Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 1 (2022) 140 - 145 ISSN Media Electronic: 2580-0760

Speaker Identification Using a Convolutional Neural Network

Suci Dwijayanti[1], Alvio Yunita Putri[2], Bhakti Yudho Suprapto[3]
[1,2,3]Department of Electrical Engineering, Faculty of Engineering, Universitas Sriwijaya
[1]sucidwijayanti@ft.unsri.ac.id*, [2]alvio.yunita1997@gmail.com, [3]bhakti@ft.unsri.ac.id

Abstract

Speech, a mode of communication between humans and machines, has various applications, including biometric systems for identifying people have access to secure systems. Feature extraction is an important factor in speech recognition with high accuracy. Therefore, we implemented a spectrogram, which is a pictorial representation of speech in terms of raw features, to identify speakers. These features were inputted into a convolutional neural network (CNN), and a CNN-visual geometry group (CNN-VGG) architecture was used to recognize the speakers. We used 780 primary data from 78 speakers, and each speaker uttered a number in Bahasa Indonesia. The proposed architecture, CNN-VGG-f, has a learning rate of 0.001, batch size of 256, and epoch of 100. The results indicate that this architecture can generate a suitable model for speaker identification. A spectrogram was used to determine the best features for identifying the speakers. The proposed method exhibited an accuracy of 98.78%, which is significantly higher than the accuracies of the method involving Mel-frequency cepstral coefficients (MFCCs; 34.62%) and the combination of MFCCs and deltas (26.92%). Overall, CNN-VGG-f with the spectrogram can identify 77 speakers from the samples, validating the usefulness of the combination of spectrograms and CNN in speech recognition applications.

Keywords: speaker identification, CNN, spectrogram, feature extraction

1. Introduction

Speech is being increasingly used in human–machine interactions for various applications, such as biometric security systems. Since unsecured systems are prone to risks, such as robbery, demolition, and misuse, security plays a major role in the lives of people. Typically, security systems implement methods that employ pattern, pin, or password locks. However, they exhibit certain flaws as they can be easily hacked. Therefore, security systems based on the identification of human physiological characteristics, namely biometrics, are preferred over the aforementioned methods. Biometric systems use pattern recognition to determine and verify the identification of a person.

Speech is a common biometric used to identify a person. In comparison with other biometrics, speech-based biometrics are more cost-effective as they do not require specific hardware; moreover, the file size requirement is comparatively small [1]. Several studies have implemented feature extraction methods to obtain accurate speech recognition results. Warohma et al. identified Indonesian dialects using Mel-frequency cepstral coefficients (MFCCs) and neural networks. However, their results indicated that the MFCC features may reduce certain important information in speech owing to dimension reduction [2]. MFCCs were also used for gender detection and speaker identification [3]. Another study utilized multiple kernel weighted Mel frequency cepstral coefficient (MKMFCCc) for feature extraction and support vector machine as classifier to perform automatic speaker identification [4]. MFCCs and the Gaussian mixture model (GMM) were used for access control [5], and timbrel features and KNN distance measure were used later [6] for speaker identification. In a recent study, MFCCs and GMM were combined with a deep neural network (DNN) to identify the speaker in an emotional talking environment. Another study utilized hidden Markov models (HMMs) and the hill climbing gradient; however, only a local optimal model was achieved [7]. Furthermore, an HMM requires a first-order Markov model, which warrants the use of “delta” and “delta-delta” coefficients to add time information in several frames. Therefore, this method may prove to be computationally expensive.

To address the aforementioned drawbacks, we implemented a simple feature, namely a spectrogram, which is an image that represents speech and is obtained

Accepted: 13-01-2022 | Received in revised: 18-02-2022 | Published: 27-02-2022

from the short-time Fourier transform (STFT). STFT provides more information in terms of both frequency and time, particularly from speech input. Raw features can be obtained directly from speech signals. In this study, we used a convolutional neural network (CNN) rather than an HMM to identify the speaker from the input spectrogram. Although CNNs are more commonly used as classifiers for image recognition, a few studies have utilized CNNs for speech recognition. Sun et al. [8] combined deep and shallow features for speech emotion recognition using a deep convolutional neural network. Fan et al. [9] proposed the use of filter banks as the input of a CNN as they considered a spectrum and sparse in lower and higher frequencies, respectively, which is suitable for the non-linear perception of the human ear. Kawamura et al. used a DNN and a CNN to identify songs and estimate the fundamental frequency. They validated that the raw waveform and filter learning combination can yield adequately accurate results [10]. Nevertheless, this feature requires complex computations. Therefore, in this study, we implemented a spectrogram, which is a less complex computation feature. As a spectrogram comprises low-level speech features and represents the image of a speech signal, the combination of CNN and spectrogram can improve the speech recognition performance.

The remainder of this paper is organized as follows. Section 2 presents the materials and methods used in the data collection for the proposed speaker identification process. The results of the experiments and the inferences drawn from them are summarized in Section 3. Finally, Section 4 concludes the paper.

2. Research Methods

2.1 Data Collection

Data used in this study were obtained from 78 speakers. Each speaker uttered the final three digits of their student identification (ID) number in the Indonesian language 10 times. Consequently, 780 speech data points were collected. The utterances were recorded in a closed room using a Rode VideoMicro microphone. The sampling frequency was 16 kHz, and each sample was recorded for 1–5 s.

2.2 Proposed Speaker Identification Process

Figure 1 presents a block diagram of the proposed process for speaker identification.

The first step, namely preprocessing, includes reduction of the noise that occurred during recording. This process involved normalization and pre-emphasis, wherein normalization was used to obtain the uniform amplitude of the speech signal by dividing the *i*-th speech data by the maximum amplitude of the speech as follows.

$$S[i] = \frac{s[i]}{\max\limits_{1 \le i \le N} |s[i]|}, i = 1,2,3, \dots, N, \quad (1)$$

where $s[i]$ denotes the speech signal, and N indicates the length of the speech signal. This was followed by pre-emphasis, wherein the noise in the speech data was reduced by retaining only the high-frequency signals from the speech signal. The pre-emphasis can be calculated using the following equation.

$$x[i] = S[i] - \alpha S[i], \ 0.9 \le \alpha \le 1, \quad (2)$$

where $x[i]$ denotes the *i*-th data of the pre-emphasized speech signal, $S[i]$ indicates the speech signal, and α represents the pre-emphasis factor.

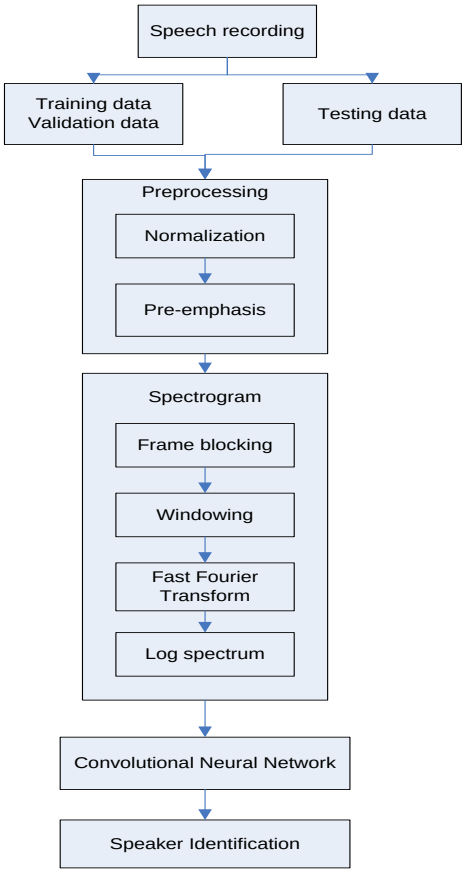

Figure 1. Block Diagram of Speaker Identification Using a Convolutional Neural Network (CNN)

After preprocessing, the speech signal was extracted into a spectrogram using the STFT algorithm. Herein, the speech signals were initially blocked into frames of 20 ms with an overlap of 10 ms. As frame blocking can result in discontinuous signals, we used a Hamming window to prevent this discontinuity. Subsequently, the product of the framed signal and Hamming window was used as the input to the Fourier transform process. This STFT was used to convert the speech signal from the time domain to a frequency domain. The process of STFT is indicated in (3).

DOI: https://doi.org/10.29207/resti.v6i1.3795

$$X(k) = \sum_{i=1}^{N} x(i)w(i-k)e^{-\frac{j2\pi n}{K}}. \quad (3)$$

Since the output of this process is a complex number, the spectrum of the signal can be calculated as follows:

$$X(k) = \log|X(k)|^2. \quad (4)$$

The output of the spectrum is a spectrogram, which is a2D image representation of the speech signal, as depicted in Figure 2. This image was inputted to the CNN in the JPEG format for the training process. The model obtained in the training stage was used to identify the speaker of the tested data. The training and validation data were in the ratio 80:20.

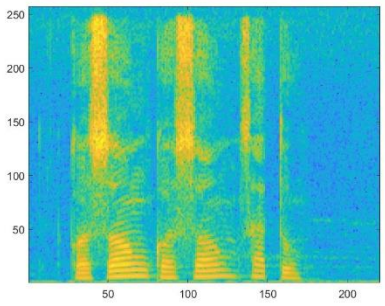

Figure 2. Spectrogram of The Speech Signal of Speaker 001

2.3. CNN

CNN is a type of artificial neural network that is typically used to identify objects [11]. It comprises several layers, including a convolution layer, rectified linear unit layer (ReLU), and pooling layer, which are used for feature learning. Furthermore, the classification is performed in the subsequent layer after flattening the input to the fully connected layer and the softmax layer. Figure 3 illustrates the architecture of CNN used for speech processing in this study.

Next, the convolution operation is performed in the convolution layer to filter the features in the images (spectrogram). An n × n matrix, referred to as the feature detector, was filtered from the input spectrogram to obtain the feature map; the collection of feature maps forms the convolution layer. The output of this process was rectified using a ReLU layer to improve the nonlinearity. Subsequently, its dimension was reduced in the pooling layer. In this study, we selected the maximum values of the feature map by using maximum pooling. As the results of the feature maps were in the matrix form, they were converted to the vector column in the flattening process. The output of this process served as the input to the fully connected layer. Finally, a softmax layer was used to classify the input. This layer calculated the probability of the target class (speaker) to identify the speaker. The calculation was performed as follows:

$$\sigma(z)_j = \frac{e^{z_k}}{\sum_{k=1}^{K} e^{z_k}}. \quad (5)$$

In this study, two CNN architectures were used: (1) a simple architecture, and (2) a VGG-f architecture, as presented in Tables 1 and 2, respectively. Although the VGG-f architecture is typically used in image processing, we verified that this type of architecture can effectively identify speakers using the information obtained from the extracted features.

Table 1. Simple CNN Architecture

Layer	Kernel Size	Stride	Padding	Bias
Convolutional_1	5 × 5	1	0	1, 5
Max Pooling	2 × 2	2	0	-
Convolutional_2	5 × 5	1	0	5, 70
Max Pooling	2 × 2	2	0	-
Convolutional_3	5 × 5	1	0	70, 78
Max Pooling	2 × 2	2	0	-
Softmaxloss	-	-	-	-

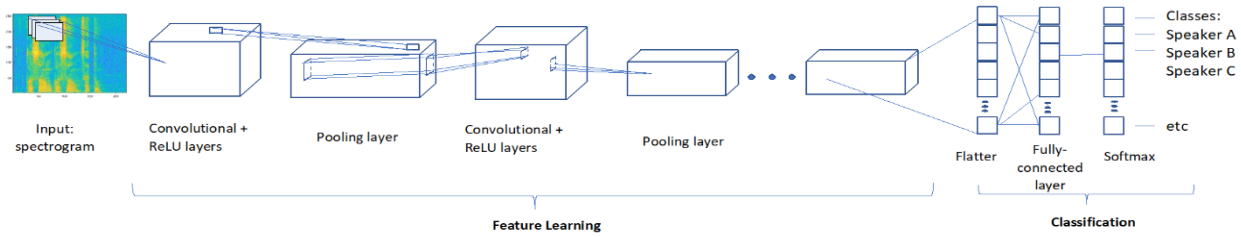

Figure 3. Architecture of the Convolutional Neural Network (CNN)

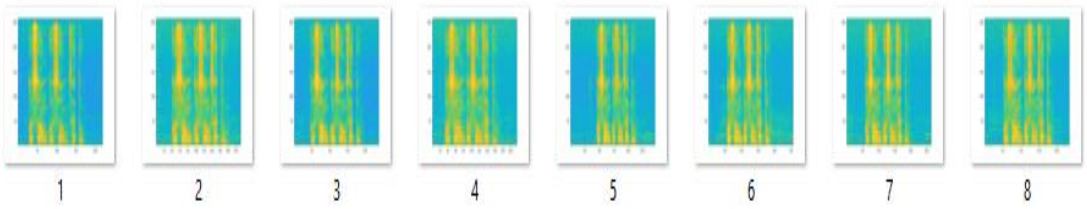

Figure 4. Samples Obtained from Spectrograms with a Size of 875 × 655 pixels for Training

DOI: https://doi.org/10.29207/resti.v6i1.3795

Table 2. Architecture of CNN-VGG-f.

Layer	Kernel Size
Convolutional_1 + ReLU	11 × 11
Max Pooling	3 × 3
Convolutional_2 + ReLU	5 × 5
Max Pooling	3 × 3
Convolutional_3 + ReLU	3 × 3
Convolutional_4 + ReLU	3 × 3
Convolutional_5 + ReLU	3 × 3
MaxPooling	3 × 3
Fully Connected_1 + Dropout	4096
Fully Connected_2 + Dropout	4096
Softmaxloss	78

3. Result and Discussion

The speech signals extracted from the spectrogram were saved in the image format with a size of 875 × 656 pixels as the input of the CNN. Figure 4 depicts the spectrogram sample used for the training process.

The training data spectrogram images were preprocessed by resizing the images to 224 × 224 pixels. The input for the CNN consisted of images of size $m \times m \times r$, where m denotes the length and width of the image, and r indicates the channel of the image in terms of RGB.

As mentioned in Section 2.3, we used two architectures in this study. In the case of the simple architecture (Table 1), the image was resized to 32 × 32 pixels. The learning rate and batch size were 0.001 and 200, respectively.

Figure 5 illustrates a graph of the objective function obtained using this architecture, wherein the number of epochs is 100. The blue and red lines represent the training and validation, respectively. At the 30th epoch, because the validation loss did not reach zero, we added a dropout. However, the results indicated that overfitting occurred in the network despite the addition of the dropout. This can be attributed to the architecture being highly shallow for generalizing the data.

Therefore, we used the VGG-f architecture. Although this architecture is generally used for facial recognition, we found that it has the capability to aid in identifying

the speaker of the speech signal since it has more layers and is deeper than the simple architecture of CNN. In the training process, the convolution layers of size 3 × 3 were stacked on top of each other, and each layer was connected with 4096 nodes followed by a classifier softmax layer with a kernel size of 78, which represents the speaker class (Table 2). The parameters used for training the VGG-f architecture are listed in Table 3. These parameters constitute 780 training data, and the class or label represents the number of classes of 78 speakers. The image size of the spectrogram after preprocessing was 224 × 224 pixels. In this architecture, the learning rate was set to 0.001 because an extremely large learning rate may cause the model to converge to a suboptimal solution and an extremely small rate may stagnate the process, which would extend the time required for computation. The batch size was set to 256 to ensure a fast-training process, and the epoch was set to 100 because the objective function could achieve the optimal loss function within 100 epochs.

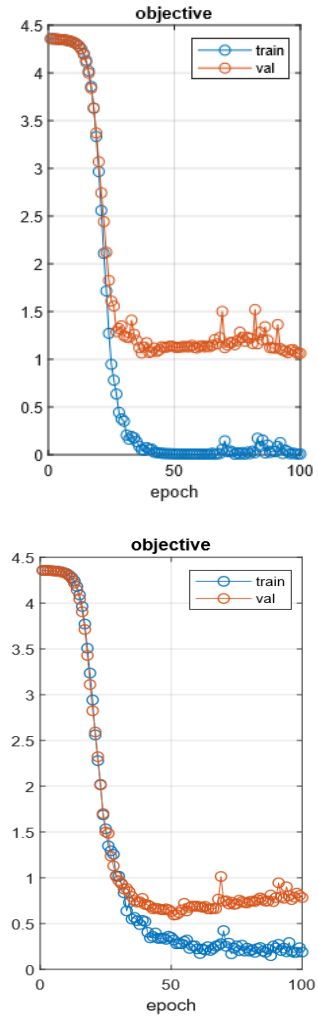

Figure 5. Training Graphs of the Objective Function for 100 Epochs Without Dropout (above) and With Dropout (below)

Table 3. Parameter Training for Identifying Speakers Using CNN.

Parameter	Value
Training data	780
Class/label	78
Image size	224 × 224
Learning rate	0.001
Batch size	256
Epoch	100

The training process was similar to that of ImageNet [12], and MatConvNet [13] was used to train this CNN for speaker identification. Figure 6 illustrates the objective of the network training, wherein we observed that CNN with a VGG-f architecture could achieve objectives within 100 epochs.

DOI: https://doi.org/10.29207/resti.v6i1.3795

As indicated in the figure, the loss in 100 epochs was close to zero, and the training and validation graphs verified that the network was appropriately generalized without overfitting or underfitting. Therefore, we used this model to identify the speakers.

To evaluate the spectrogram performance in terms of the input features of CNN, we evaluated the performance of MFCCs and the combination of MFCCs with their delta and delta-delta features, which have previously exhibited satisfactory results in speech recognition [14]. The obtained results were compared to verify the effectiveness of the proposed method. The testing involved data of 78 samples that were not included in the training process. Table 4 summarizes the results obtained for 10 samples of speakers.

As indicated in the table, the spectrogram exhibits better accuracy in identifying speakers compared to those of MFCCs and MFCCs combined with their delta and delta-delta features. These results imply that the spectrogram is the most suitable feature to be used with CNN to identify the speaker. Moreover, the spectrogram comprises raw features that contain important information about the speech uttered by the speaker. Conversely, MFCCs may result in information loss owing to the dimension reduction caused by the computation process. Furthermore, the results verify that the combination of MFCC and its delta features was ineffective in improving the accuracy of identifying the speakers.

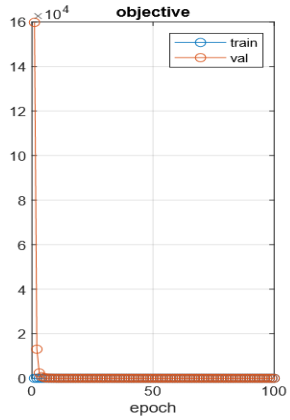

Figure 6. Training Graph of the Objective Function for 100 Epochs

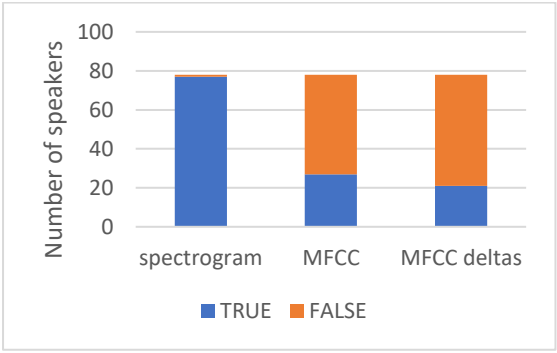

Figure 7. Training Graph of The Objective Function for 100 Epochs

Table 4. Results of Ten Samples of Speakers Obtained Using Different Feature Extraction Methods.

No.	Student ID	STFT			MFCC		MFCC with deltas	
		Sample	Identified	Accuracy (%)	Identified	Accuracy (%)	Identified	Accuracy (%)
1	001	Sample 1	True	88.380	False	32.393	False	25.448
2	002	Sample 2	True	91.788	True	75.986	False	49.522
3	003	Sample 3	True	98.188	True	94.358	True	68.784
4	004	Sample 4	True	83.595	False	33.863	False	31.666
5	005	Sample 5	True	92.781	False	46.553	False	24.279
6	006	Sample 6	True	90.582	True	50.563	False	21.625
7	007	Sample 7	True	83.536	False	26.278	False	38.136
8	008	Sample 8	True	84.872	False	45.278	False	40.693
9	009	Sample 9	True	72.881	False	30.324	True	51.835
10	010	Sample 10	True	75.952	False	33.773	False	33.601

The testing results from 78 speakers validated that the combination of spectrogram and CNN-VGG could identify 77 speakers or 98.78% of the speakers. Conversely, MFCC and MFCC with delta features could only identify 34.62% and 26.92%, respectively, of the speakers. Figure 7 compares the number of speakers identified using STFT (spectrogram), MFCC, and MFCC with deltas. The results of the study validate that the CNN can be applied suitably for both image recognition and speech recognition. Furthermore, the CNN-VGG architecture was found to be effective in identifying speakers.

4. Conclusion

This study verified the application of CNN in speech recognition. Speech recognition has various applications, including human–machine interaction for biometrics. We determined that the VGG-f architecture with a learning rate of 0.001 and batch size of 256, which is deeper and has more layers than a simple CNN architecture, is suitable for speaker identification; therefore, it can effectively extract features from spectrograms. Furthermore, the experimental results validate that, compared to MFCC and MFCC with delta features, the spectrogram exhibits a high accuracy of 98.78% in identifying the speakers. Thus, the spectrogram is more beneficial for speech recognition

DOI: https://doi.org/10.29207/resti.v6i1.3795

using CNN as it comprises raw features and does not require complex computations. In the future, we intend to investigate the potential of using the spectrogram for dialect identification, which can be implemented in various applications of human–machine interactions.

References

[1] T. O. Majekodunmi and F. E. Idachaba, "A review of the fingerprint, speaker recognition, face recognition and iris recognition based biometric identification technologies," in *Proceeding of World Congress in Engineering. 2011, WCE 2011*, vol. 2, pp. 1681–1687, 2011.

[2] A. M. Warohma, P. Kurniasari, S. Dwijayanti, and B. Y. Suprapto, "Identification of Regional Dialects Using Mel Frequency Cepstral Coefficients (MFCCs) and Neural Network," in *2018 Int. Semin. Appl. Technol. Inf. Commun.*, pp. 522–527, 2018.

[3] O. Mamyrbayev, A. Toleu, G. Tolegen, and N. Mekebayev, "Neural architectures for gender detection and speaker identification," *Cogent Eng.*, vol. 7, no. 1, 2020, doi: 10.1080/23311916.2020.1727168.

[4] O. S. Faragallah, "Robust noise MKMFCC–SVM automatic speaker identification," *Int. J. Speech Technol.*, vol. 21, no. 2, pp. 185–192, 2018, doi: 10.1007/s10772-018-9494-9.

[5] J. C. Liu, F. Y. Leu, G. L. Lin, and H. Susanto, "An MFCC-based text-independent speaker identification system for access control," *Concurr. Comput. Pract. Exp.*, vol. 30, no. 2, pp. 1–16, 2018, doi: 10.1002/cpe.4255.

[6] V. M. Sardar and S. D. Shirbahadurkar, "Speaker identification of whispering speech: an investigation on selected timbrel features and KNN distance measures," *Int. J. Speech Technol.*, vol. 21, no. 3, pp. 545–553, 2018, doi: 10.1007/s10772-018-9527-4.

[7] X. Lu, S. Li, and M. Fujimoto, "Automatic speech recognition," *SpringerBriefs Comput. Sci.*, pp. 21–38, 2020, doi: 10.1007/978-981-15-0595-9_2.

[8] L. Sun, J. Chen, K. Xie, and T. Gu, "Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition," *Int. J. Speech Technol.*, vol. 21, no. 4, pp. 931–940, 2018, doi: 10.1007/s10772-018-9551-4.

[9] R. Fan and G. Liu, "CNN-Based Audio Front End Processing on Speech Recognition," *2018 Int. Conf. Audio, Lang. Image Process.*, pp. 349–354, 2018.

[10] T. Kawamura, A. Kai, and S. Nakagawa, "Noise Robust Fundamental Frequency Estimation of Speech using CNN-based Discriminative Modeling," in *2018 5th Int. Conf. on Adv. Informatics: Concept Theory and Applications (ICAICTA*), Krabi, Thailand, 2018, pp. 60–65. https://doi.org/10.1109/ICAICTA.2018.8541328

[11] S. Albawi, T. A. M. Mohammed, and S. Alzawi, "Understanding of a Convolutional Neural Network,"in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1-6, 2017.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks." Advances in neural information processing systems 25, pp. 1097-1105, 2012.

[13] A. Vedaldi and C. V May, "MatConvNet Convolutional Neural Networks for MATLAB." in *Proceedings of the 23rd ACM international conference on Multimedia*. 2015.

[14] M. Parchami, W. P. Zhu, B. Champagne, and E. Plourde, "Recent developments in speech enhancement in the short-time fourier transform domain," *IEEE Circuits Syst. Mag.*, vol. 16, no. 3, pp. 45–77, 2016, doi: 10.1109/MCAS.2016.2583681.

DOI: https://doi.org/10.29207/resti.v6i1.3795