(MIJ) 2023, Vol. No. 9, Special Issue **e-ISSN: 2454-924X; p-ISSN: 2454-8103**

Real-time Voice Recognition and Modification By Convolutional Neural Network[1]

***Tameem Hameed Obaida, **Abeer Salim Jamil, #Nidaa Flaih Hassan**
**Computer Systems Techniques Department, Al-Furat Al-Awsat Technical, Najaf Technical Institute AL Najaf, Iraq*
***Department of Computer Technology Engineering, Al-Mansour University College Baghdad, Iraq*
#Department of Computer Science, University of Technology , Baghdad, Iraq

التعرف على الصوت وتعديله في الوقت الفعلي بواسطة الشبكة العصبية التلافيفية

تأميم حميد عبيده[1]*, عبير سالم جميل[2], نداء فليح حسن[3]
[1]قسم تقنيات انظمة الحاسوب, المعهد التقني النجف, جامعة الفرات الاوسط التقنية, محافظة النجف الاشرف, العراق
[2]قسم هندسة تقنيات الحاسوب، كلية المنصور الجامعة، بغداد، العراق
[3]قسم علوم الحاسوب, , الجامعة التكنولوجية, بغداد, العراق

ABSTRACT

Voice recognition is an interesting topic, as many researchers have sought to use different applications and devices in the field of voice recognition. But more advanced solutions use artificial intelligence techniques for their great ability to simulate the human mind. Voice recognition is useful in many aspects, especially in security fields, for example, witness protection. In this paper, a new model is designed to protect the witness and ensure the Safety and security of witnesses in order to ensure their cooperation and testimony. By recognizing his/her voice in real-time, then changing it by applying by applying some filters and manipulating the sampling frequency and time to amplify the sound. The method relied on the salient features of sound, and the Mel-Frequency Cepstral Coefficients is one of the important and effective techniques on feature extraction (MFCC) approach, it was used as a first step in the proposed model. The model then used a convolutional neural network (CNN), due to its high classification and prediction accuracy, to recognize the witness's voice. Then his/her voice was changed as a last step. The proposed model achieved a classification accuracy of about 99% in distinguishing the witness's voice.

Keywords: *Voice Recognition; CNN; Convolutional Neural Networks; MFCC, Mel-Frequency Cepstral Coefficients.*

الخلاصة

يعد التعرف على الصوت موضوعًا مثيرًا للاهتمام ، حيث سعى العديد من الباحثين إلى استخدام تطبيقات وأجهزة مختلفة في مجال التعرف على الصوت. لكن الحلول الأكثر تقدمًا تستخدم تقنيات الذكاء الاصطناعي لقدرتها الكبيرة على محاكاة العقل البشري. التعرف على الصوت مفيد في العديد من الجوانب ، لا سيما في المجالات الأمنية ، على سبيل المثال ، حماية الشهود. في هذه الورقة ، تم تصميم نموذج جديد لحماية الشاهد وضمان سلامة وأمن الشهود من أجل ضمان تعاونهم وشهادتهم. من خلال التعرف على صوته / صوتها في الوقت الفعلي ، ثم تغييره عن طريق تطبيق بعض المرشحات ومعالجة تردد أخذ العينات والوقت لتضخيم الصوت. اعتمدت الطريقة على السمات البارزة للصوت ، وتعد معاملات Cepstral ذات التردد الميل من التقنيات المهمة والفعالة في نهج استخراج الميزات (MFCC) ، وقد تم استخدامها كخطوة أولى في النموذج المقترح. استخدم النموذج بعد ذلك شبكة عصبية تلافيفية (CNN) ، نظرًا لتصنيفها العالي ودقتها في التنبؤ ، للتعرف على صوت الشاهد. ثم تم تغيير صوته كخطوة أخيرة. حقق النموذج المقترح دقة تصنيف بلغت حوالي 99٪ في تمييز صوت الشاهد.

[1] *How to cite the article:* Obaida T.H., Jamil A.S., Hassan N.F. (2023), Real-time Voice Recognition and Modification By Convolutional Neural Network, *Multidisciplinary International Journal*, Vol 9 (Special Issue), 107-118

Presented at 2nd International Conference on Research in Multidisciplinary Studies (ICRMS-2023)
Held on 06th August, 2023 through Google Meet

INTRODUCTION

The process of distinguishing between human voices is a very difficult task, because many factors affect the voice such as illness, emotion, aging and background noise, etc. Therefore, it is not possible to repeat the speech with the same accuracy twice [1, 2]. Comparing voices is very necessary because it is used in many areas, including the security field, for example, recognizing the witness's voice[3, 4]. Speaker recognition is divided into two categories: the first is the identification of the speaker, and the second is the verification or authentication of the speaker. The first category is to identify the speaker by matching the voice with other voices through the use of a database containing the voices of the speakers that have been previously recorded and saved, and the second category can be implemented by taking a sample of the speaker's voice to be matched [5, 6]. Researchers have used many methods to recognize the voice, including traditional and deep learning methods. Using an image-based spectral approach, the recorded speech is converted into speech images, called a spectrogram, which reflects the frequency spectrum and is also called the voiceprints. Spectrograms are used as input for Convolutional Neural Networks(CNN) [7, 8].

The great progress in the field of deep learning, especially in (CNN), has shifted from designing features to learning features [9]. Features extraction is a very important step in speech recognition, to extract useful features from raw data, one example of a feature extraction approach is Mel Frequency Cepstrum Coefficient (MFCC) [10]. The MFCC approach is used to extract characteristics from a voice stream. It determines the value or vector in this method to identify items associated with these vectors or values. In contrast to many audio samples, MFCC is a lossy representation of data rates. Additionally, not all data rates are appropriate for categorization and identification. On the other hand, in the classification application of the generative network, audio representations such as Mel-spectrogram eliminate lossily. In addition, except for pre-emphasized and DCT phases, the M-S computation follows the same procedures as the MFCC [11].

In this paper, Convolutional Neural Network (CNN) is used to recognize the witness's voice in real-time, then the sound is amplified to ensure that the witness's voice is not recognized, and this is for protection purposes. The method proved to have a great ability to distinguish sounds despite the occurrence of delays due to continuous speech in real-time.

The remaining parts of the paper are arranged as follows: Section 2 contains related works. Section 3 contains a description of the proposed model. Section 4 contains the experimental results, and Section 5 represents the conclusion.

RELATED WORK

Automatic recognition of speakers is an important technique , because it is through which the identity of the speaker is determined. CNN is one of the most used methods in this field, it is a type of deep learning algorithm that eliminates the need for manual feature extraction on inputs having local correlation structure, such as images or spectrograms of spatiotemporal connections [7].

Lukic et al. [7] used a simple spectrogram as input to CNN and studied of optimized design of those networks for speaker identification and clustering. Also how to move a network that has been trained for speaker recognition to speaker clustering. The training was carried out using the well-known TIMIT dataset, without the need to specify local features. Wang et al. [12] introduced a comparative for small (CNN) and evaluate speaker recognition effectiveness, they use the transfer learning technique to address the issue of limited training data by creating a mechanism that allows inference to be run locally on edge devices. To overcome well-known cloud computing problems such as network latency and data privacy, etc. Results achieved ~84% accuracy for speaker classification in time less than 60 ms on mobile using the Atom Cherry Trail processor. Totakura et al. [13] presented a method for developing voice-guided self-driving cars using CNN, Asphalt-8 game data was used. CNN was trained to predict the voices of three different persons (man, woman, and child). The results obtained for this method proved that it achieved an accuracy of 99%.

Lee et al.[14] Introduced a method that uses CNN to recognize the emotion of speech. The dataset used audio recordings including utterances of varying lengths. Deep learning techniques such as a multi-layer perceptron (MLP)

and a convolutional neural network were used to extract one-dimensional data from audio recordings and train two-dimensional Mel-spectrogram images (CNN). After reducing audio files to less than two seconds to improve accuracy using CNN, the results of the experiment got an accuracy up to 60%.

Dey et al. [15] developed a safe voice communication system by combining a voice-over-internet protocol system with a trained model based on a deep convolutional neural network (DCNN). The DCNN-based developed model's voice recognition accuracy in a noiseless environment was 93.7% percent, according to experimental results. The existing support vector machine (SVM) algorithms and K-nearest-neighbour (KNN), on the other hand, had 79% and 82.1% percent accuracy, respectively. The DCNN, KNN, and SVM algorithms have response times of 178, 220, and 228 milliseconds, respectively, for voice recognition.

THE PROPOSED MODEL

The goal of the proposed model is to design a system that provides witness protection through modification of his voice. The proposed model consists of three basic stages, the first stage includes the process of extracting features from the voice of the witness that was previously recorded, to reduce the dimensions. As for the second stage, use CNN for training, to determine the identity of the speaker, so that in the last step the voice of the witness is changed to prevent identification of him for his protection. Figure 1. shows the stages of the proposed model

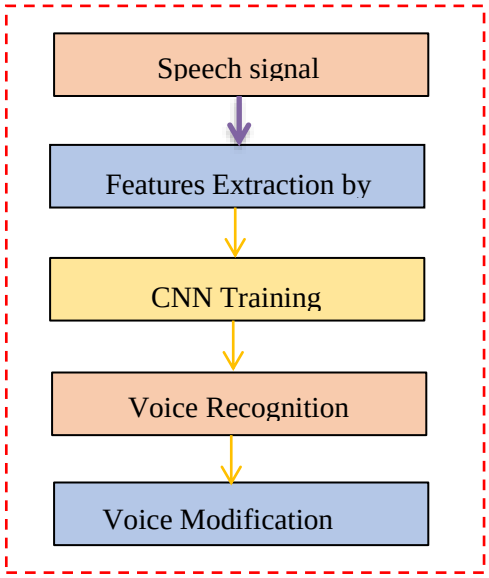

Figure -1 The proposed model.

Description of Dataset

The CNN needs a dataset for training purpose on the witness's voice, so the first step in this proposal was to record audio segments of the witness of different lengths, to create a dataset that fits our work. The reason for performing this step is that there is no dataset containing a large number of voices for one person available on the Internet. Then, the distinctive features in the witness's voice are extracted and converted into spectrogram images, in order to train the CNN on them, to distinguish the witness's voice among a group of sounds.

Feature Extraction

The features are key portions that enable the representation of an essential part of data. Feature extraction is a crucial step in converting a speaker's voice into a stream of a feature vector containing only the information needed to recognize a specific speech. Mel-Frequency Cepstral Coefficients is one example of a feature extraction approach (MFCC)[11, 16]. which was used in the paper as a first step, and the following is a description of this technique :

Mel-Frequency Spectrum Coefficients (MFCCs)

The MFCC approach is used to extract features from a voice stream. It determines the value or vector in this method to identify items associated with these vectors or values. MFCCs are a feature widely used in automatic speech and speaker recognition[17]. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. When applying classification to the generative network, audio representations, such as Mel-spectrogram, , are useful.[18]. In addition, except for pre-emphasized and DCT phases, the Mel-spectrogram computation follows the same procedures as the MFCC. As a result, enhancing the Signal to Noise Ratio necessitates the pre-processing(pre-emphasized) phase (SNR) [19]. Figure 2 describes the basic steps of MFCC.

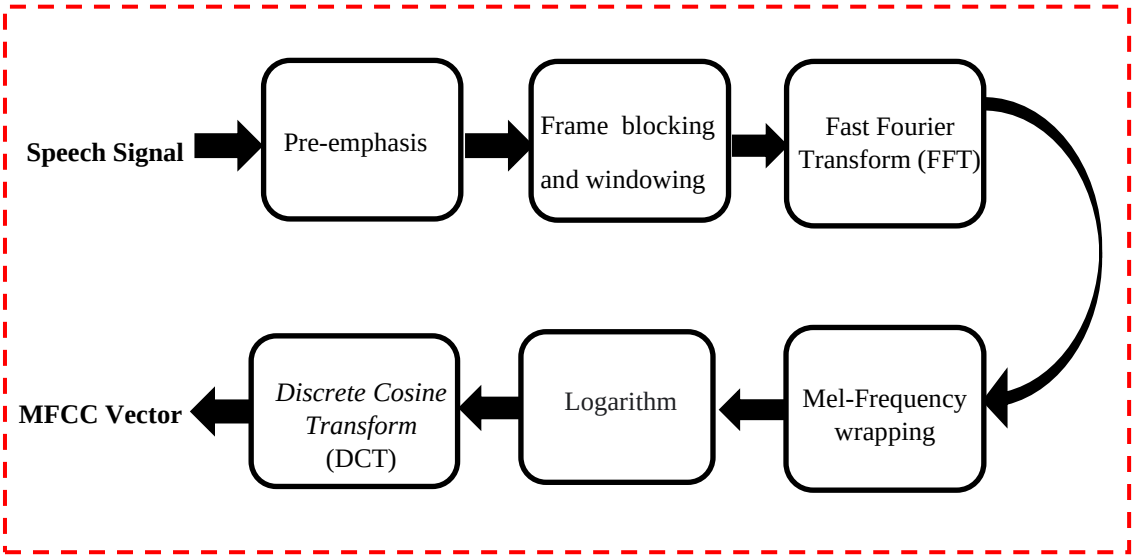

Figure -2 Diagram of MFCC [20].

1. Pre-emphasis

This step is interesting with filtering that emphasizes values for high frequencies. Its purpose is to compensate for the range of steep slope of speech sounds in the high-frequency area. Eq. (1) represents the commonly used pre-emphasis filter.

$$H(x) = 1 - b.\, x^{-1} \quad (1)$$

where x refers to the speech signal, and b represents the value that controls the slope of the filter [19].

2. Frame Blocking

The voice signal is broken down into frames. Each frame has N data samples, and there are M data samples between each frame and the another frame. Eq. (2) depicts separate lengthy voice signal.

$$X[n] = N_2 + M_2 \quad (2)$$

Where $X[n]$: represents the input of a lengthy voice signal. M_2: denote to overlapping between one frame with another (data samples number). N_2: represent a data sample in one frame.

3. Windowing

At this phase, Hamming window is applied to each frame, to eliminate cutouts in the edges by Eq. (3).

$$y_1(n) = x_1(n)w(n),\ 0 \le n \le N - 1 \quad (3)$$

Where $x_1(n)$: denote to input signal. $y_1(n)$: represent windowing signal. $w(n)$: while windowing function

(n)=0.54 − 0.46. cos $(\frac{2\pi n}{N})$, 0 ≤ n ≤ N , N: represents the total length of the window.

4. Fast Fourier Transform (FFT)

To obtain a frequency spectrum, the Fast Fourier Transform is employed. Thus, converting every sample to frame from time domain into the frequency domain. As shown in Eq. (4).

$$X(k) = \sum_{n=0}^{N-1} (X(n).e^{\frac{-j2\pi nk}{N}} \quad 0 \leq k \leq N-1 \tag{4}$$

where *N* illustrate the size of the FFT.

5. Mel spectrum

A Mel spectrum is computed when an FFT signal is sent across the series of band-pass filters called the Mel-filter bank. A Mel is a unit of measurement based on the frequency perception of the human ear. To emulate the experiment of human hearing to get coefficients of Mel-spectral have to multiplying a coefficient of power spectrogram with a triangular filter by Eq. (5). In the last step, Discrete cosine transform (DCT) is calculated through Eq. (6) [21].

$$(5) mel\_f = 2595\, log_{10}\left(1 + \frac{F_{HZ}}{700}\right)$$

$$(6) c(n) = \sum_{m=0}^{M-1} log_{10}\big(s(m)\big)\, cos\,(\frac{\pi n(m-0.5)}{M})\,;\; n = 0,1,2,\dots\dots\dots, C-1$$

The spectrogram images generated by the MFCC are used as input to the CNN network to perform the classification, as shown in Figure 3. After running the CNN practically and the witness speaking in real-time by the microphone, his or her voice was classified with high accuracy and distinguished from among the sounds, including the noise.

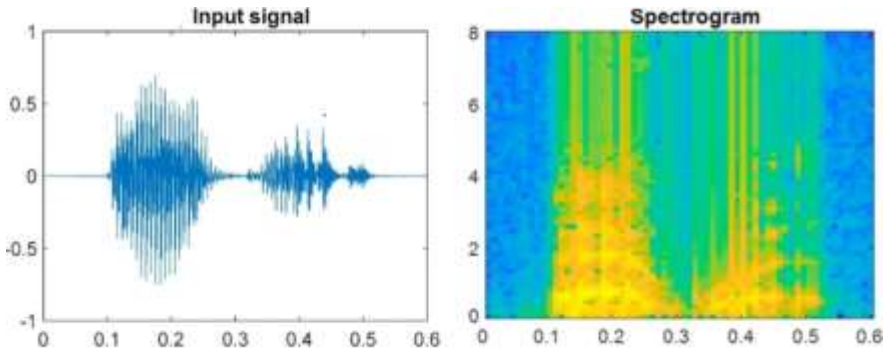

Figure -3 Converting voice signal into a spectrogram image [22].

Apply Convolution Neural Network (CNN)

The second part of the proposed model represents the training process by inserting the spectrogram images into the CNN network to obtain the classification of the voice (witness voice or not witness voice). CNN gives high classification accuracy, because it has strong advantages, as it stimulates the human brain. The network structure consists of four layers [23]. It is the Convolution layer, Activation function, Pooling layer, and fully connected layer. Each layer has a specific function to perform [24].

1. Convolution Layer

This step is extracting the features from input image, which maintains the spatial relation between pixels through learning these features of images by use small squares from input data[25].

2. Activation Function

Choosing an activation function for the neural network is a significant consideration because it may impact the way must format input data [26]. Rectified Linear Unit (ReLU) is applied as one of the important functions. Eq. (7). represents this function.

$$f(x) = \max(0, x) \tag{7}$$

3. Pooling Layer

There are a few pooling options (Average Pooling, max pooling, sum pooling, and L2_Norm Pooling), Max-Pooling, however, is the most widely utilized and popular. Eq. (8) can be used to express the output elements of the pooling layers[25].

$$y_{i,j}^{k} = \frac{1}{M_H\, M_W} \sum_{u=1}^{M_H} \sum_{v=1}^{M_W} x_{2i-\left[\frac{M_H}{2}\right]+u,2j-\left[\frac{M_W}{2}\right]+v}^{k} \tag{8}$$

4. Fully Connected Layer

In a similar approach to a normal Neural Network, the entire architecture thereafter trained through updating and adjusting filters/weights in the Neural Network by a training process named back-propagation [27].
Each layer learns the features locally by region. Filters are applied to each area. The size of the filters is usually smaller than the actual image. Each filter convolves with the image and creates an activation map. The proposed model featured a 3×3 filter size and number of filters (12, 24, 48). The specified number of filters in each layer is given in Figure 4. Also, Table 1 illustrates the architecture hyper-parameters and their values.

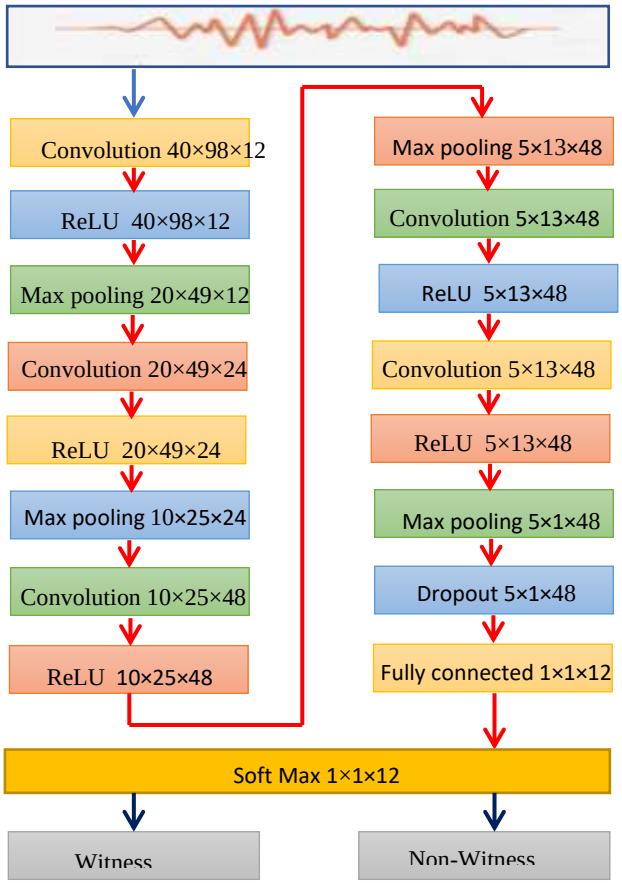

Figure -4 Diagram of Proposed CNN Architecture for Voice Recognition.

Table 1- The specified hyper-parameters of CNN

Phase	hyper-Parameters	Value
Initialization	Bias	0.1
	Weights	Random
	Padding	same
Dropout	P	0.2
	Maximum epochs	20
Training	Batch	40×98

RESULT AND DISCUSSION

In this work, we need more than one audio clip recorded for the same person and under certain circumstances, to form a dataset for our work. So a voice segments of the witness was recorded using the "WavePad Sound Editor" program, and the following properties were defined: Bit rate (256 bps), Channel (Mono), Sample rate (8000 HZ), and Sample size (32 Bit), to create our dataset. The data was divided into 20% for testing and 60% for training and 20% for validation. In addition, creating a dataset for noise, which is another type of sound, was used to verify the accuracy of classification. The results have achieved high accuracy. But these voices contain a lot of data, most of which are unimportant, so the features were extracted from them by the MFCC technique to reduce the dimensionality.

Figure 5 shows the confusion matrix obtained as a result of training on the witness's voice and other voices that represent noise, which is classified into (12) classes representing different voices collected from different places, in order to distinguish the voice of the witness between those voices, the voices were classified into (x, c, n, etc), as shown in the figure, and the figure also contains the precision values for each class. Table 2 represents the number of audio data to be trained and tested.

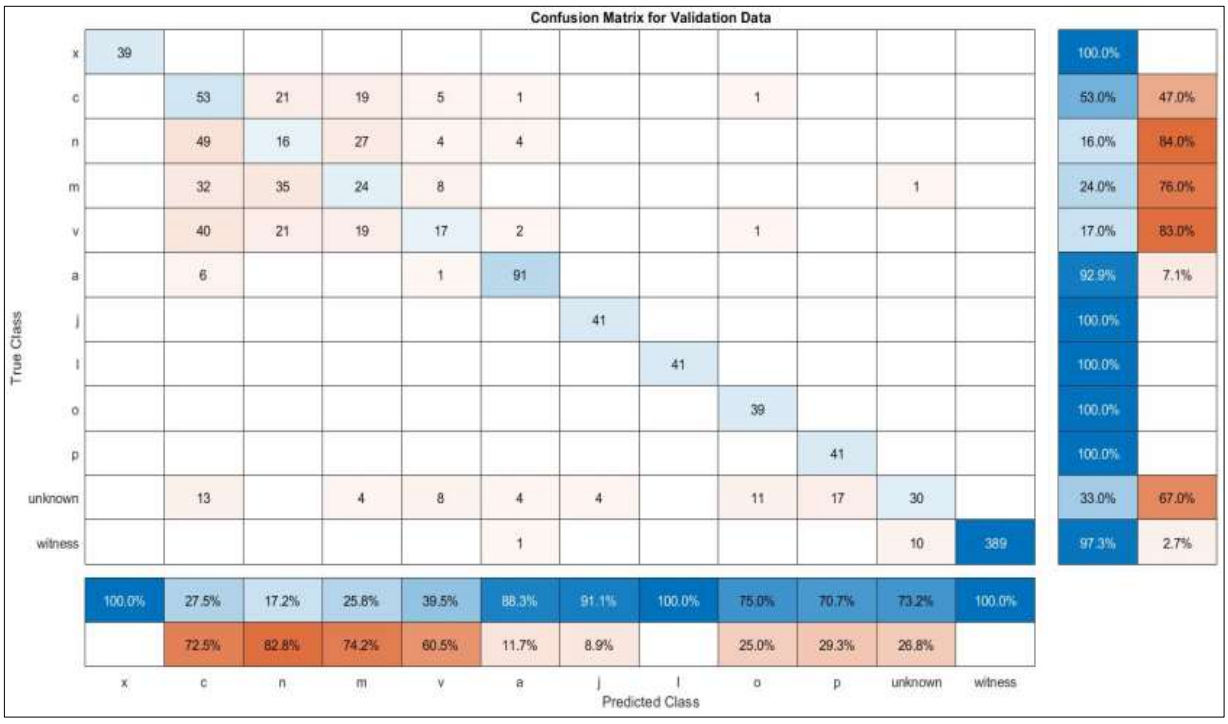

Figure -5 Performance chart showing loss and accuracy of CNN.

Table 2- Number of voice segments

Label	Count of voice segments
a, c, j, l, m, n, o, p, v, x	911
unknown	1400
witness	6899

Table 2 shows a comparison of some of the methods that are used for the recognition of voices and the accuracy of each method. The comparison proved that the proposed model achieved the highest accuracy.

No.	Author	Method	Dataset	Accuracy
1	Lukic et al [102]	Optimize the design of networks for speaker identification and clustering. without the need to specify local features. Using (CNNs).	TIMIT	97.0%
2	Wang et al [103]	use the transfer learning technique to address the issue of limited training data by creating a mechanism that allows inference to be run locally on edge devices. Using (CNNs).	LibriSpeech and Their dataset	~84%
3	Totakura et al [104]	developing voice-guided self-driving cars using CNN, Asphalt-8 game data was used. CNN was trained to predict the voices of three different persons (man, woman, and child).	Use 3 different person voices (Kid, Man, Woman) only	99%
4	Lee et al [105]	uses CNN to recognize the emotion of speech. The dataset used audio recordings including utterances of varying lengths to extract one-dimensional data from audio recordings and train two-dimensional Mel-spectrogram images.	Recorded voices	60%
5	The Proposed Method	Voice Recognition and Modification. By Convolutional Neural Network and MFCC	Our dataset,(12) classes	99% For witness

Table 2- A comparison of the methods currently used

Figure 6 shows the implementation of the proposed model for distinguishing and recognizing the witness's voice. And depicts the case of hearing strange and high sounds unknown to CNN or known and it's were classified according to their type, such as the variable x, which represents background noise that was defined for CNN in advance.

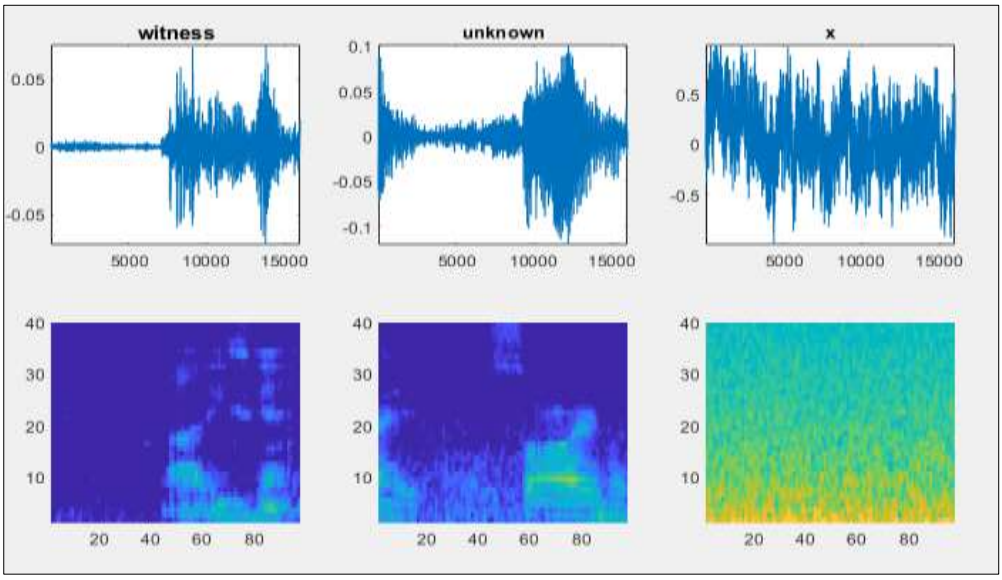

Figure -6 Recognition of Sounds and Classify them in CNN

Figure 7 also represents some other results for distinguishing the sounds by using CNN, as (c) appears one of the sounds that the network has been trained on previously.

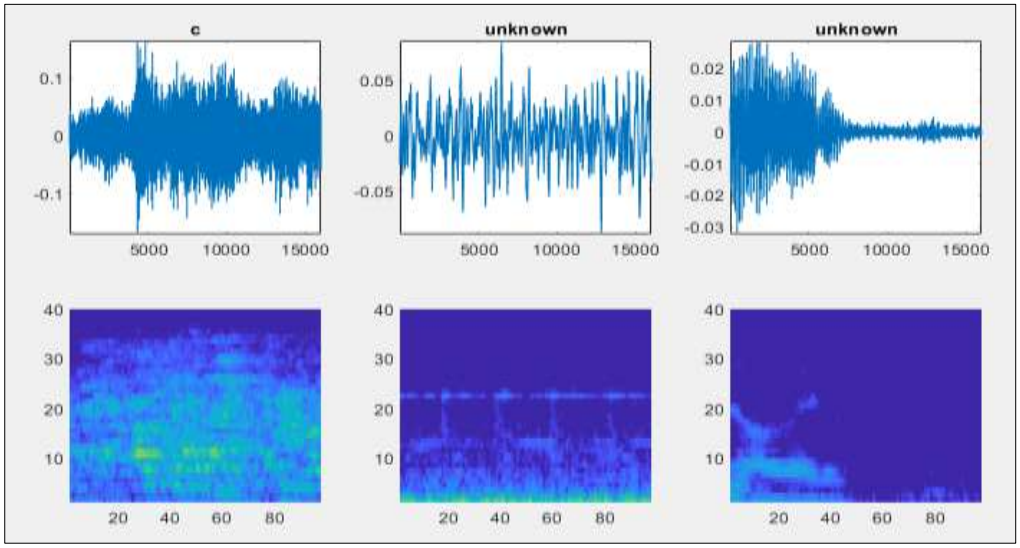

Figure -7 CNN's Classification for (c and unknown) Voices.

Finally, after applying the process of extracting the voice features by MFCC and converting them into spectrogram images, to be entered into CNN to perform the classification process and recognize the witness's voice, the last phase is to change the voice to protect the witness, which that is applied in simple steps by manipulating the sound wave frequency and time, using the MATLAB 2018b program. After determining the value for the sampling frequency, the sound passed through the low pass filter to eliminate the noise. In Figure 8, the red color represents the

original signal and the blue color represents the filtered signal. Satisfactory results for the voice change were obtained by setting the sampling frequency to the value 6000, as shown in Figure 9.

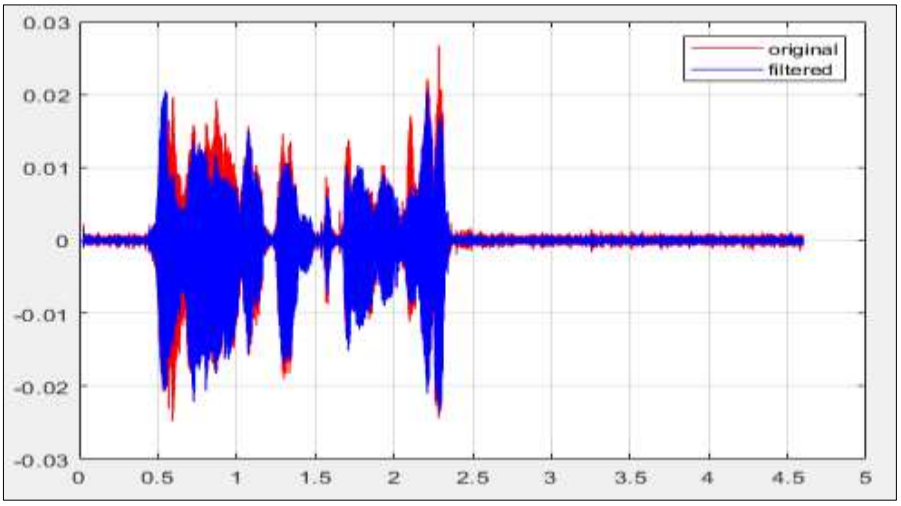

Figure -8 The Original and Filtered Signal.

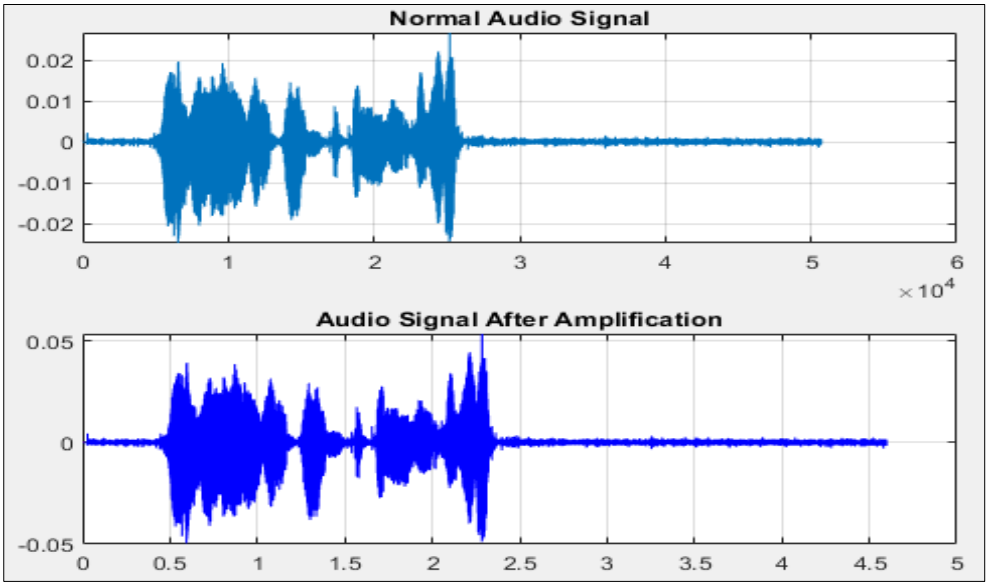

Figure -9 The Original and Amplified Signal.

CONCLUSION

Recognizing voices is a very big challenge due to the sensitivity of sound toward many influences, including illness, emotion, and others. Due to its great importance in protecting people, because of voice it is possible to determine the identity of the speaker. In this paper, a new model is presented to recognize the voice of the witness after recording many audio clips for him /her, so the features are extracted by MFCC and converted into spectrogram images used as input to the CNN, in addition to other sounds that are classified as noise. After CNN has been trained on the sounds in the dataset. The witness must speak in real time in order to distinguish his/her voice from the rest of the voices. In the last phase, after determining the voice of the witness, it is changed by amplifying the sound to prevent people from recognizing him. The experimental results gave a high classification accuracy, as it was able to distinguish the voice of the witness from the non-witness. But the process was slow due to the continuous speech of the witness, which causes some words to be lost, so it's necessary to stop between one sentence and another to perform the process of amplifying the voice. In the future, it's possible to improve the proposed model by using other techniques that address the slowness.

REFERENCES

1. Tandel, N.H., H.B. Prajapati, and V.K. Dabhi. Voice recognition and voice comparison using machine learning techniques: A survey. in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). 2020. IEEE.
2. Hassan, N.F. and H.B.A. Wahab, Proposed a new approach for voiced/unvoiced decision of speech file using lagrange technique. Telecommunications and Radio Engineering, 2013. 72(6).
3. Variani, E., et al. Deep neural networks for small footprint text-dependent speaker verification. in 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). 2014. IEEE.
4. Hassan, N.F., A. Aladhami, and M.S. Mahdi, Digital Speech Files Encryption based on Hénon and Gingerbread Chaotic Maps. Iraqi Journal of Science, 2022: p. 830-842.
5. Lee, H., et al., Unsupervised feature learning for audio classification using convolutional deep belief networks. Advances in neural information processing systems, 2009. 22.
6. Chakroborty, S. and G. Saha, Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter. International Journal of Signal Processing, 2009. 5(1): p. 11-19.
7. Lukic, Y., et al. Speaker identification and clustering using convolutional neural networks. in 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP). 2016. IEEE.
8. Bansal, M. and D.T. Thivakaran, Analysis of Speech Recognition using Convolutional Neural Network. Journal of Engineering Sciences, 2020. 11(1): p. 285-291.
9. Khan, S., et al. Facial recognition using convolutional neural networks and implementation on smart glasses. in 2019 International Conference on Information Science and Communication Technology (ICISCT). 2019. IEEE.
10. Curelaru, F. Front-End Factor Analysis For Speaker Verification. in 2018 International Conference on Communications (COMM). 2018. IEEE.
11. Muda, L., M. Begam, and I. Elamvazuthi, Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083, 2010.
12. Wang, M., et al. Speaker recognition using convolutional neural network with minimal training data for smart home solutions. in 2018 11th International Conference on Human System Interaction (HSI). 2018. IEEE.
13. Totakura, V., B.R. Vuribindi, and E.M. Reddy. Improved Safety of Self-Driving Car using Voice Recognition through CNN. in IOP Conference Series: Materials Science and Engineering. 2021. IOP Publishing.
14. Lee, K.H. Design of a convolutional neural network for speech emotion recognition. in 2020 International Conference on Information and Communication Technology Convergence (ICTC). 2020. IEEE.
15. Dey, P., et al., Deep convolutional neural network based secure wireless voice communication for underground mines. Journal of Ambient Intelligence and Humanized Computing, 2021. 12(10): p. 9591-9610.
16. Abdel-Hamid, O., et al., Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 2014. 22(10): p. 1533-1545.
17. Tiwari, V., MFCC and its applications in speaker recognition. International journal on emerging technologies, 2010. 1(1): p. 19-22.
18. Hasan, M.R., M. Jamil, and M. Rahman, Speaker identification using mel frequency cepstral coefficients. variations, 2004. 1(4): p. 565-568.
19. Rao, K.S. and K. Manjunath, Speech recognition using articulatory and excitation source features. 2017: Springer.
20. Li, H.-C., et al., Make patient consultation warmer: A clinical application for speech emotion recognition. Applied Sciences, 2021. 11(11): p. 4782.
21. Singh, P.P. and P. Rani, An approach to extract feature using MFCC. IOSR Journal of Engineering, 2014. 4(8): p. 21-25.
22. Sharan, R.V., H. Xiong, and S. Berkovsky, Benchmarking Audio Signal Representation Techniques for Classification with Convolutional Neural Networks. Sensors, 2021. 21(10): p. 3434.
23. Chauhan, R., K.K. Ghanshala, and R. Joshi. Convolutional neural network (CNN) for image detection and recognition. in 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). 2018. IEEE.

24. Obaida, T.H., A.S. Jamil, and N.F. Hassan, Real-time face detection in digital video-based on Viola-Jones supported by convolutional neural networks. International Journal of Electrical & Computer Engineering (2088-8708), 2022. 12(3).
25. Aghdam, H.H. and E.J. Heravi, Guide to convolutional neural networks. New York, NY: Springer, 2017. 10(978-973): p. 51.
26. Deng, L. and D. Yu, Foundations and Trends in Signal Processing: DEEP LEARNING–Methods and Applications. 2014.
27. Wu, J., Introduction to convolutional neural networks. National Key Lab for Novel Software Technology. Nanjing University. China, 2017. 5(23): p. 495.
28. Lukic, Y., Carlo Vogt, Oliver Durr, Thilo Stadelmann. Speaker identification and clustering using convolutional neural networks. in 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP). 2016. IEEE.
29. Wang, M., Tejaswini Sirlapu, Alicja Kwasniewska, Maciej Szankin, Marko Bartscherer, and Rey Ni. Speaker recognition using convolutional neural network with minimal training data for smart home solutions. in 2018 11th International Conference on Human System Interaction (HSI). 2018. IEEE.
30. Totakura, V., B.R. Vuribindi, and E.M. Reddy. Improved Safety of Self-Driving Car using Voice Recognition through CNN. in IOP Conference Series: Materials Science and Engineering. 2021. IOP Publishing.
31. Lee, K.H. Design of a convolutional neural network for speech emotion recognition. in 2020 International Conference on Information and Communication Technology Convergence (ICTC). 2020. IEEE.