# Deep Learning Algorithms based Voiceprint Recognition System in Noisy Environment



# Deep Learning Algorithms based Voiceprint Recognition System in Noisy Environment

**Hajer Y. Khdier, Wesam M. Jasim, Salah A. Aliesawi,**

College of Computer Science and Information Technology, University of Anbar, Iraq.

hajer91.hajer@gmail.com
wmj_r@yahoo.com
co.wesam.jasim@uoanbar.edu.iq
salah_eng1996@yahoo.com

**Abstract**. Voiceprint Recognition (VPR) is the mechanism by which a user's so-called identity is determined using characteristics taken from their voice, where this- technique is one of the world's most useful and common biometric recognition techniques particularly the fields-relevant to security. These can be used for authentication, monitoring, forensic identification of speakers, and a variety of related activities. In this work, an attempt is applied to create a system that recognizes human speaker identity using Convolutional Neural Network (CNN). Two methods are used in this work which are MFCC-CNN and RW-CNN. The first method is standard method using MFCC, to use the features in the audio, where these features are will be entered into CNN to perform a process. The training CNN will take input as a picture and then the process of training via the proposed CNN is beginning. The second method, RW-CNN, the same steps as the first method, but without going through the MFCC phases where direct entry to CNN. In which, the same CNN structure was used in both methods. In this work, a 96% accuracy gained for both RW-CNN and MFCC-CNN. Both methods are similar in their results, either with or without noise, but the performance is mixed. This system can deep learn a large amount of human voices with high accuracy and minimum processes requirement.

**Keywords.** Convolutional Neural Network, Voiceprint Recognition System, Deep Learning.

## 1. Introduction

Machine learning (ML) technology supports the modern society in many ways. It has become more and more popular in research and has been incorporated in a large number of applications, including multimedia concept retrieval, image classification, video recommendation, social network analysis, text mining. It is also used in other applications such as cameras, smart phones, visual data processing, speech and audio processing, and many other applications[1][2].

Deep learning (DL) is a part of a wider family of machine learning methods focused on representations of the learning data, as opposed to task-specific algorithms. Compared to shallow learning deep learning has the advantage of building deep architectures to learn more abstract

information. The most important property of deep learning methods is that it can automatically learn feature representations thus avoiding a lot of time-consuming. Traditional ML relies on shallow networks which are composed of one input and one output layer, and no more than one hidden layer between input and output layers. While, DL is qualified when more than three layers exist in a network including input and output layers. Therefore, the more the number of hidden layers is increased, the more the network gets deeper[1][2].

Deep learning has been growing very fast, several new networks and new structures appear every few months, currently, some popular DL algorithms are Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Restricted Boltzmann Machine (RBM),and Autoencoders algorithms.

CNN was firstly introduced by Kunihiko Fukushima. It was later proposed by Yann LeCun. He combined CNN with back-propagation theory to recognize handwritten digits and document recognition. His system was eventually used to read hand-written checks and zip codes. CNN uses convolutional layers and pooling layers. Convolutional layers filter inputs for useful information. They have parameters that are learned so that filters are adjusted automatically to extract the most useful information for a certain task. Multiple convolutional layers are used that filter images for more and more abstract information after each layer. Pooling layers are used for limited translation and rotation invariance. Pooling also reduces memory consumption and thus allows for the usage of more convolutional layers [3][4].

Nowadays, several researchers were involved in the voiceprint recognition area of research. Some of them are; J.Lee, et al. applied two types of sample-level deep convolutional neural networks that take raw waveforms as input and uses filters with small granularity. The first one is a basic model that consists of convolution and pooling layers. The second one is an improved model that additionally has residual connections, squeeze-and-excitation modules and multi-level concatenation. It shows that the sample-level models reach state-of-the-art performance levels for the three different categories of sound. The authors also visualized the filters along layers and compared the characteristics of learned filters. The results show the possibility that they can be applied to different audio domains as a true end-to-end model [5].

Authors in [6] were proposed a SincNet, as a novel CNN that encourages the first layer to discover meaningful filters by exploiting parametrized sinc functions. In contrast to standard CNNs, which learn all the elements of each filter, only low and high cutoff frequencies of band-pass filters are directly learned from data. This inductive bias offers a very compact way to derive a customized front-end, that only depends on some parameters with a clear physical meaning. The experiments, conducted on both speaker and speech recognition. It shows that the proposed architecture converges faster, performs better, and more computationally efficient than standard CNNs. The SincNet outperforms other systems on both TIMIT (462 speakers) and Librispeech (2484 speakers). In [7] authors were studyed the convolutional neural networks (CNNs) on audio classification. They compared between CNN based wavelet and CNN based spectrogram on audio classification and they proposed a sample CNN based wavelet audio. They used three different audio domains: music, speech and acoustic scene sound. The language used in this work TensorFlowand Keras. The best performance was obtained by Sample CNNs when the Sample CNNs have the smallest filter and stride sizes, raw waveform (RW) end-to-end computational scheme for speaker identification based on CNNs with noise and reverberation data augmentation (DA). The CNN is designed for a frame-to-frame analysis to handle the variable length signals.

This paper proposed a deep learning strategy, which will provide a way to implicitly learn the voice print recognition from a raw waveform in noisy environments. For this purpose, a data set of 400 different speakers was chosen and recorded 10 samples of the different text speech from each speaker. Next, the noise on these speaker's data was added randomly. In the first step of this work, the Mel-Frequency Cepstral Coefficients (MFCC) was applied to the raw waveform to extract the features (cepstral coefficients). Then, an appropriate architecture of the DL algorithm which is in particular CNN algorithm was applied. The main obtained results from the MFCC approaches were compared with that

obtained from applying the CNN directly on the raw waveform with and without adding noise. Comparing these results was illustrated to prove the efficiency of DL on recognition of a person from his/her voice in noisy environments.

We analyze the identification performance with simulated experiments in noisy and reverberation conditions comparing the proposed RW-CNN with the mel-frequency cepstral coefficients (MFCCs) features. The results show that the method offers robustness to adverse conditions. The RW-CNN outperforms the MFCC-CNN in noise conditions, and they have similar performance in reverberant environments were introduced in [8].

The rest of the paper is organized as follows: In section 2, the voiceprint recognition and generic method for recognizing VP is demonstrated. The architectures of the proposed CNN is described in detail in Section 3. Section 4 describes the used dataset. Section 5 discusses the obtained results from the implementation of the proposed system. Section 6 concludes the detailed work on voiceprint recognition system.

## 2. Voiceprint Recognition

Voice is probably the most important mode of contact with humans. Human voice or speech is an information-rich signal that transmits a wide range of information such as language material, emotions of the speaker and tone of speech. The VPR seeks to separate, identify, and recognize a speaker based on speech characteristics. Several methods can simplify the process of speaker recognition. Such systems typically involve two phases of extraction the features and matching or classifying of the features, where the element of classification has two components: the pattern matching and the decision. Figure 1 depicts a generic voiceprint recognition system. The feature extraction module estimates a collection of speech signal features that reflect some speaker-specific information, where the voice(s) of each speaker is collected and used to construct the corresponding speaker model. The compilation of voice models for all speakers is called the voice dataset [9][10].

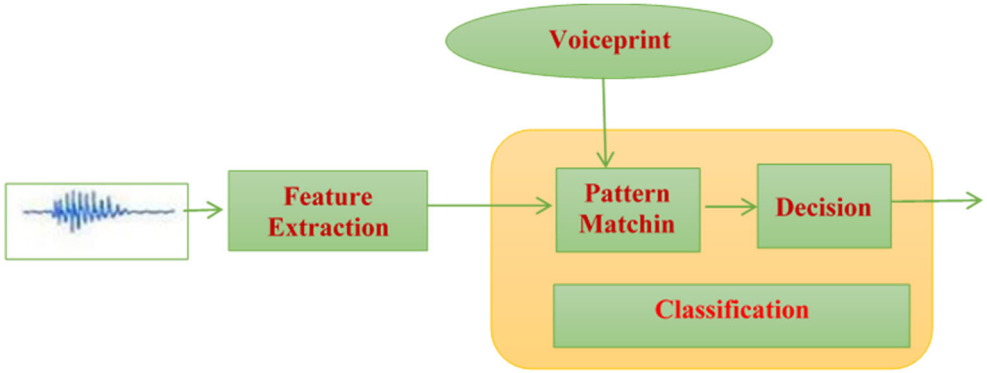



**Figure 1.** Generic method for recognizing VP [10].

Feature matching is responsible for matching the approximate features of the speaker versions. There are many types of pattern matching methods used in speaker recognition and the corresponding models [11]. Some of these methods include Hidden Markov Models (HMM), Dynamic Time Warping (DTW), Vector Quantization (VQ), Artificial Neural Networks (ANNs), and deep learning algorithms [10].

## 3. CNN Architecture

In this section, the architectures of the proposed CNN will be described in detail. The network consists of one two-dimensional convolutional layers, one fully connected layers, and a classification layer with softmax function. The kernel size and the number of filters of convolutional layers were carefully tuned with the max pooling operation, and the output of intermediate fully connected layers to obtain an optimization performance. After the convolutional layer the activation with the ReLU are computed. Then a max pooling layer operates a dimensionality reduction. Each kernel of the convolutional layers has dimension 3×3 with stride of 1. In the convolutional layer, the number of filters is 2. The max

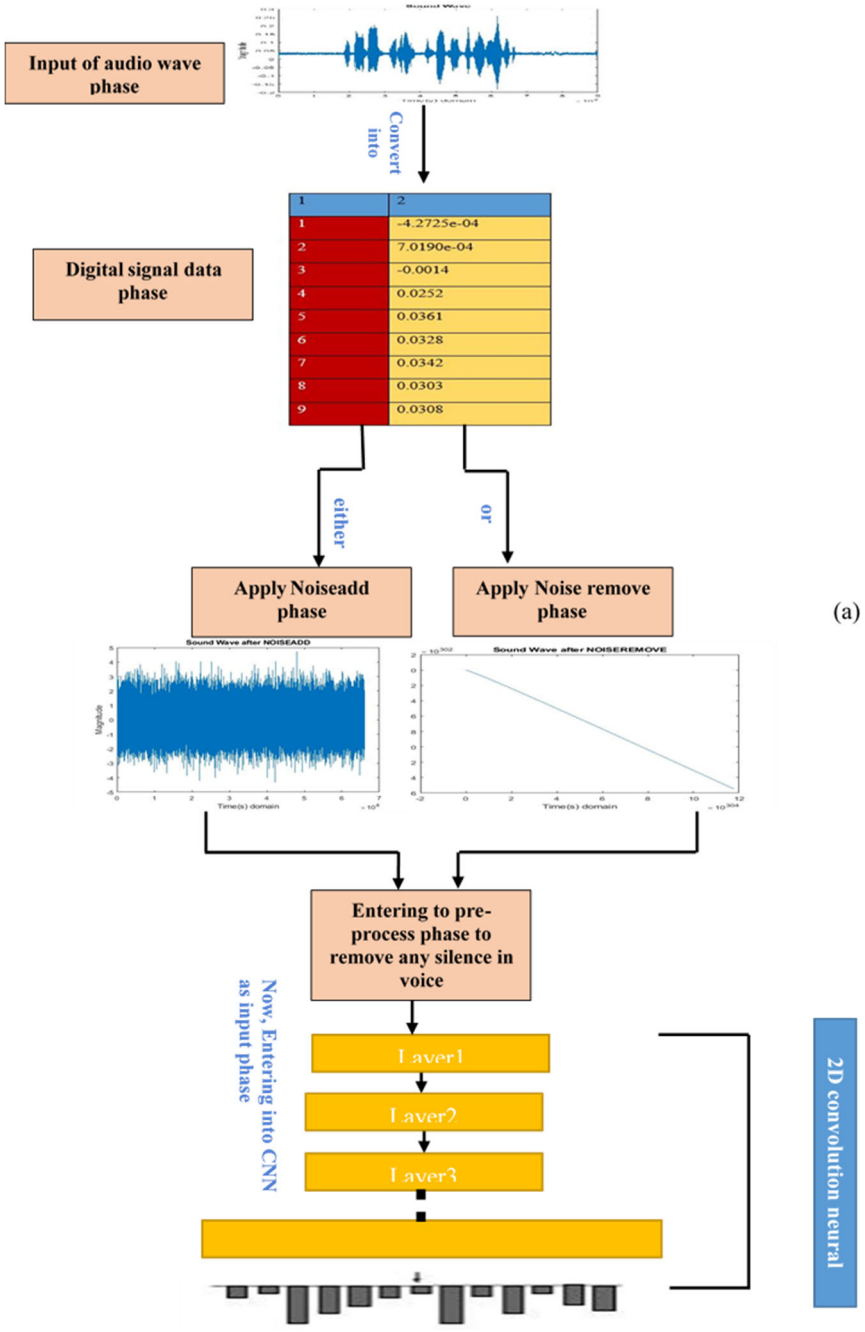
pooling layers are of 2×2 dimensions with stride 2. To enhance the nonlinearity and to reduce the over fitting, one fully connected layer is used. The fully connect layer has 10 neurons. The network is thus composed by 7 layers (input, convolution, pooling, Stack2line, ReLU, fully connect, classification).

The CNN based training is utilized from scratch. Therefore, the appropriated CNN model for the studied dataset is generated using an empirical experiment. The model was implemented in MATLAB 2018b program. The model trains on 100000 epochs with a learning rate of 1e-1. The stochastic gradient descent with momentum (SGDM) optimization algorithm is employed. The momentum term is set to 0.95.

The architecture of the proposed CNN is demonstrated in Table 1. It consists of an input layer with size of 28*28, one convolution layer which contains 2 filters with size of 3*3 and uses relu function in this layer, one pooling layer which contains of 2*2 max-pooling, After convolution and pooling, the multi-dimension "outputs" usually are converted to a vector to be used as the inputs of the fully connected non-linear layers. The stack2line layer is used to indicate this converting. The last layer is the fully connected layer which contains nonlinear type as softmax and the number of classifications is 10.

**Table 1.** Arictecture Of The Proposed CNN Model

| Layer No. | Layer Name | Description |
|---|---|---|
| 1 | Input | 28×28×1 |
| 2 | Convolution | 3×3, 2 filters, pad 1 |
| 3 | Relu | |
| 4 | Maxpooling | 2×2,stride 2 |
| 5 | Stack2line | |
| 6 | Relu | |
| 7 | Relu | |
| 8 | FullyConnectedLayer | 10 |
| 9 | SoftMax | |
| 10 | ClassificationLayer | 10 |


Input of audio wave phase
Convert into
Digital signal data phase
1 2
1 -4.2725e-04
2 7.0190e-04
3 -0.0014
4 0.0252
5 0.0361
6 0.0328
7 0.0342
8 0.0303
9 0.0308
either
or
Apply Noiseadd phase
Apply Noise remove phase
Sound Wave after NOISEADD
Sound Wave after NOISEREMOVE
Time(s) domain
Entering to pre-process phase to remove any silence in voice
Now, Entering into CNN as input phase
Layer1
Layer2
Layer3
2D convolution neural

(a)

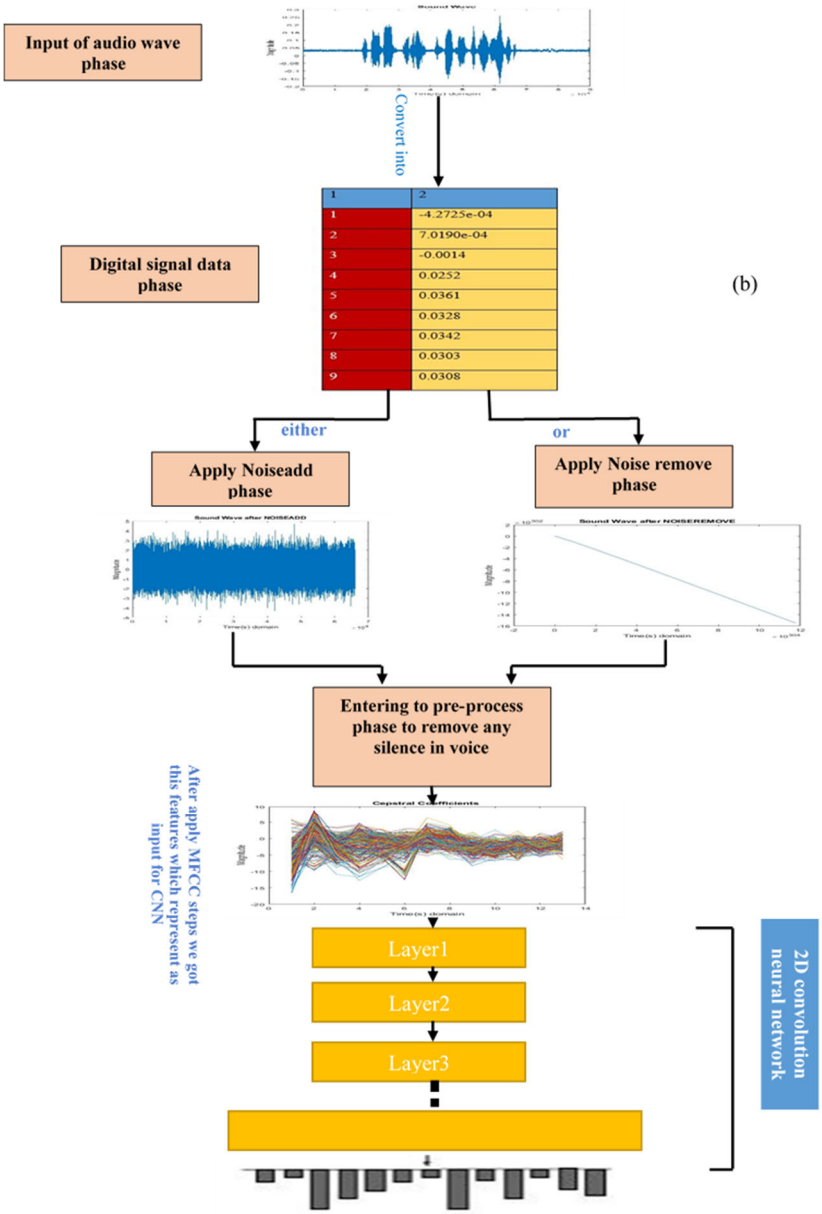



**Figure 2.** (a) represent RW-CNN steps
(b) represent MFCC-CNN steps

Two methods are used in this work which are MFCC-CNN and RW-CNN. The same CNN structure was used in both methods. The first method is called the standard method and includes the following: reading the audio files and then these audio files pass in a phase of removing or adding random noise to the audio files. Then the system goes through the pre-processing stage to remove any silence that exists in the sound. Next, the audio files are passed in the MFCC to extract the features. These features is used in the audio and before entering the voice for recognition stage using the CNN the features are stored in the 2D matrix because the proposed CNN is 2D. Therefore; these features are converted and stored in a two-dimensional matrix and then these extracted features will be entered into CNN to perform a process The training .CNN will take the input as a picture and then begin the process of training via the proposed CNN. In the second method, RW-CNN, the same steps as the first method, but without going through the MFCC phases where the audio files are read and then these audio files are passed in the phase of removing or adding random noise to the audio files. Then, the system goes through the pre-processing

stage to remove any silence in the sound After that, the audio files are stored in the 2D matrix. This work are illustrated in figure 2.

## 4. Datasets

The used dataset is a VoxForge Speech Audio files. Audio can be encoded at different sampling rates (i.e. samples per second - the most common being: 8kHz, 16kHz, 32kHz, and so on), and different bits per sample (the most common being: 8-bits, 16-bits or 32-bits). The dataset contains different ages and people of all groups speak the English language in various words. The speaker characteristics are: (Gender: Male & female, Age Range: Adult, Language: EN, Pronunciation dialect: New Zealand English).

The recording information are: (Microphone make: n/a, Microphone type: Headset mic, Audio card make: unknown, Audio card type: unknown, Audio Recording Software: VoxForge Speech Submission Application O/S. The file information are: (File type: wav, Sampling Rate: 48000, Sample rate format: 16, Number of channels: 1). The dataset size used for this work is 400 classes, each class has 10 voice audio files. The train set taken from each class is 8 voices and the rest are for test set [15].

## 5. Simulation Results

In this section, the obtained results from the implementation of the proposed system will be discussed. Two approaches were implemented for VPR. The first one is by applying MFCC as feature extraction for the audio data and then applying the CNN for training and recognition. The second one is by applying CNN directly to the raw data. The CNN is designed as a feature extraction and training method.

Adding some noise before applying the proposed approach to conclude more results that shows the obtained results from the proposed method is more accurate in different circumstances. The most important part of each ML or DL system is how this system is accurate in his process. As it is known, no system that uses massive data is 100% accurate, since no algorithm is perfect so far, but increased the accuracy as high as possible. Therefore; the first step was to take a look at the accuracy calculation process to understand how the accuracy has been calculated. The meaning of accuracy is the difference between the test set and the predict set for the same samples and it can be calculated as shown in Equation (1).

Accuracy = (Number of correct predictions / Total number of predictions) * 100 (1)

It has been considered that the accuracy can be changed from system to another according to its internal design and circumstances. If some voices have noise in its background, then it should be removed. In this case, some audio data will be corrupted and the accuracy is affected. Moreover, the effect of using the preprocessing of mel-frequency and mel-filter should be considered as it can affect the system accuracy. Using these processes, the proposed system accuracy was evaluated when the audio without any noise and the mel-frequency and mel-filter were considered.

The obtained results were explained in table 2. In which, the accuracy of each category of data for both CNN-MFCC and with or without noise CNN-RW were shown. The results showed that the accuracy for both methods and for different data sizes are the same, because in both ways the input process for the CNN network is the same as in the first way using CNN -MFCC features are used and entered into the network. The CNN network also applies some filters to extract the features and train the network based on the extracted features. As for the second method, the sound is directly entered into the network, but the network in both ways is similar to the input, which was taken as an image and by applying some network filters on it. The features have come out of the sound, and thus we have similar results, regardless of the increase or decrease in the volume of data.

In the first stage, MFCC-CNN was implemented without noise on 68 audio files, which means 544 voiceprint and the accuracy was 0.96. It was applied to 88 audio files, 704 voiceprints, and the accuracy was equal to 0.96. It was also applied to 400 audio files, means 3200 voiceprint and the accuracy was 0.96. The CNN-MFCC was also used with noiseadd on 68 audio files means 544 voiceprint and the accuracy was 0.96. It was also applied to 88 audio files means 704 voiceprint and the accuracy was 0.96. It was also applied to 400 audio files means 3200 voiceprint and the accuracy was 0.96.

The second method CNN-RW was applied with and without noise on 68 audio files means 544 voiceprints. The accuracy is equal to 0.96, as well as it was applied to 88 audio files, means 704 voiceprints. An audio file, means, on 544 voiceprints, the accuracy was equal to 0.96, as well as it was applied to 88 audio files, means 704 voiceprints, and the accuracy was equal to 0.96. It was also applied to 400 audio files, means 3200 voiceprints, and the accuracy was equal to 0.96.

There is another evaluation criterion called MSE (Mean Square Error), which calculates the error occurred in each epoch and then find its mean. It measures the common of the squares of the errors, that is, the average squared difference among the estimated values and the real value. MSE is a threat function, similar to the predicted fee of the squared mistakes loss. The truth that MSE is almost continually strictly positive (and not zero) because of the randomness or because the estimator does not account for statistics that would produce a more accurate estimationt. The MSE can be calculated as in equation 2.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \widehat{y}\right)}^{2}$$

The square of the difference between actual and predicted

The MSE was calculated for the same above four cases. In our system, the MSE has reached its minimum with using noise. Table 2 shows the MSE of each subset of data in the four circumstances.

**Table 2.** The MSE and Accuracy of Proposed CNN

| Technique | Dataset size | | | Number of TrainingData | | | MSE | | | Accuracy | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CNN-MFCC Without Noise | 68 | 88 | 400 | 544 | 704 | 3200 | 3.2000e-08 | 3.2000e-08 | 3.2000e-08 | 0.96 | 0.96 | 0.96 |
| CNN-MFCC With Noise | 68 | 88 | 400 | 544 | 704 | 3200 | 3.2000e-08 | 3.2000e-08 | 3.2000e-08 | 0.96 | 0.96 | 0.96 |
| CNN-RW Without Noise | 68 | 88 | 400 | 544 | 704 | 3200 | 3.2000e-08 | 3.2000e-08 | 3.2000e-08 | 0.96 | 0.96 | 0.96 |
| CNN-RW With Noise | 68 | 88 | 400 | 544 | 704 | 3200 | 3.2000e-08 | 3.2000e-08 | 3.2000e-08 | 0.96 | 0.96 | 0.96 |

In Table 2, MSE also shows for each category of data for both CNN-MFCC and with or without noise CNN-RW as shown in the table where the results showed that MSE for both methods and for different data sizes are the same, because in both ways the input process for the CNN network is the same as in the first way using CNN -MFCC features are used and entered into the network. The CNN network also applies some filters to extract the features and train the network based on the extracted features. As for the second method, the sound is directly entered into the network, but the network in both ways is similar to the input, which is taken as an image and by applying some network filters on it is used The features have come out of the sound, and thus we have similar results, regardless of the increase or decrease in the volume of data.

**Table 3.** Compare The Classification Accuracy From Each Method [12]

| Method | Accuracy (%) |
|---|---|

| | Max | Average |
|---|---|---|
| MFCC | 91.30 | 91.26 |
| CNN of signal wave | 54.00 | 49.77 |
| CNN of Spectrogram(proposed) | 99.00 | 95.83 |

| Method | Accuracy (%) | |
|---|---|---|
| | Max | Average |
| MFCC | 91.30 | 91.26 |
| CNN of signal wave | 54.00 | 49.77 |
| CNN of Spectrogram(proposed) | 99.00 | 95.83 |

Table 3, shows the accuracy result from the classification when the proposed method is compared to MFCC method and CNN of signal wave method. The experiments conduct on 50 times to the testing set for each method. Maximum and average of the classification accuracy show for comparison. It reveals the proposed CNN based method trains on spectrogram image of voice is the best compared with the other two methods. The average classification results of the testing set by the proposed method is 95.83% of accuracy. For MFCC based method is 91.26% and for CNN trained on image of raw signal wave is only 49.77%. The proposed method is very efficient for text-independent approach where only short utterance of voice is needed as an input.

**Table 4.** The frame identification performance FIA (%)[14]

| | Clean | Noise (SNR=5dB) | Reverb (RT60=0.5 s) |
|---|---|---|---|
| MFCC-CNN | 45.72 | 16.85 | 37.32 |
| RW-CNN | 64.01 | 46.24 | 38.51 |

In Table 4, The FIA is the percentage of identification accuracy by considering each frame individually. The capability of the RW-CNN to correctly identify the speaker is widely greater for clean and noisy signals with respect to MFCC features. In the reverberation case, the RW and MFCC have similar identification performance. It was demonstrated that the RW-CNN is more robust to noise if compared to the MFCC-CNN trained with the same DA dataset, and they have similar performance in reverberant environments.

## 6. Conclusions and Future work

In this paper, a convolution neural network-based Voiceprint Recognition System in Noisy Environment was presented. The CNN architecture is designed to operate for both MFCC-CNN and RW-CNN. The obtained results show that both methods are similar in their results, either with or without noise, but the performance is mixed. The proposed CNN inputs are images in both cases, i.e. the network deals with images, so the result was the same. The system shows high accuracy and minimum mean square error for both methods. The audio files and the recording by microphone files were used to test the system and the system in most cases were positive. Future works include added some real noise and try to eliminate it using a very huge dataset of human audio (millions of audio). Connection between traditional and modern methods for the voiceprint recognition system and use CNN directly on digital data of Audio is another step in the future.

## References

[1] J. Choudhary, “Survey of Different Biometrics Techniques,” vol. 2, 2012.
[2] J. Lee, J. Park, and T. Kim, “Raw Waveform-based Audio Classification Using Sample-level CNN Architectures arXiv : 1712 . 00866v1 [ cs . SD ] 4 Dec 2017,” no. Nips, 2017.
[3] S. E. E. Profile, “An overview of popular deep learning methods,” no. December, 2017.
[4] I. Namatēvs, “Deep Convolutional Neural Networks : Structure , Feature Extraction and Training,” vol. 20, no. December, pp. 40–47, 2017.

[5] J. Lee, J. Park, and T. Kim, “Raw Waveform-based Audio Classification Using Sample-level CNN Architectures arXiv : 1712 . 00866v1 [ cs . SD ] 4 Dec 2017,” no. Nips, 2017.
[6] M. Ravanelli and Y. Bengio, “SPEECH AND SPEAKER RECOGNITION FROM RAW WAVEFORM WITH SINCNET,” no. 1,2019.
[7] T. Kim, J. Lee, and J. Nam, “Comparison and Analysis of SampleCNN Architectures for Audio Classification,” IEEE J. Sel. Top. Signal Process., vol. PP, no. 8, p. 1, 2019.
[8] D. Salvati, C. Drioli, and G. L. Foresti, “End-to-End Speaker Identification in Noisy and Reverberant Environments Using Raw Waveform Convolutional Neural Networks,” pp. 4335–4339, 2019.
[9] K. N. Van, T. P. Minh, T. N. S. B, and M. H. L. B, “Text-dependent Speaker Recognition System Based on Speaking Frequency Characteristics Text-dependent Speaker Recognition System Based on Speaking Frequency Characteristics,” no. January, 2018.
[10] Q. Jin, “Robust Speaker Recognition,” no. January, 2007.
[11] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, “Very Deep Convolution Neural Networks For Rae Waveforms,” pp. 421–425, 2017. [12] M. Shyu, S. Chen, and S. S. Iyengar, “A Survey on Deep Learning : Algorithms , Techniques ,” vol. 51, no. 5, 2018.
[12] S. Bunrit, T. Inkian, N. Kerdprasop, and K. Kerdprasop, “Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network,” no. August, 2019.
[13] D. Salvati, C. Drioli, and G. L. Foresti, “End-to-End Speaker Identification in Noisy and Reverberant Environments Using Raw Waveform Convolutional Neural Networks,” pp. 4335–4339, 2019.
[14] A. Raji and V. I. Nnebedum, “IDENTITY AUTHENTICATION USING VOICE BIOMETRICS TECHNIQUE,” pp. 130–136, 2015.
[15] http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/16kHz_16bit/