dense_1 (Dense)	10	104550

The output of this process served as the input to the fully connected layer. Finally, a softmax layer was employed to classify the input. This layer computed the probability of the target (speaker) being identified. The calculations were performed as follows:

$$\sigma(z)_j = \frac{e^{zk}}{\sum_{k=1}^{K} e^{zk}} \quad k = 0, 1, 2, ..., t \tag{3}$$

In this study, three architectures were investigated: (1) VGG16, (2) AlexNet, and (3) a self-modelled architecture, as presented in Tab.1 to Tab. 3, respectively. Although the VGG16 architecture is widely used in image processing, we verified that it can effectively identify voices using the information obtained from the extracted features.

The modified VGG16 architecture was designed to align with the limited size of the dataset. The hyperparameters were carefully optimized to ensure compatibility with the data characteristics. In this

modification, the number of convolutional layers was reduced by three and the number of dense layers by one, with the remaining dense layer containing 1024 neurons. This reduction in convolutional layers lowers the overall model complexity and computational cost while enhancing generalization performance under limited data conditions. Similarly, decreasing the number of dense layers reduces training time and helps mitigate overfitting, contributing to a more efficient and robust model.

3 RESULTS AND DISCUSSION

3.1 Training Results of the 3 CNN Architectures

In this study, training was conducted both with and without the transfer learning method. For the transfer learning method, the Keras library was customized to accommodate the 102 existing classes and retrained using ImageNet weights to incorporate visual knowledge from previously trained datasets. Transfer learning was applied using the VGG16 model pretrained on ImageNet as the base architecture. The original fully connected layers were removed and replaced with a new dense layer containing 1024 neurons in the modified VGG16, followed by a softmax output layer corresponding to the number of voice classes. The input spectrogram images were resized to 224 × 224 × 3 to match the model's input requirements. During training, the convolutional layers were initialized with pretrained weights and partially fine-tuned, allowing the network to adapt to the spectrogram-based voice recognition task while retaining the general feature extraction capability of the pretrained model.

Training was performed to generate three models from three different architectures: Models A, B, and C. Each architecture used the same hyperparameters, including a learning rate of 0.0001 and a batch size of 256. The Adam optimizer was employed, and two epoch values, 50 and 100, were tested. The accuracy results for the 50-epoch training of the three architectures are illustrated in Fig. 3. It was observed that Model A achieved initial and final accuracies of 0.0144 and 0.9824, respectively; Model B achieved initial and final accuracies of 0.0111 and 0.9209, respectively; and Model C obtained initial and final accuracies of 0.0124 and 1.00.

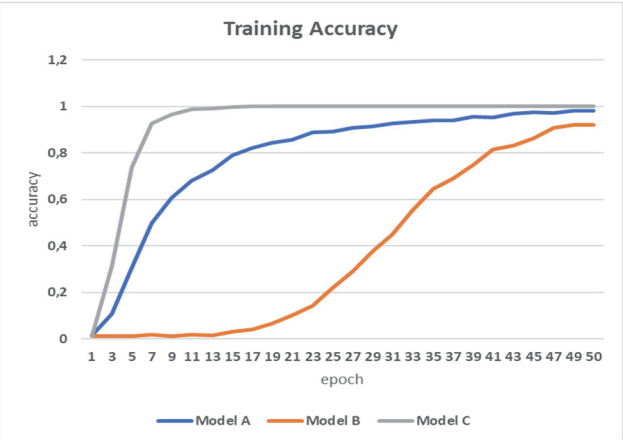


Figure 3 Training accuracies with 50 epochs of voice recognition

Subsequently, training continued with an increased number of epochs to 100 in order to assess the potential for improving accuracy. The graphs for 100 epochs are presented in Fig. 4. The initial and final accuracies of Model A were 0.0216 and 0.9993, respectively; for Model B, they were 0.0059 and 0.9915, respectively; and for Model C, the initial and final accuracies were 0.0150 and 1.00, respectively.

Figure 4 Training accuracies with 100 epochs of voice recognition

Based on Fig. 3 and Fig. 4, it can be concluded that increasing the number of epochs in the training process improved the accuracy to a value closer to 1. From the obtained data, the increase in accuracy became smaller or insignificant as the accuracy approached 1.0.

A comparison of the training accuracies of Models A, B, and C with 50 and 100 epochs is summarized in Tab. 4. As shown in this table, the accuracy was higher for all three architectures at 100 epochs compared to 50 epochs. Furthermore, Model C exhibited the best accuracy at both epoch values, with a final accuracy of 1.

3.2 Testing Using Test Data

The models resulting from the training process were then tested to assess their performance, and each architecture was used to recognize the trained facial dataset. Testing was performed using voice data that had not been previously obtained during the training process. Based on testing conducted on 102 samples from 102 classes, the following accuracies were achieved with 50 and 100 epochs, as listed in Tab. 4.

Table 4 Comparison of training accuracies of the three architectures for speech recognition

Epoch	Best training accuracy		
	Model A	Model B	Model C
50	0.9824	0.9209	1
100	0.9993	0.9915	1

Table 5 Comparison of testing accuracies among the three architectures for speech recognition using 102 samples.

Epoch	Average accuracy / %		
	Model A	Model B	Model C
50	22.54	81.37	94.11
100	62.74	94.11	95.09

As shown in Tab. 5, Model C, with 100 epochs, achieved the highest accuracy of 95.09% during offline testing, followed closely by Model C, with 50 epochs, which reached an accuracy of 94.11%. In contrast, Model A recorded the lowest accuracy, achieving 22.54% with 50 epochs and 62.74% with 100 epochs. Conversely, Model B attained an accuracy of 81.37% with 50 training epochs and 94.11% with 100 epochs. Among the three models, Model

B was the fastest to train, utilizing the fewest layers, while Model A, with the most layers, took the longest to train.

3.3 Biometrics Security System Box Prototype

To devise a room security system, the required coding, components, and microcontroller (Arduino) were integrated, with the microcontroller connected to a laptop via serial communication. Additionally, to facilitate the opening and closing of a door, a door-lock solenoid was connected to a relay module and a 12-V adapter. A 16 × 2 LCD, equipped with an I2C module, was also included to provide feedback from the system, such as predicted voice name information and solenoid status. All the components were connected through a breadboard to simplify the wiring process. The prototype and its wiring circuit are illustrated in Fig. 5 and Fig. 6, respectively.

Fig. 6 shows a laptop device connected to a microphone, which serves as the input to record the incoming sound. The feature extraction process was carried out using the STFT method, and spectrogram images were generated. These images were processed using the best-trained model, which could then predict the registered and trained voices. Furthermore, the laptop would send byte 1 to the Arduino via the serial cable if the sound was recognized and byte 0 if the sound was not recognized, in which case the solenoid would close. Subsequently, the 16 × 2 I2C LCD would display the predicted name and the status of the door-lock solenoid.

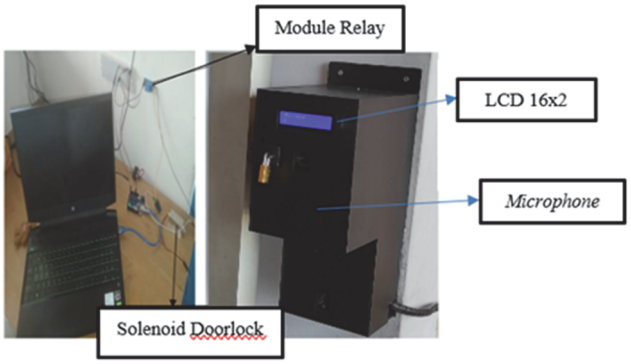


Figure 5 Prototype of the voice recognition security system

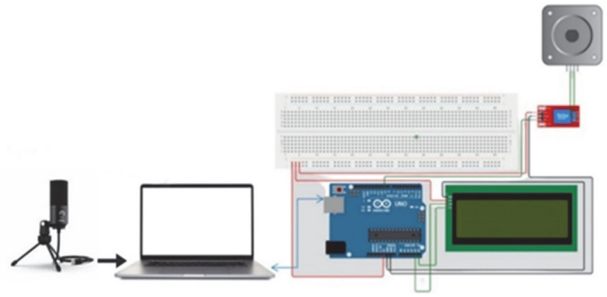


Figure 6 Wiring diagram of the prototype voice recognition security system

3.4 Real-Time Testing

Next, online testing was performed in real-time using prototypes and microphones as voice inputs to determine whether the system as a whole could function properly and effectively. The voice recognition system was tested by asking the respondent to say the number code (the class year plus the last three digits of the respondent's student identification number). The distance between the mouth and the microphone inside the prototype box was approximately 5 cm. The positioning of the respondent during the testing is illustrated in Fig. 7.

Figure 7 Respondent positioning during testing

Table 6 Online testing – spectrogram results

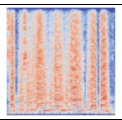
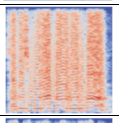
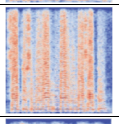
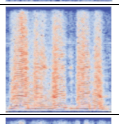
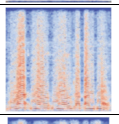
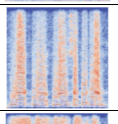
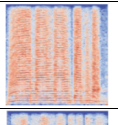
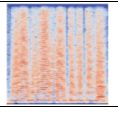
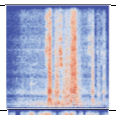
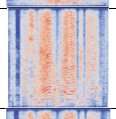
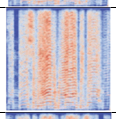
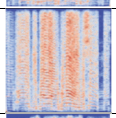
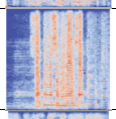
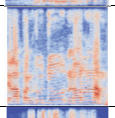
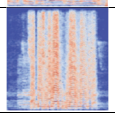
No	Name	Image spectrogram	Model C 100 epochs
1	Sample 1		Recognized
2	Sample 2		Recognized
3	Sample 3		Unrecognized
4	Sample 4		Unrecognized
5	Sample 5		Recognized
6	Sample 6		Recognized
7	Sample 7		Recognized
8	Sample 8		Unrecognized

Table 6 Online testing – spectrogram results - continuation

No	Name	Image spectrogram	Model C 100 epochs
9	Sample 9 (out of dataset)		Unrecognized
10	Sample 10 (out of dataset)		Unrecognized
11	Sample 11 (out of dataset)		Unrecognized
12	Sample 12 (out of dataset)		Unrecognized
13	Sample 13 (out of dataset)		Unrecognized
14	Sample 14		Recognized
15	Sample 15 (out of dataset)		Unrecognized
Accuracy Rate For 15 Samples			80%

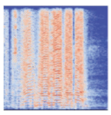
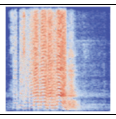
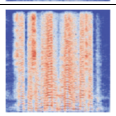
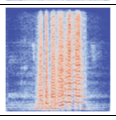
The online speech recognition testing was conducted with a real-time confidence threshold of 70%, using the best-performing model (Model C) trained for 100 epochs. The test results showed 80% accuracy across 15 tested classes, as presented in Tab. 6. This testing was performed under environmental noise and spontaneous speech variations, which differ significantly from the controlled conditions used in offline testing. These results are significant because the testing also included speech samples from students who were not registered in the dataset. This indicates that the system is capable of correctly identifying unknown individuals (samples outside the dataset) as "unrecognized", as shown in Samples 9-13 and Sample 15. Overall, these findings demonstrate that the voice recognition accuracy in identifying speakers is relatively high. Furthermore, the results in Tab. 6 show that Model C can effectively recognize authorized speakers for room access, confirming the system's feasibility as a real-time security application.

3.4.1 Testing with Machine Noise

Besides testing without noise, a noisy environment was also considered in the real-time test. This test was conducted with machine noise present around the installed prototype area to determine whether the voice recognition system was sufficiently capable of recognizing voices in noisy conditions. Tab. 7 lists the test results under machine noise conditions.

From Tab. 7, it can be observed that testing in the presence of machine noise was successful, as it correctly recognized three of the four respondents during voice recognition. The noise level during testing was set to 64-70 dB, as measured using a sound meter.

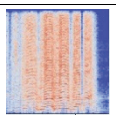
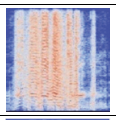
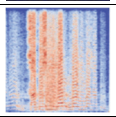
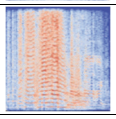
Table 7 Testing with machine noise - spectrogram results

No	Name	Image spectrogram	Model C, 100 epochs *Machine noise*
1	Sample 1		Recognized
2	Sample 2		Recognized
3	Sample 3		Unrecognized
4	Sample 14		Recognized

3.4.2 Testing with Bubble Noise

The test was also conducted with bubble noise present in the area around the installed prototype. The goal was to determine whether the voice recognition system could effectively recognize voices in noisy conditions, where multiple people talk simultaneously through the bubble noise. Tab. 8 presents the test results under bubble noise conditions.

Table 8 Testing with bubble noise - spectrogram results

No	Name	Image spectrogram	Model C, 100 epochs *Bubble noise*
1	Sample 1		Recognized
2	Sample 2		Recognized
3	Sample 3		Unrecognized
4	Sample 14		Unrecognized

From Tab. 8, it can be observed that testing under bubble noise conditions was ineffective, as it successfully recognized only two of the four respondents. The noise level during testing ranged from 71 to 76 dB, as measured using a sound meter. During testing in bubble noise, the voices could not be recognized because the system was disrupted by various vocal sounds emanating from the surrounding people, preventing the voices of the respondents from being clearly captured. Nevertheless, these results demonstrate that the proposed model is still able to recognize some speakers effectively.

In this study, recognition performance was evaluated based on true positive (TP) and false negative (FN) outcomes, as the dataset primarily focused on known speaker identification for voice-based access control. The

true positive rate (TPR), or recall, and the false negative rate (FNR), or miss rate, were calculated to assess how effectively the system recognized target voices. The comparison between machine and bubble noise conditions is presented in Tab. 9. As shown in the table, the proposed system achieved a TPR of 0.75 under machine noise, which is higher than that obtained under bubble noise. This is because bubble noise consists of overlapping speech signals, making it more difficult for the system to accurately recognize individual voices.

Table 9 TPR and FNR for machine and bubble noise

No	Performance	Machine noise	Bubble noise
1	$TPR = \frac{TP}{TP + FN}$	0.75	0.50
2	$FNR = 1 - TPR$	0.25	0.50

4 CONCLUSIONS

In addition, the robustness of the proposed system in the variation of the system to variations in microphone distance, intonation, speech rate, and speaker gender must be explored in the future work, including the number of dataset. This study successfully implemented a security system using a CNN algorithm to identify authorized voices and operate a door access mechanism. Among the three CNN architectures tested in the voice recognition-based security prototype, Model C performed the best, achieving a training accuracy of 1.0 and a training loss of 0.0045 after 50 epochs. With 100 epochs, Model C reached a training accuracy of 1.0 and a training loss of 0.0014. The test accuracy using 102 test samples was 94.11% (50 epochs) and 95.09% (100 epochs). However, the fastest training time was achieved by Model B, followed by Models C and A, which required longer training durations.

For online testing, the model with the best offline accuracy (Model C) was used. The results showed that the system achieved an accuracy of 80% for nine users registered in the database and six users not included in the dataset. Further development is needed to obtain a more optimal and robust model, particularly by updating the dataset with current user information to maintain system performance. In addition, the robustness of the proposed system to variations in microphone distance, intonation, speech rate, speaker gender, and dataset size should be further explored in future work.

Acknowledgements

The authors would like to thank Universitas Sriwijaya for the funding to conduct the research. This research was funded by DIPA of the Public Service Agency of Universitas Sriwijaya (No. SP DIPA-023.17.2.677515/2023), in accordance with the Rector's Decree No. 0118/UN9.3.1/SK/2023, dated April 18, 2023.

5 REFERENCES

[1] Munir, A., Ehsan, S. K., Raza, S. M. M., & Mudassir, M. (2019). Face and speech recognition based smart home. *Proceedings of the 2019 International Conference on Engineering and Emerging Technologies (ICEET)*, 1-5. https://doi.org/10.1109/CEET1.2019.8711849

[2] Hasan, Y., Abdurrahman, Y., Wijanarko, S., Muslimin, S., & Maulidda, R. (2020). The automatic door lock to enhance security in RFID system. *Journal of Physics: Conference Series*, *1500*(1), 012132. https://doi.org/10.1088/1742-6596/1500/1/012132

[3] Ross, A., Banerjee, S., Chen, C., Chowdhury, A., Mirjalili, V., Sharma, R., Swearingen, T., & Yadav, S. (2019). Some research problems in biometrics: The future beckons. *Proceedings of the 12th IAPR International Conference on Biometrics (ICB)*, 1-8. https://doi.org/10.1109/ICB45273.2019.8987307

[4] Jahnavi, S. & Nandini, C. (2019). Smart anti-theft door locking system. *Proceedings of the 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, 205-208. https://doi.org/10.1109/ICATIECE45860.2019.9063836

[5] Pawar, S., Kithani, V., Ahuja, S., & Sahu, S. (2018). Smart home security using IoT and face recognition. *Proceedings of the 4th International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1-6. https://doi.org/10.1109/ICCUBEA.2018.8697695

[6] Surantha, N. & Wicaksono, W. R. (2018). Design of smart home security system using object recognition and PIR sensor. *Procedia Computer Science*, *135*, 465-472. https://doi.org/10.1016/j.procs.2018.08.198

[7] Taiwo, O., Ezugwu, A. E., Oyelade, O. N., & Almutairi, M. S. (2022). Enhanced intelligent smart home control and security system based on deep learning model. *Wireless Communications and Mobile Computing*, *2022*, 1-22. https://doi.org/10.1155/2022/9307961

[8] Majekodunmi, T. O. & Idachaba, F. E. (2011). A review of the fingerprint, speaker recognition, face recognition, and iris recognition based biometric identification technologies. *Proceedings of the World Congress on Engineering (WCE)*, *2*, 1681-1687.

[9] Irugalbandara, C., Naseem, A. S., Perera, S., Kiruthikan, S., & Logeeshan, V. (2023). A secure and smart home automation system with speech recognition and power measurement capabilities. *Sensors*, *23*(13), 5784. https://doi.org/10.3390/s23135784

[10] Ge, Y., Ansari, S., Abdulghani, A., Imran, M. A., & Abbasi, Q. H. (2020). Intelligent instruction-based IoT framework for smart home applications using speech recognition. *Proceedings of the 2020 IEEE International Conference on Smart Internet of Things (SmartIoT)*, 197-204. https://doi.org/10.1109/SmartIoT49966.2020.00037

[11] Xin, Z., Liu, L., & Hancke, G. (2020). AACS: Attribute-based access control mechanism for smart locks. *Symmetry*, *12*(6), 1050. https://doi.org/10.3390/sym12061050

[12] Sudharsan, B., Corcoran, P., & Ali, M. I. (2022). Smart speaker design and implementation with biometric authentication and advanced voice interaction capability. *arXiv preprint arXiv:2207.10811*.

[13] Muhammad, H. Z., Nasrun, M., Setianingsih, C., & Murti, M. A. (2018). Speech recognition for English to Indonesian translator using hidden Markov model. *Proceedings of the 2018 International Conference on Signals and Systems (ICSigSys)*, 255-260. https://doi.org/10.1109/ICSIGSYS.2018.8372768

[14] Winursito, A., Hidayat, R., & Bejo, A. (2018). Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition. *Proceedings of the 2018 International Conference on Information and Communications Technology (ICOIACT)*, 379-383. https://doi.org/10.1109/ICOIACT.2018.8350748

[15] Fan, R. & Liu, G. (2018). CNN-based audio front end processing on speech recognition. *Proceedings of the 2018*

International Conference on Audio, Language and Image Processing, 349-354. https://doi.org/10.1109/ICALIP.2018.8455731

[16] Ismail, A., Abdlerazek, S., & El-Henawy, I. M. (2020). Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping. *Sustainability*, *12*(6), 1-15. https://doi.org/10.3390/su12062403

[17] Chandankhede, H., Titarmare, A. S., & Chauhvan, S. (2021). Voice recognition based security system using convolutional neural network. *Proceedings of the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 738-743. https://doi.org/10.1109/ICCCIS51004.2021.9397151

[18] Luque-Suárez, F., Camarena-Ibarrola, A., & Chávez, E. (2019). Efficient speaker identification using spectral entropy. *Multimedia Tools and Applications*, *78*(12), 16803-16815. https://doi.org/10.1007/s11042-018-7035-9

[19] Chauhan, N., Isshiki, T., & Li, D. (2019). Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database. *Proceedings of the IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 130-133. https://doi.org/10.1109/CCOMS.2019.8821751

[20] Dwijayanti, S., Putri, A. Y., & Suprapto, B. Y. (2022). Speaker identification using a convolutional neural network. *Jurnal Resti (Rekayasa Sistem dan Teknologi Informasi)*, *6*(1), 140-145. https://doi.org/10.29207/resti.v6i1.3795

[21] Gunawan, T. S., Mokhtar, M. N., Kartiwi, M., Ismail, N., Effendi, M. R., & Qodim, H. (2020). Development of voice-based smart home security system using google voice kit. *2020 6th International Conference on Wireless and Telematics (ICWT)*, 1-4. https://doi.org/10.1109/ICWT50448.2020.9243633

[22] Budiyanto, S., Silalahi, L. M., Simanjuntak, I. U. V., Silaban, F. A., Osman, G., & Rochendi, A. D. (2022). Smart Door Lock Prototype Design at Internet of Things-Based Airport. *2022 5th International Conference of Computer and Informatics Engineering (IC2IE)*, 331-334). https://doi.org/10.1109/IC2IE56416.2022.9970074

[23] Hamsa, S., Shahin, I., Iraqi, Y., Damiani, E., & Bou, A. (2023). Speaker identification from emotional and noisy speech using learned voice segregation and speech VGG. *Expert Systems with Applications*, *224*, 119871. https://doi.org/10.1016/j.eswa.2023.119871

[24] Rudregowda, S., Kulkarni, S. P., Gururaj, H. L., Ravi, V., & Krichen, M. (2023). Visual Speech Recognition for Kannada Language Using VGG16 Convolutional Neural Network. *Acoustics*, *5*(1), 343-353. https://doi.org/10.3390/acoustics5010020

Contact information:

Suci DWIJAYANTI
(Corresponding author)
Department of Electrical Engineering, Universitas Sriwijaya,
Jl. Raya Palembang Prabumulih KM 32, Indralaya, Indonesia 30662
E-mail: sucidwijayanti@ft.unsri.ac.id

Bhakti Yudho SUPRAPTO
Department of Electrical Engineering, Universitas Sriwijaya,
Jl. Raya Palembang Prabumulih KM 32, Indralaya, Indonesia 30662
E-mail: bhakti@ft.unsri.ac.id

Adji SULTHONI
Department of Electrical Engineering, Universitas Sriwijaya,
Jl. Raya Palembang Prabumulih KM 32, Indralaya, Indonesia 30662
E-mail: ajisutoni@gmail.com

Hera HIKMARIKA
Department of Electrical Engineering, Universitas Sriwijaya,
Jl. Raya Palembang Prabumulih KM 32, Indralaya, Indonesia 30662
E-mail: herahikmarika@ft.unsri.ac.id