Journal of Theoretical and Applied Information Technology
15th September 2022. Vol.100. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPTIMIZING SPEAKER RECOGNITION USING D-VECTOR AND X-VECTOR SPEECH EMBEDDINGS AND DEEP LEARNING

[1]KEVIN KURNIAWAN, [2]AMALIA ZAHRA

[12]Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480

E-mail: [1]kevin.kurniawan007@binus.ac.id, [2]amalia.zahra@binus.edu

ABSTRACT

In speaker recognition, a way to identify individual's identity is to identify the characteristics of the individual voice, which can be identified using speaker's speech embedding. In this paper, we propose a method using x-vector and d-vector speech embedding feature extraction and ResNet50 model on identifying user's identity. This paper uses part of *VoxCeleb2* dataset with total train data of 12.207 utterances, total validation data of 3.618 utterances, and total testing data of 2.068 utterances from 552 speakers. Research done in this paper shows ResNet50 model with x-vector achieved accuracy, recall, precision, and f-1 score of 74.12%, 73.42%, 78.02%, and 0.72 respectively, whereas ResNet50 model with d-vector achieved accuracy, recall, precision, and f-1 score of 36.17%, 35.77%, 36.86%, and 0.32 respectively, whereas the current state-of-the-art model achieved accuracy, recall, precision, and f-1 score of 10.97%, 10.61%, 6.64%, and 0.07 respectively.

Keywords: *Speaker Recognition, Text Independent Speaker Recognition, ResNet50 Model, X-Vector Speech Embeddings, D-Vector Speech Embeddings, MFCC*

1. INTRODUCTION

There are few traditional approaches for verifying identities like using a Personal Identification Number (also known as PIN) and passwords, but those approaches are very risky since it doesn't really verify an individual characteristic, meaning everyone can access the protected resource if the password or PIN is stolen or leaked [1][2]. Biometric system gives a solution to that problem by using individual characteristic in order to authenticate or verify the identity [3] since biometric system focuses on statistical analysis of biological characteristics [1][2] such as fingerprints, hand geometry, voice identification, and retina identification.

Speaker recognition is process of recognizing a person, who is speaking, by obtaining characteristics or speech wave parameters [4]. It enables system to recognize the voice of known speakers. Speaker identification (generally also known as speaker recognition) is part of speaker recognition [5]. Speaker recognition is a process of identifying speaker's identity through their voice characteristic. There are two types of speaker recognition: text-dependent speaker recognition and text-independent speaker recognition [4].

The difference between text-dependent speaker recognition and text-independent speaker recognition is the utterance the speaker. Text-dependent speaker recognition uses same utterance for training and testing, whereas text-independent speaker recognition uses different utterance for training and testing. There are two main steps for text-independent speaker recognition, which are training and decision making. Training process in text-independent speaker recognition requires the characteristic of the speaker's voice, which can be achieved by using an audio feature extraction algorithm (mel frequency cepstral coefficient (MFCC) or speech embeddings algorithm (such as *x-vector*)) and decision-making process decides which of the listed speakers have the closest characteristic of the given input voice. The current state-of-the-art

of speaker verification model using *VoxCeleb2* dataset is ResNet50 with MFCC feature extraction where the model successfully produces EER of 3.95% [6]. Although the current state-of-the-art produces rather low EER of 3.95%, it can be further improved by upgrading the feature extraction aspect of the system by using a speech embedding features rather than pure MFCC features as the representative of the speaker's characteristic.

X-vector speech embedding is an output vector from a deep neural network (DNN) to capture the characteristic of the voice [7], and d-vector speech embedding is an output vector from an LSTM network to capture the characteristic of the voice [8]. We experiment with ResNet50 based architecture using x-vector and d-vector speech embedding as the input to compare two embeddings to further improve the quality of the model. The rest of the paper is organized in the following manner. Review from previous work in Section 2, followed by methodology in Section 3. In Section 4, we discuss the experiment result of ResNet50 using x-vector and d-vector speech embedding features and finally Section 5 concludes the paper.

2. RELATED WORKS

Research by [7] proposes a method using i-vectors from feed neural network to extract the characteristic of speaker's voice which can be used to represent a speaker's characteristic for speaker verification and recognition system. The research shows that a text independent speaker verification system using i-vectors from feed forward neural network using *NIST SRE* dataset yields EER of 8.1%, 6.8%, 4.3%, 2.8%, 2.1%, 1.8%, 6.3%, 15.4%, and 11.3% for utterance with durations of 10 (with training of 10 seconds utterances), 5, 10, 20, 60 seconds, full duration, cantonese language, tagalog language, and mixed language respectively. Research performed by [9] propses a method that extracts speech features that utilizes CNN to extract a frame level speaker embedding. The proposed method successfully returns an EER of 6.61% with a *VoxCeleb1* dataset. There is also another text independent speaker recognition research performed by [10] by using CNN deep learning model for image processing for audio's spectrogram image. The proposed method returns accuracy of 95.83%.

Research performed by [11] proposed a text independent speaker recognition method based on MFCC with Gaussian Mixture. This research have a total of 5 experiments to test the model to identify speakers, whereas the first experiment is text dependent speaker recognition, second experiment is text independent speaker recognition, third and fourth experiment is a clone of first and second experiment but with Chinese language, and the last experiment is a comparison experiment between MFCC and MFCC delta. The proposed method produces accuracy range of 88%-95%, 84%-91%, 88%-92%, and 60%-80% for the first, second, third, and fourth experiments. There is also an experiment proposed by [12] that optimizes MFCC in extracting speaker's characteristic by combining MFCC with time-based features for speaker recognition. The research proposes a system with 3 sub-systems, which are gender model (that predicts whether the voice is a male or female), male model (that predicts the voice to listed male voices), and female model (that predicts the voice to listed female voices). The proposed model have accuracy of 92.9%, 88.5%, and 83.5% for gender, male, and female models.

Research by [13] focuses on text independent speaker recognition on short utterances by utilizing ResCNN model. This experiment trains the model with a total of 5, 3, and 2 utterances and have produced EER of 5.40%, 5.98%, and 6.83% respectively. Research done by [14] focuses on the decision making for text independent speaker recognition based on the closest feature vector. The proposed decision-making model have produced accuracy of 88.6%.

Research by [6] compares text independent speaker recognition model, such as VGG-M and ResNet, using 3 different methods. The 3 different methods consist of the following: (1) baseline method, (2) 10 sampling for 3 seconds temporal crops for each test segment and utilize mean from feature, and lastly (3) 10 sampling 3 seconds temporal crops for each test segment, compute distance between pair of crops from 2 speech segments and utilizes mean from the computed distance. Result from the experiment produces EER of 5.94%, 5.04%, 5.11%, 4.83%, 4.19%, 4.43%, and 3.95% for VGG-M(1), ResNet34(1), ResNet34(2), ResNet34(3), ResNet50(1), ResNet50(2), and ResNet50(3). Research by [15] proposes an approach for text independent speaker recognition by utilizing MFCC and probabilistic neural network. This research combines MFCC and delta derivative

from MFCC (DMFCC and DDMFCC) that obtained from mel spaced Gaussian filter banks calculation for the input of text independent speaker recognition system. Best result from this experiment have achieved an accuracy of 94% by using feature vector size of 18.

Current state-of-the-art for *VoxCeleb2* dataset on identifying speaker's utterance is ResNet50 with MFCC feature extraction. The feature extraction of the proposed model can be further improved by using more advance way to determine the speaker's characteristics such as using x-vector or d-vector speech embeddings; therefore, we propose a new method of identifying utterance by using x-vector or d-vector speech embedding for representing the characteristic of the speaker's voice and ResNet50 model for verifying the speaker's utterance.

3. METHODOLOGY

This section will explain the steps necessary to build x-vector and d-vector speech embeddings feature extraction speaker recognition system using ResNet50 as shown in Figure 1.

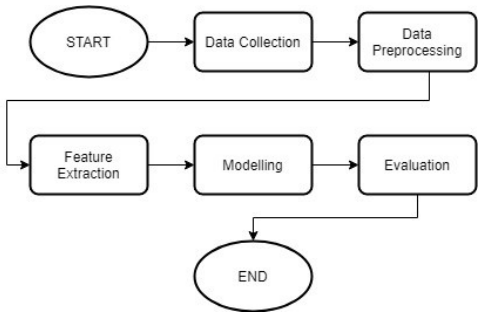

Figure 1: Speaker Recognition System Research Process

3.1 Data Collection

Data used in this study is called *VoxCeleb2* which consists of 1.128.246 total speakers and 6.112 total voices. *VoxCeleb2* consists of 2 parts, namely: *vox2-dev* and *vox2-test*, where *vox2-dev* part have 1.092.009 total voices from 5.994 different speakers and *vox2-test* have 36.237 total voices from 118 total speakers. We use *vox2-dev* part of *VoxCeleb2* to do hard negative mining for the system and *vox2-test* will be partitioned into 70% training data, 20% validation data, and for the rest 10% of the data will be used for testing. We also only select speakers with total number of utterances more than 25 and less than 40 utterances to minimize error caused by imbalance number of utterance used for training the model, which brings us to total of 552 speakers, 12.207 training datas, 3.618 validation datas, and 2.068 testing datas.

3.2 Feature Extraction

3.2.1 Mel-Frequency Cepstral Coefficient (MFCC)

After the data has been gathered, feature extraction step begins. The goal of feature extraction step is to extract feature (such as the characteristic of the voice). Feature extraction step in this paper uses mel frequency cepstral coefficient algorithm, also known as MFCC, to obtain a vector of voice features for the recordings. MFCC algorithm first frames the audio signal into 20 milliseconds to 40 milliseconds frames then calculates the Discrete Fourier Transform, also known as DFT, on each frame using the formula as shown in Eq. (1):

$$S_i(k) = \sum_{n=1}^{N} S_i(n)h(n)e^{-j2\pi kn/N} \quad 1 \leq k \leq K \qquad (1)$$

In Eq. (1), k represents the length of DFT, S(n) represents domain signal, which Si(n) represents the domain signal of each i^{th} frame where n ranges from 1 to number of samples, and h(n) represents N sample long analysis window. After the DFT of each frame calculated, we then able to calculate Periodogram estimate of the power spectrum using the formula as shown in Eq. (2):

$$P_i(k) = \frac{1}{N}|S_i(k)|^2 \qquad (2)$$

After power spectrum is calculated, we then able to apply mel-spaced filterbank to the power spectrum to obtain the sum of energy in each filter, which then we can obtain the logarithm of all filterbank energies to obtain the Discrete Cosine Transform of the respective log energies which results in features called mel frequency cepstral coefficients.

3.2.2 Speech Embeddings

3.2.2.1 X-Vector Speech Embedding

After MFCC has been gathered, the feature vector from the MFCC will be then goes speech activity detection algorithm to filter out empty frames (no speech activity frames) and then the filtered frames will be used as the input for x-vector speech embedding model. DNN architecture to extract the x-vector is as follows:

Table 1: X-Vector DNN Architecture

Layer	Layer context	Total context	Input x output
frame1	[t-2, t+2]	5	120x512
frame2	{t-2, t, t+2}	9	1536x512
frame3	{t-3, t, t+3}	15	1536x512
frame4	{t}	15	512x512
frame5	{t}	15	512x1500
stats pooling	[0, T)	*T*	1500*T*x3000
segment6	{0}	*T*	3000x512
segment7	{0}	*T*	512x512
softmax	{0{	*T*	512x*N*

The input segment for this DNN has T frames. The first 5 layers operate on speech frames, with a centered small temporal context in the current frame denoted as t. For example, the input for layer frame3 is linked to the output of layer frame2, at frames t-3 and t+3. This operation will create a temporal context from the previous layer, so frame3 will have 15 contexts. The statistics pooling layer aggregates all T frame-level outputs of frame5 and computes its mean and standard deviation. Statistics are 1500 dimensional vectors, which are computed once for each input segment. This process aggregates information on the time dimension so that the next layer operates on all segments. The mean and standard deviation are combined and propagated into segment-level layers and finally to the softmax output layer using the activation function rectified linear unit (ReLU).

3.2.2.2 D-Vector Speech Embedding

Similar to how x-vector speech embedding works, the input for d-vector speech embedding model will be MFCC feature vector. One of the main differences between x-vector and d-vector is x-vector utilizes DNN for the model whereas d-vector utilizes LSTM network for its model. D-vector speech embedding utilizes sliding window algorithm where the audio will be partitioned into many parts of 3 seconds audio with the slide or hop length of 1 second, which then each partition will undergo MFCC feature extraction then will be fed to LSTM network. Embedding vectors from the LSTM network then will undergo normalization using formula as shown in Eq. (3):

$$e_{ji} = \frac{f(x_{ji}; w)}{||f(x_{ji}; w)||_2} \quad (3)$$

e_{ji} represents embedding vector speaker j from i utterance that has been normalized using L2 normalization, where x_{ji} is the extracted MFCC feature for speaker j on i utterance and $f(x_{ji};w)$ represents the output of the LSTM neural network where w represents all of the parameter in the neural network (including linear layer).

3.3 Modelling

In order to identify the speaker, *ResNet50* model will be used to determine which of the registered speakers have the closest characteristic of the test voice. *ResNet50* model architecture is as follows:

Table 2: ResNet50 Architecture

Layer Name	Layer
conv1	7x7, 64, stride 2
pool1	3x3, max pool, stride 2
conv2_x	[1x1, 64 3x3, 64 1x1, 256] x 3
conv3_x	[

	1x1, 128 3x3, 128 1x1, 512] x 4
conv4_x	[1x1, 256 3x3, 256 1x1, 1024] x 6
conv5_x	[1x1, 512 3x3, 512 1x1, 2048] x 3
fc1	9x1, 2049, stride 1
apool	1xN, avg pool, stride 1
fc2	1x1, 2

The output from the proposed *ResNet50* is a multi-class, which is one of the registered speaker's label.

3.4 Evaluation

Since the speaker recognition system outputs a multi-class, we utilize multi-class confusion matrix to analyze system's performance on accepting and rejecting voices. We use true positive (TP), false positive (FP), false negative (FN), true negative (TN), accuracy, recall, precision, and f-1 score for the calculation.

4. RESULTS

Our goal in this experiment is to create a better speaker recognition system that is able to identify speakers with text independent utterances using part of VoxCeleb2 dataset. Table 2 shows the partition of the dataset used in this experiment. Our experiment begins with training the model with 70% total utterances for each speaker, validating the model with 20% total utterances for each speaker, and the rest 10% is for testing the model. We also only take speakers who have more than 25 utterances and less than 40 utterances to minimize error caused by number of utterance imbalance between speakers.

Table 2: Voice Data Distribution

Purpose	# of speakers	# of utterances
Training	552	12207
Validation		3618
Testing		2068
		17893

4.1 X-Vector and ResNet50 Evaluation

Validation accuracy for ResNet50 model and x-vector speech embedding can be seen in Figure 2. Testing prediction output from the model can be turned into a multi-class confusion matrix that provides the necessary information in order to calculate accuracy, recall, precision, and f-1 score that are needed to evaluate the model performance. Accuracy, average recall, precision, and f-1 score for ResNet50 model with x-vector speech embedding feature testing are 74.12%, 73.42%, 78.02%, and 0.72.

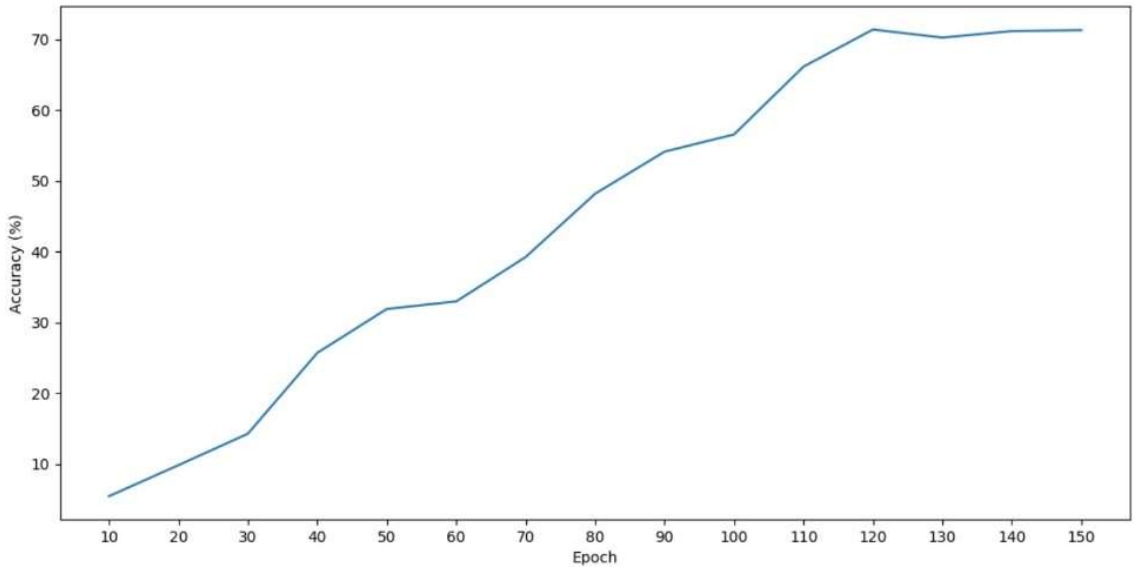

Figure 2: Model with x-vector validation accuracy for 150 epochs

4.2 D-Vector and ResNet50 Evaluation

Validation accuracy for ResNet50 model and d-vector speech embedding can be seen in Figure 3. Testing prediction output from the model can be turned into a multi-class confusion matrix that provides the necessary information in order to calculate accuracy, recall, precision, and f-1 score that are needed to evaluate the model performance. Accuracy, average recall, precision, and f-1 score for ResNet50 model with d-vector speech embedding feature testing are 36.17%, 35.77%, 36.86%, and 0.32.

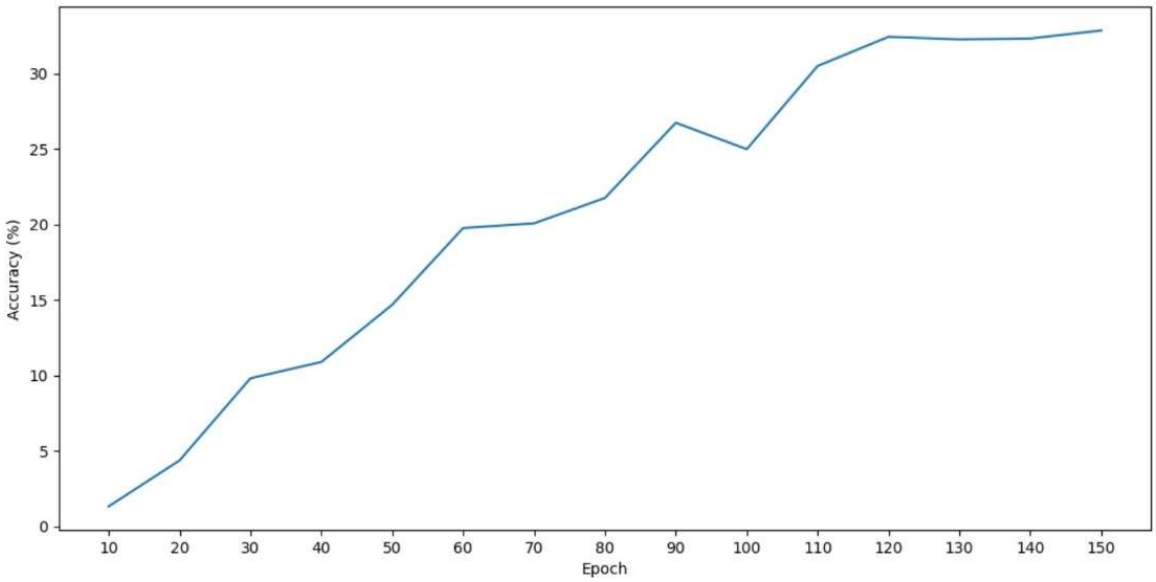

Figure 3: Model with d-vector validation accuracy for 150 epochs

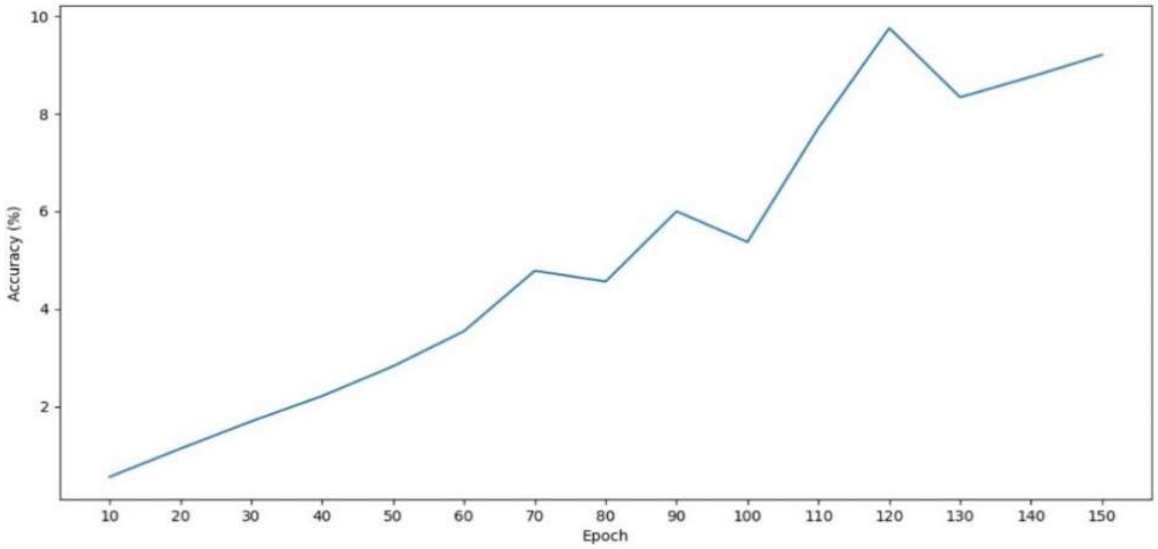

Figure 4: Model with MFCC validation accuracy for 150 epochs

4.2 MFCC and ResNet50 evaluation

Validation accuracy for ResNet50 model and MFCC can be seen in Figure 4. Testing prediction output from the model can be turned into a multi-class confusion matrix that provides the necessary information in order to calculate accuracy, recall, precision, and f-1 score that are needed to evaluate the model performance. Accuracy, average recall, precision, and f-1 score for ResNet50 model with MFCC feature testing are 10.97%, 10.61%, 6.64%, and 0.07.

5. Comparison to Previous Work

Compared to the previous work that has been done by [6], our deep learning model that utilizes x-vector speech embeddings with ResNet50 yields a much better results in all aspects (accuracy, average recall, precision, and f-1 score). This basically means that our proposed model is better at recognizing valid voices, recognizing invalid voices, and have lower error rate for recognizing voices compared to the current state-of-the-art model. The drastic or significant comparison of model quality can be determined with 2 possible reasons, which are the difference of the number of quantity used for the experiment and how the feature is extracted and then fed into the model (as explained in chapter 3.2.2). This experiment is using roughly 2% of the original dataset used by [6] which can greatly affect the model performance.

Aside from the accuracy performance, our proposed model is running significantly longer than the model proposed by [6] due to the extra steps needed in the feature extraction step.

6. Conclusion

Our study focuses on implementing x-vector and d-vector speech embedding into the current state-of-the-art speaker recognition model. From the results discussed in Section 5, it can be concluded that the proposed method with x-vector speech embedding is better than the model with d-vector speech embedding and the current state-of-the-art. One of the main reasons why x-vector speech embedding achieved better performance compared to d-vector speech embedding is mainly because utilization of speech activity detection, in which d-vector is lacking of. There are several reasons on why our proposed model, both x-vector and d-vector speech embeddings, achieves much better result. Which are the number of quantity used for the experiment and how the features from the audio is extracted.

REFERENCES

[1] S. Kumar, P. Tiwari, M. Zymbler, Internet of things is a revolutionary approach for future technology enhancement: a review, *Journal of Big Data 6*, 2019

[2] B. H. Prasetio, D. Syauqy, Design of speaker verification using dynamic time warping (dtw) on graphical programming for authentication

process, *Journal of Information Technology and Computer Science 2*, 2017.

[3] B. Soewito, Y. Marcellinus, Iot security system with modified zero knowledge proof algorithm for authentication, *Egyptian Informatics Journal*, 2020.

[4] M. Faundez-Zanuy, Biometric security technology, *Aerospace and Electronic Systems Magazine, IEEE*, vol. 21, no. 1 2006.

[5] F. Orság, Speaker recognition in the biometric security systems, *Computing and Informatics*, vol. 25, 2006, pp. 369–391.

[6] J. S. Chung., A. Nagrani., & A. Zisserman. (2018). "VoxCeleb2: Deep Speaker Recognition", INTERSPEECH.

[7] Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). "Deep Neural Network Embeddings for Text-Independent Speaker Verification". Interspeech.

[8] Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018). "Generalized end-to-end loss for speaker verification". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.

[9] Shon, S., Tang, H., & Glass, J. (2018). "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model". IEEE.

[10] Bunrit, S., Inkian, T., Kerdprasop, N., & Kerdprasop, K. (2019). "Text-independent speaker identification using deep learning model of convolution neural network". International Journal of Machine Learning and Computing.

[11] Liu, J. C., Leu, F. Y., Lin, G. L., & Susanto, H. (2018). "An MFCC-based text-independent speaker identification system for access control". Concurrency and Computation: Practice and Experience.

[12] Jahangir, R., Teh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., ... & Ali, I. (2020). "Text-independent speaker identification through feature fusion and deep neural network". IEEE Access.

[13] Ji, R., Cai, X., & Xu, B. (2018). "An End-to-End Text-Independent Speaker Identification System on Short Utterances". Interspeech.

[14] Soleymanpour, M., & Marvi, H. (2017). "Text-independent speaker identification based on selection of the most similar feature vectors". International Journal of Speech Technology.

[15] Ahmad, K. S., Thosar, A. S., Nirmal, J. H., & Pande, V. S. (2015). "A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network". 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR). IEEE.

[16] Desylvia, S. N., Buono, A., & Silalahi, B. P. (2017). "Modeling text independent speaker identification with vector quantization". Telkomnika,