Website : jurnal.umj.ac.id/index.php/icecream

Accurate Speech Recognition AI using Model of Deep Learning for Security Access

Haris Isyanto[1*], Wahyu Ibrahim[1], Riza Samsinar[1], Wiwik Sudarwati[2]
[1]*Electrical Engineering Department / Faculty of Engineering, Universitas Muhammadiyah Jakarta, Indonesia*
[2]*Industrial Engineering Department / Faculty of Engineering, Universitas Muhammadiyah Jakarta, Indonesia*
**Email address of corresponding author: haris.isyanto@umj.ac.id*

ABSTRACT

Identity theft poses a significant threat to data privacy and online transactions in cybercrime. A voice recognition approach was created to prevent this issue in security access. Every person possesses distinct and varied voice characteristics. Speech recognition is the device's capacity to identify spoken words. This speech recognition research utilizes artificial intelligence through deep learning models that are built on the Convolutional Neural Network (CNN) algorithm. CNN can accurately process vast quantities of data. The testing yielded a training accuracy of 99.8304% and a validation accuracy of 99.4001%. Testing the keywords "Welcome" and "Hello" yielded optimal results with a 100% accuracy rate. The keyword "Hello" was tested and resulted in the fastest response time of 0.64 seconds. This project aims to enhance the accuracy and speed of speech recognition, with potential applications in banking security.

Keywords: Speech recognition, deep learning, security access, accuracy, response time

1. Introduction

Every person has a distinct vocal characteristic. Differences in the size and composition of human vocal cords contribute to variations in vocal characteristics. Volume, pitch, and timbre are three key factors that influence an individual's voice. This volume pertains to amplitude, which indicates the intensity of a person's voice. This pitch is defined by a certain voice frequency. Timbre is the unique quality of a voice that sets it apart from others. People are distinguished by their distinct timbre. An individual's vocal timbre is greatly influenced by the attributes of their vocal cords [1].

Speech recognition is the ability of a machine or software to identify and convert spoken words or phrases from spoken language into a machine-readable format. Use a microphone to capture audio. Speech algorithms are used to capture and analyze vocal sounds. This study includes transforming voice input from a recorded application into a WAV file format. This enables the examination of various voice content utterances to verify voice content or keywords for speech recognition [2]–[5].

Automatic speech recognition (ASR) is the process of converting unknown voice waveforms into the correct written representation of a language. The voice signal uses a synchronized analysis technique that matches the tone frequency. Speech recognition technology is used to transcribe a set of speech files in wav format into text format, resulting in text output [6]–[9].

The current problem with speech is transmitting information to a device by manually entering text data. By using speech recognition, users do not need to enter commands using the keyboard. Information can be conveyed through voice by following the orders inputted into the device's system. Research has shown

that voice communication results in quicker response times compared to typing text while transmitting information [10]–[13].

Artificial intelligence (AI) is the study and creation of extremely intelligent systems, especially complex computer programs. AI encompasses the disciplines of machine learning and deep learning. The difference between these algorithmic models is in the amount of data they process. Machine learning usually works with a smaller dataset [12], [14], but deep learning requires a larger dataset to detect and classify items in different formats like video, text, images, and audio, without depending on predetermined rules or specific domain knowledge [3], [14]–[16]. The work utilizes a deep learning model to analyze input speech data containing keywords, particularly for speech recognition.

Convolutional neural networks are a type of deep learning algorithms known for their ability to effectively and accurately handle large amounts of collected data [13], [17]–[20].

This project aims to create a framework to enhance the accuracy and speed of speech recognition in security access verification applications. Speech recognition employs an AI system utilizing a deep learning model to authenticate spoken content, facilitating user access to system security without the need for typing. This study utilizes a voice recognition method, eliminating the need for extra hardware like retina and fingerprint scanners. The Convolutional Neural Network (CNN) method in this deep learning model may effectively and reliably train extensive speech samples with a high level of success. Speech recognition research aims to attain a training and validation accuracy of above 90% to be very effective and valuable in validating access to banking security.

2. Material and Methods

2.1. Speech Recognition

Voice recognition can be divided into two main categories: speech recognition and speaker recognition. Speech recognition is the capacity to recognize and understand spoken words. Input voice data is captured during the user identification process using a capture mechanism, which produces a speech signal. Afterward, it goes through digital signal processing. Speech data is analyzed, and then feature extraction is carried out to retain the pertinent information from the input data. The final stage is to create a template that simplifies the registration of voices, which will be saved in the database. The verification process authenticates the voice user by comparing the input speech data with a template, similar to the identification process. This enables the system to recognize the speech and produce relevant keyword parameters [2], [3], [6], [7], [21], [22].

2.2. Speech to Text

Speech-to-text recognition is also known as automatic voice recognition or computer speech recognition. The telematics interface function converts spoken language into printed text. Occasionally, the word 'voice recognition' is used to refer to speech recognition systems that are specifically trained to identify a particular speaker, like software created for personal computers. Speaker recognition is a key aspect that tries to precisely identify the speaker and improve the system's capacity to understand their speech. Speech recognition is a type of input that allows individuals to interpret spoken words. Speech recognition is the computational process of identifying spoken voices without taking into account the speaker's identity. Utilizing speech recognition technology, such issuing vocal commands to control computer applications, The parameter being compared is the voice emphasis level,

which will then be matched with the current database template [2], [8], [23], [24].

2.3. Convolutional Neural Network (CNN)

CNN are a deep learning technique primarily employed for the analysis of picture data. CNNs have been extensively utilized for the syntactic and semantic representation of text in many NLP tasks. The CNN architecture comprises two primary layers: the feature extraction layer and the classification layer. CNNs are an advancement of artificial neural networks that are designed with multiple layers, known as deep feed-forward artificial neural networks. The CNN architecture includes an input layer, convolutional layers as hidden layers, pooling layers, ReLU layers, fully connected layers, and an output layer [13], [17], [25], [26]. Figure 1 illustrates the many layers present in the convolutional neural network technique.

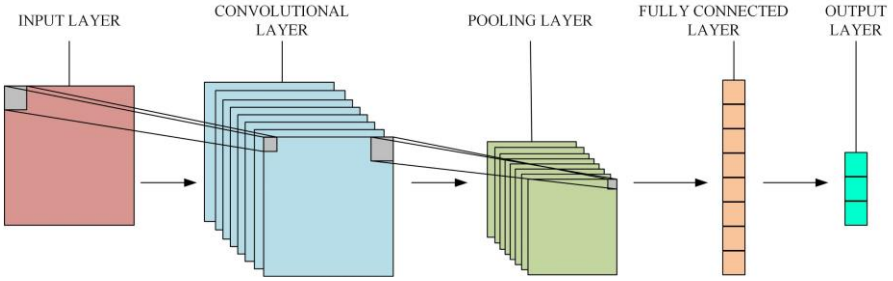

Figure 1: Convolutional Neural Network [21], [25]

2.4. Flowchart of Speech Recognition System

Speech recognition research in Figure 2 entails inputting voice data with speech and utilizing feature extraction to retain important voice information. The next stage is to train the voice data by applying convolutional neural network technology to assign labels. The training voice data produces classification results that allow for the tagging or registration of keywords. Training audio data with labels results in a speech recognition model, which is then stored in the database. Analyze the voice by administering an auditory content assessment test. This exam requires the ability to articulate accurate or inaccurate language. If a keyword match is found, the entered content is authorized or recognized if it is appropriate. If it is unsuitable, it is not acknowledged. Information is presented using corresponding terms in these instances. Keyword testing is performed by using a threshold value to identify the exact threshold level.

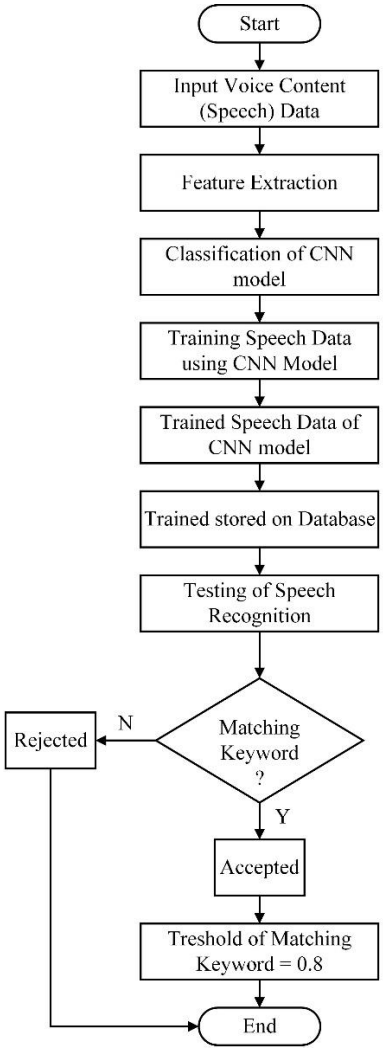

Figure 2: Flowchart of Speech Recognition System

The following Pseudocode of create speech recognition system, as described in algorithm 1:

Algorithm 1: Create speech recognition system
1. Input Voice Content (Speech) Data
2. Feature extraction voice content
3. Classification of CNN model
4. Training Speech Data using CNN Model
5. Trained Speech Data of CNN model
6. Trained stored on Database
7. Testing of Speech Recognition
8. If Matching Keyword
9. Then Accepted
10. Else Rejected (End)
11. Treshold of Matching Keyword = 0.8 (Then)
12. End

2.5. Design of Block Diagrams

Figure 3 shows the writing divided into two separate processes: keyword tagging and matching. The technique includes collecting

voice and then converting the analog signal into a digital signal through a preprocessing stage. Subsequently, feature extraction is carried out. A keyword dataset is generated and utilized to train a Convolutional Neural Network (CNN) algorithm for keyword labeling. Following the training process, a model is developed to recognize keywords. The matching method is akin to the labeling process, but it does not require a keyword dataset. The system evaluates keywords based on voice input and shows the highest score earned when the material is recognized, assuring successful matching of the keywords for the desired output. This stage comprises many processes such as data collection, voice data processing, speech template generation, training, and testing. The goal is to capture audio content by utilizing a convolutional neural network algorithm model.

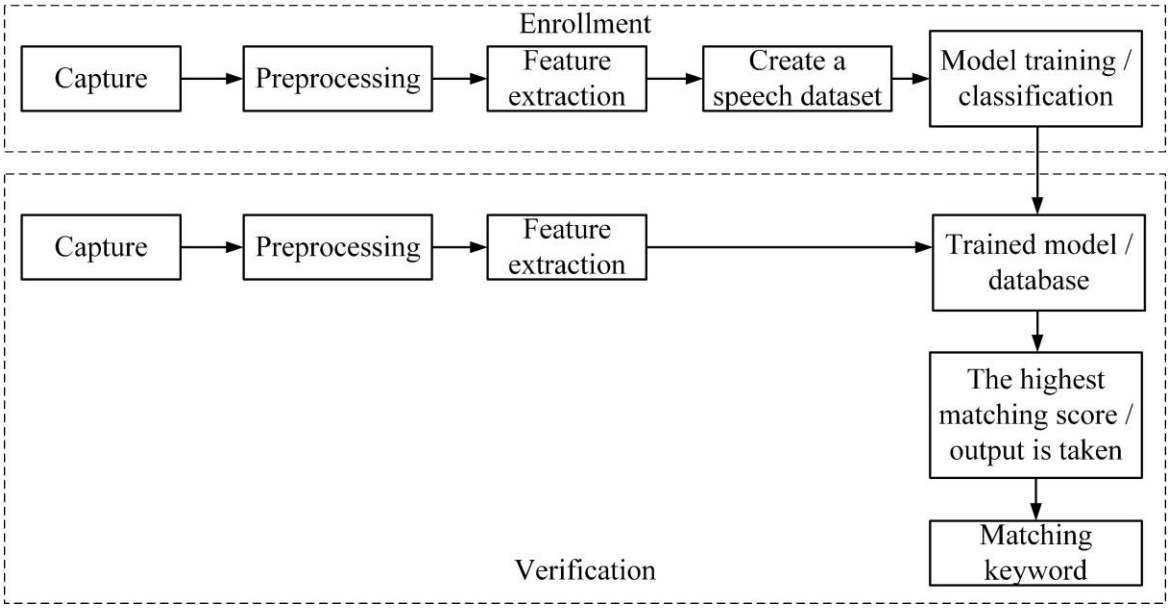

Figure 3: Design of Block Diagrams

2.6. Collecting of Data

Four individuals spoke into a microphone to record voice data. The recorded speech was processed to eliminate gaps and noise in the pronunciation and to transform the analog signal into digital format. This stage is known as preprocessing. Table 1 displays the subsequent vote data.

Table 1: Voice Data Collection

Voice Data	Sample rate (Hz)	Types of speech	Keywords
Audio	16000	*Microphone*	*Welcome*
			Open Access
			Hello
Audio	32000	*Microphone*	*Speech-to-Text could not properly transcribe*

			speech audio

Table 1 displays the keywords collected by a recording application on a smartphone. A voice microphone was employed with sampling rates of 16 KHz and 32 KHz. The sampling rate, often known as sampling frequency, is the number of audio samples taken per second, measured in Hz. A sampling rate of 16,000 Hz is recommended as a baseline requirement for speech-to-text services. For higher sample frequency audio of 32.000 Hz, speech-to-text could not properly transcribe speech audio, because the standard sample frequency in audio of telephony and telecommunications to be within the 16 kHz frequency. The data gathering procedure includes inputting voice data for voice retrieval, which involves keywords (speech), from a total of 4 voice users. After the collection process is finished, the voice data file format is transformed into WAV. Change the sample rate frequency in the speech data format to 16 KHz to adjust the voice and eliminate pauses or silence. Afterward, the keyword is tested to identify any remaining deficiencies in the voice data. After the keyword testing step, the input voice data is saved on the laptop for the purpose of segmenting or extracting voice files. The results obtained from segmenting the voice files are then used for training and testing using a convolutional neural network method. The picture shows the sequential procedure of collecting voice data. Figure 4 displays the data gathering techniques.

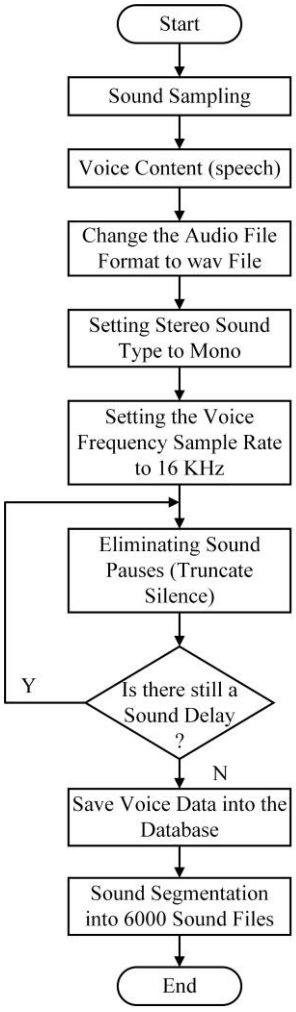

Figure 4: Data Collection Procedures

2.7. Processing of Voice Data

The quiet portion of the speech data is removed in this phase to reduce the chance of similarities in digital signals among different voice data sets. The cutting process was executed using the voice editing software Audacity.

2.8. Templates for Speech

A template is a predefined function within a library that retains extracted features from voice input, used as a reference for speech analysis. The speech recognition system will next use the

template to match features with the test voice data that has previously undergone feature extraction. The template will save voice data for up to four people, recorded by a microphone.

2.9. *Training*

Voice input data is analyzed during training to categorize speech content and then saved in a database by the system using a convolutional neural network (CNN) approach. The CNN design includes an input layer with voice datasets and a convolutional layer for training the voice data. A pooling layer is utilized to stabilize the size of the voice data. A completely linked layer is used to create links between voice actions. An output layer is used to assign a label to the keyword. Using the term "data training" results in a trained model that should be stored in the database, indicating the start of the following testing phase. Figure 5 illustrates the CNN architecture.

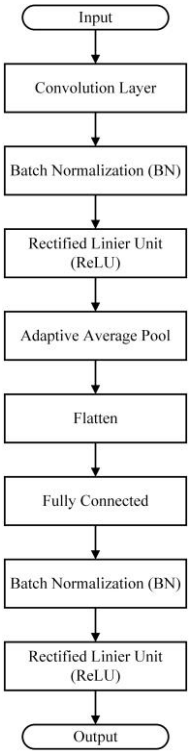

Figure 5: Architecture of CNN

2.10. Testing

During the testing stage, the keyword is assessed by analyzing the coding system's capability to conduct voice recognition on audio data. The speech recognition system is integrated into the Python programming language environment. After completing the coding process, a testing procedure is carried out utilizing different template data and testing data that consist of long and wrong spellings. This is done to assess the appropriateness of the user's keyword. Once the keyword is recognized, the system will display "true." If the keyword is not recognized, the system will display "false."

3. Results and Discussions

This study seeks to evaluate the efficiency of employing a convolutional neural network (CNN) algorithm for analyzing voice samples collected by sampling. After training, the speech data samples will create an algorithm that represents the input data, allowing for the registration of voice sample data. The next testing phase involves using a web server-based application that can be accessed via cellphones or PCs. This program is utilized for administering tests with new voice input data as well as pre-registered voice input data. The goal is to create a comparison between the two datasets and then verify if they are legitimate or invalid. Figure 6 displays an illustration of the voice recognition test.

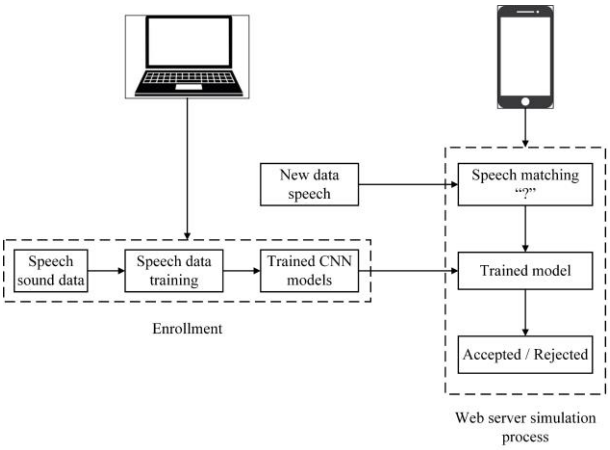

Figure 6: Testing Ilustration

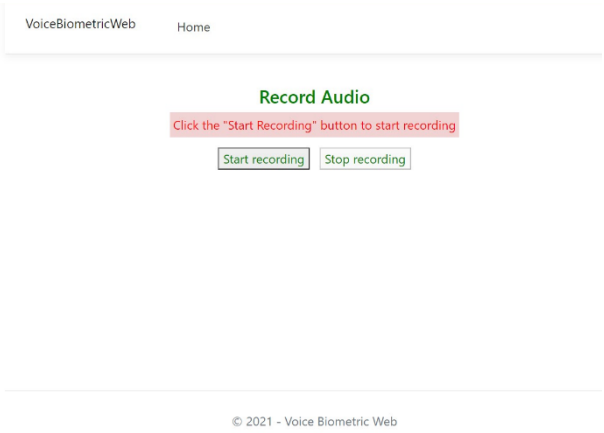

Figure 7: Speech Recognition Simulation

Figure 6 shows a testing approach where a convolutional neural network algorithm is trained using voice input data. After training the algorithm, a simulation is performed to assess the accuracy of the findings generated by the speech data. Not valid. Figure 7 depicts Simulation Testing.

3.1. Train a CNN Model using Test Data.

The CNN model training test aimed to determine the amount of speech input data required to authenticate voice users. The test results will be based on the quantity of sound data sample files, namely 6,000 sound sample files. Hence, the results will be presented as training accuracy and validation accuracy. Upon completion of the training procedure, a model is generated in the CNN algorithm. The model will be evaluated using a confusion matrix to assess the accuracy of predicting actual values. Testing the sound training file data involves conducting 40 epochs with a validation ratio of 10%. The findings include training accuracy, validation accuracy, and the duration required to train the CNN model for each epoch up to 40. Table 2 displays the training and testing outcomes of the CNN model.

Table 2: Testing of Training and Validation Accuracy

Epoch	Training accuracy (%)	Validation accuracy (%)	Training Time (M)
5	97.1794	96.1614	4.53 minute
10	99.5785	98.2350	
15	99.8493	98.8570	
20	99.6251	98.6993	
25	99.7177	98.4076	
30	99.5850	98.6844	
35	99.7876	98.4076	
39	99.8304	99.4001	
40	99.5317	98.4076	

Table 2 displays the results of training accuracy and validation accuracy testing. Training testing involves extracting 10% of the whole voice sample data quantity. 6,000 voice samples were trained. The CNN model achieved outstanding results with a training accuracy of 99.8304% and a validation accuracy of 99.4001%. The optimal outcome was obtained on period 39 out of a total of 40 epochs. The training exam necessitates 4.53 minutes of sound sample training time. The CNN model method efficiently trains on

extensive user speech samples with high accuracy. Figure 8 displays a graph comparing training accuracy with validation accuracy findings.

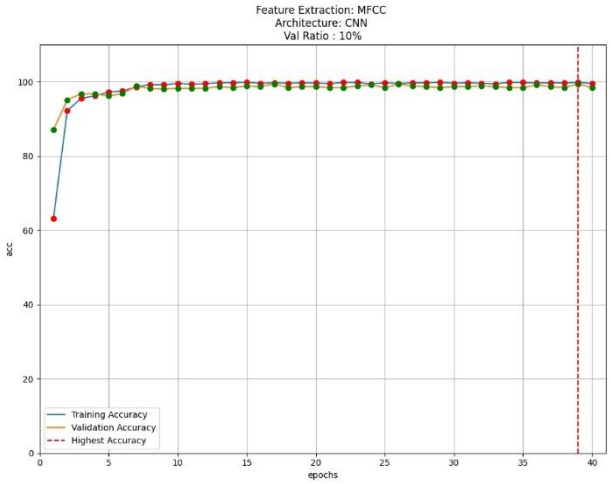

Figure 8: Graph of Training and Validation Accuracy

3.2. *Keyword Testing for Check Pronunciation*

Keyword testing Speech recognition is the assessment of accurately pronouncing keywords. The keywords examined in this pronunciation test are "Open Access, Hello, and Welcome." The speech recognition system will compare the utterances of registered words with those of tested words using certain keywords. Correct pronunciation results in approval, while incorrect pronunciation leads to rejection. Table 3 displays the outcomes of speech recognition keyword testing.

Table 3: Testing of Keyword for Check Pronunciation

Number of Testing	Open Access		Welcome		Hello	
	Check Pronunciation					
	A	R	A	R	A	R
1	A	-	A	-	A	-
2	A	-	A	-	A	-
3	A	-	A	-	A	-
4	A	-	A	-	A	-
5	A	-	A	-	A	-
6	A	-	A	-	A	-
7	A	-	A	-	A	-
8	A	-	A	-	A	-
9	A	-	A	-	A	-
10	A	-	A	-	A	-
11	A	-	A	-	A	-
12	A	-	A	-	A	-
13	A	-	A	-	A	-
14	A	-	A	-	A	-
15	-	R	A	-	A	-
16	A	-	A	-	A	-
17	A	-	A	-	A	-
18	A	-	A	-	A	-
19	A	-	A	-	A	-
20	A	-	A	-	A	-
Accuracy	95 %		100 %		100 %	

Table 3 displays the examination of keywords to verify the accuracy of pronunciation. The testing process was repeated 20 times with word utterances from 4 users. This experiment on word utterance involves three keyword phrases: "Open Access," "Welcome," and "Hello." The keywords "Welcome" and "Hello" achieved a 100% accuracy rate in 20 tests, while the keyword "Open Access" had a 95% accuracy rate with 5% rejection. Precise pronunciation is essential for it to be acknowledged and authenticated by security access systems.

3.3. Testing of Response-time for Speech Recognition

We are currently assessing the system's response time for speech recognition. Speech recognition is used to test the system for each user individually. This test is undertaken to determine the results of system validation or lack thereof. The system will provide

information in the form of confirmation or negation if there is a match in the voice input data.

Testing is performed to assess the acoustic quality of speech. When a voice is input, verification information will be created based on the voice's content. The voice will be processed to allow the algorithm to comprehend it and provide a score value. The matching score is a numerical value that establishes the threshold for identifying the test value in voice recognition. The threshold value under consideration is 0.8, which is equal to 80%. The threshold is established to prevent any parallels or resemblances between distinct users in this situation. The test results include the reaction time statistics presented in Table 4.

Table 4: Testing of Speech Recognition Response Time

Response time (seconds)			
Speech recognition			
Speech user	Open Access	Hello	Welcome
SR0	0,75	0,61	0,65
SR1	0,77	0,67	0,69
SR2	0,70	0,63	0,65
SR3	0,69	0,65	0,67
Average	0,73	0,64	0,67

Table 4 displays the evaluation of speech recognition reaction time with the phrases "Open Access," "Hello," and "Welcome." The user will say the phrase "utterance" for each instance to measure the response time. The response time for the term "Hello" was 0.64 seconds, the fastest recorded. The response time for the term "Welcome" is 0.67 seconds, while for "Open Access" is 0.73 seconds. More vocabulary lead to slower response time performance. Testing the term "Open Access" yielded slower response times in comparison to the keywords "Welcome" and "Hello." Figure 9 illustrates a comparison of testing visuals for speech recognition response time.

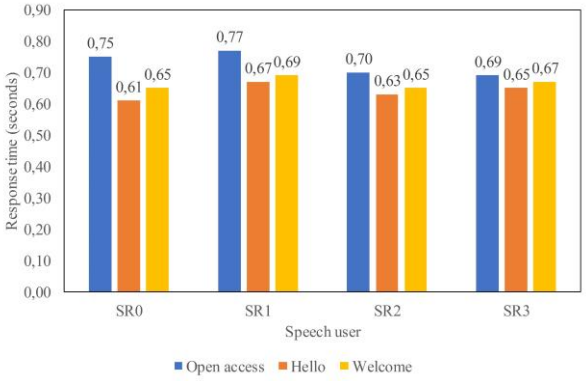

Figure 9: Comparison of Testing Graphic for Speech Recognition Response Time

Figure 9 illustrates a negative association between the number of words used and reaction time. The word "open access" is spoken more slowly than "welcome" and "hello."

4. Conclusion

The CNN model achieved a training accuracy of 99.8304% and a validation accuracy of 99.4001%. The optimal outcomes were obtained in epoch 39 out of a total of 40 epochs. Testing the keywords "Welcome" and "Hello" yielded the highest accuracy rate of 100%, but the keyword "Open Access" resulted in an accuracy rate of 95%. Precise pronunciation is crucial for recognition and verification by security access systems. Accurate results in testing the pronunciation system's performance can be acquired by verifying keyword pronunciation. Accurate pronunciation leads to approval of the voice recognition system, whereas bad pronunciation results in denial of the system. The response time for the term "Hello" was 0.64 seconds, the fastest recorded. The response time for the term "Welcome" is

0.67 seconds, and for "Open Access" it is 0.73 seconds. More vocabulary lead to decreased response time and performance. Testing the term "Open Access" yielded slower response times in comparison to the keywords "Welcome" and "Hello." Deep learning-based speech recognition AI offers advantages such as not requiring additional hardware like retina and fingerprint scanners, enabling system security access verification without typing, and enhancing the accuracy and speed of speech recognition, making it suitable for applications in banking security.

Acknowledgement

This research was supported by the Department of Electrical Engineering and the Faculty of Engineering, Universitas Muhammadiyah Jakarta.

References

[1] A. Sholokhov, T. Kinnunen, V. Vestman, and K. A. Lee, "Voice biometrics security: Extrapolating false alarm rate via hierarchical Bayesian modeling of speaker verification scores," *Comput. Speech Lang.*, vol. 60, p. 101024, 2020, doi: https://doi.org/10.1016/j.csl.2019.101024.

[2] G. Kapyshev, M. Nurtas, and A. Altaibek, "Speech recognition for Kazakh language: a research paper," *Procedia Comput. Sci.*, vol. 231, no. 2023, pp. 369–372, 2024, doi: 10.1016/j.procs.2023.12.219.

[3] A. Alsobhani, H. M. A. Alabboodi, and H. Mahdi, "Speech Recognition using Convolution Deep Neural Networks," *J. Phys. Conf. Ser.*, vol. 1973, no. 1, 2021, doi: 10.1088/1742-6596/1973/1/012166.

[4] A. Baevski, W.-N. Hsu, A. CONNEAU, and M. Auli, "Unsupervised Speech Recognition," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 27826–27839. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/ea159dc9788ffac311592613b7f71fbb-Paper.pdf

[5] D. S. Park *et al.*, "Improved noisy student training for automatic speech recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, no. Lm, pp. 2817–2821, 2020, doi: 10.21437/Interspeech.2020-1470.

[6] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimed. Tools Appl.*, vol. 80, no. 6, pp. 9411–9457, 2021, doi: 10.1007/s11042-020-10073-7.

[7] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Comput. Speech Lang.*, vol. 84, no. March 2022, p. 101567, 2024, doi: 10.1016/j.csl.2023.101567.

[8] S. Singh, F. Hou, and R. Wang, "Real and synthetic Punjabi speech datasets for automatic speech recognition," *Data Br.*, vol. 52, p. 109865, 2024, doi: 10.1016/j.dib.2023.109865.

[9] S. Alharbi *et al.*, "Automatic Speech Recognition: Systematic Literature Review," *IEEE Access*, vol. 9, pp. 131858–131876, 2021, doi: 10.1109/ACCESS.2021.3112535.

[10] H. Isyanto, A. S. Arifin, and M. Suryanegara, "Design and Implementation of IoT-Based Smart Home Voice Commands for disabled

people using Google Assistant," in *2020 International Conference on Smart Technology and Applications (ICoSTA)*, 2020, pp. 1–6. doi: 10.1109/ICoSTA48221.2020.1570613925.

[11] H. Isyanto, A. S. Arifin, and M. Suryanegara, "Performance of Smart Personal Assistant Applications Based on Speech Recognition Technology using IoT-based Voice Commands," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 2020, pp. 640–645. doi: 10.1109/ICTC49870.2020.9289160.

[12] B. A. Alsaify, H. S. A. Arja, B. Y. Maayah, and M. M. Al-Taweel, "A dataset for voice-based human identity recognition," *Data Br.*, vol. 42, p. 108070, 2022, doi: 10.1016/j.dib.2022.108070.

[13] M. Wang, H. Ma, Y. Wang, and X. Sun, "Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion," *Appl. Acoust.*, vol. 218, no. January, p. 109886, 2024, doi: 10.1016/j.apacoust.2024.109886.

[14] D. O'Shaughnessy, "Trends and developments in automatic speech recognition research," *Comput. Speech Lang.*, vol. 83, no. June 2022, p. 101538, 2023, doi: 10.1016/j.csl.2023.101538.

[15] M. . Taye, "Theoretical Understanding of Convolutional Neural Network :," *Computation*, vol. 11, 2023.

[16] J. Wu, E. Yılmaz, M. Zhang, H. Li, and K. C. Tan, "Deep Spiking Neural Networks for Large Vocabulary Automatic Speech Recognition," *Front. Neurosci.*, vol. 14, no. March, pp. 1–14, 2020, doi: 10.3389/fnins.2020.00199.

[17] J. Boyd, M. Fahim, and O. Olukoya, "Voice spoofing detection for multiclass attack classification using deep learning," *Mach. Learn. with Appl.*, vol. 14, no. August, p. 100503, 2023, doi: 10.1016/j.mlwa.2023.100503.

[18] D. Nagajyothi and P. Siddaiah, "Speech recognition using convolutional neural networks," *Int. J. Eng. Technol.*, vol. 7, no. 4.6 Special Issue 6, pp. 133–137, 2018, doi: 10.14419/ijet.v7i4.6.20449.

[19] A. M S and S. P S, "Classification of Pitch and Gender of Speakers for Forensic Speaker Recognition from Disguised Voices Using Novel Features Learned by Deep Convolutional Neural Networks," *Trait. du Signal*, vol. 38, pp. 221–230, Feb. 2021, doi: 10.18280/ts.380124.

[20] R. Shashidhar, S. Patilkulkarni, V. Ravi, H. L. Gururaj, and M. Krichen, "Audiovisual speech recognition based on a deep convolutional neural network," *Data Sci. Manag.*, vol. 7, no. 1, pp. 25–34, 2023, doi: 10.1016/j.dsm.2023.10.002.

[21] H. Isyanto, A. S. Arifin, and M. Suryanegara, "Voice Biometrics for Indonesian Language Users using Algorithm of Deep Learning CNN Residual and Hybrid of DWT-MFCC Extraction Features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 5, pp. 622–634, 2022, doi: 10.14569/IJACSA.2022.0130574.

[22] S. T. Abate, M. Y. Tachbelie, and T. Schultz, "Deep Neural Networks Based Automatic Speech Recognition for Four Ethiopian Languages," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8274–8278. doi: 10.1109/ICASSP40776.2020.9053883.

[23] X. Lu, S. Li, and M. Fujimoto, "Automatic Speech Recognition," in *Speech-to-Speech Translation*, Y.

Kidawara, E. Sumita, and H. Kawai, Eds., Singapore: Springer Singapore, 2020, pp. 21–38. doi: 10.1007/978-981-15-0595-9_2.

[24] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised Automatic Speech Recognition: A review," *Speech Commun.*, vol. 139, no. March, pp. 76–91, 2022, doi: 10.1016/j.specom.2022.02.005.

[25] W. Ibrahim, H. Candra, and H. Isyanto, "Voice Recognition Security Reliability Analysis Using Deep Learning Convolutional Neural Network Algorithm," *J. Electr. Technol. UMY*, vol. 6, no. 1, pp. 1–11, 2022, doi: 10.18196/jet.v6i1.14281.

[26] M. M. Taye, "Theoretical understanding of convolutional neural network: concepts, architectures, applications, future directions," *Computation*, vol. 11, no. 3, p. 52, 2023, doi: https://doi.org/10.3390/computation11030052.