International Journal on Advanced Science Engineering Information Technology

Vol.14 (2024) No. 2
ISSN: 2088-5334

Improvement Model for Speaker Recognition using MFCC-CNN and Online Triplet Mining

Ayu Wirdiani [a,*], Steven Ndung'u Machetho [b], I Ketut Gede Darma Putra [a], Made Sudarma [c], Rukmi Sari Hartati [c], Henrico Aldy Ferdian [a]

[a] *Information Technology, Udayana University, Kampus Bukit Jimbaran, Badung, 80361, Indonesia*
[b] *Computer Science Department, University of Groningen, Nijenborgh 9, 9747 AG Groningen, Netherlands*
[c] *Electrical Engineering, Udayana University, Kampus Bukit Jimbaran, Badung, 80361, Indonesia*
Corresponding author: [*]ayuwirdiani@unud.ac.id

***Abstract*—Various biometric security systems, such as face recognition, fingerprint, voice, hand geometry, and iris, have been developed. Apart from being a communication medium, the human voice is also a form of biometrics that can be used for identification. Voice has unique characteristics that can be used as a differentiator between one person and another. A sound speaker recognition system must be able to pick up the features that characterize a person's voice. This study aims to develop a human speaker recognition system using the Convolutional Neural Network (CNN) method. This research proposes improvements in the fine-tuning layer in CNN architecture to improve the Accuracy. The recognition system combines the CNN method with Mel Frequency Cepstral Coefficients (MFCC) to perform feature extraction on raw audio and K Nearest Neighbor (KNN) to classify the embedding output. In general, this system extracts voice data features using MFCC. The process is continued with feature extraction using CNN with triplet loss to obtain the 128-dimensional embedding output. The classification of the CNN embedding output uses the KNN method. This research was conducted on 50 speakers from the TIMIT dataset, which contained eight utterances for each speaker and 60 speakers from live recording using a smartphone. The accuracy of this speaker recognition system achieves high-performance accuracy. Further research can be developed by combining different biometrics objects, commonly known as multimodal, to improve recognition accuracy further.**

***Keywords*— KNN; MFCC; ResCNN; speaker recognition; triplet loss.**

Manuscript received 15 Aug. 2023; revised 9 Sep. 2023; accepted 12 Mar. 2024. Date of publication 30 Apr. 2024.
IJASEIT is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.

I. INTRODUCTION

The need for digital data security is growing along with technological advances. Various digital data security systems have been created to enhance data security, such as using an encrypted password-based encryption system in the authentication process. Passwords have a higher level of protection if they use a combination of numbers, letters, and characters long enough. Humans have been proven only to be able to remember short and easy passwords [1]. In addition, passwords are vulnerable to theft and loss. A better security system needs to be developed so that it can recognize everyone's characteristics. One solution that can be applied to overcome this is to use biometric technology. Biometrics is the study of distinctive patterns that may be used to distinguish or identify people based on one or more aspects of their physical or behavioral makeup, such as their faces, fingerprints, voices, hand geometry, palmprints, or iris [2], [3], [4], [5]. Recognition with this method is more reliable when compared to conventional methods because apart from being easy to process, it is also hard to falsify.

Identification and verification are two crucial tasks carried out by recognition using a biometric system. A person's identity is to be ascertained via an identifying system. The verification system seeks to accept or reject the asserted identification of someone [6]. Generally, biometric technology works by using pattern recognition techniques or pattern recognition. The patterns learned typically are fingerprint patterns, palm lines, faces, irises, and voice recognition [7].

Apart from being a communication medium, the human voice is also a form of biometrics that can be used for identification [8], [9]. Voice has unique characteristics that can be used as a differentiator between one person and another. A sound speaker recognition system must be able to pick up the features that characterize a person's voice. These features represent and describe the voice signal [10]. Human

speech patterns can be recognized using algorithms such as Mel Frequency Cepstral Coefficients (MFCC) [11].

Modern technology development has given rise to various new methods of pattern recognition, one of which is the Convolutional Neural Network (CNN). This method has the advantage of using input in the form of raw waveforms, and all parameters can be learned by the algorithm in the training process to adapt to specific use case studies.

This research was conducted by Rashid Jahangir using a new combination of MFCC and time domain (MFCCT) features. This combination improves the accuracy of text-independent speaker recognition (SI) systems. These features are used as input to a deep neural network (DNN) to create a speaker recognition model. This study uses male and female voices of the LibriSpeech dataset. The accuracy of this research system is 89%. [12]. Other speech recognition research uses the LDA, MLLT, and MFCC algorithms for feature extraction and the CNN method. The training dataset comes from the TID corpus of 8623 utterances uttered by 111 male and 114 female adult speakers, while the test data used is from the Aurora-5 corpus. This corpus comprises speech data collected from 24 speakers at the International Institute of Computer Science at Berkeley, resulting in 2400 utterances for each microphone. The results obtained from this study show that the use of LDA and MLLT transformations can reduce the relative error by 20% compared to DNN-based speech recognition, which uses the same number of hidden layers [13].

Jagiasi et al. [14] also researched human speech recognition with text-free voice samples using the Convolutional Neural Network (CNN) method. Text-independent samples are recorded voice samples that are not limited to the pronunciation of specific words or sentences. This study features an extraction process for voice samples combined with the Mel Frequency Cepstral Coefficients (MFCC) and CNN methods. A sampling rate of 44.1 kHz is used to optimize voice quality and provide better results. The CNN architecture consists of 3 convolution layers and one fully connected layer. The results obtained from this study are that using CNN for 5-10 voice classes recorded in real conditions provides an accuracy of 75 to 80 percent. As for voice samples taken in laboratory conditions with 50 classes, an accuracy of up to 70 percent was obtained.

A study by [15] also researched human speech recognition using K-Nearest Neighbor (KNN), MFCC, and Formant. The KNN method is used to classify objects based on training data with the closest distance value to the tested object. MFCC and Forman are used for feature extraction. Voice sample data was obtained from 5 speakers, three males and two females. Ten voice samples were taken from each person with the pronunciation of "login." This study found that the KNN and Formant methods produced an accuracy of 85 percent. Another study used MFCC and CNN to perform speech emotion recognition, where the accuracy obtained was 83.69% [16]. A study by [17] proposed MFCC for feature extraction and Multi-Layer Perceptron (MLP) to automatically recognize speakers. She uses an IITG Multivariability Speaker Recognition database as a trained and tested dataset. This proposed method provides an accuracy of 94.44 %. Based on the above background, this study aims to develop a speaker recognition algorithm to determine a person's identity using Mel-frequency Cepstral Coefficients and Convolutional Neural Networks. Algorithm development uses transfer learning from previously studied speech recognition models. This model was developed by making improvements to the previously developed architecture. The results of this study are expected to be the basis for related research in developing speech recognition systems to obtain better accuracy.

II. Materials and Methods

A. Dataset

This research used a spectrogram to train learning models to perform voice recognition. The dataset used is the Texas Institute/Massachusetts Institute of Technology (TIMIT) speech corpus to refine the model. TIMIT's speech corpus was designed in 1993 as a data source for speech and acoustic-phonetic studies and has been used for various studies [18]. TIMIT contains recordings from 630 speakers speaking American English. The corpus includes orthography, phonetics, time-aligned word transcriptions, and 16-bit 16 kHz speech waveform files for each utterance, where each speaker has eight recorded audio waveforms [19].

This research improved the model by using 50 speaker classes from the TIMIT database, 60 speakers from live recordings using a smartphone, and eight audio waves. The data used is divided into two parts: 5 audios as a training dataset and three as a test dataset to evaluate the accuracy and performance of the model. These voice samples are stored in the training and test folders. The data assumptions used in this research were carried out in healthy conditions.

B. Proposed Method

This section will describe the outline of the method used in this research. This research generally uses a combination of MFCC and CNN as feature extractors and KNN as a classification algorithm. The speaker recognition process begins by preprocessing the raw audio to remove noise and silence in the raw waveform. The process is continued by extracting the MFCC feature from the voice signal, which produces the MFCC feature output, a 2-dimensional matrix measuring n x 64. The output feature length will depend on a filtered raw waveform. This feature is used as input to the ResCNN feature extraction model, which accepts an input dimension of 160 x 64. The MFCC output will be resized to 160 x 64 to cope with the input dimension by adding silence or selecting random samples. The production of the ResCNN model is fine-tuned to improve recognition accuracy with triplet loss and calculated using online triplet mining to optimize the weight value of the ResCNN feature extraction model. The output of this model is a 1-dimensional matrix measuring 128, which is used as base data for speaker classification using the KNN method. The production of the KNN model is a single class of the recognized voice.

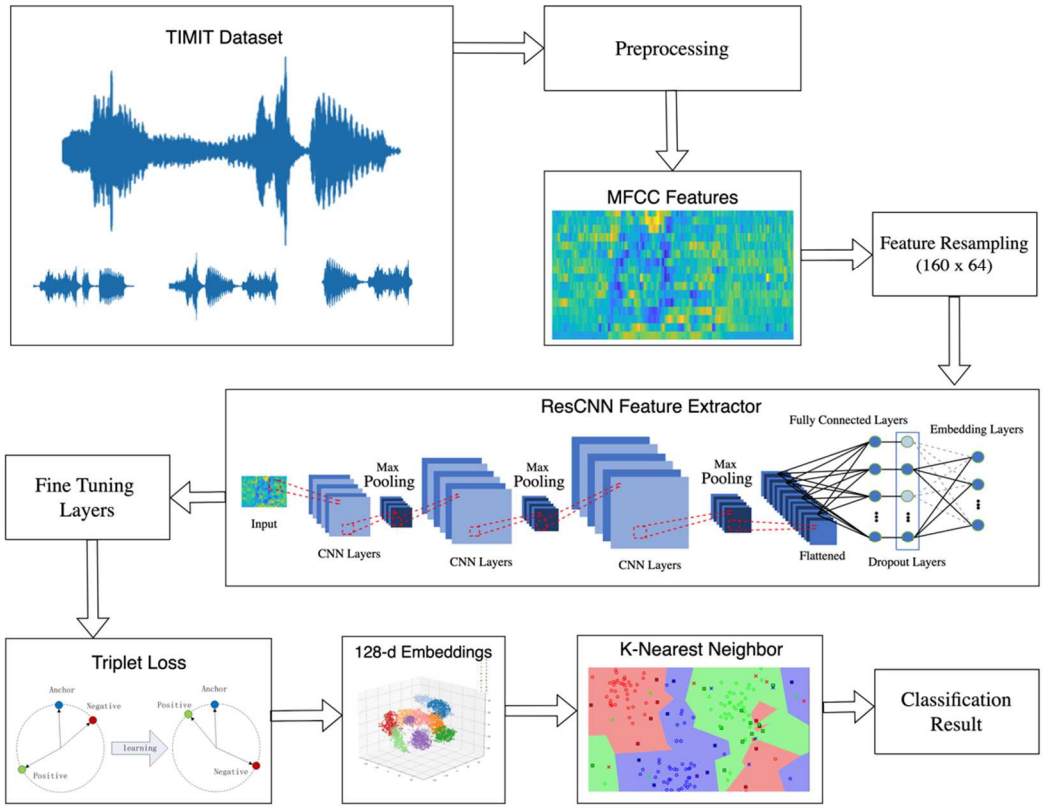

Fig. 1 Block Diagram of the Proposed Approach

C. *Preprocessing*

All the raw audio data is preprocessed before the MFCC feature extraction process. Preprocessing removes noise and silence from the audio spectrum by calculating the 95th percentile to retrieve the noise threshold. The percentile calculation can be described with the following formula.

$$R = \frac{P}{100}(N) \tag{1}$$

where R is the percentile rank, P is the desired percentile, and N is the number of data points. The audio spectrum that falls between these thresholds will be removed. This method can reduce the hiss and hum generated by the surroundings. [17].

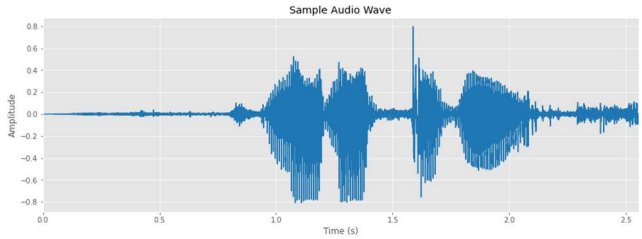

Fig. 2 Raw Audio Spectrum

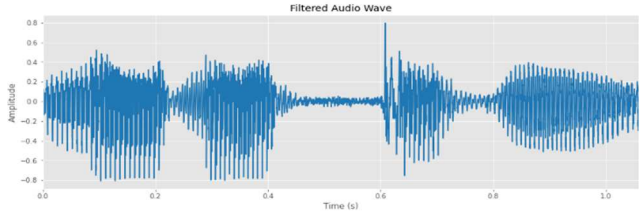

Fig. 3 Filtered Audio Spectrum

Fig. 2 shows the raw audio spectrum before preprocessing occurs. The duration of the raw waveform is 2.6 seconds, with the noise and silence part at the start and end of the waveform. These parts will be removed in preprocessing to filter out the necessary parts of the human voice. Fig. 3 shows the filtered audio spectrum, with the noise and silence part removed. The duration of the filtered spectrum will vary depending on the noise level of the raw waveform. The filtered spectrum will be used as an input for the MFCC feature extraction process.

D. *Mel Frequency Cepstral Coefficients*

Mel-frequency Cepstral Coefficients (MFCC) is a technique widely used in audio feature extraction processes. This is due to variations in the known human ear frequency bandwidth. This coefficient is created by embellishing the filter bank's output log energy, which comprises triangle filters positioned linearly on the Mel frequency scale [20], [21]. The MFCC coefficient is obtained by decorating the output log energy of a filter bank consisting of triangular filters, which are linearly spaced on the Mel frequency scale [22].

The audio is passed through a filter that can increase the frequency. The voice signal is divided into small parts known as frames. Each frame is multiplied with the Hamming window in the signal join order. Next, the Fast Fourier Transform (FFT) process is performed on each frame to get the frequency scale. The selected feature will be reduced by multiplying the frequency with a triangular band filter [23].

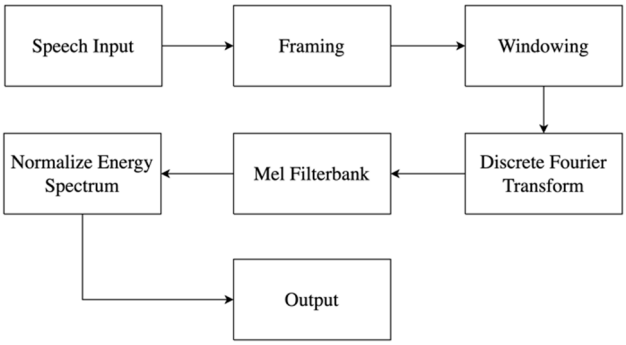

Fig. 4 Block Diagram of MFCC Feature Extraction

This research uses MFCC as a first-stage feature extraction method to extract features from raw waveforms. It accepts input from any dimension of the raw waveform and gives the output a two-dimensional array of size n x 64. The output size will depend on the duration of the raw waveform. This feature will be used for further extraction with the ResCNN deep learning method.

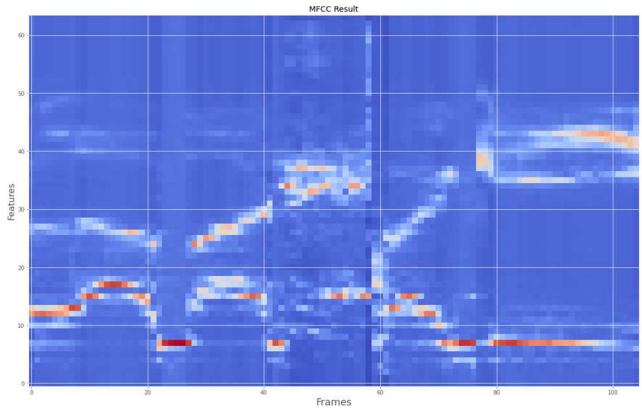

Fig. 5 MFCC Feature

E. ResCNN Architecture

This research uses the Residual Convolutional Neural Network (ResCNN) architecture, which Li researched as the second stage of the feature extraction method. This network will be connected to a fully connected neural network with a 128-dimensional embedding output [9]. The ResCNN model used in this research was trained on the LibriSpeech ASR research by John Hopkins University, which contained 2484 speakers and 1000 speech hours [24]. This model is trained with a SoftMax activation function and triplet loss at the output layer.

In this research, architectural development is carried out by modifying the architectural model by adding two new fully connected layers. The developed architecture freezes the first 28 layers to be non-trainable in the training process. This is an integral part of feature extraction, and the refinement is done in the embedding part. Table 1 lists the details of the development of the ResCNN architecture that we made in this study.

As described in Table 1, two fine-tuning layers are added in this research. This layer is fully connected, with dimensions of 512 x 256 and 256 x 128. We also use L2 normalization to calculate the vector coordinate's distance from the vector space's origin. The 128-dimension output from the network will be fine-tuned using triplet loss with online triplet mining.

TABLE I
RESCNN ARCHITECTURE

Layer	Structure	Stride	Dim	#Params
Konv64-s	5x5, 64	2x2	2048	6K
res64	3x3, 64 3x3, 64	1x1	2048	41K x 2
res64	3x3, 64 3x3, 64	1x1	2048	41K x 2
res64	3x3, 64 3x3, 64	1x1	2048	41K x 2
Konv128-dtk	5x5, 128	2x2	2048	209K
Res128	3x3, 128 3x3, 128	1x1	2048	151K x 2
Res128	3x3, 128 3x3, 128	1x1	2048	151K x 2
Res128	3x3, 128 3x3, 128	1x1	2048	151K x 2
Konv256-dtk	5 x 5, 256	2x2	2048	823K
Res256	3x3, 256 3x3, 256	1x1	2048	594K x 2
Res256	3x3, 256 3x3, 256	1x1	2048	594K x 2
Res256	3x3, 256 3x3, 256	1x1	2048	594K x 2
Konv512-s	5 x 5, 512	2x2	2048	3.3M
Res512	3x3, 512 3x3, 512	1x1	2048	2,4M x 2
Res512	3x3, 512 3x3, 512	1x1	2048	2,4M x 2
Res512	3x3, 512 3x3, 512	1x1	2048	2,4M x 2
Average	-	-	2048	0
affine	2048 x 512	-	512	1M
iFinetune-1	512x256	-	256	65K
Finetune-2	256x128	-	128	32K
L2-Norma	-	-	128	0
Total				25M

F. Online Triplet Mining

The approach used in this research will incorporate neural network embeddings for classification purposes. A matrix with 128-dimensional floating-point values in Euclidean space represents the embedding. To fine-tune the embedding space, we will use triplet loss with online triplet mining to generate pairs of data to be fed into our network. Triplet loss minimizes the distance between anchors and positive points with the same class label and maximizes the distance between anchors and negative points with different class labels [25] [26].

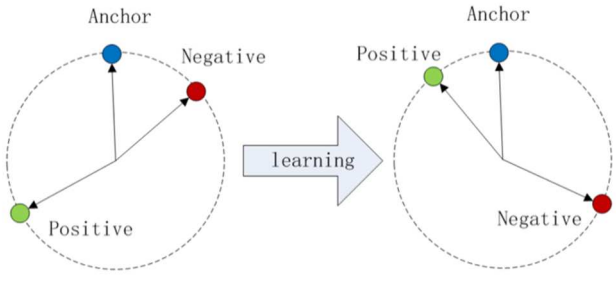

Fig. 6 Triplet Loss

The triplet loss function above can be described as a Euclidean Distance function as follows:

$$\mathcal{L}(A,P,N) = \max\left(\|f(A) - f(P)\|^2 - \|f(A) - (N)\|^2 + \alpha, 0\right) \quad (2)$$

where A is the anchor input, P is the positive sampled input, N is the negative sampled input, and alpha is some margin used to determine when a triplet pair becomes too "easy" and there is no need to adjust the weight values of the artificial neural network [27] [28].

III. Result and Discussion

A. Training Result

The training process for the Convolutional Neural Network (CNN) model is carried out by providing input data resulting from extracting the Mel Frequency Cepstral Coefficients (MFCC) feature from 550 sound samples divided into 110 speaker classes. The output of the MFCC feature extraction process is a 2-dimensional matrix measuring 160 x 4 which is further processed using the CNN feature extraction method. The CNN model training process was carried out for 25 epochs using the triplet loss parameter and Adam's optimizer.

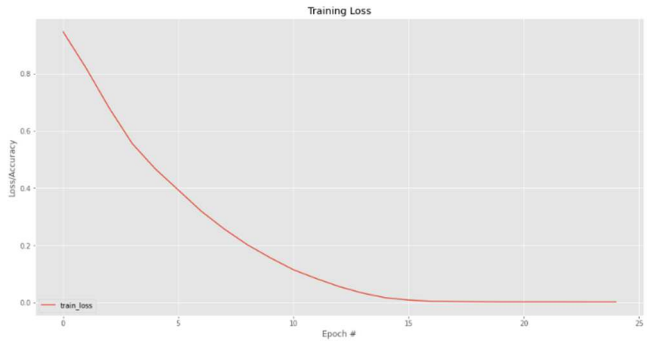

Fig. 7 Training Loss per Epochs

Figure 7 shows the results of the ResCNN model training in a loss value matrix for each epoch, visualized in a line chart. The figure shows that the loss value for each epoch decreases significantly and gets closer to zero, and it can be concluded that the learning model is quite good on the test dataset [29].

B. Embedding Output

To determine the success rate of the ResCNN model in learning, a comparison process of embedding results before and after training was conducted to assess the convergence of embedding in each class of speakers. The result of embedding this speech recognition architecture is 128 dimensions, which are further processed using the t-SNE algorithm to visualize it in 2-dimensional space.

Fig 8 shows a 128-dimensional embedding visualization of CNN model output before fine-tuning. Embedding has been processed using the t-SNE dimensionality reduction algorithm to be visualized in 2-dimensional space. In the figure, classes are still spread over a vast space. This could reduce the classification accuracy, as the close embedding between different courses can be challenging to classify. Embedding is further optimized to converge through the training process using triplet loss.

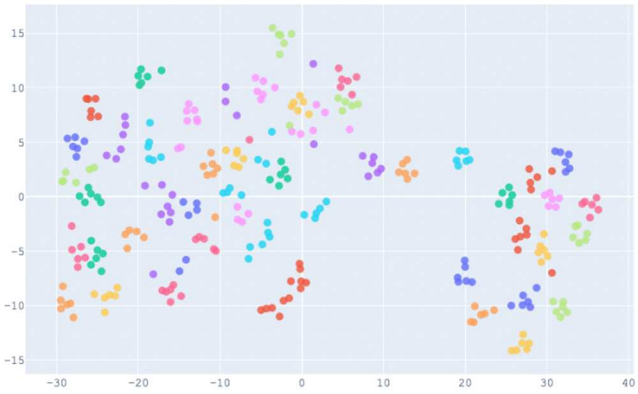

Fig. 8 Embedding before Fine Tuning with Triplet Loss

Fig 9 shows a 128-dimensional embedding visualization of the CNN model after the fine-tuning process. The embedding has successfully converged and approached the same class in the figure. The distance between the same classes is significantly reduced, and the distance between opposite classes is further maximized. The optimized embedding is further processed for classification using the K-nearest neighbor (KNN) algorithm.

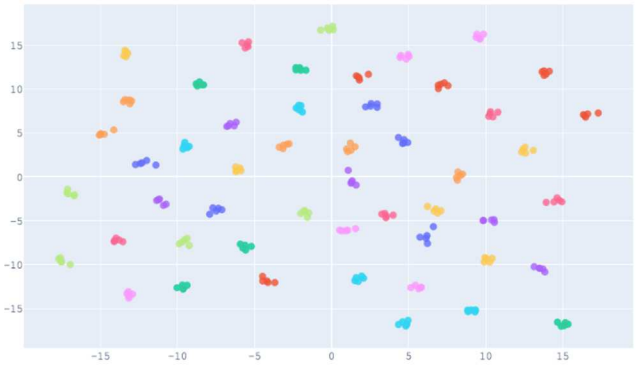

Fig. 9 Embedding after Fine Tuning with Triplet Loss

C. Classification Result

The classification process is carried out using the K-Nearest Neighbor (KNN) method on the embedding output provided by the ResCNN model. The classification result matrix is presented as a confusion matrix, as shown in Figure 10. Fig 10 shows the confusion matrix of classification results on our test datasets. The figure shows that, from 50 test data, the recognition model managed to classify most of the classes correctly.

We can gather information on our model performance from this confusion matrix by measuring its accuracy, precision, recall, and f1 score with the following formula [30] [31].

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (6)$$

where TP is the prediction of the class True Positive, TN True Negative, FN False Negative, and FP False Positive.

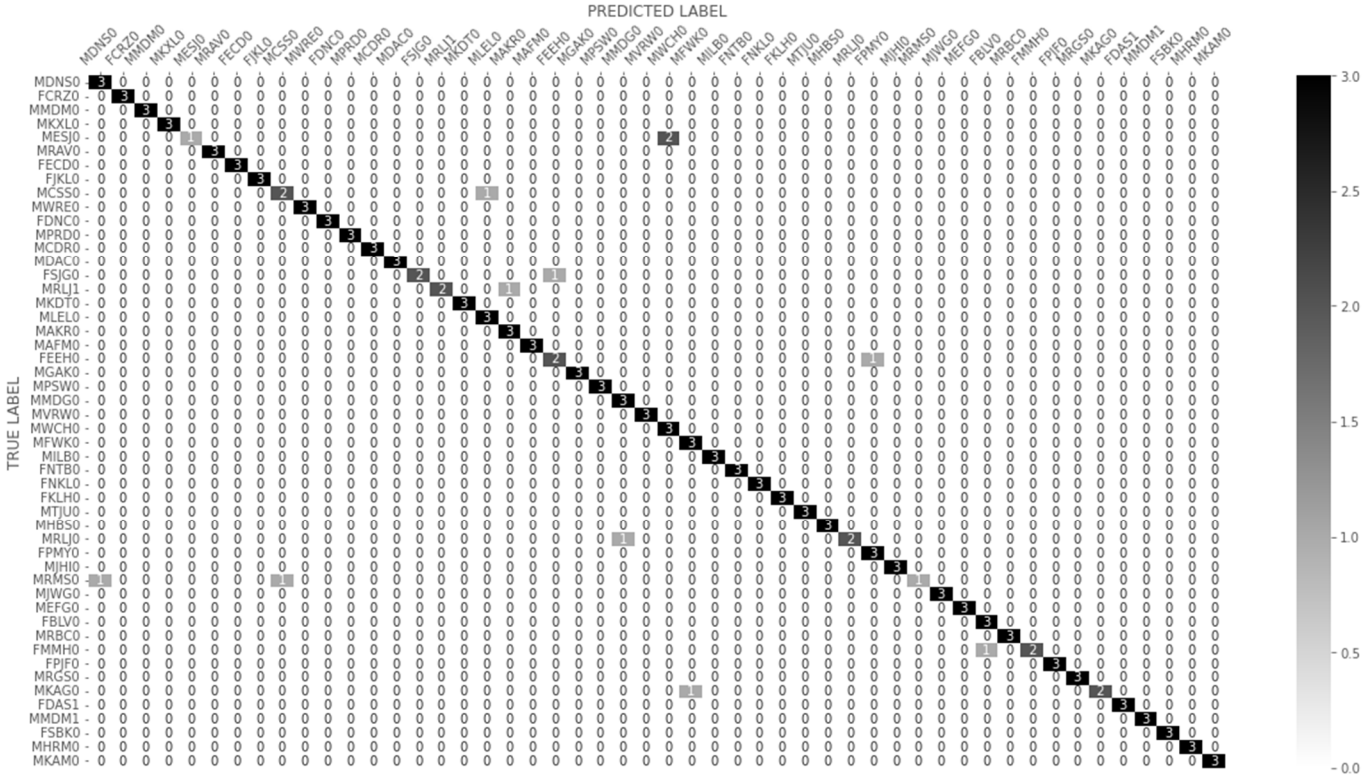

Fig. 10 Confusion Matrix of Classification Result on Test Dataset

This research uses testing scenarios to recognize TIMIT test data without fine-tuning and fine-tuning ten epochs, 15 epochs, 20 epochs, and 25 epochs. The second scenario is live speaker recognition with recording from a smartphone. The test results are compared to find out the best scenario of the epoch value and the effect of the fine-tuning process on the accuracy of results.

TABLE II
THE COMPARISON TEST RESULT OF THE TIMIT DATASET

No.	Test Scenario	Accuracy	Precision	Recall	f1 Score
1	Without fine-tuning	89,33%	88,45%	89,33%	88,92%
2	Fine-tuning 10 Epoch	89,33%	93,25%	89,33%	88,71%
3	Fine-tuning 15 Epoch	92%	94%	92%	91,73%
4	Fine-tuning 20 Epoch	90%	91,33%	90%	88,93%
5	**Fine-tuning 25 Epoch**	**94.66%**	**94%**	**94.66%**	**93.51%**

Table II is a comparison test result of the speaker recognition system using the TIMIT dataset. The model with fine-tuning of 25 epochs has the highest accuracy, recall, and f1 score with an accuracy value of 94.66 percent, a recall of 94 percent, and an f1 score of 93.51 percent. A high precision value also indicates that the model has good recognition accuracy. A good recall value suggests that the model can recognize the class well. Meanwhile, a high f1 score can indicate a pretty good alignment between precision and recall values [32]. The results of this test showed that the fine-tuning process can significantly improve the performance of the recognition model using a parameter of 25 epochs.

The second scenario tests live speaker recognition using a smartphone recording and a model with 25 epochs.

TABLE III
COMPARISON TEST RESULT OF LIVE SPEAKER RECOGNITION

No.	Test Scenario	Accuracy	Precision	Recall	f1 Score
1	Without fine-tuning	89,16%	92,10%	89,16%	88,47%
2	**Fine-tuning**	**96%**	**96.74%**	**95.99%**	**95.78%**

Tables II and III show that architecture improvements can improve recognition systems' accuracy.

D. Comparison of Result with Previous Studies

Comparison with previous research was conducted to find out and compare the performance of the recognition model from the results of research conducted with previous research. Comparisons were made with the research referred to in the introduction section. The results of the comparison can be seen in Table IV.

TABLE IV
COMPARISON OF RESULTS WITH PREVIOUS STUDIES

No.	Method	Author	Dataset	Conditions	Accuracy
1	MFCC-MLP	K. J. Devi, A. A. Devi, and K. Thongam [17]	IITG-MV phase-IV Part-III	Different sentences	94.44 %
2	MFCC – CNN – DNN	R. Jagiasi, S. Ghosalkar, P. Kulal, and A. Bharambe [14]	Develop 50 classes with lab conditions and ten classes with actual conditions	Records of 50 class lab conditions	70 %

No.	Method	Author	Dataset	Conditions	Accuracy
3	MFCC – ResCNN – KNN	This Research	TIMIT 50 classes, 8 for each class, and live condition record 60 class	Real live condition record10 class	75%
				Different sentences	94.66 %
				Live condition record 60 class	96%

Based on Table IV, a combination of the MFCC method for feature extraction and ResCNN architectural improvement can provide high accuracy.

IV. Conclusion

The speaker recognition system developed in this research achieved a high accuracy of 96%. This result is strongly influenced by the improvements made to the ResCNN model to help extract further voice spectrum from the MFCC feature extraction process and the combination of triplet loss used in the recognition process. Additional research can be carried out by combining different biometrics objects commonly known as multimodal, to be able to improve recognition accuracy further.

References

[1] S. H. Moi *et al.*, "An Improved Approach to Iris Biometric Authentication Performance and Security with Cryptography and Error," *Int. J. Informatics Vis.*, vol. 6, no. August, pp. 531–539, 2022.

[2] V. M. Arun Ross, Sudipta Banerjee, Cunjian Chen, Anurag Chowdhury, "Some Research Problems in Biometrics: The Future Beckons," in *IAPR International Conference on Biometrics (ICB)*, 2019.

[3] C. Medjahed, A. Rahmoun, C. Charrier, and F. Mezzoudj, "A deep learning-based multimodal biometric system using score fusion," *IAES Int. J. Artif. Intell.*, vol. 11, no. 1, pp. 65–80, 2022, doi:10.11591/ijai.v11.i1.pp65-80.

[4] I. K. G. D. Putra, D. Witarsyah, M. Saputra, and P. Jhonarendra, "Palmprint Recognition Based on Edge Detection Features and Convolutional Neural Network," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 11, no. 1, pp. 380–387, 2021, doi: 10.18517/ijaseit.11.1.11664.

[5] R. Blanco-Gonzalo *et al.*, "Biometric Systems Interaction Assessment: The State of the Art," *IEEE Trans. Human-Machine Syst.*, vol. 49, no. 5, pp. 397–410, 2019, doi: 10.1109/THMS.2019.2913672.

[6] A. Pradhan, J. He, and N. Jiang, "Score, Rank, and Decision-Level Fusion Strategies of Multicode Electromyogram-Based Verification and Identification Biometrics," *IEEE J. Biomed. Heal. Informatics*, vol. 26, no. 3, 2022, doi: 10.1109/JBHI.2021.3109595.

[7] S. Shakil, D. Arora, and T. Zaidi, "An optimal method for identification of finger vein using supervised learning," *Meas. Sensors*, vol. 25, Feb. 2023, doi: 10.1016/j.measen.2022.100583.

[8] A. Sithara, A. Thomas, and D. Mathew, "Study of MFCC and IHC feature extraction methods with probabilistic acoustic models for speaker biometric applications," *Procedia Comput. Sci.*, vol. 143, pp. 267–276, 2018, doi: 10.1016/j.procs.2018.10.395.

[9] D. Cai, Z. Cai, and M. Li, "Deep Speaker Embeddings with Convolutional Neural Network on Supervector for Text-Independent Speaker Recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1478–1482. doi: 10.23919/APSIPA.2018.8659595.

[10] S. Hidayat, M. Tajuddin, S. A. Alodiayusuf, J. Qudsi, and N. N. Jaya, "Wavelet Detail Coefficient as A Novel Wavelet-MFCC Features in Text-Dependent Speaker Recognition System," *IIUM Eng. J.*, vol. 23, no. 1, 2022, doi: 10.31436/IIUMEJ.V23I1.1760.

[11] A. Ashar, M. S. Bhatti, and U. Mushtaq, "Speaker Identification Using a Hybrid CNN-MFCC Approach," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, IEEE, Mar. 2020, pp. 1–4. doi: 10.1109/ICETST49965.2020.9080730.

[12] R. Jahangir *et al.*, "Text-Independent Speaker Identification through Feature Fusion and Deep Neural Network," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2973541.

[13] H. F. Pardede, A. R. Yuliani, and R. Sustika, "Convolutional Neural Network and Feature Transformation for Distant Speech Recognition," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 6, pp. 5381–5388, 2018, doi: 10.11591/ijece.v8i6.pp5381-5388.

[14] R. Jagiasi, S. Ghosalkar, P. Kulal, and A. Bharambe, "CNN based Speaker Recognition in Language and Text-independent Small Scale System," in *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, 2019, pp. 176–179. doi:10.1109/I-SMAC47947.2019.9032667.

[15] A. Maurya, D. Kumar, and R. K. Agarwal, "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach," in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 880–887. doi:10.1016/j.procs.2017.12.112.

[16] F. Reggiswarashari and S. W. Sihwi, "Speech emotion recognition using 2D-convolutional neural network," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 6, pp. 6594–6601, 2022, doi:10.11591/ijece.v12i6.pp6594-6601.

[17] K. J. Devi, A. A. Devi, and K. Thongam, "Automatic Speaker Recognition using MFCC and Artificial Neural Network," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 1S, pp. 39–42, 2019, doi:10.35940/ijitee.a1010.1191s19.

[18] V. Z. John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, "TIMIT Acoustic-Phonetic Continuous Speech Corpus." 1993.

[19] J. Villalba *et al.*, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Comput. Speech Lang.*, vol. 60, p. 101026, Mar. 2020, doi: 10.1016/j.csl.2019.101026.

[20] H. R. Yulianto and Afiahayati, "Fighting COVID-19 : Convolutional Neural Network for Elevator User ' s Speech Classification Neural in Bahasa Indonesia," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 84–91. doi: 10.1016/j.procs.2021.05.079.

[21] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," *IEEE Access*, vol. 10, no. November, pp. 122136–122158, 2022, doi:10.1109/ACCESS.2022.3223444.

[22] Heriyanto, S. Hartati, and A. P. Eko, "Ekstraksi Ciri Mel Frequency Cepstral Coefficient (MFCC) dan Rerata Coefficient untuk Pengecekan Bacaan Al-Quran," *Telematika*, vol. 15, no. 02, pp. 99–108, 2018.

[23] M. Altayeb and A. Al-ghraibah, "Classification of three pathological voices based on specific features groups using support vector machine," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 1, pp. 946–956, 2022, doi: 10.11591/ijece.v12i1.pp946-956.

[24] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, and T. Johns, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[25] S. Ozturk and T. Cukur, "Deep Clustering via Center-Oriented Margin Free-Triplet Loss for Skin Lesion Detection in Highly Imbalanced Datasets," *IEEE J. Biomed. Heal. Informatics*, vol. 26, no. 9, 2022, doi: 10.1109/JBHI.2022.3187215.

[26] C. Zhang, S. Ranjan, and J. H. L. Hansen, "An Analysis of Transfer Learning for Domain Mismatched Text-independent Speaker Verification An analysis of transfer learning for domain mismatched text-independent speaker verification," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 181–186. doi:10.21437/Odyssey.2018-26.

[27] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet : A Unified Embedding for Face Recognition and Clustering".

[28] Z. Ren, Z. Chen, and S. Xu, "Triplet Based Embedding Distance and Similarity Learning for Text-independent Speaker Verification".

[29] L. Zhang, Z. Cheng, Y. Shen, and D. Wang, "Palmprint and palmvein recognition based on DCNN and a new large-scale contactless palmvein dataset," *Symmetry (Basel).*, vol. 10, no. 4, pp. 1–15, 2018, doi: 10.3390/sym10040078.

[30] H. Rahmat, S. Wahjuni, and H. Rahmawan, "Performance Analysis of Deep Learning-based Object Detectors on Raspberry Pi for Detecting Melon Leaf Abnormality," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 12, no. 2, pp. 572–579, 2022, doi: 10.18517/ijaseit.12.2.13801.

[31] I. P. A. Dharmaadi, D. Witarsyah, I. P. A. Bayupati, and G. M. A. Sasmita, "Face Recognition Application Based on Convolutional Neural Network for Searching Someone's Photo on External Storage," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 12, no. 3, pp. 1222–1228, 2022, doi: 10.18517/IJASEIT.12.3.11666.

[32] M. K. Nandwana *et al.*, "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings," no. September, pp. 1106–1110, 2018.