Voice Biometrics for Indonesian Language Users using Algorithm of Deep Learning CNN Residual and Hybrid of DWT-MFCC Extraction Features

Haris Isyanto, Ajib Setyo Arifin, Muhammad Suryanegara
Department of Electrical Engineering
Universitas Indonesia, Depok, Indonesia

***Abstract*—This research develops a Voice Biometrics model for the Indonesian language users by using deep learning algorithm of CNN Residual and Hybrid of DWT-MFCC Feature Extraction. The voice dataset of Indonesian speakers were created with a duration of 5, 10, 15, 20, and 25 minutes. The testing phase of speaker recognition and speech recognition were carried out by comparing the model of CNN Residual with CNN Standard. In the phase of speaker recognition, CNN Residual model has obtained the best results with the highest precision percentage of 99.91% and the highest accuracy of 99.47% at 25 minutes voice samples, compared to the CNN Standard obtaining precision of 96.83% and accuracy of 99.00%. In the phase of speech recognition, CNN Residual model has reached the best performance at 100% accuracy during 20 trials, while CNN Standard only gave 95% accuracy. CNN Residual Model provides a better performance for its accuracy and precision, but it is slightly slower than the CNN Standard, with a time difference of 0.03 – 1.28 seconds.**

Keywords*—*Voice biometric; deep learning; CNN; DWT-MFCC; security

I. Introduction

The crime of fraud and identity theft has become a crucial threat in cybercrime. It can be associated with the excessive use of the Internet for miscellaneous activities, including online transactions, social networking, and the storage of personal information. To minimize these problems, a biometric identification method was developed, especially for high-level security entry and privacy of sensitive data access in banking transactions [1-3].

The biometric-based personal identification method is one of the alternatives developed especially for high-level security entry, such as government or military buildings, access to sensitive data or information, and theft prevention. Voice biometrics is a biometric technology that utilizes the biological characteristics of the human voice for the identification and authentication of unique patterns for each individual [4-6]. Voice biometrics includes the voice commands, allowing devices such as smartphones, computers, or laptops to receive what the user has spoken and translated into certain electronic commands. The communication of voice commands between the user voice and the device is also known as human-machine interaction [7-9]. The development of the implementation of voice biometrics technology is a solution to maintain the privacy and security of individual identity data and to avoid frauds.

The voice biometric has been perceived providing a more secure and a more reliable identification and authentication process. In principle, the authentication mechanism can be conducted remotely using a common device such as smartphones and laptops, while the cost of implementing voice biometrics is lower than other biometric solutions because it does not require special devices, such as fingerprint readers or retina scanners. It also has higher security, easy to operate, and accurate identification method to identify a person [10-13].

Currently, voice object recognition research is being developed using the CNN deep learning model. Deep learning Convolutional Neural Network (CNN) technology is one of the neural network algorithms that can assist in solving problems with large amounts of data and data complexity in the object identification process [14-16]. However, most of the previous research provides a separate discussion between the performance of speaker recognition [17-19] and speech recognition [20-22]. Meanwhile, most of the paper discussions on voice biometrics still use machine learning methods, in which it has the disadvantage of not being able to process large amounts of data and not being able to handle the complexity of large data in the identification process of voice biometrics.

In this paper, a Deep Learning voice biometric model was developed using CNN Residual and Hybrid of DWT-MFCC Extraction Feature. The use of CNN Residual is done to simplify the training and validation process, as well as to improve classification accuracy [23, 24]. Meanwhile, the hybrid extraction feature, the Discrete Wavelet Transform (DWT) [25, 26], and Mel-Frequency Cepstral Coefficients (MFCC) extraction features are used to eliminate noise interference, recognize the shape of the voice pattern from a person's characteristics and select the required voices [27, 28].

The test was carried out on 2 security system processes that apply to voice biometrics [29, 30], namely the speaker recognition security system to detect "Whose voice is the person speaking?" [31, 32]. And a speech recognition security system to detect "What keywords are spoken?" [33, 34]. If both securities are successfully accessed, then the system will be "Accepted". But if these two securities fail to be accessed, then the system will be "Rejected". Furthermore, testing is also carried out by measuring the processing time required to carry out a voice biometrics process.

To test the model, this paper compares the performance of the proposed model with the CNN Standard. The comparison is essential to see how the performance of the CNN Residual model may significantly improve the performance of voice biometrics, especially on its accuracy, which lead to better security system.

The remainder of the paper presents Underlying Theories in Section II, elaborates the theory of voice biometric, its relevant studies, and the theory of DWT and MFCC. Section III presents the architecture of the deep learning model, Section IV presents the results and analysis, while Section V concludes this paper.

II. Underlying Studies

A. Voice Biometrics

As shown in Fig. 1, the voice biometrics system consists of 2 (two) processes, namely the user enrollment and user verification/authentication [35-38]. The user enrollment is the process of identifying the user's voice identification for registration of the user's voice data into the database. The user enrollment process begins with a capturing process where the user's voice as input is captured by the microphone as a voice sensor. The user's voice input contains the speaker's voice and speech content. Preprocessing is the process of converting analog user input voice signals into digital ones. The process of creating this template is a user identification process that is carried out to register the identity of the user's voice which is a unique individual characteristic, which registers the speaker's voice (speaker recognition) and the speech recognition content which is stored in the database [36]. The workflow of the user enrollment and verification process can be shown in the first line of Fig. 1.

The second process, namely user verification or authentication, is the process of verifying the user's voice by matching the user identification between the incoming voice and the voice that has been registered in the database. In this process, there is a template match process that is intended to verify the voice data by matching the user identification between the incoming voice data and the voice sample data template that has been registered and stored in the previous database. The output is Voice Biometrics Authentication with validation of the user's voice data (accepted/rejected).

As shown in Fig. 1, in the context of verification, basically the core of the voice biometrics security system works on 2 phases [29, 30], i.e. speaker recognition to detect "Whose voice is the person speaking?" [31, 32], and speech recognition to detect "What keywords are spoken?" [33, 34]. If the voice recognition on both phases are successful, then the system user will be verified, otherwise, it will be rejected. However, most of previous studies discussed the phase of speaker recognition and speech recognition separately.

B. Relevant Studies

The relevant studies of Speaker Recognition, Speech Recognition and their combination on build up the voice biometrics are shown in Table I.

In the previous research, several papers have discussed voice biometric by using machine learning methods, including machine learning k-Nearest Neighbors (k-NN) [35], SVM machine learning, and MFCC extraction features [36], GMM machine learning and MFCC extraction features [37]. However, it is rare for papers to discuss using deep learning algorithms.

The weakness of using deep learning methods ANN and DNN is less reliable than RNN and CNN [48, 49]. While RNN is a sequential data modeling unit, RNN includes less feature compatibility when compared to CNN. The weakness of this RNN has a gradient loss problem. To avoid the problem of disappearing gradients, the RNN is combined with the LSTM. However, this LSTM has the disadvantage that it requires more memory to train [46, 50-52]. CNN is considered more reliable than ANN and RNN. And the weakness of using machine learning methods GMM, GMM-UBM, GMM-HMM, and HMM is only able to process data in smaller amounts and is less able to process complex data [53, 54]. Therefore, the advantages of CNN are having reliable computing capabilities, having a high accuracy, having the ability to process large training data, having an ability to automatically detect important features without human supervision, and being able to classify data complexity in the voice identification process [17-22].

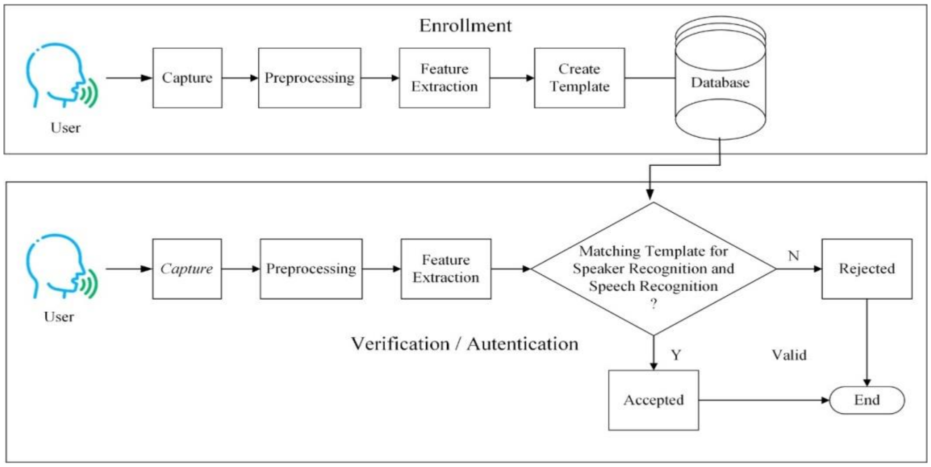

Fig. 1. Process Workflow of user Enrollment and user Verification [36].

TABLE I. RELEVANT STUDIES

Studies of	Research by	Methods	Explanation
Speaker Recognition	[55]	i-vector with machine learning GMM, PNCC, and RASTA PLP feature extraction	PNCC, RASTA PLP and MFCC are sensitive to noise and Machine Learning (ML) is only able to process smaller amounts of data and is less capable of processing complex data
	[56]	i-vector with machine learning GMM-UBM, and MFCC extraction features	
	[57]	i-vektor with deep learning DNN and MFCC extraction features	DNN is less reliable in computing capabilities
	[17-19]	deep Learning CNN	No extraction feature
	[39]	deep-learning CNN and MFCC extraction features	MFCC is sensitive to noise
Speech Recognition	[40]	i-vector with machine learning GMM	ML is only able to process smaller amounts of data and is less capable of processing complex data
	[41]	machine-learning GMM and HMM	
	[42];	machine-learning HMM, MFCC extraction feature	The explanation of ML and MFCC is the same as above
	[43];	machine-learning HMM and deep learning ANN	ANN and DNN are less reliable in computing capabilities
	[44]	deep-learning DNN	
	[45];	deep-learning RNN	RNN only has less feature compatibility compared to CNN
	[46];	deep-learning RNN and LSTM	
	[20-22],	deep-learning CNN	No extraction feature
	[47].	deep-learning CNN and LSTM	
Voice Biometrics	[35];	machine-learning k-Nearest Neighbors (k-NN)	MFCC is sensitive to noise and ML is only able to process smaller amounts of data and is less capable of processing complex data
	[36];	machine learning SVM and MFCC extraction features	
	[37].	machine-learning GMM and MFCC extraction features	
	The Model developed in this research	deep-learning CNN Residual and Hybrid of DWT-MFCC extraction features	Deep-Learning CNN Residual is able to solve problems of large amounts of data, to process complex data, to reduce the number of parameters and arithmetic operations in convolution operations and is able to simplify the training and validation process, thus may increase the classification accuracy. In addition, the Hybrid of DWT-MFCC extraction feature can remove noise, recognize the shape of a voice pattern from a person's characteristics, and select the required voices.

From the feature extraction point of view, studies in [55, 56] discussed the performance of speaker recognition, while studies in [40-43] discussed the performance of speech recognition with the i-vector extraction features, PNCC, RASTA PLP, and MFCC. They had performed an accuracy of about 76%. Research on speaker recognition [17-19, 39] and speech recognition [20-22, 44-47, 57] by using a deep learning model with i-vector extraction features and MFCC, have obtained an accuracy of around 71-90%.

C. *Deep Learning CNN Standard*

This CNN standard is a deep learning algorithm technology that has high performance and has been used for database training and testing. With the performance of the CNN standard algorithm, it is expected to improve higher performance compared to using the previous machine learning algorithm.

The architecture of CNN standard is shown in in Fig. 2. It consists of 1 input layer, 5 layers 3x3 Convolution 16 Filters, 1 layer 3x3 Convolution 32/2 Filters, 3 layers 3x3 Convolution 32 Filters, 1 layer 3x3 Convolution 64/2 Filters, 3 layers 3x3 Convolution 64 Filters, 1 layer 3x3 Convolution layer 128/2 Filter, 3 layer 3x3 Convolution 128 Filter, 1 layer Adaptive Average pool, 1 layer Flatten, 1 layer Fully connected and 1 layer output layer.

In CNN Standard, the input layer is a layer for processing data as a set of features. In each Convolution layer, there is a filter/kernel size (filter size) convolution matrix of 3x3 with some filters 16, 32, 64, and 128. This convolutional layer carries the main part of the network computing load, which does most of the heavy computational work. The convolution layer is needed to speed up the extraction of spatial features in the data so that the number of parameters that need to be used to extract features can be reduced, and in the end, will speed up runtime training. Each Convolution layer contains Batch Normalization and ReLu. The batch normalization layer is a normalization technique performed between the layers of the Neural Network, which standardizes the input to the layer for each mini-batch. This is done with mini-batches instead of full datasets. This serves to speed up the training process and uses a higher learning rate, making learning easier. Batch normalization is also able to solve the main problem called internal covariate shift.

ReLU (Rectified Linear Unit) is a node or unit that implements the network layer activation function. ReLU is useful for helping prevent the exponential growth in the computations required to operate neural networks. The adaptive average pooling layer is an easy average pooling operation layer, which gives the input and output dimensions, to calculate the correct kernel size required to produce an output of the given dimensions from the given input. The flatten layer is a layer that involves taking the combined feature maps generated in the pooling step and converting the data into one-dimensional vectors, to be inserted into the next layer; by flattening the output of the convolution layer to create one long feature vector. And it is connected to the final classification model, which is called the fully-connected layer. Furthermore, a Fully Connected Layer is a layer where all inputs from one layer are connected to each activation unit of the next layer. The last few layers are full connected layers that compile the data extracted by the previous layer to form the final result. The fully connected layer is a full process of Batch Normalization and ReLU data input.

D. MFCC Method for Extraction

As shown in Fig. 1, feature extraction plays an important role to provide good accuracy. Mel Frequency Cepstral Coefficients (MFCC) is believed to be a method that has the highest level of accuracy with speech recognition rates and the fastest feature extraction time compared to other voice feature extraction methods [58]. It is so that the MFCC method is good in accuracy for feature extraction in speech recognition processing in voice biometrics. MFCC is one of the feature extraction methods and methods that are most often used in various fields of voice processing, because it is considered very good in presenting the characteristics of a signal, such as in speech recognition technology, both voice biometrics, speaker recognition, and speech recognition. MFCC is used to recognize the shape of the voice pattern from the extraction of a person's characteristics and choose only the voices that are needed from other voices that are not needed. The feature extraction process with MFCC is a process of taking from feature extraction using a discrete Fourier transform. The Fourier transform can only determine the frequency that appears in a signal, but cannot determine when that frequency appears. The sequence process for the MFCC block diagram can be shown in Fig. 3 [59].

The following is the sequencing process for the MFCC block diagram:

1) Pre-emphasis: Used for the filtering process which compensates for the high-frequency portion of the voice signal that is suppressed during the voice production mechanism. The pre-emphasis process is following Equation (1) [28, 59].

$$y(n) = s(n) – a\ s(n -1) \tag{1}$$

where *y(n)* = signal from the calculation result of pre-emphasis process, *n* = serial number of voice signal, *s(n)* = voice signal before pre-emphasis process, = constant of pre-emphasis filter, with a value of 0.9 α 1.0 and *s* = voice signal.

2) Framing and windowing: In the framing process, analyze the speech signal of the voice in the form of frames. The signal is divided into several pieces, to facilitate the calculation and analysis of voice signals. Each frame is represented with an interval of 20-40 ms and the signal is continued every 10 ms, which overlaps the previous signal and the next signal [60]. Windowing is used to avoid discontinuity between signals. The most widely used type of window is the hamming window [28, 59, 61].

3) Fast Fourier Transform (FFT): In the Fourier transform, the digital voice signal is transformed into a frequency signal. FFT is an algorithm that has a very fast calculation to perform Fourier transforms in the discrete domain. The results of the FFT process produce detection of frequency domain waves in discrete form [28, 59].

4) Mel filterbank: Filterbank is used to determine the energy size of a certain frequency band in a voice signal. Filterbanks are overlapping bandpass filters. Mel is a unit of measure based on the frequency perceived by the human ear. Based on the Mel scale, it is linear below the 1 kHz frequency and logarithmic above it. Mel scaling process according to Equation (2) as follows [27, 59, 60]:

$$mel = 2595 \log_{10} (1+ f / 700) \tag{2}$$

where Mel is the output of the Mel filterbank, and *f* is the input of the filterbank. While 2595 and 700 are fixed values that have been widely used in the MFCC method in many studies. Mel spaced filterbank as in Fig. 4, the filter bandwidth below 1 kHz is linear while above 10 kHz is logarithmic [59].

5) Discrete Cosine Transform (DCT): DCT is used to calculate the MFCC of a single frame. DCT aims to produce a Mel spectrum to improve recognition quality. The DCT process is following Equation (3) [59].

$$C_m = \sum_{k=1}^{K} (log_{10} Y[k]\ cos \left[m(k - \tfrac{1}{2}) \tfrac{\pi}{K}\right];\ m = 1,2,\dots. \tag{3}$$

In this case, C_m = Coefficient, where *Y[k]* = the output of the filterbank process on the index, *m* = the number of coefficients, and *K* is the expected number of coefficients.

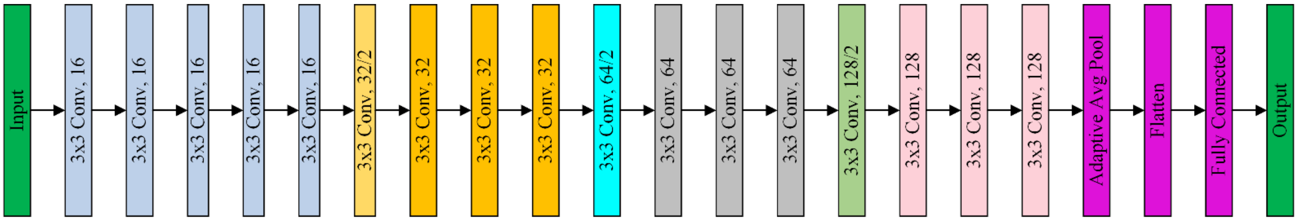

Fig. 2. The Architecture of CNN Standard (CNN No Residual).

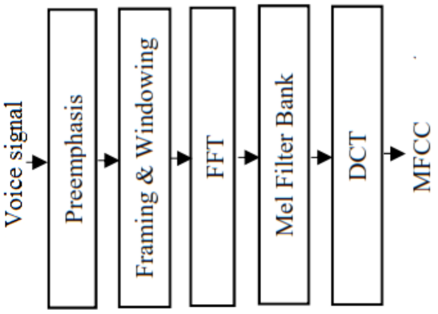

Fig. 3. MFCC Block Diagram [59].

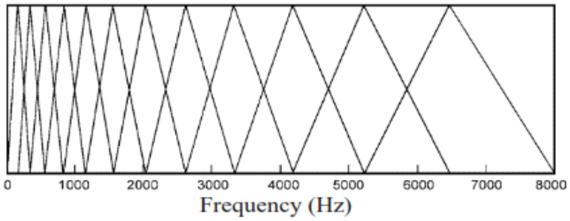

Fig. 4. Example of Mel Spaced Filterbank [59].

6) Delta coefficients: Delta coefficients have been used mostly, in addition to the MFCC extraction method. The accuracy of the speech recognition system can be improved by adding the time derivative to obtain stable basic parameters. The equation for calculating the delta can be seen in Equation (4) [59].

$$dt = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \qquad (4)$$

where *dt* is the delta coefficients of the *t* frame. In general, the value of *N* is 2. The data for the sum of the delta coefficients is the same as the MFCC, the number of coefficients is 13. The sum of the MFCC data plus the delta coefficient is equal to 26 features of the data dimensions [28, 59, 61].

E. DWT Method for Hybrid Extraction MFCC

This MFCC method has drawbacks, where the feature extraction method of voice signals is sensitive to noise [61]. From several previous studies, there is still a need to improve the performance of MFCC. To improve the performance of MFCC on the voice biometrics identification system, a method that can eliminate noise frequencies is needed. There is a need to develop a hybrid method which will help to provide better performance solutions. It is signified that the voice biometrics with a Hybrid DWT-MFCC extraction feature can be used to eliminate noise interference, recognize the shape of the voice pattern from a person's characteristics and choose only the voices that are needed from other voices that are not needed. With the Hybrid MFCC-DWT feature extraction method, it is hoped that reliable features can be formed and produce a high level of accuracy and are better than before [62, 63].

Based on previous research, Discrete Wavelet Transform (DWT) is a good method to eliminate noise (denoising) in signal processing so that the voice quality in voice biometrics is better. The wavelet signal processing is suitable for nonstationary signals, whose spectral content changes over time. Each wavelet transforms measurement according to a fixed parameter will provide information about the time-temporal range of the signal and information about the frequency spectrum of the signal. The wavelet transform provides an approach to multi-analytical signal resolution and this technique has been used to identify voice signal features. The wavelet transform is an integral part of the raw signal *x(t)* multiplied by the scale, type shift of the basic wavelet function *ψ(t)*.

Continuous wavelet transform (CWT) is calculated in Equation (5) as follows [26, 63, 64]:

$$CWT(a,b) = \int_R x(t)\frac{1}{\sqrt{a}}\psi^*(\frac{t-b}{a})dt \qquad (5)$$

where a is the scaling parameter and b is the time localization parameter. DWT is often more efficient than CWT to avoid counting on each CWT scale.

With parameter changes, DWT is defined in Equation (6) as follows [62]:

$$DWT(j,k) = 2^{-j/2}\int_R x(t)\frac{1}{\sqrt{a}}\psi^*(2^{-j}\,t-k)dt \qquad (6)$$

After changing from CWT to DWT, the original continuous wavelet function becomes a discrete wavelet function and a scaling function. The discrete wavelet function and the scaling function as Low Pass Filter (LPF) and High Pass Filter (HPF). The DWT process works like Fig. 5 [63].

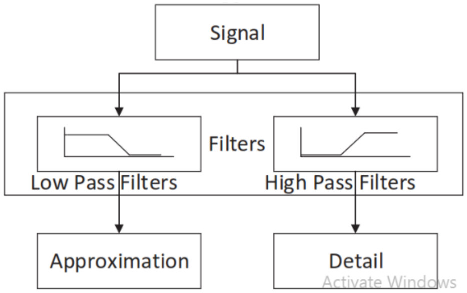

Fig. 5. Process of DWT Signal Filtering [63].

Thus, architecturally it can be described, Referring to Fig. 1, the hybrid extraction process is carried out on the processing results (signal in Fig. 1) which is then carried out with the extra feature DWT-MFCC process to help eliminate noise interference, recognize the shape of the voice pattern from someone and selects the required voices.

III. The Framework

A. The Architecture

Fig. 6 shows the architecture of the voice biometric model which is developed in this research. In principle, the user enrollment/training is processed by using the DWT-MFCC for the part of feature extraction and CNN Residual model for the part of training process. This training process is a capability learning process where the CNN model is trained to identify user voice datasets using large GPU and CPU computing devices. In this training process, the user identification process is carried out to register the user's voice identity which is stored in the database. After completing the training process, the CNN model that has been trained will produce a Trained CNN Model. Such a trained CNN model will be subsequently used for the user verification process.

This user verification process is the process of classifying and authenticating voice datasets. This user verification process will directly apply the new voice data to the Trained CNN Model and use it to conclude the output. So, when a new user's voice data is entered into the Trained CNN Model, the system will verify the voice data by matching the user identification between the new voice data and the voice data that has been registered in the database. Next, the system will issue predictions based on the prediction accuracy of the data that has been trained on the Trained CNN Model. The Trained CNN Model classification is optimized to maximize prediction performance to achieve high accuracy. The output of the trained CNN model classification is user voice authentication, in the form of data validation (valid / not) or (accepted/rejected) of the user's voice data.

B. *CNN Residual as Deep Learning used in this Research*

In this research, the architecture of CNN Residual is shown in Fig. 7. The architecture of CNN Residual lies in the implementation of the Residual 8 Shortcut layer by stepping over every 2 layers, which consists of 1 input layer, 5 layers 3x3 Convolution 16 Filters, 1 layer 3x3 Convolution 32/ 2 Filters, 3 layers 3x3 Convolution 32 Filters, 1 layer 3x3 Convolution 64/2 Filters, 3 layers 3x3 Convolution 64 Filters, 1 layer 3x3 Convolution layer 128/2 Filters, 3 layers 3x3 Convolution 128 Filters, 1 layer Adaptive Average pool, 1 Flatten layer, 1 fully connected layer, and 1 output layer.

For CNN Residual, Residual Shortcut is a branching technique for CNN layers, where one branch is a shortcut over 1 or several other branch layers. Initially, the CNN Residual technique was intended to deal with the problem of saturation by increasing the number of layers. Difficult iteration problems and a large number of layers tend to cause a decrease in the quality of classification in terms of speed and accuracy. With the increasing amount of large data, it will affect the increasing capacity of the CNN model, on the number of parameters, filters, and layers. By using the residual technique, the iteration training can be shorter, and the accuracy value will increase, along with the increase in the number of parameters, filters, and layers. The following is the general equation of the shortcut residual identity function, which can be seen in Equations (7) and (8) [24].

$$y=F(x,\{W_i\})+x \tag{7}$$

$$y=F(x,\{W_i\})+W_s x \tag{8}$$

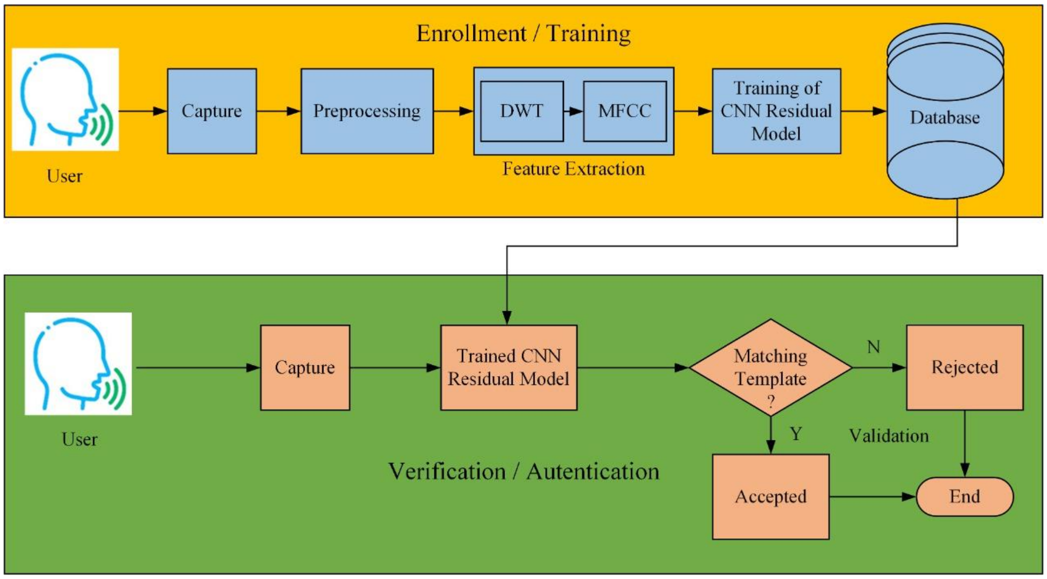

Fig. 6. Framework of Voice Biometrics Model for user Enrollment/Training and user Verification/Authentication Processes, based on DWT-MFCC and CNN Residual.

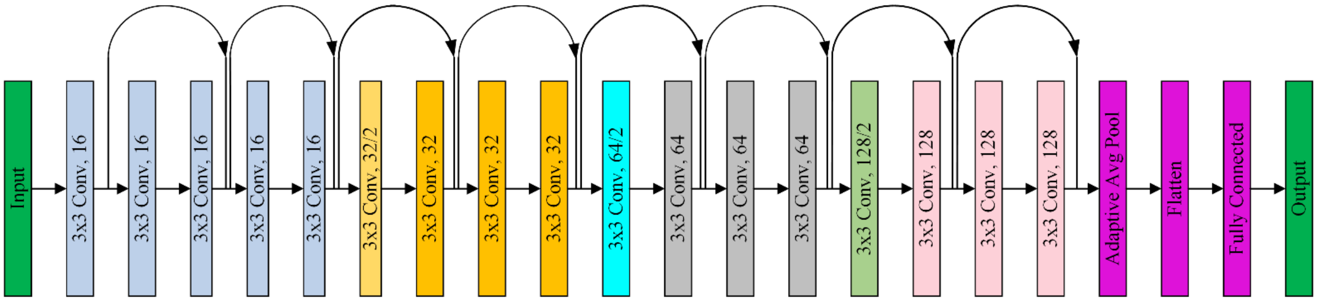

Fig. 7. The Architecture of CNN Residual.

y is a feature map after residual, *F(x,{W_i})* is a filter (residual mapping) whose optimal value is determined, and *x* is a feature map input. *W_i* is the layer group that is skipped, and *W*_s is a linear projection in adjusting the dimensions for *x* and *y* when performing shortcuts such as downsampling or upsampling. Although there is almost no change in arithmetic operations and the number of parameters, the addition operation performed can be neglected for the computational load. The application of this residual technique will result in a shorter iteration process and affect the classification results for the better [65, 66]. To further improve the performance of the voice biometrics system, it is proposed to optimize CNN using a CNN residual model. The optimization of this CNN residual is needed to simplify the training and validation process, as well as increase the classification accuracy.

C. CNN Standard as a Comparison of Performance

The performance analysis of CNN Residual model is conducted by comparing with CNN Standard. The essential differences between them are about the Total Parameters and Parameter Size, in which the parameters on CNN Residual Model are greater than the CNN Standard Model. This will affect the working process of the CNN Residual Model which is longer than the CNN Standard Model. CNN Standard Model Parameters and CNN Residual can be seen in Table II.

TABLE II. PARAMETERS OF CNN STANDARD AND CNN RESIDUAL MODEL

No.	Parameter	CNN Standard Model	CNN Residual Model
1.	Total Parameters	707,386	718, 586
2.	Parameter Size (MB)	2.70	2.74

D. Data Set of Indonesian Language

In this research, the original voice dataset of Indonesian language speaker was created. It is essentially used on training the CNN Model algorithm. The creation of the voice dataset begins with the user's voice input via the microphone on the smartphone. The making of this voice dataset involved 10 users, starting from Voice Biometric0 to Voice Biometric9 (VB0 - VB9). Each VB user input contains the unique speaker and speech. Each VB user contributes the voice sample by speaking in Indonesian language for 50 minutes duration.

To make a uniform voice sample files in the dataset, it is necessary to set the following parameters: First, changing the stereo voice to mono voice; Second, changing the frequency of the voice sample rate to 16,000 Hz; Third, truncating the silence to eliminate the pause in the user's voice, so that the result is that every VB user is sampled for 25 minutes, without any pauses; Fourth, segmenting the voice samples for each VB user into 1500 files each; Fifth, changing the voice sample file in the form of a WAV file type format. Finally, with the number of 10 users, a voice dataset is obtained with a total number of voice samples being 15,000 files.

Furthermore, the voice dataset is processed with the DWT-MFCC extraction feature so that it can recognize the shape of the voice pattern from a person's characteristics, can choose only the voices that are needed, and can eliminate noise disturbances. After completing the feature extraction process, the voice dataset is ready to be trained with the CNN model algorithm.

E. Testing

In this research, the system's performance was tested by conducting a performance assessment.

1) The first phase of Performance Testing, namely Speaker Recognition with the CNN Residual Model Algorithm using DWT-MFCC, (compared to CNN Standard).

2) The second phase of Performance Testing is the performance of Voice Biometric from Speech Recognition with the CNN Residual Model Algorithm using DWT-MFCC, (compared to CNN Standard).

3) Performance Testing of Training Process Time on Voice Biometric with Algorithm CNN Residual Model using DWT-MFCC, (compared to CNN Standard).

Each test was carried out for a sample duration of 5 minutes, 10 minutes, 15 minutes, 20 minutes, and 25 minutes.

IV. RESULT AND DISCUSSION

A. Performance Testing of Speaker Recognition ("Whose Voice is the Person Speaking?")

Performance testing of speaker recognition on the CNN Model is to test the performance of speaker recognition with the CNN Residual Model Algorithm using DWT-MFCC, (compared to CNN Standard). This performance measurement uses the confusion matrix which is a machine learning classification method. This confusion matrix provides information on the comparison of the classification results carried out by the CNN Training model system with the actual classification results. From the results of the CNN Trained Model, it will be used to measure performance with the Confusion Matrix [67, 68]. In this research, a classification system for identifying voice datasets was carried out, where the input data were grouped into 10 VB users to classify the VB voice datasets. In determining the best model, the confusion matrix method becomes important to consider in choosing the best model between deep learning CNN Residual models using DWT-MFCC (compared to CNN Standard).

This performance measurement uses a confusion matrix, which is divided into 4 (four) combinations representing the results of the classification process, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). From the values of TN, FP, FN, and TP, the accuracy and precision of speaker recognition performance with the CNN Standard Model and CNN Residual algorithms are obtained at 5, 10, 15, 20, and 25 minutes of voice sample duration. The analysis data on speaker recognition performance testing with the CNN Standard and CNN Residual Model algorithms can be seen in Table III and IV, and Fig. 8 and 9.

Based on the comparison of accuracy on speaker recognition performance with CNN Residual model using DWT-MFCC, the best results were obtained with the highest percentage accuracy value of 99.47% for the 25 minutes duration voice sample. Accordingly, the larger the number of voice sample durations or the larger the number of voice sample files executed, the higher the percentage of accuracy performance in prediction. It can be shown in Table III and Fig. 8.

Based on the comparison data of precision on speaker recognition performance with CNN Residual model using DWT-MFCC, the best results were obtained with the highest percentage precision value of 99.91% for the 25 minutes duration voice sample. Based on data analysis, the larger the number of voice sample durations or the larger the number of voice sample files executed, the higher the percentage of precision performance in prediction. It can be shown in Table IV and Fig. 9.

By comparing the performance of Speaker Recognition between CNN Residual Models and CNN Standard, the main results are as follows:

1) The accuracy of CNN Residual is higher than CNN Standard, in which CNN Residual is about 96.10% - 99.47% while the later is 95.80% - 99.00%; as can be seen in Table III and Fig. 8.

2) The precision of CNN Residual is higher than CNN Standard, in which CNN Residual is about 80.05% - 99.91% while the later is 78.85% - 96.83%; as can be seen in Table IV and Fig. 9.

3) The CNN Residual's best results or the highest percentage value are 99.91% precision and 99.47% accuracy for 25 Minutes voice sample duration. The same condition is also applied to the CNN Standard of 96.83% precision and 99.00% accuracy.

It can be implied that the greater the number of voice sample files and the more voice sample training carried out, the higher the level of precision and accuracy in prediction performance will be. By looking at the comparison results, the highest percentage value shows the best value for the precision and accuracy of Speaker Recognition on CNN Residual. It can be concluded that the speaker recognition performance of the CNN Residual model is better than the CNN Standard.

B. *Performance Testing of Speech Recognition ("What Keywords are Spoken?")*

Performance testing of speech recognition on the CNN Model Algorithm aims to test the accuracy of speech recognition performance with the CNN Residual Model Algorithm using DWT-MFCC (compared to CNN Standard) at 5, 10, 15, 20, and 25 minutes of voice sample duration. This is done by matching keyword speech or matching speech content.

TABLE III. COMPARISON OF ACCURACY ON SPEAKER RECOGNITION PERFORMANCE WITH CNN STANDARD AND CNN RESIDUAL USING DWT-MFCC

Duration of Voice Samples (Minutes)	**Accuracy of Speaker Recognition Performance (%)**	
	CNN Standard	***CNN Residual***
5 Minutes	95,80	96,10
10 Minutes	96,33	96,58
15 Minutes	96,76	97,05
20 Minutes	97,25	97,95
25 Minutes	99,00	99,47

TABLE IV. COMPARISON OF PRECISION ON SPEAKER RECOGNITION PERFORMANCE WITH CNN STANDARD AND CNN RESIDUAL USING DWT-MFCC

Duration of Voice Samples (Minutes)	**The precision of Speaker Recognition Performance (%)**	
	CNN Standard	***CNN Residual***
5 Minutes	78,85	80,05
10 Minutes	81,02	82,63
15 Minutes	84,52	86,12
20 Minutes	89,74	93,18
25 Minutes	96,83	99,91

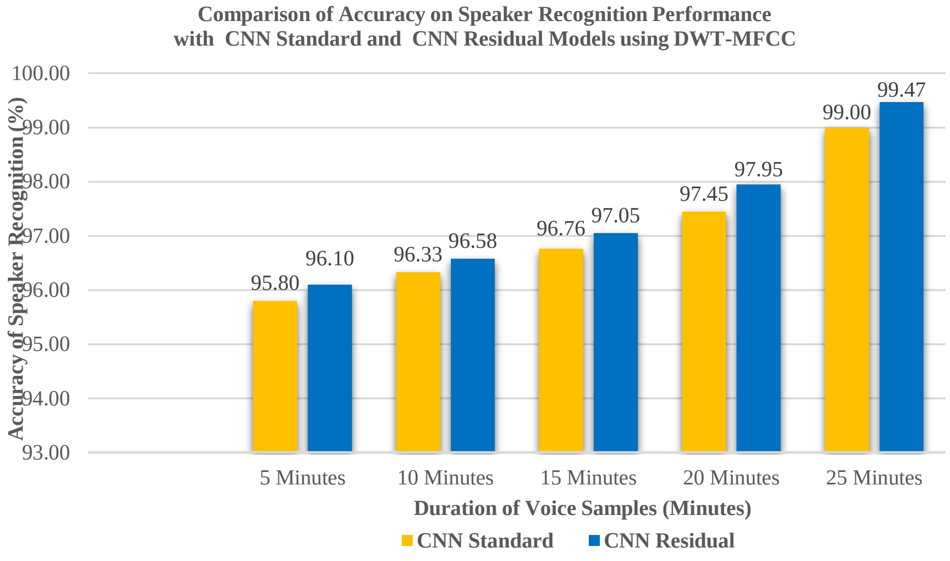

Fig. 8. Comparison of Accuracy on Speaker Recognition Performance with CNN Standard and CNN Residual using DWT-MFCC.

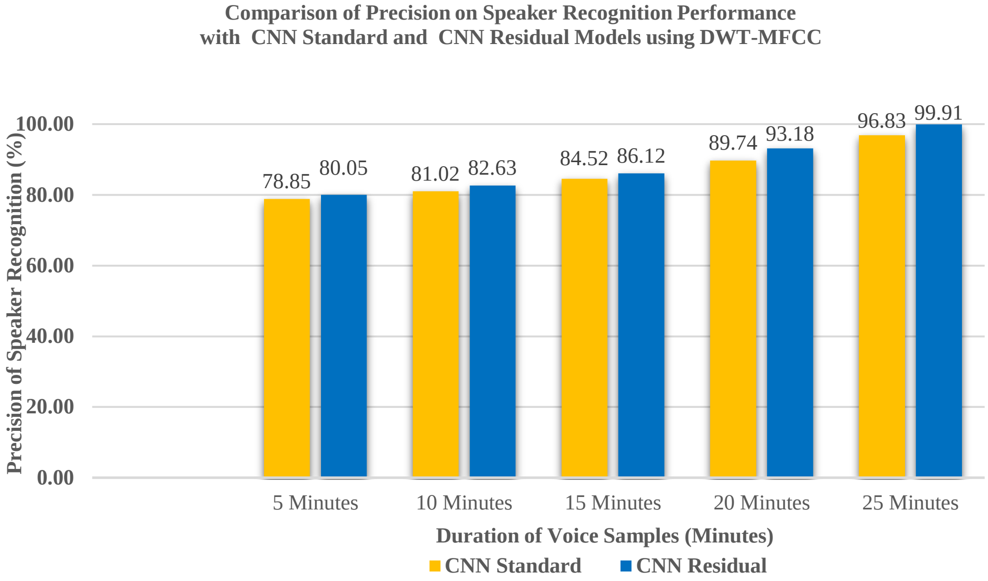

Fig. 9. Comparison of Precision on Speaker Recognition Performance with CNN Standard and CNN Residual using DWT-MFCC.

This test uses a speech content of keyword "Open Access", spoke by the Indonesian users. If the statement is match or correct (True), it will be accepted, while if the speech is wrong or unclear (False), then it is rejected.

Fig. 10 shows that speech recognition performance test has been carried out with the CNN Standard and CNN Residual. It was tested by 20 voice pronunciations, with a total of 10 VB users saying "Open Access". The results show that the percentage of Speech Recognition accuracy performance obtains the best results with the highest percentage value in the 100% CNN Residual, which is higher than the 95% CNN Standard.

The testing has signified that CNN Residual model is better than the CNN Standard. Optimizing the CNN Residual model can improve the validation performance of voice biometric training accuracy, speaker recognition accuracy, and speech recognition accuracy. This is because the CNN Residual model can simplify the training and validation process, as well as increase accuracy in voice biometric classification.

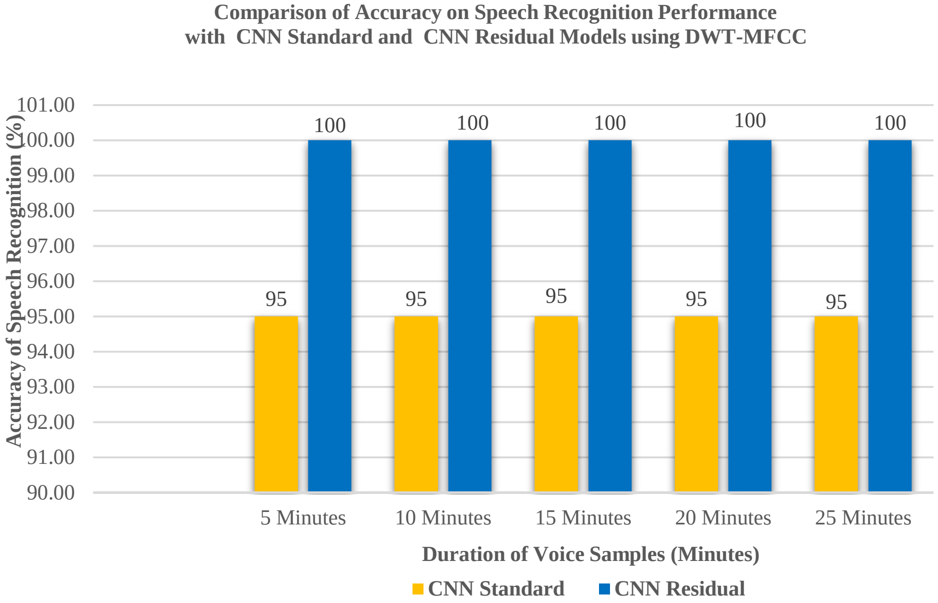

Fig. 10. Comparison of Accuracy of Speech Recognition Performance on Standard CNN Model Algorithm and CNN Residual using DWT-MFCC.

C. Analysis of Time Process for Training

The performance testing of training process time on voice biometrics with CNN Model Algorithm is to test the performance of training process time on the voice biometrics with Algorithm CNN Residual Model using DWT-MFCC, (compared to CNN Standard). This test is to determine how long the processing time is needed for 40 epochs in each running of voice biometrics training on CNN Standard and CNN Residual on voice sample durations of 5, 10, 15, 20 and 25 minutes.

From Fig. 11, the comparison of the performance testing of the voice biometrics training process shows that the CNN Standard training process time performance results are faster than the CNN Residual training process time. This happens because the total number of parameters and the parameter size of the CNN Residual Model is more than the Standard CNN Model, so it requires a longer processing time, with a time difference of 0.03 – 1.28 seconds. It can be implied that the more training time and the more voice sample files are performed, it will result a higher level of accuracy in prediction. It is also indicated that the larger the file duration, the higher the processing time but with a not too big difference.

By analyzing the performance of training process on voice biometrics for a sample duration of 5, 10, 15, 20, and 25 minutes, it can be signified that the accuracy value are consistently above 95%. Accordingly, it can be concluded that by only using the sample of 5 minutes, the voice biometrics system can recognize and identify the speaker with a decent performance.

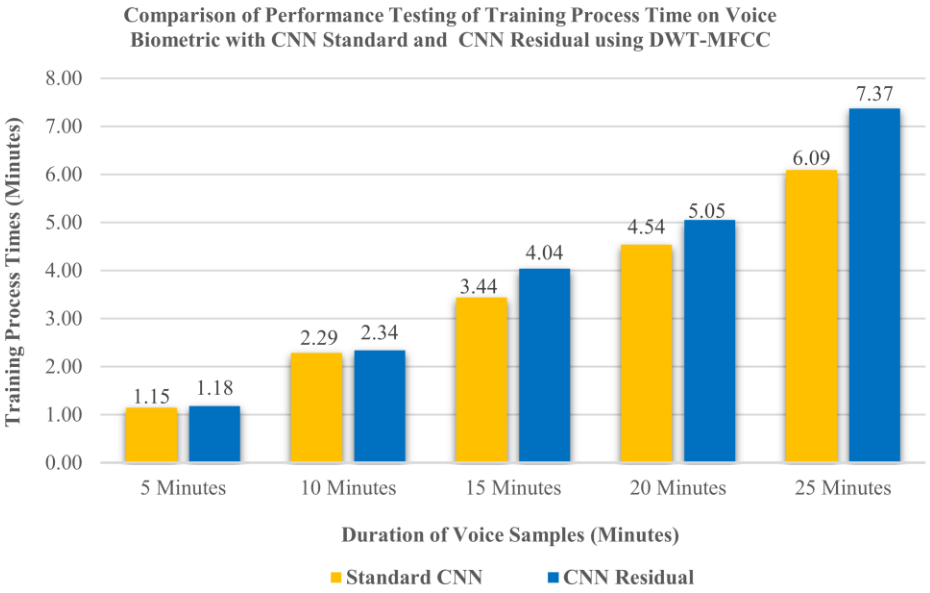

Fig. 11. Comparison of Performance Testing of Training Process Time on Voice Biometric with CNN Standard and CNN Residual Model using DWT-MFCC.

V. Conclusion

This paper has developed a Voice Biometric research model for Indonesian language speaker using the CNN Residual Deep Learning algorithm, which uses Hybrid Feature Extraction DWT-MFCC. Testing is done by comparing the model with the CNN Standard. In this study, a voice dataset was created with 10 users (VB0 – VB9). Each VB is a unique speakers who speak in Indonesian language, resulting a total number of 15,000 voice samples with a voice sample duration of 5, 10, 15, 20, and 25 minutes.

The testing was conducted in the phase of speaker recognition and speech recognition. For the speaker recognition phase, the CNN Residual model has obtained the best results with the highest percentage value of 99.91% precision and 99.47% accuracy at a voice sample duration of 25 minutes, compared to Standard CNN of 96.83% precision and 99.00% accuracy. For the speech recognition phase, CNN Residual has achieved the best results of accuracy which is 100% accurate in 20 trials, while Standard CNN only gave 95% accurate results.

From the results of performance testing of training process time for a sample duration of 5, 10, 15, 20, and 25 minutes, the accuracy value has been consistently above 95%. It can be implied that by only using 5 minutes voice data set, this developed system is able to recognize who is the speaker as well as to identify what keywords are spoken.

Optimizing the CNN Residual model can improve the validation performance of voice biometric training accuracy, speaker recognition and speech recognition accuracy as well as its precision. However, CNN Residual is slightly slower than the CNN Standard, with a time difference of 0.03 – 1.28 seconds.

It can be concluded that the performance of the CNN Residual model provides better results for its accuracy and precision. This research is expected to assist in developing a new model that is able to apply an accurate and efficient individual voice identification and authentication algorithm for voice biometrics systems for security and privacy systems to access sensitive data in banking transactions.

ACKNOWLEDGMENT

Haris Isyanto is in PhD program funded by Beasiswa Pendidikan Pascasarjana Dalam Negeri (BPPDN) Ministry of Education and Culture Republic of Indonesia. Dr. Muhammad Suryanegara is main supervisor, and Dr. Ajib Setyo Arifin is co-supervisor as well as the corresponding author. The voice data set is built by the support of Electrical Engineering - Faculty of Engineering, Universitas Muhammadiyah Jakarta. This publication is supported by Research Grant Universitas Indonesia.

REFERENCES

[1] Z. Rui and Z. Yan, "A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification," IEEE Access, vol. 7, pp. 5994-6009, 2019, doi: 10.1109/ACCESS.2018.2889996.

[2] S. K. Choudhary and A. K. Naik, "Multimodal Biometric Authentication with Secured Templates — A Review," in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 23-25 April 2019 2019, pp. 1062-1069, doi: 10.1109/ICOEI.2019.8862563.

[3] S. Safavi, H. Gan, I. Mporas, and R. Sotudeh, "Fraud Detection in Voice-Based Identity Authentication Applications and Services," in 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 12-15 Dec. 2016 2016, pp. 1074-1081, doi: 10.1109/I CDMW.2016.0155.

[4] N. A. Kulkarni and L. J. Sankpal, "Efficient Approach Determination for Fake Biometric Detection," in 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), 17-18 Aug. 2017 2017, pp. 1-4, doi: 10.1109/ICCUBEA.2017.8463715.

[5] R. Devi and P. Sujatha, "A study on biometric and multi-modal biometric system modules, applications, techniques and challenges," in 2017 Conference on Emerging Devices and Smart Systems (ICEDSS), 3-4 March 2017 2017, pp. 267-271, doi: 10.1109/ICEDSS. 2017.8073691.

[6] A. Tyagi, Ipsita, R. Simon, and S. K. khatri, "Security Enhancement through IRIS and Biometric Recognition in ATM," in 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 21-22 Nov. 2019 2019, pp. 51-54, doi: 10.1109/ISCON47742.2019.9036156.

[7] C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang, and W. Xu, "The Feasibility of Injecting Inaudible Voice Commands to Voice Assistants," IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 3, pp. 1108-1124, 2021, doi: 10.1109/TDSC.2019.2906165.

[8] S. Sachdev, J. Macwan, C. Patel, and N. Doshi, "Voice-Controlled Autonomous Vehicle Using IoT," Procedia Computer Science, vol. 160, pp. 712-717, 2019/01/01/ 2019, doi: https://doi.org/10.1016/j.procs. 2019.11.022.

[9] 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), 20-20 Feb. 2020 2020, pp. 1-6, doi: 10.1109/ICoSTA48221.2020.1570613925.

[10] A. Boles and P. Rad, "Voice biometrics: Deep learning-based voiceprint authentication system," in 2017 12th System of Systems Engineering Conference (SoSE), 18-21 June 2017 2017, pp. 1-6, doi: 10.1109/SYSOSE.2017.7994971.

[11] S. Duraibi, F. Sheldon, and W. Alhamdani, "Voice Biometric Identity Authentication Model for IoT Devices," International Journal of Security, Privacy and Trust Management, vol. 9, pp. 1-10, 05/31 2020, doi: 10.5121/ijsptm.2020.9201.

[12] A. Kamalu, A. Raji, and V. I. Nnebedum, "Identity Authentication using Voice Biometrics Technique U," 2015.

[13] C. Supeshala, "Speaker Recognition using Voice Biometrics," 08/28 2017.

[14] J. Wang, Y. Zheng, M. Wang, Q. Shen, and J. Huang, "Object-Scale Adaptive Convolutional Neural Networks for High-Spatial Resolution Remote Sensing Image Classification," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 283-299, 2021, doi: 10.1109/JSTARS.2020.3041859.

[15] P. Fang and Y. Shi, "Small Object Detection Using Context Information Fusion in Faster R-CNN," in 2018 IEEE 4th International Conference on Computer and Communications (ICCC), 7-10 Dec. 2018 2018, pp. 1537-1540, doi: 10.1109/CompComm.2018.8780579.

[16] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation With Noisy Labels," IEEE Transactions on Information Forensics and Security, vol. 13, no. 11, pp. 2884-2896, 2018, doi: 10.1109/TIFS.2018.2833032.

[17] S. Hourri, N. S. Nikolov, and J. Kharroubi, "Convolutional neural network vectors for speaker recognition," International Journal of Speech Technology, vol. 24, no. 2, pp. 389-400, 2021/06/01 2021, doi: 10.1007/s10772-021-09795-2.

[18] H. Salehghaffari, "Speaker Verification using Convolutional Neural Networks," 03/14 2018.

[19] M. Wang, T. Sirlapu, A. Kwasniewska, M. Szankin, M. Bartscherer, and R. Nicolas, "Speaker Recognition Using Convolutional Neural Network with Minimal Training Data for Smart Home Solutions," in 2018 11th International Conference on Human System Interaction (HSI), 4-6 July 2018 2018, pp. 139-145, doi: 10.1109/HSI.2018.8431363.

[20] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533-1545, 2014, doi: 10.1109/TASLP.2014.2339736.

[21] A. Alsobhani, H. Alabboodi, and H. Mahdi, "Speech Recognition using Convolution Deep Neural Networks," Journal of Physics: Conference Series, vol. 1973, p. 012166, 08/01 2021, doi: 10.1088/1742-6596/1973/1/012166.

[22] N. Dimmita and P. Siddaiah, "Speech Recognition Using Convolutional Neural Networks," International Journal of Engineering and Technology(UAE), vol. 7, pp. 133-137, 09/25 2018, doi: 10.14419/ijet.v7i4.6.20449.

[23] S. T. Seydi, M. Hasanlou, M. Amani, and W. Huang, "Oil Spill Detection Based on Multiscale Multidimensional Residual CNN for Optical Remote Sensing Imagery," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 10941-10952, 2021, doi: 10.1109/JSTARS.2021.3123163.

[24] E. Ihsanto, K. Ramli, D. Sudiana, and T. S. Gunawan, "Fast and Accurate Algorithm for ECG Authentication Using Residual Depthwise Separable Convolutional Neural Networks," Applied Sciences, vol. 10, no. 9, p. 3304, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/9/3304.

[25] Y. Cengiz and Y. D. U. Arıöz, "An Application for speech denoising using Discrete wavelet transform," in 2016 20th National Biomedical Engineering Meeting (BIYOMUT), 3-5 Nov. 2016 2016, pp. 1-4, doi: 10.1109/BIYOMUT.2016.7849377.

[26] M. F. Pouyani, M. Vali, and M. A. Ghasemi, "Lung sound signal denoising using discrete wavelet transform and artificial neural network," Biomedical Signal Processing and Control, vol. 72, p. 103329, 2022/02/01/ 2022, doi: https://doi.org/10.1016/j.bspc.2021.103329.

[27] S.-Y. Jung, C.-H. Liao, Y.-S. Wu, S.-M. Yuan, and C.-T. Sun, "Efficiently Classifying Lung Sounds through Depthwise Separable CNN Models with Fused STFT and MFCC Features," Diagnostics, vol. 11, no. 4, p. 732, 2021. [Online]. Available: https://www.mdpi.com/ 2075-4418/11/4/732.

[28] A. Antony and R. Gopikakumari, "Speaker identification based on combination of MFCC and UMRT based features," Procedia Computer Science, vol. 143, pp. 250-257, 2018/01/01/ 2018, doi: https://doi.org/10.1016/j.procs.2018.10.393.

[29] K. Khotimah et al., "Validation of Voice Recognition in Various Google Voice Languages using Voice Recognition Module V3 Based on Microcontroller," in 2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE), 3-4 Oct. 2020 2020, pp. 1-6, doi: 10.1109/ICVEE50212.2020.9243184.

[30] S. Fegade, D. Chaturvedi, and D. Agarwal, "Voice Recognition Technology : A Review," International Journal of Advanced Research in Science, Communication and Technology, pp. 31-34, 08/03 2021, doi: 10.48175/IJARSCT-1807.

[31] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," Neural Networks, vol. 140, pp. 65-99, 2021/08/01/ 2021, doi: https://doi.org/10.1016/j.neunet.2021.03.004.

[32] I. Bisio, C. Garibotto, A. Grattarola, F. Lavagetto, and A. Sciarrone, "Smart and Robust Speaker Recognition for Context-Aware In-Vehicle Applications," IEEE Transactions on Vehicular Technology, vol. 67, no. 9, pp. 8808-8821, 2018, doi: 10.1109/TVT.2018.2849577.

[33] Đ. T. Grozdić and S. T. Jovičić, "Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2313-2322, 2017, doi: 10.1109/TASLP.2017.2738559.

[34] 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), 21-23 Oct. 2020 2020, pp. 640-645, doi: 10.1109/ICTC49870.2020.9289160.

[35] L. Moreno, "The Voice Biometrics Based on Pitch Replication," International Journal for Innovation Education and Research, vol. 6, pp. 351-358, 10/31 2018, doi: 10.31686/ijier.Vol6.Iss10.1201.

[36] S. Duraibi, F. T. Sheldon, and W. Alhamdani, "Voice Biometric Identity Authentication Model for IoT Devices," International Journal of Security, Privacy and Trust Management, vol. 9, pp. 1-10, 05/31 2020, doi: 10.5121/ijsptm.2020.9201.

[37] S. al, "Voice Biometric: A Novel and Realistic Approach," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, pp. 5684-5694, 04/10 2021, doi: 10.17762/turcomat.v12i3.2243.

[38] X. Zhang, Q. Xiong, Y. Dai, and X. Xu, "Voice Biometric Identity Authentication System Based on Android Smart Phone," in 2018 IEEE 4th International Conference on Computer and Communications (ICCC), 7-10 Dec. 2018 2018, pp. 1440-1444, doi: 10.1109/CompComm.2018. 8780990.

[39] A. Chowdhury and A. Ross, "Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals," IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1616-1629, 2020, doi: 10.1109/TIFS.2019.2941773.

[40] J. Gomes and M. El-Sharkawy, i-Vector Algorithm with Gaussian Mixture Model for Efficient Speech Emotion Recognition. 2015, pp. 476-480.

[41] D. Hoesen, C. Satriawan, D. Lestari, and M. L. Khodra, "Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models," Procedia Computer Science, vol. 81, pp. 167-173, 12/31 2016, doi: 10.1016/j.procs.2016.04.045.

[42] S. Gholamdokht-Firooz, F. Almasganj, and Y. Shekofteh, "Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals," Computers & Electrical Engineering, vol. 58, pp. 215-226, 02/01 2017, doi: 10.1016/j.compeleceng.2016.07.006.

[43] Q. Liu, Z. Chen, H. Li, M. Huang, Y. Lu, and K. Yu, "Modular End-to-End Automatic Speech Recognition Framework for Acoustic-to-Word Model," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. PP, pp. 1-1, 07/15 2020, doi: 10.1109/TASLP.2020. 3009477.

[44] Y.-f. Liao and Y.-R. Wang, Some Experiences on Applying Deep Learning to Speech Signal and Natural Language Processing. 2018, pp. 83-94.

[45] M. Elmahdy and A. Morsy, Subvocal Speech Recognition via Close-Talk Microphone and Surface Electromyogram Using Deep Learning. 2017, pp. 165-168.

[46] Z. Ma, Y. Liu, X. Liu, J. Ma, and F. Li, "Privacy-Preserving Outsourced Speech Recognition for Smart IoT Devices," IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8406-8420, 2019, doi: 10.1109/JIOT.2019. 2917933.

[47] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 Conversational Speech Recognition System," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 15-20 April 2018 2018, pp. 5934-5938, doi: 10.1109/ICASSP.2018.8461870.

[48] N. T. Babu, A. Aravind, A. Rakesh, M. Jahzan, R. D. Prabha, and M. Ramalinga Viswanathan, "Automatic fault classification for journal bearings using ANN and DNN," Archives of Acoustics, vol. 43, pp. 727-738, 01/01 2018, doi: 10.24425/aoa.2018.125166.

[49] O. I. Abiodun et al., "Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition," IEEE Access, vol. 7, pp. 158820-158846, 2019, doi: 10.1109/ACCESS.2019.2945545.

[50] S. A. Syed, M. Rashid, S. Hussain, and H. Zahid, "Comparative Analysis of CNN and RNN for Voice Pathology Detection," BioMed Research International, vol. 2021, p. 6635964, 2021/04/15 2021, doi: 10.1155/2021/6635964.

[51] I. Banerjee et al., "Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification," Artificial Intelligence in Medicine, vol. 97, pp. 79-88, 2019/06/01/ 2019, doi: https://doi.org/10.1016/j. artmed.2018.11.004.

[52] C. Zhao, J. Han, and X. Xu, "CNN and RNN Based Neural Networks for Action Recognition," Journal of Physics: Conference Series, vol. 1087, p. 062013, 09/01 2018, doi: 10.1088/1742-6596/1087/6/062013.

[53] D. Chauhan et al., "Comparison of machine learning and deep learning for view identification from cardiac magnetic resonance images," Clinical Imaging, vol. 82, pp. 121-126, 2022/02/01/ 2022, doi: https://doi.org/10.1016/j.clinimag.2021.11.013.

[54] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," Electronic Markets, vol. 31, no. 3, pp. 685-695, 2021/09/01 2021, doi: 10.1007/s12525-021-00475-2.

[55] P. K. Nayana, D. Mathew, and A. Thomas, "Comparison of Text Independent Speaker Identification Systems using GMM and i-Vector Methods," Procedia Computer Science, vol. 115, pp. 47-54, 12/31 2017, doi: 10.1016/j.procs.2017.09.075.

[56] A. Poddar, M. Sahidullah, and G. Saha, "Speaker Verification with Short Utterances: A Review of Challenges, Trends and Opportunities," IET Biometrics, vol. 7, 10/03 2017, doi: 10.1049/iet-bmt.2017.0065.

[57] J. Zhong, W. Hu, F. Soong, and H. Meng, DNN i-Vector Speaker Verification with Short, Text-Constrained Test Utterances. 2017, pp. 1507-1511.

[58] S. Tantisatirapong, C. Prasoproek, and M. Phothisonothai, "Comparison of Feature Extraction for Accent Dependent Thai Speech Recognition System," in 2018 IEEE Seventh International Conference on Communications and Electronics (ICCE), 18-20 July 2018 2018, pp. 322-325, doi: 10.1109/CCE.2018.8465705.

[59] A. Winursito, R. Hidayat, and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," in 2018 International Conference on Information and Communications Technology (ICOIACT), 6-7 March 2018 2018, pp. 379-383, doi: 10.1109/ICOIACT.2018.8350748.

[60] N. Chauhan, T. Isshiki, and D. Li, "Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database," in 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 23-25 Feb. 2019 2019, pp. 130-133, doi: 10.1109/CCOMS.2019.8821751.

[61] U. Bhattacharjee, S. Gogoi, and R. Sharma, "A statistical analysis on the impact of noise on MFCC features for speech recognition," in 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 23-25 Dec. 2016 2016, pp. 1-5, doi: 10.1109/ICRAIE.2016.7939548.

[62] A. K. H. Al-Ali, D. Dean, B. Senadji, V. Chandran, and G. R. Naik, "Enhanced Forensic Speaker Verification Using a Combination of DWT and MFCC Feature Warping in the Presence of Noise and Reverberation Conditions," IEEE Access, vol. 5, pp. 15400-15413, 2017, doi: 10.1109/ACCESS.2017.2728801.

[63] M. Abdalla, H. Abobakr, and T. Gaafar, "DWT and MFCCs based Feature Extraction Methods for Isolated Word Recognition," International Journal of Computer Applications, vol. 69, pp. 21-25, 05/17 2013, doi: 10.5120/12087-8165.

[64] N. Mukherjee, A. Chattopadhyaya, S. Chattopadhyay, and S. Sengupta, "Discrete-Wavelet-Transform and Stockwell-Transform-Based Statistical Parameters Estimation for Fault Analysis in Grid-Connected Wind Power System," IEEE Systems Journal, vol. 14, no. 3, pp. 4320-4328, 2020, doi: 10.1109/JSYST.2020.2984132.

[65] M. Ogawa and Y. Yang, "Residual-Network -Based Deep Learning for Parkinson's Disease Classification using Vocal Datasets," in 2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech), 9-11 March 2021 2021, pp. 275-277, doi: 10.1109/LifeTech52111 .2021.9391925.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 June 2016 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[67] P. Shih, J. Wang, H. Lee, H. Kai, H. Kao, and Y. Lin, "Notice of Violation of IEEE Publication Principles: Acoustic and Phoneme Modeling Based on Confusion Matrix for Ubiquitous Mixed-Language Speech Recognition," in 2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (sutc 2008), 11-13 June 2008 2008, pp. 500-506, doi: 10.1109/SUTC.2008.78.

[68] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Information Sciences, vol. 507, pp. 772-794, 2020/01/01/ 2020, doi: https://doi.org/10.1016/j.ins. 2019.06.064.