Fast and Accurate Voice Biometrics with Deep Learning Algorithm of CNN Depthwise Separable Convolution Model and Fusion of DWT-MFCC Methods

Haris Isyanto, Ajib Setyo Arifin, Muhammad Suryanegara
Dept. of Electrical Engineering, Universitas Indonesia, Pondok Cina Beji, Depok, 16424, Indonesia

ARTICLE INFO

Article history:

Received July 23, 2022
Revised August 29, 2022
Published September 14, 2022

Keywords:

Voice Biometrics;
Security;
Artificial Intelligence;
Deep Learning;
CNN;
DWT-MFCC

ABSTRACT

Theft of private data became a threat to crime in cyberspace. This issue was in line with the rapid development of data technology, especially online transactions. To attenuate this problem, voice biometrics was developed as an answer to keep up security identity. This paper develops the voice biometric framework based on the Convolutional Neural Network (CNN) Depthwise Separable Convolution (DSC) model and the fusion of Discrete Wavelet Transform (DWT) and Mel Frequency Cepstral Coefficients (MFCC). Such a scheme has targeted to increase the high accuracy, reduce the burden of high computational costs, and speed up the performance of classification process time. We conduct three testing performances, i.e., Voice Biometric Training Performance, Speaker Recognition Performance (" Who is speaking?"), and Speech Recognition performance (" What keyword is uttered?"). For each of the testing, the results are compared with CNN Standard performance. The training results have shown that the CNN DSC model has reduced the number of training parameters to 364,506, leading to accelerating the performance of the training process time to 5.12 minutes. The results of speaker recognition performance have attained the best performance with an accuracy of 99.25%, precision of 97.14%, recall of 98.17%, and F1-score of 97.28%. The results of speech recognition performance have been able to improve the best performance with an accuracy of 100%. It can be concluded that CNN DSC has outperformed the CNN Standard. The contribution of this paper is to provide a new voice biometric framework that works on the basis of the new model that is lighter on its computational load and better in its accuracy performance. The framework can be applied for the identification and verification of user voices accurately, quickly, and efficiently for any applications requiring better security performance.

This work is licensed under a Creative Commons Attribution-Share Alike 4.0

Corresponding Author:

Haris Isyanto, Department of Electrical Engineering, Universitas Indonesia, Pondok Cina Beji, Depok, 16424, Indonesia
Email: haris.isyanto@ui.ac.id, haris.isyanto@ftumj.ac.id

1. INTRODUCTION

There is an increasing problem of stolen private data, which could be a threat to crime in cyberspace. Such a phenomenon is in line with the faster growth of increasingly sophisticated digital and computerized information, especially online transaction activities through the internet network. To overcome this criminal problem, a biometric method was developed to identify individual-supported biological characteristics by utilizing each individual's human body [1, 2]. The biometric method is also an alternative that will be developed for security access applications.

The development of voice biometric technology is an answer to make sure the protection of individuals to avoid theft or fraud of private data, personalize and keep up security, prevent fraud and protect the privacy of one's identity. This biometric method functions to enhance the security classes and user authentication methods. Voice biometrics uses the characteristics of biological and unique patterns of voice to identify the characteristics of somebody's voice-supported spoken voice input. Humans can recognize an individual through his voice, like the identity of the speaker, speech style, accent, and accent [3, 4]. Voice biometrics uses voice commands in sending voice messages through laptops and smartphones. The device will receive the user's voice command. Then the incoming voice command is verified to be matched with the voice contained within the database [5, 6].

This voice biometrics technology operates with 2 (two) security process systems [7], namely the primary speaker recognition security process by verifying "Who's talking?" [8, 9]. And also the second is the speech recognition security process by verifying "What was said?" [10, 11]. If both system accesses are successfully opened, the verification result will be accepted. If both access systems fail, the verification result will be rejected. Voice biometrics was chosen because this method is safe and reliable within the process of voice identification and authentication, accurate and simple to work in identifying someone. The implementation of voice biometrics does not require special devices, and the cost is less than the biometric methods for fingerprint readers or retina scanners [12].

This paper develops the voice biometric framework based on the Convolutional Neural Network Depthwise Separable Convolution (DSC) model and the fusion of Discrete Wavelet Transform (DWT) and Mel Frequency Cepstral Coefficients (MFCC). DSC can significantly solve the problem of high computational costs and speed up the prediction processing time performance in voice biometric classification compared to CNN Standard. This advantage is because DSC operates by reducing the training parameters and reducing arithmetic operations on the operation of convolution, and reducing the computational training burden. Further, we use the fusion of the DWT and MFCC feature extractions to obtain its performance on the signal denoising performance and recognize the feature extraction of somebody's voice.

Some previous research on the theme of voice biometrics has discussed voice biometrics using machine learning methods, such as voice biometrics with machine learning k-Nearest Neighbors (k-NN) [13], voice biometrics with machine learning SVM, and feature extraction MFCC [12, 14] and voice biometrics with GMM machine learning and feature extraction MFCC [15]. The use of machine learning and MFCC feature extraction has given an accuracy performance of about 76% [12-15]. Such a performance is because machine learning methods have limitations that can only process small amounts of data and are less able to process complex data. In this paper, the proposed voice biometrics framework is utilizing the deep learning method of CNN-DSC. Thus, the testing is expected to provide an accuracy performance to be more than 90%. We implement three testing performances, i.e., Voice Biometric Training Performance, Speaker Recognition Performance, and Speech Recognition Performance.

The contribution of this paper is to provide a new voice biometric framework that works on the basis of a new model combining the CNN-DSC and DWT-MFCC. It eventually brings 2 significant research contributions a new biometric framework is lighter on its computational load and better on its accuracy performance. Further, an applied benefit of this research is that Voice Biometrics using the CNN DSC model and the DWT-MFCC method can be applied to the algorithm of identification and verification/authentication of the user's voice accurately, quickly, and efficiently for access to banking security systems. This research is also part of our work on developing voice biometrics for Indonesian language users. In [16], we have reported the algorithm development of Deep Learning CNN Residual and hybrid MFCC-DWT and compared its testing results with CNN Standard.

The remainder of the paper presents the theories and relevant studies in Section 2. Section 3 presents the Method, focusing on the proposed framework model development of Voice Biometrics using CNN DSC and MFCC- DWT. Section 4 elaborates on the testing results and discussion, while Section 5 concludes the paper.

2. THEORIES AND RELEVANT STUDIES

2.1 The Main concept of Voice Biometrics Systems

Voice biometrics technology is an advanced technology developed for data access security. Voice biometrics is applied to the recognition process of a user's voice pattern based on the biological characteristics of each individual. This is done by verifying the user's voice identity data correctly. Voice biometrics is divided into 2 (two) processes, including user identification and user verification [12], as seen in Fig. 1.

User identification is a process of recognizing a person's voice identity by registering a person's voice identity in the voice database. This user identification starts from the process of capturing voice from the user's voice as input, which consists of the speaker and speech. This voice capture process is able to capture voice characteristics that are important in the speech recognition process and is able to capture important information contained in the voice signal. The incoming voice input is still in analog form and is processed to be converted into digital. Furthermore, the audio feature extraction process is a process that can recognize a person's voice pattern from the user's audio feature extraction. The voice patterns template process is the process of recognizing a person's voice identification by registering a person's voice identity, which is the unique character of each individual. The voice patterns template process registers and stores speaker and speech data in the database [12, 14]. The user identification process is in the first line. This can be shown in Fig. 1.

The user verification process is a process where a person's voice is verified by comparing the new user identification voice and matching it with the voice registered in the voice model database. This user verification process for the beginning of the process is exactly the same as the user identification process. Everyone has a unique and different voice pattern. When someone accesses data in the system, the system will perform a pattern-matching process in which the identity of a person's voice is verified. The system will compare and verify the identity of a person's voice pattern according to the voice samples enrolled and stored in the database. The result is the validation of someone's voice with decision logic (accepted or rejected). The user identification process is in the second line. This can be shown in Fig. 1.

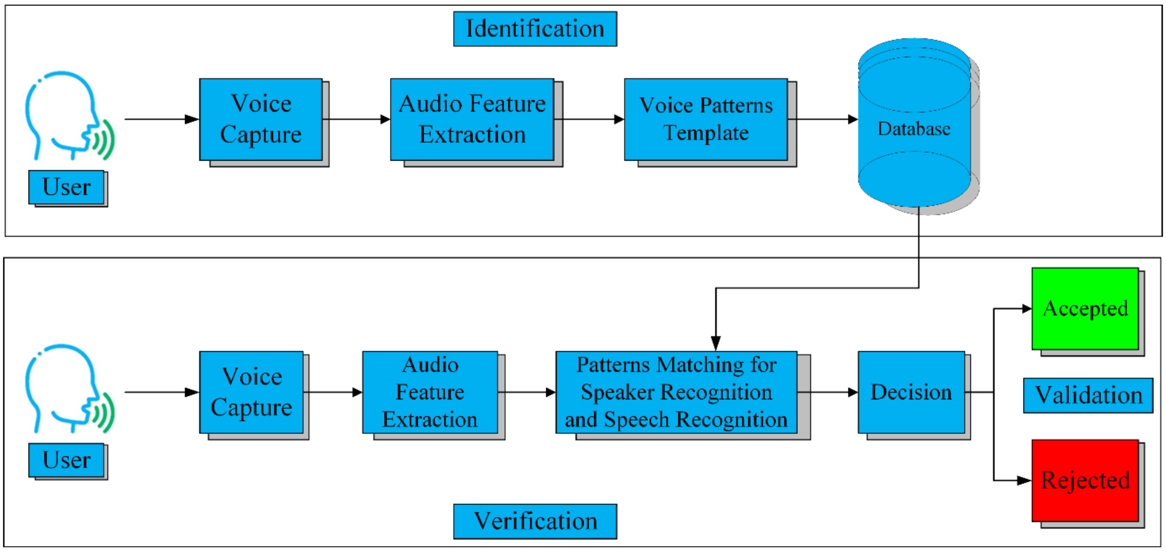


Fig. 1. Block Diagram of the user identification and user verification process

Fig. 1 shows that the basic principle, the most important in the security system of voice biometrics, operates in two stages [7], namely the speaker recognition stage to verify "Who is speaking?" [8, 9], and speech recognition stage to verify "What keyword is uttered?" [10, 11]. If both system accesses are successfully opened, the verification result will be accepted. If it fails, then the result will be rejected. Meanwhile, largely the previous papers examined the two themes separately between speech and speaker recognition.

2.2 Relevant Studies

The Deep Learning algorithm is incredibly effective and makes it easier to identify patterns from the entered object data [9, 17]. CNN is in a position to resolve issues in large data and complicated data for the identification process. This CNN is believed to have this high performance which has been widely used for object identification training and testing. Research to identify voice objects is currently being evolved with the CNN model. The implementation of recognizing a personality's voice in voice biometrics with the CNN model has been developed for speech recognition and speaker recognition. In previous related research, only a few papers discussed the theme of paper voice biometrics by combining speech recognition [18] and speaker recognition [19]. Generally, speech recognition and speaker recognition themes are discussed separately. Previous studies associated with the themes of speech recognition and speaker recognition using machine learning methods are discussed separately, shown in Table 1.

Referring to previous studies that use machine learning and feature extraction methods, the accuracy performance is around 76% [20-23, 31, 32]. The constraints of the machine learning methods for SVM, GMM, and HMM could only process smaller amounts of data, and less could process complicated data [17]. In comparison with other feature extraction, the MFCC method is a feature extraction method with the best results in extracting human voice features. This MFCC could acknowledge the form of the voice from somebody's characteristics and process a voice in a short time and is able to pick only the voices that are needed [37]. However, this MFCC extraction feature has the disadvantage that it is not resistant to noise [38].

Table 1. Related Research

Patterns of Voice	Models of algorithm	Audio Feature Extractions	Descriptions	Ref.
Speech Recognition	Machine Learning (ML) SVM	MFCC	ML can process data only for small amounts of data and cannot process complex data, and MFCC has the problem of not being resistant to noise	[20]
	Machine Learning on GMM	X	The descriptions are the same as above. No Audio Feature Extractions	[21]
	Machine Learning on GMM/HMM	X	The descriptions are the same as above. No Audio Feature Extractions	[22]
	Machine Learning on HMM	X	The description is the same as above. No Audio Feature Extractions	[23]
	Machine Learning on HMM	MFCC	The descriptions are the same as those above	[24]
	Deep Learning on (DL) ANN	MFCC	DL is not reliable for computational capabilities.	[25]
	Deep Learning DNN	MFCC	The descriptions are the same as above.	[26]
	Deep Learning RNN	X	DL RNN features less compatibility than CNN, and RNN has a gradient loss problem. Then No Audio Feature Extractions	[27]
	Deep Learning RNN and LSTM	X	DL RNN features less compatibility than DL CNN and No Audio Feature Extractions	[28]
	Deep Learning CNN	X	The complex structure makes the CNN workload increase. This results in the problem of high computational costs and slow system work. Then No Audio Feature Extractions	[18]
	Deep Learning CNN and LSTM	X	The descriptions are the same as those above	[29]
Speaker Recognition	Machine Learning on GMM	MFCC	ML can process data only for small amounts of data and cannot process complex data. Then MFCC has the problem of not being resistant to noise.	[30]
	Machine Learning on SVM-GMM	X	The descriptions are the same as above. Then No Audio Feature Extractions	[31]
	Machine Learning HMM	X	The descriptions are the same as above. Then No Audio Feature Extractions	[32]
	Deep Learning ANN	LPC, MFCC, ZCR	DL is not reliable for computational capabilities. Then these Feature Extractions have the problem of not being resistant to noise.	[33]
	Deep Learning DNN	X	The descriptions are the same as above. Then No Audio Feature Extractions	[34]
	Deep Learning RNN	X	DL RNN features less compatibility than DL CNN, and DL RNN has a gradient loss problem. Then No Audio Feature Extractions	[35]
	Deep Learning CNN	X	The complex structure makes the CNN workload increase. This results in the problem of high computational costs and slow system work. Then No Audio Feature Extractions.	[19]
	Deep Learning CNN	MFCC	The descriptions are the same as above. Then MFCC has the problem of not being resistant to noise.	[36]
Voice Biometrics	Machine Learning k-NN	X	ML can process data only for small amounts of data and cannot process complex data. Then No Audio Feature Extractions.	[13]
	Machine learning SVM	MFCC	ML can process data only for small amounts of data and cannot process complex data. Then MFCC has the problem of not being resistant to noise.	[12, 14]
	Machine learning GMM	MFCC	The descriptions are the same as above.	[15]
	Deep-learning CNN Depthwise Separable Convolution (DSC) Model	Fusion of DWT-MFCC Methods	Optimizing this CNN DSC can effectively reduce the number of parameters and arithmetic operations in the convolution operation and reduce the computational training burden. So that it can significantly solve the problem of high computational costs and speed up the prediction processing time performance in voice biometrics classification compared to the CNN Standard	Proposed

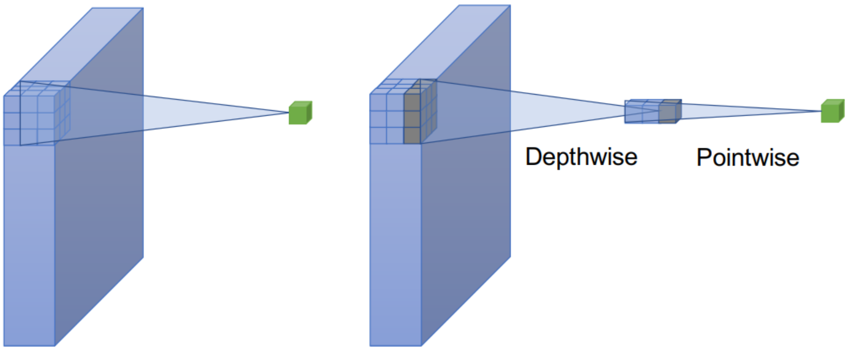
Furthermore, previous studies related to the themes of speech recognition and speaker recognition using deep learning methods also are discussed separately, as shown in Table 1. From the two paper themes that use deep learning and feature extraction methods, the accuracy performance is around 71-90% [25- 28, 33-35].

The restriction of the deep ANN or DNN method is less dependable in computational ability than RNN and CNN [33, 34]. As for this RNN, there's an issue with gradient loss. To overcome this gradient problem, it combines RNN and LSTM. But LSTM has the disadvantage that it requires more computational memory to train [28, 39]. This CNN model is more reliable than the ANN, DNN, and RNN models. The benefits of CNN have dependable computational abilities, high accuracy, can process large data, could automatically operate important features on neural networks which do not require human supervision, then could perform complex data classifications within the process of voice identification [18, 19, 29, 36]**.**

2.3 CNN Depthwise Separable Convolution (DSC)

The deep learning CNN algorithm is in a position to resolve issues in large data and complicated data for the identification process [40]. The basics of CNN, without any modification, the so-called CNN Standard, has succeeded in carrying out its object classification tasks. However, the performance of the CNN standard model will increase along with a large amount of data and complex data structures in the voice identification process. This complex structure makes the CNN Standard workload increase, which results in the problem of high computational costs and slow system work. To overcome this problem, a network framework for voice biometrics technology was designed using Deep Learning with the CNN Depthwise Separable Convolution (DSC) model developed in this research. Continuous improvement and optimization in the development of standard convolution (Conv layer), one of which is the development of Depthwise Separable Convolution (DSC) [41]. Based on previous research, the optimization of the DSC model is believed to effectively solve the problem of reducing the number of training parameters, reducing computational load, speeding up the process, and reducing computational costs, which are higher in convolution operations compared to standard convolution [42].

Another advantage of DSC is that it may compensate for the drawback of CNN Standard, which can only present the channel-wise and spatial-wise calculation processes in one way and uses the same spatial location filter in all channels. The DSC can present the division of calculations in two ways, namely depthwise convolution presents a filter of single convolution in the input of each channel and utilizes features with the same spatial place but with different channels, and pointwise convolution to plots the resulting feature map and performs linear combinations of the output depthwise convolutions on the channel space. A pointwise convolution is a standard convolution that has a convolution kernel size of 1×1 Conv, where the dimensions of the output channel are from the previous layer. Fig. 2 shows a comparison between standard and depthwise separable convolutions [41].



(a) (b)

Fig. 2. (a) Standard Convolution, (b) Depthwise Separable Convolution [41].

DSC is a method that is able to reduce the computational load by making some slight changes to the calculation technique, where the results [43] of the convolution calculations are almost the same as before. There are 2 examples of DSC implementation, namely a single layer and a group of layers (3 layers). Specifically, DSC divides standard convolution calculations into depthwise and pointwise convolutions. For a single layer, the standard convolution calculation is presented, which can be shown in Equation (1), whereby separation, which consists of 2 types of calculations, it can be shown in Equation (2) and Equation (3). For Equation (1), three accumulation layers are implemented, including k, l, and m which are used to provide standard convolution calculations. For Equation (2), one accumulation layer is m, and for Equation (3), two accumulation layers are executed, including k and l. As for the DSC calculation, which is to combine calculations between pointwise convolution and depthwise convolution, it can be shown in Equation (4) [41, 43]. Although DSC is able to reduce the number of arithmetic operations and could reduce the computational load, the results of the feature map calculation the end result is almost the same as the standard convolution process [43, 44].

$$Conv(\omega, y)_{(i,j)} = \sum_{k,l,m}^{K,L,M} \omega_{(k,l,m)} \cdot y_{(i+k,j+l,m)} \tag{1}$$

$$PointwiseConv(\omega, y)_{(i,j)} = \sum_{m}^{M} \omega_m . y_{(i,j,m)} \tag{2}$$

$$DepthwiseConv(\omega, y)_{(i,j)} = \sum_{k,l}^{K,L} \omega_{(k,l)} \odot y_{(i+k,j+l)} \tag{3}$$

$$SepConv(\omega_p, \omega_d, y)_{(i,j)} = PointwiseConv_{(i,j)}\left(\omega_p, DepthwiseConv_{(i,j)}(\omega_d, y)\right) \tag{4}$$

2.4 The Fusion of MFCC and DWT Method for Improving Accuracy Performance of Feature Extraction

As seen in Fig. 1, the audio feature extraction of MFCC has an important role in improving the performance of high accuracy and shortness of voice recognition when compared to other methods [37]. The MFCC is the right method for recognizing the feature extraction of the user's voice and could only perform the necessary voice processing. MFCC feature extraction using a discrete Fourier transform process. This Fourier transform is a transformation that only determines the frequency domain of the signal and does not determine the time domain. The block diagram of the MFCC can be seen in Fig. 3 [45].

The MFCC process starts from the voice signal input process. Pre-emphasis is a filter process that is used to increase higher frequencies and increase the amount of energy in high frequencies in signal processing. Framing and windowing, Framing could process for dividing the voice signal within several frames, which simplifies the calculation and analysis of the voice signal, where each frame consists of several voice samples and their sampling frequency. The length of the frame is affected by the success of the spectral analysis. And the windowing used is windowing hamming. This type of windowing hamming is the most widely used to prevent signal discontinuities in the framing process. Fast Fourier transform (FFT) is a method that is able to perform calculations and solve discrete Fourier transforms quickly for voice recognition and extract voice signals in the frequency domain without losing relevant information, so that voice processing becomes easier. Mel Filterbank is a filter bank for energy measurement that responds to triangular bandpass frequencies in speech recognition which operate at frequencies audible to humans. Mel Filterbank is used to provide good resolution when the frequency is a low but low resolution when the frequency is high. Discrete Cosine Transform (DCT) for processing human voice signal information. DCT is a Mel spectrum that is converted to a time domain to improve voice recognition performance [36, 45].

Discrete Wavelet Transform (DWT) is a wavelet transformation method that is able to remove voice noise to improve voice signal quality in a better voice biometric system. This DWT signal processing is very suitable for use on non-stationary signals or non-periodic signals where the frequency signal varies in time and period of time. The frequency is not constant. Measurements in the wavelet transformation parameters are fixed. It will present information on the signal time range and information in the signal frequency range. Wavelet transform presents the approach for resolution of signal in multi-analysis, then the method could be applied in identifying voice signals. Actually, a change from CWT to DWT wavelet function with this scaling function aims to process signal filtering in Low Pass Filter and High Pass Filter. Signal filtering processing on DWT can be shown in Fig. 4 [46].

3. METHODS

3.1. The Development of Voice Biometrics framework using CNN Depthwise Separable Convolution Model and DWT-MFCC Methods

Fig. 5 depicts the proposed Framework Model Development of Voice Biometrics using CNN Depthwise Separable Convolution and DWT-MFCC methods. It refers mainly to Fig.1, which illustrates the main concept of the voice biometrics system.

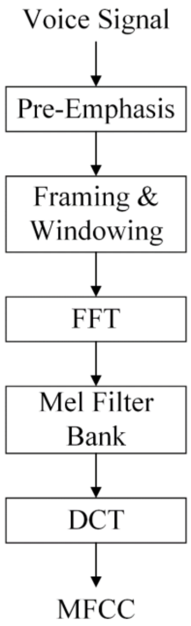


Fig. 3. Block Diagram of MFCC [45]

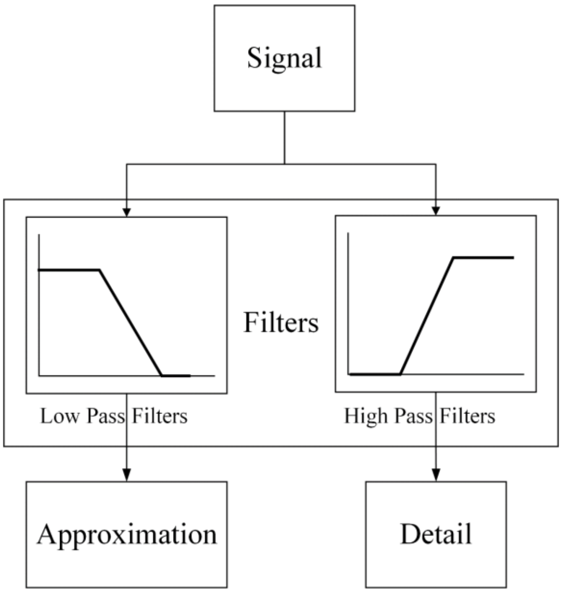


Fig. 4. Signal filtering processing on DWT [46]

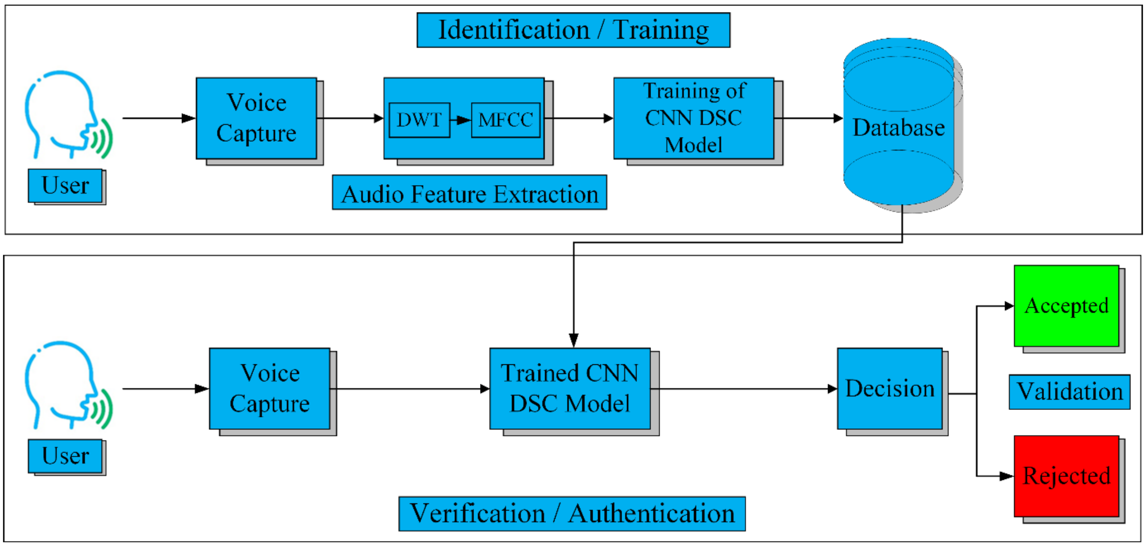


Fig. 5. Voice Biometrics framework using CNN Depthwise Separable Convolution Model and DWT-MFCC Methods

Basically, the block diagram of the voice biometrics study shown in Fig. 5 is divided into 2 (two) processes, i.e., user identification/training and user verification/authentication. The user identification/training process starts with the process of capturing user input voice samples, signifying the uniqueness of the speaker and his/her speech. The sample is processed with the development of the DWT-MFCC fusion method used to carry out the feature extraction process and improve the performance of feature extraction accuracy to be higher.

The MFCC Feature Extraction mentioned in Section 2.4 still has a weakness, which is not resistant to noise in the voice feature extraction process [38]. Based on previous research, there is a gap in improving the performance of extraction features. Thus, this research proposed the fusion of DWT-MFCC, which aims to denoise voice signal performance. With the development of the Fusion of DWT-MFCC, it is hoped that it can solve the noise problem for better voice quality so that reliable voice feature extraction can be created. It will greatly affect the improvement of MFCC's feature extraction accuracy performance which is higher than before [47, 48].

Subsequently, it runs the training of the CNN DSC model to solve the problem of high computational costs and speed up processing time performance. The training process is the ability of the CNN DSC model learning process to be trained in the process of identifying the user's voice using large computing on the GPU and CPU. This training process also performs the registration process for the voice of the user's identity, which will be stored in the database.

Furthermore, the CNN DSC process that has been completed will produce a Trained CNN DSC, and this Trained CNN DSC will be applied to the user verification/authentication process. The voice verification/authentication process will carry out the voice classification process and voice authentication. The user voice verification process can be directly implemented through the Trained CNN DSC, which is used to produce decision outputs. The live voice of a new user will be verified by the system by comparing their identity and matching it with the identity of the user's voice sample registered in the database. Then the system will generate prediction output which aims to determine the success rate of prediction accuracy. The result of voice data can be trained by CNN DSC. To increase the maximum prediction performance, the voice classification optimization was carried out using the Trained CNN DSC Model to get high accuracy. Voice classification will produce a decision output in the user authentication process by validating the data (accepted/rejected) of the voice of the user's identity.

3.2. Model of CNN Depthwise Separable Convolution (DSC) used in this research

In this research, the CNN DSC architecture was designed as shown in Fig. 7 (with Fig. 6 Standard CNN architecture as a comparison). The proposed CNN DSC architecture contains 22 layers, of which there are eight DSC layers, one input layer, three 16 kernel Conv layers, two 16 kernel DSC layers, and one 32/2 conv layer. Kernel, one 32-kernel Conv layer, two 32-kernel DSC layers, one 64/2 kernel Conv layer, one 64-kernel Conv layer, two 64-kernel DSC layers, one 128/2 kernel Conv layer, one 128 kernel Conv layer, two layers DSC 128 kernel, one layer of adaptive avg pool, one layer of flattened, one layer of fully connected and output layer.

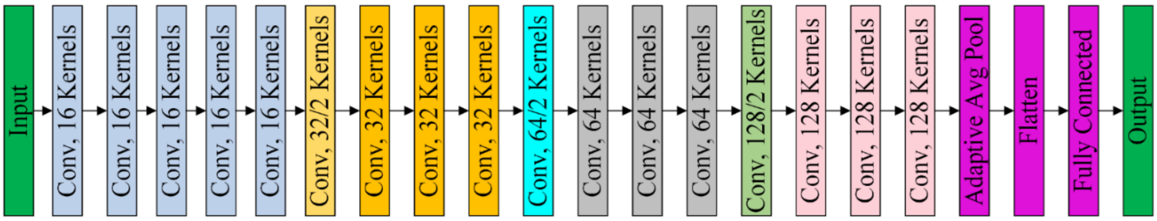


Fig. 6. Architecture of CNN Standard (CNN Without DSC) 707,386 parameters

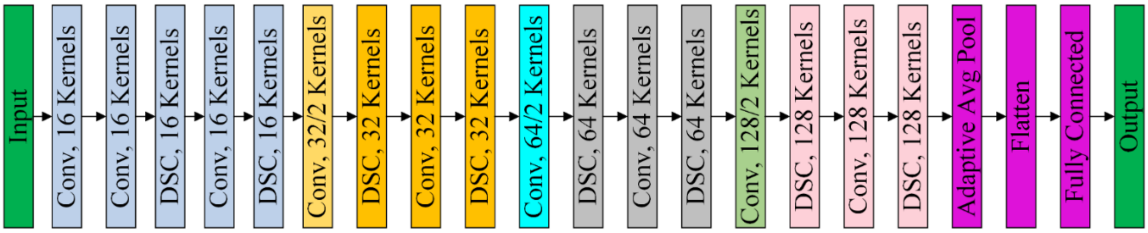


Fig. 7. Architecture of CNN Depthwise Separable Convolution 364,506 parameters

For increasing the performance of CNN Standard, CNN DSC architecture was developed in this study, namely by optimizing the deep learning of the CNN standard model by replacing eight Conv layers with eight DSC layers. In the standard convolution structure on the conv layer in the architecture Fig. 6, in which there is a 3×3 Conv layer process, batch normalization, and activation layer. Furthermore, the conv layer structure was replaced with a DSC layer in the CNN DSC architecture Fig. 7, by developing DSC, especially making the configuration of the convolution layer size smaller to a 1×1 Conv layer, whereas, in the DSC, there is a 3×3 Conv layer process, batch normalization layer, activation layer, 1×1 layer Conv, batch normalization layer and activation layer. From Fig. 7, it can be shown that the number of parameters for the CNN DSC model is 364,506, which is lower than the standard CNN model 707,386 (Fig. 6).

With the development of this CNN DSC model, it is hoped that it can effectively help to solve problems by reducing the number of training parameters and reducing arithmetic on convolution operations, then reducing computational complexity. So significantly, the CNN DSC model is able to reduce the size of network parameters, reduce the number of memory operations, reduce computational load, and can speed up the user voice classification process time. Thus, it is very helpful in reducing the cost of higher computational loads.

3.3. Testing and Voice Data Set for Indonesian Language Speaker

To test the proposed framework, we conduct a performance appraisal for Indonesian language speaker with the mechanism as follow:
1) Testing of Voice Biometric Training Performance
2) Testing of Speaker Recognition Performance (” Who is speaking?”)
3) Testing of Speech Recognition (” What keyword is uttered?”)

Each of the performance testing results of CNN DSC and DWT-MFCC model algorithms is compared with CNN Standard. In this research, each voice sample was tested for 25 minutes. We used the data set of Indonesian language speakers we created in [16]. This voice dataset is trained by the deep learning algorithm of the CNN model. This voice dataset was created by starting with the user sample voice input process on the smartphone microphone, which included 10 users as Voice User0 - Voice User9 (VU0 - VU9). Each voice user fills out a voice sample within 25 minutes with an Indonesian speaker, where each voice user input contains the user's unique voice data, namely speech and speaker. The user's voice dataset will be processed using the extraction feature of the Fusion of DWT-MFCC method for denoising voice signals, recognizing the feature extraction of somebody's voice pattern, but processing only for necessary voice. Then the results of the voice extraction feature can be trained with the CNN deep learning model algorithm.

4. RESULTS AND DISCUSSION

4.1. Testing of Voice Biometric Training Performance

Voice Biometrics Model CNN training is a learning process for the capability of training a user's voice dataset with CNN DSC and DWT-MFCC method (with CNN Standard as a comparison). CNN model could be trained in identifying the voice user dataset. In this study, we tested the performance of voice biometrics with the CNN algorithm, where there were 15,000 voice sample files trained on CNN Standard and CNN DSC Model.

4.1.1 Analysis of Training Parameters Performance for CNN Depthwise Separable Convolution Model

This parameter training performance is for testing performance on the process of parameter training in voice biometrics with CNN DSC and the DWT-MFCC method (with CNN Standard as a comparison). This test is to determine the comparison of training parameters between the CNN DSC model and the CNN Standard model. Results of testing the training parameters for the two CNN models that have been carried out, it is obtained that there are differences in the results of the process in data trainable parameters, then data size parameters, where the results of the training performance on CNN DSC, number of parameters is 364,506, that is less than CNN standard of 707,386. So, the difference in the total number of parameters is almost doubled, with a difference of 342,880 parameters. Based on these performance results will help speed up the parameter training process for the CNN DSC model. Parameters Comparison of CNN DSC and CNN Standard can be shown in Table 2.

Table 2. Parameters Comparison of Training on Voice Biometrics between CNN Depthwise Separable Convolution and CNN Standard

Parameters	CNN DSC	CNN Standard
Total Parameters	364,506	707,386
Trainable Parameters	364,506	707,386
Non-trainable Parameters	0	0
Parameter Size (MB)	1.39	2.70

4.1.2 Analysis of Training Process Times Performance for CNN Depthwise Separable Convolution Model

This training process time test is for testing performance on the time of training process for voice biometrics on CNN DSC and DWT-MFCC method (with CNN Standard as a performance comparison). This test is to compare the length of time required for each training operation process with the voice sample duration of 25 minutes between the CNN DSC model and the CNN Standard model. The results of the training process performance are shown in Table 3.

Based on Table 3 and Fig. 8, it is shown that the comparison of performance on the time of training process for voice biometrics using two CNN models that have been carried out, it is obtained that the performance on the time of training process in CNN DSC is 5.12 minutes faster than CNN Standard 6.39 minutes. This is caused by the computational load of the total parameters, and the parameter size on CNN DSC is smaller than CNN Standard. The comparison results show that CNN DSC time performance of the training process is quicker than CNN Standard, with a time difference of 1.27 minutes. Reducing the number of training parameters, then the faster time of the training process for the CNN DSC model will reduce the amount of computational load on the training process. So, this will have an impact on reducing high computational operational costs compared to the CNN standard model.

Table 3. Performance Comparison of Training Process Times on Voice Biometrics between CNN Depthwise Separable Convolution and CNN Standard Model with the duration of voice samples in 25 minutes.

Number of Experiments	**Training Process Times (Minutes)**	
	CNN DSC	**CNN Standard**
1	5.45	6.51
2	5.16	6.36
3	5.05	6.39
4	5.07	6.37
5	5.09	6.37
6	5.08	6.37
7	5.06	6.38
8	5.07	6.38
9	5.07	6.37
10	5.09	6.37
Averages	**5.12**	**6.39**

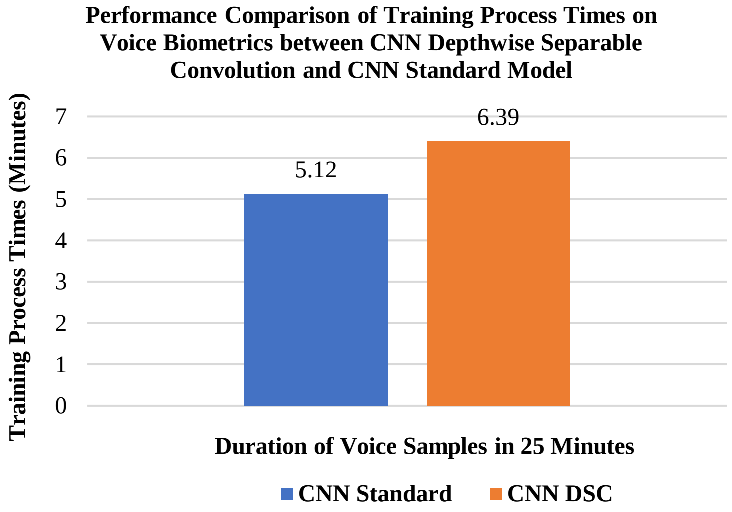


Fig. 8. Performance Comparison of Training Process Times for Voice Biometrics between CNN Depthwise Separable Convolution and CNN Standard Model with a duration of voice samples in 25 minutes.

Analysis of the performance testing on the time of training process for voice biometrics, it can be concluded that the longer the voice sample time in 25 minutes in the training process that are running, this will increase the predictive accuracy value to be higher in recognizing and identifying the user's voice.

4.2. Testing of Speaker Recognition Performance (" Who is speaking?")

Testing of speaker recognition performance uses CNN DSC deep learning and DWT-MFCC method (with CNN Standard as a comparison). To evaluate performance metrics on the CNN model, a confusion matrix is used for the classification algorithm. This confusion matrix could be used to find out the results of the comparison of the classification values between the CNN training classification and the actual classification. Trained CNN could produce for testing of speaker recognition performance on CNN DSC using the Confusion Matrix method. The classification algorithm is used to identify 10 user voices. To find out the best CNN model performance, this confusion matrix needs to be considered in determining the best CNN model by comparing the CNN DSC model and the DWT-MFCC (with the CNN Standard model as a comparison) [49, 50].

In measuring the performance of this CNN model by applying the confusion matrix method, where the confusion matrix consists of 4 predictions that become the results of the classification algorithm, as follows True Positive, False Positive, True Negative, and False Negative. Based on the calculation values, the Accuracy, Recall, Precision, then F1-Score are obtained. The performance of speaker recognition measurement uses CNN DSC (with CNN Standard as a comparison), where the voice sample is 25 minutes long. The results of the comparison of test performance on speaker recognition can be seen in Table 4 and Fig. 9.

Table 4. Comparison of Testing on Speaker Recognition Performance with Duration of Voices Sample 25 Minutes using CNN Standard and CNN Depthwise Separable Convolution with DWT-MFCC

Voice Users	Comparison of Performance Testing on Speaker Recognition (%)							
	Accuracy		Precision		Recall		F1-Score	
	CNN Standard	CNN DSC	CNN Standard	CNN DSC	CNN Standard	CNN DSC	CNN Standard	CNN DSC
VU0	98.33	98.60	97.27	95.85	97.83	97.31	96.83	96.57
VU1	99.27	99.33	96.12	96.84	98.58	97.87	97.35	97.35
VU2	98.60	99.20	95.72	96.88	96.67	97.88	97.46	96.88
VU3	99.73	99.07	99.62	97.34	96.83	97.42	91.54	96.32
VU4	99.27	99.47	98.26	99.46	96.98	97.53	97.07	97.88
VU5	98.67	99.33	98.57	97.4	97.40	97.49	94.92	97.4
VU6	99.13	99.80	97.66	98.46	97.31	99.98	96.57	99.22
VU7	99.67	99.60	97.37	97.94	98.67	98.96	96.46	98.45
VU8	98.80	99.00	93.66	96.34	99.48	97.83	95.50	96.08
VU9	98.50	99.13	94.00	94.91	99.46	99.46	96.85	96.61
Averages	99.00	99.25	96.83	97.14	97.92	98.17	96.06	97.28

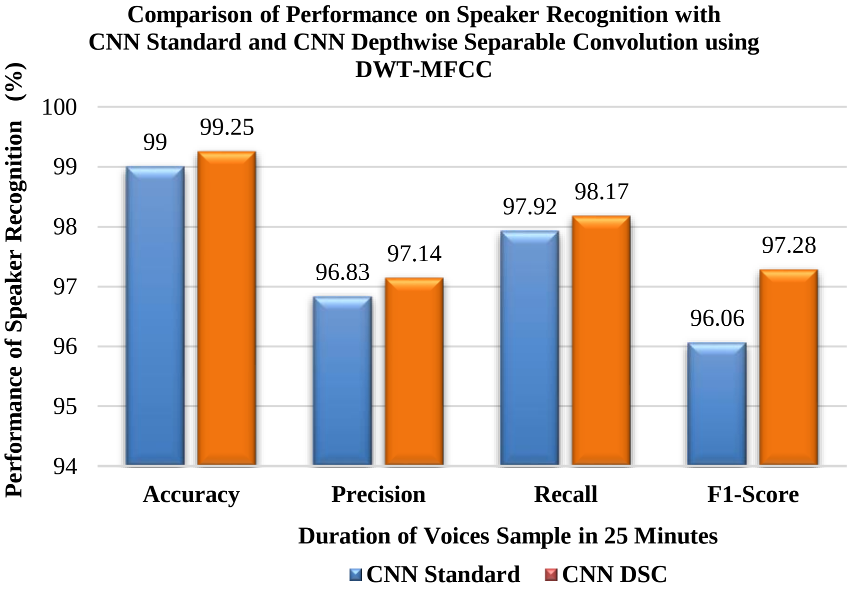


Fig. 9. Comparison of Performance Testing on Speaker Recognition with Duration of Voices Sample 25 Minutes using CNN Standard and CNN Depthwise Separable Convolution with DWT-MFCC

From the comparative analysis of test performance on speaker recognition with the duration of voice sample 25 minutes using the CNN DSC model and the DWT-MFCC method (with CNN Standard as a comparison), the best results obtained are percentage values for Accuracy 99.25%, Precision 97.14%, Recall 98.17% and F1-Score 97.28%. So, with the increasing amount of voice sample files or longer duration of the running voice sample, this will result in a higher percentage value of speaker recognition performance in classification prediction decisions.

By referring back to Table 1, it has been indicated that the accuracy performance of other deep learning models (ANN, DNN and feature extraction MFCC, deep learning RNN with LSTM, deep learning CNN and feature extraction MFCC) is around 71-90% [25-28, 33-35]. Thus, the proposed model has been able to improve performance on the Accuracy of Speaker recognition by more than 90%.

4.3. Testing of Speech Recognition Performance (” What keyword is uttered?”)

Testing of speech recognition performance is for testing the performance accuracy on speech recognition using CNN DSC deep learning and DWT-MFCC method (with CNN Standard as a comparison) with the duration of voice sample 25 minutes. Testing this speech recognition by matching or verifying the spoken keywords (speech content) of the user.

The keyword spoken in this test is "Open Access," where the Indonesian user pronounces it. This test is done by verifying and matching the voice utterances in accordance using voice patterns template data saved in the database. If it is verified that the statement is "true," then it is "accepted." But if the verified statement does not match the words "false," then it is "rejected."

Fig. 10 shows the performance testing of keyword speech on speech recognition with both CNN models, where this test was carried out 20 times by ten voice users with the keyword "Open Access." From the analysis of performance testing that has been performed, the best performance is obtained with CNN DSC speech recognition accuracy with a percentage value of 100%, which is better than the CNN Standard percentage value of 95%. The overall results have indicated that the proposed framework model can increase the high accuracy value, which is better in the classification of voice biometrics compared to CNN Standard.

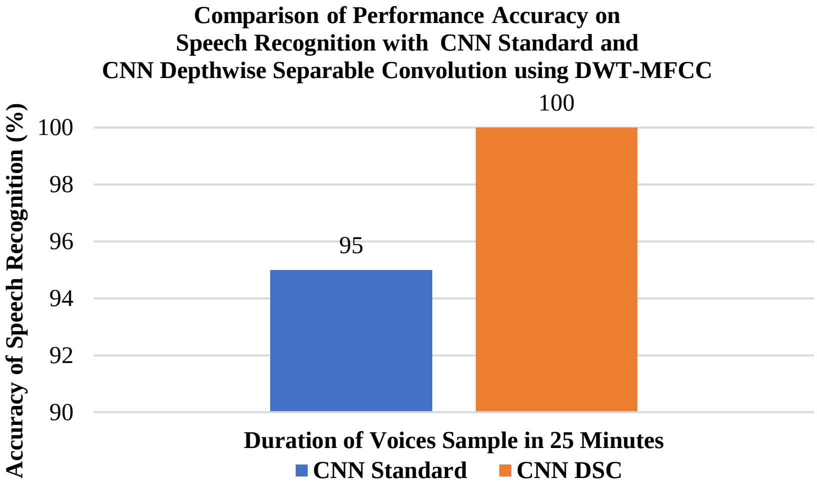


Fig. 10. Comparison of Performance Accuracy of Speech Recognition using CNN Standard and CNN Depthwise Separable Convolution with DWT-MFCC

5. CONCLUSION

This paper has developed the voice biometric framework based on the CNN Depthwise Separable Convolution (DSC) model and the fusion of Discrete Wavelet Transform (DWT) and Mel Frequency Cepstral Coefficients (MFCC). It is targeted to increase the high accuracy, reduce the burden of high computational costs, and speed up the performance of classification process time. We conduct three testing performances, i.e., voice Biometric Training Performance, speaker Recognition Performance (” Who is speaking?”), and Speech Recognition performance (” What keyword is uttered?”). For each of the testing, the results are compared with CNN Standard performance. The results of the training performance testing that has shown that CNN DSC is better than CNN Standard. This CNN DSC performance optimization is able to reduce the number of parameters, where the total CNN DSC training parameters become shorter, only 364,506 when compared total parameters of CNN training Standard 707,386. The reduced number of training parameters will have an effect in accelerating the performance on the time of the training process with CNN DSC to 5.12 minutes faster than CNN Standard 6.39 minutes. Analysis of the performance testing of the training process time on voice biometrics, it can be concluded that the longer the voice sample time in 25 minutes that are running, this will increase the predictive accuracy value to be higher in recognizing and identifying the user's voice.

By optimizing the performance of this DSC model, it is expected to be able effectively to reduce the number of training parameters, reduce the amount of arithmetic in convolution operations, reduce computational complexity, reduce the network parameters size, reduce the amount of operations memory, and reducing the amount of training computational load. So that the optimization of this DSC model can help in solving the problem of higher computational costs and slow system performance. Thus, the development of the research results of this CNN DSC is able significantly in reducing the higher computational costs burden, then accelerate the performance of the prediction processing time in the classification of voice biometrics. The results for testing the performance on speaker and speech recognition are that CNN DSC is better than CNN Standard. The results are obtained by implementing CNN DSC performance optimization on voice biometrics. For speaker recognition performance testing, the results we're able to improve the best performance with the highest percentage value at 99.25% accuracy, 97.14% precision, 98.17% recall, and 97.28% F1-score. And for speech recognition performance testing, the results we're able to improve the best performance with a percentage value at 100% accuracy with 20 trials. So with the development of the research results on the CNN DSC, it could be shown that results of the performance testing of speaker and speech recognition can increase high accuracy value, which is better in the classification of voice biometrics.

The future work is to build up the actual application of the proposed voice biometrics framework, to be implemented for the real-time identification and verification/authentication of the user's voice directly for security access to the Internet banking service.

Acknowledgments

Haris Isyanto is in the Ph.D. program funded by Beasiswa Pendidikan Pascasarjana Dalam Negeri (BPPDN) Ministry of Education and Culture, Republic of Indonesia. Dr. Muhammad Suryanegara is the main supervisor, and Dr. Ajib Setyo Arifin is the co-supervisor as well as the corresponding author. The voice data set is built with 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, https://doi.org/10.1109/ACCESS.2018.2889996.

[2] X. Zhang, D. Cheng, P. Jia, Y. Dai, and X. Xu, "An Efficient Android-Based Multimodal Biometric Authentication System With Face and Voice," *IEEE Access,* vol. 8, pp. 102757-102772, 2020, https://doi.org/10.1109/ACCESS.2020.2999115.

[3] A. Gumaei, R. Sammouda, A. M. S. Al-Salman, and A. Alsanad, "Anti-spoofing cloud-based multi-spectral biometric identification system for enterprise security and privacy-preservation," *Journal of Parallel and Distributed Computing*, vol. 124, pp. 27-40, 2019, https://doi.org/10.1016/j.jpdc.2018.10.005.

[4] M. Vijay and G. Indumathi, "Deep belief network-based hybrid model for multimodal biometric system for futuristic security applications," *Journal of Information Security and Applications*, vol. 58, p. 102707, 2021, https://doi.org/10.1016/j.jisa.2020.102707.

[5] 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, https://doi.org/10.1109/TDSC.2019.2906165.

[6] 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)*, pp. 1-6, 2020, https://doi.org/10.1109/ICoSTA48221.2020.1570613925.

[7] 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," *Computer Speech & Language*, vol. 60, p. 101024, 2020, https://doi.org/10.1016/j.csl.2019.101024.

[8] 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, https://doi.org/10.1109/TVT.2018.2849577.

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

[10] 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)*, pp. 640-645, 2020, https://doi.org/10.1109/ICTC49870.2020.9289160.

[11] L. Chai, J. Du, Q. F. Liu, and C. H. Lee, "A Cross-Entropy-Guided Measure (CEGM) for Assessing Speech Recognition Performance and Optimizing DNN-Based Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 106-117, 2021, https://doi.org/10.1109/TASLP.2020.3036783.

[12] S. Shakil, D. Arora, and T. Zaidi, "Feature based classification of voice based biometric data through Machine learning algorithm," *Materials Today: Proceedings*, vol. 51, pp. 240-247, 2022, https://doi.org/10.1016/j.matpr.2021.05.261.

[13] L. Moreno, "The Voice Biometrics Based on Pitch Replication," *International Journal for Innovation Education and Research*, vol. 6, pp. 351-358, 10/31 2018, https://doi.org/10.31686/ijier.vol6.iss10.1201.

[14] 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, 2020, https://doi.org/10.5121/ijsptm.2020.9201.

[15] 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)*, pp. 1440-1444, 2018, https://doi.org/10.1109/CompComm.2018.8780990.

[16] 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," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, 2022, https://doi.org/10.14569/IJACSA.2022.0130574.

[17] 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, doi: https://doi.org/10.1016/j.clinimag.2021.11.013.

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

[19] 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, https://doi.org/10.1007/s10772-021-09795-2.

[20] G. C. Batista, D. L. Oliveira, O. Saotome, and W. L. S. Silva, "A low-power asynchronous hardware implementation of a novel SVM classifier, with an application in a speech recognition system," *Microelectronics Journal*, vol. 105, p. 104907, 2020, https://doi.org/10.1016/j.mejo.2020.104907.

[21] J. Pribil, A. Pribilova, and J. Matousek, "Artefact Determination by GMM-Based Continuous Detection of Emotional Changes in Synthetic Speech," in *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, pp. 45-48, 2019, https://doi.org/10.1109/TSP.2019.8768826.

[22] C. S. Manasa, K. J. Priya, and D. Gupta, "Comparison of acoustical models of GMM-HMM based for speech recognition in Hindi using PocketSphinx," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 534-539, 2019, https://doi.org/10.1109/ICCMC.2019.8819747.

[23] M. Hamidi, H. Satori, O. Zealouk, K. Satori, and N. Laaidi, "Interactive administration service based on HMM speech recognition system," *International Journal of Computer Aided Engineering and Technology*, vol. 16, no. 2, pp. 266-282, 2022, https://doi.org/10.1504/IJCAET.2022.120819.

[24] K. Naithani, V. M. Thakkar, and A. Semwal, "English Language Speech Recognition Using MFCC and HMM," in *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*, pp. 1-7, 2018, https://doi.org/10.1109/RICE.2018.8509046.

[25] M. D. Shakil, M. A. Rahman, M. M. Soliman, and M. A. Islam, "Automatic Isolated Speech Recognition System Using MFCC Analysis and Artificial Neural Network Classifier: Feasible For Diversity of Speech Applications," in *2020 IEEE Student Conference on Research and Development (SCOReD)*, pp. 300-305, 2020, https://doi.org/10.1109/SCOReD50371.2020.9250964.

[26] M. A. Haque, J. S. R. Alex, and N. Venkatesan, "Evaluation of Modified Deep Neural Network Architecture Performance for Speech Recognition," in *2018 International Conference on Intelligent and Advanced System (ICIAS)*, pp. 1-5, 2018, https://doi.org/10.1109/ICIAS.2018.8540636.

[27] N. H. Ho, H. J. Yang, S. H. Kim, and G. Lee, "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network," *IEEE Access*, vol. 8, pp. 61672-61686, 2020, https://doi.org/10.1109/ACCESS.2020.2984368.

[28] 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, https://doi.org/10.1109/JIOT.2019.2917933.

[29] W. Wang, X. Yang, and H. Yang, "End-to-End Low-Resource Speech Recognition with a Deep CNN-LSTM Encoder," in *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 158-162, 2020, https://doi.org/10.1109/ICICSP50920.2020.9232119.

[30] R. A. Malik, C. Setianingsih, and M. Nasrun, "Speaker Recognition for Device Controlling using MFCC and GMM Algorithm," in *2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, pp. 1-6, 2020, https://doi.org/10.1109/ICECIE50279.2020.9309603.

[31] N. Sen, M. Sahidullah, H. A. Patil, S. K. Das Mandal, K. S. Rao, and T. K. Basu, "Utterance partitioning for speaker recognition: an experimental review and analysis with new findings under GMM-SVM framework," *International Journal of Speech Technology*, Article vol. 24, no. 4, pp. 1067-1088, 2021, https://doi.org/10.1007/s10772-021-09862-8.

[32] N. Maghsoodi, H. Sameti, H. Zeinali, and T. Stafylakis, "Speaker Recognition With Random Digit Strings Using Uncertainty Normalized HMM-Based i-Vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1815-1825, 2019, https://doi.org/10.1109/TASLP.2019.2928143.

[33] 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)*, pp. 130-133, 2019, https://doi.org/10.1109/CCOMS.2019.8821751.

[34] Z. Hu, Y. Fu, X. Xu, and H. Zhang, "I-Vector and DNN Hybrid Method for Short Utterance Speaker Recognition," in *2020 IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, vol. 1, pp. 67-71, 2020, https://doi.org/10.1109/ICIBA50161.2020.9277099.

[35] M. B. Andra and T. Usagawa, "Improved Transcription and Speaker Identification System for Concurrent Speech in Bahasa Indonesia Using Recurrent Neural Network," *IEEE Access*, vol. 9, pp. 70758-70774, 2021, https://doi.org/10.1109/ACCESS.2021.3077441.

[36] 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, https://doi.org/10.1109/TIFS.2019.2941773.

[37] 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)*, pp. 322-325, 2018, https://doi.org/10.1109/CCE.2018.8465705.

[38] O. L. Baroi, A. Niaz, M. J. Islam, M. S. A. Kabir, E. Islam, and M. J. Rahimi, "Effects of Different Environmental noises and Sampling Frequencies on the performance of MFCC and PLP based Bangla Isolated Word Recognition System," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pp. 1-6, 2019, https://doi.org/10.1109/ICASERT.2019.8934462.

[39] 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, https://doi.org/10.1155/2021/6635964.

[40] 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, https://doi.org/10.1109/JSTARS.2020.3041859.

[41] W. Shan *et al.*, "A 510-nW Wake-Up Keyword-Spotting Chip Using Serial-FFT-Based MFCC and Binarized Depthwise Separable CNN in 28-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 151-164, 2021, https://doi.org/10.1109/JSSC.2020.3029097.

[42] W. Shan *et al.*, "14.1 A 510nW 0.41V Low-Memory Low-Computation Keyword-Spotting Chip Using Serial FFT-Based MFCC and Binarized Depthwise Separable Convolutional Neural Network in 28nm CMOS," in *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*, pp. 230-232, 2020, https://doi.org/10.1109/ISSCC19947.2020.9063000.

[43] P. Pyykkönen, S. I. Mimilakis, K. Drossos, and T. Virtanen, "Depthwise Separable Convolutions Versus Recurrent Neural Networks for Monaural Singing Voice Separation," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1-6, 2020, https://doi.org/10.1109/MMSP48831.2020.9287169.

[44] l. Gangzhao, W. Zhang, and Z. Wang, "Optimizing Depthwise Separable Convolution Operations on GPUs," *IEEE Transactions on Parallel and Distributed Systems*, vol. 1, pp. 70-87, 2021, https://doi.org/10.1109/TPDS.2021.3084813.

[45] 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)*, pp. 379-383, 2018, https://doi.org/10.1109/ICOIACT.2018.8350748.

[46] 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, https://doi.org/10.1109/JSYST.2020.2984132.

[47] H. M. Naing, R. Hidayat, R. Hartanto, and Y. Miyanaga, "Discrete Wavelet Denoising into MFCC for Noise Suppressive in Automatic Speech Recognition System," *International Journal of Intelligent Engineering and Systems*, vol. 13, pp. 74-82, 2020, https://doi.org/10.22266/ijies2020.0430.08.

[48] F. Amelia and D. Gunawan, "DWT-MFCC Method for Speaker Recognition System with Noise," in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 2019, pp. 1-5, https://doi.org/10.1109/ICSCC.2019.8843660.

[49] 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, https://doi.org/10.1016/j.ins.2019.06.064.

[50] F. J. Ariza-Lopez, J. Rodriguez-Avi, and M. V. Alba-Fernandez, "Complete Control of an Observed Confusion Matrix," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1222-1225, 2018, https://doi.org/10.1109/IGARSS.2018.8517540.

BIOGRAPHY OF AUTHORS

Haris Isyanto (Member, IEEE) received a Bachelor’s degree in Electrical Engineering from the Universitas Muhammadiyah Jakarta in 1996 and a Master’s degree from the Universitas Trisakti, Jakarta, in 2000. He works as a Senior Lecturer at the Department of Electrical Engineering, Universitas Muhammadiyah Jakarta in Indonesia. Since 2019, he has been studying Ph.D. in the Department of Electrical Engineering - at Universitas Indonesia. His research interests are mobile and wireless communications, IoT, sensors, electronics, and implementations of artificial intelligence and machine learning algorithms. Email: haris.isyanto@ui.ac.id, haris.isyanto@ftumj.ac.id, and Orcid: 0000-0001-6723-1506.

Ajib Setyo Arifin (Member, IEEE) received a Bachelor’s in Electrical Engineering and a Master’s Degree from the Universitas Indonesia in 2009 and 2011, respectively. He got a Ph.D. degree in Telecommunications in 2015 from Keio University, Japan. He is an Assistant Professor at Universitas Indonesia. His research areas include Wireless Sensor Networks, Wireless Communication, and signal processing for communication. Email: ajib@eng.ui.ac.id and Orcid: 0000-0002-4648-6347.

Muhammad Suryanegara (Senior Member, IEEE) received the Bachelor's degree in Electrical Engineering from the Universitas Indonesia in 2003, the Master's degree from University College London, London, U.K., in 2004, and a Ph.D. degree from the Tokyo Institute of Technology, Japan, in 2011. He is currently an Associate Professor of telecommunications management with the Department of Electrical Engineering, Universitas Indonesia. He is the principal investigator in the research area of ICT policy and technology management, the IoT, 4G/5G, and wireless communication technology, concerning both technical research and regulatory aspects, and he also engaged in international regulatory activities. Email: m.suryanegara@ui.ac.id and Orcid: 0000-0003-0488-3931.