Convolutional Neural Network for Arabic Speech Recognition

Engy R. Rady*[1], A. Hassen**[2], N.M. Hassan*[3], M. Hesham***[4]

[*] *Basic Science Department, Faculty of Computers and Information, Fayoum University, El Fayoum, Egypt.*
[1]era00@fayoum.edu.eg
[3]nmh00@fayoum.edu.eg

[**] *Physics Department, Faculty of Sciences, Fayoum University, El Fayoum, Egypt.*
[2]ash02@fayoum.edu.eg

[***] *Engineering Math. & Physics Department, Faculty of Engineering, Cairo University, Giza, Egypt.*
[4]mhesham@eng.cu.edu.eg

Abstract: ***This work is focused on single word Arabic automatic speech recognition (AASR). Two techniques are used during the feature extraction phase; Log frequency spectral coefficients (MFSC) and Gammatone-frequency cepstral coefficients (GFCC) with their first and second-order derivatives. The convolutional neural network (CNN) is mainly used to execute feature learning and classification process. CNN achieved performance enhancement in automatic speech recognition (ASR). Local connectivity, weight sharing, and pooling are the crucial properties of CNNs that have the potential to improve ASR. We tested the CNN model using an Arabic speech corpus of isolated words. The used corpus is synthetically augmented by applying different transformations such as changing the pitch, the speed, the dynamic range, adding noise, and forward and backward shift in time. It was found that the maximum accuracy obtained when using GFCC with CNN is 99.77 %. The outcome results of this work are compared to previous reports and indicate that CNN achieved better performance in AASR.***

Keywords*: Arabic automatic speech recognition (AASR), Log frequency spectral coefficients (MFSC), gammatone-frequency cepstral coefficients (GFCC), convolutional neural network (CNN), isolated words.*

1 INTRODUCTION

ASR is a basic component of a virtual assistant. It works by processing a human voice and training a system to recognize vocabulary in that voice. ASR has many applications ranging from speech-based controls to online gaming to deliver commands to IoT devices. In the last five decades, ASR became an active research area. ASR is important for human-human and human-machine communication [1]. Single-word speech recognition can be used in voice interfaces for applications with keyword detection, which can be useful on mobile and embedded devices.

Some of the most common approaches of ASR systems are hidden Markov models (HMMs) [2], Gaussian mixture model HMMs (GMM-HMMs). Alternatively, it is known that the CNN model is a deep learning algorithm that can perform complex tasks with images, videos, texts, and sounds that are inspired by the human visual system [3]. CNN's achieved great success in image recognition [4], and recently they are widely adopted in ASR systems [5]–[9]. Most leading technology companies like Google, Facebook, Microsoft, IBM, Yahoo!, Twitter and Adobe, have initiated research and development projects [10]–[13] which employs CNN for image recognition products and services. Different works concerning the convolutional neural network (CNN) were found. Haque et al. [14] proposed three convolutional layer architecture in which the extracted features from the audio samples are MFCCs. The TIDIGITS corpus was used to assess the proposed models. They proved that a three-layer CNN gives the best classification accuracy of 97.46%. An end-to-end (E2E) speech recognition model, Jasper, was also reported [15], where the authors used 1D convolutional and introduced a new layer-wise optimizer called NovoGrad. The beam-search decoder with an external neural language model achieves 2.95 % WER while the greedy decoder achieves 3.86 % on LibriSpeech test-clean.

Besides, another language was previously applied for the CNN model. Nagajyothi et al. [16] presented an ASR using the airport inquiry system for the Telugu language. The raw speech signal was used as input to CNN. The results confirmed that the proposed system achieved better performance than the conventional Neural Network Techniques. An isolated digits recognition for the Pashto language was developed [17], using deep CNN. The data was collected from 25 male and 25 female Pashto native speakers. Each speaker utters digits from zero (Sefar) to nine (Naha), a total of 10 digits uttered by 50 speakers. The MFCCs features were extracted from data. The researchers constructed a CNN architecture with four hidden layers. The accuracy of recognition for all digits was 84.17 %.

Rajagede et al. [18] presented a CNN system for speech recognition of Arabic letters. The used dataset is composed of ten Arabic letters by which every pair of letters has similarities in sound. The MFSCs are used as input to CNN. They proved that CNN with a convolutional layer and one fully-connected layer gives an accuracy of up to 80.75%, which is better than the Multilayer Perceptron (MLP) with an accuracy of up to 72.0%. A framework for AASR was presented by [19] that uses Long Short-Term Memory (LSTM) and Multi-Layer Perceptron (MLP) classifiers. They used two feature extraction methods, (1) static and dynamic MFCCs features, and (2) the Filter Banks (FB) coefficients. They tested their

work on a spoken Arabic digit and a TV command data set. They compared their experiment with some previously published approaches and also using different encoders.

Boussaid and Hassine [20] performed a speaker-dependent ASR model for isolated Arabic words. They used a database of 11 standard Arabic isolated words by 10 male and 10 female speakers. Three types of datasets were used for the experiments. First corpus: it consists of 10 male and 10 female speakers. Each speaker uttered each word 5 times. Second corpus: the database was divided into four subcorpora of five speakers each. Third corpus: the database was also divided into two subcorpora. Each corpus contains the speech signal of ten speakers. Three methods were used for feature extraction MFCC, PLP, and relative perceptual linear prediction and their first-order temporal derivatives. PCA was applied to the extracted features to reduce the dimension of the feature matrix using two different methods. The extracted features are fed into a feed-forward backpropagation neural network (FFBPNN) using two learning algorithms; the scaled conjugate gradient "Trainscg" and the Levenberg–Marquardt "Trainlm". The hybrid techniques for feature extraction from the above-mentioned techniques were used and examined on the two learning algorithms.

A challenging issue is introduced to recognize three Arabic letters sa (س), sya (ش), and tsa (ث), which have identical pronunciation [21]. These letters were pronounced by Indonesian speakers but had different makhraj in Arabic. He extracted 13 MFCC feature vector from the used dataset of size 738 (248 samples of sa (س), 254 samples of sya (ش), and 236 samples of tsa (ث)). He used the backpropagation ANN model as a classifier with a sigmoid activation function. He obtained an accuracy of 92.42 %. The model of LSTM to recognize digits in the Arabic language was designed [22]. The MFCC approach was used to extract the distinctive features from the input signals. A dataset of different dialects of Arabic digits was used in this case. The dataset of 1040 samples is used, where 840 trained the network and 200 were used for testing. After testing this model, the results show that the LSTM model achieved an accuracy of 69% for spoken Arabic digits. The highest precision for such a model is 80% for recognizing the zero digit.

Alalshekmubarak and Smith [23] introduced a noise-robust system using Echo State Networks and Extreme Kernel machines, which we call ESNEKM. He examined different feature extraction methods: MFCCs, PLP, and RASTA-PLP. They compared his models with the HMM. Such a model was examined with different types and noise levels. The best accuracy was obtained when RASTA-PLP combined with ESNEKM.

2 BACKGROUND

A. Automatic Speech Recognition

The traditional ASR system has four main elements: signal processing and feature extraction, language model (LM), acoustic model (AM), and hypothesis search [1]. Speech recognition can be achieved at a variety of levels of speech (Phone/ Phoneme/ grapheme, syllable, word, phrase, etc.). It was found that combining the four components of the ASR system into end-to-end (E2E). E2E ASR system is a dominating topic in the recent research scope [24]–[27] E2E system is meant that the four elements could be learned simultaneously that dispose of any intermediate components. Traditional ASR systems require separate training of the four components because of the complexity. On the other side, E2E ASR is a single integrated approach with a much simpler training pipeline to reduce the decoding time and training time [28].

B. The Use of CNN in ASR

The CNN model has four main processes, that is, convolution, non-Linearity (ReLU), pooling, and classification (fully connected layer). The convolutional and pooling layers in CNNs are directly inspired by the classic concepts of simple cells and complex cells in visual neuroscience [29]. The convolutional layer uses filters (kernels) that perform convolution operations as it is scanning the input image to create a feature map that indicates the presence of detected features in the input. For an input image *I*, the output feature map is estimated as follows [30]:

$$\boldsymbol{X^{(k)} = f(\sum_{i=0}^{n}\sum_{i=0}^{n} W_{i,j}^{(k)} I_{l+i,m+j}^{(k-1)} + b^{(k)})} \qquad (1)$$

where $X^{(k)}$ is the output of the featured map, $\boldsymbol{I_{l+i,m+j}^{(k-1)}}$ is the upper layer of the featured map, $\boldsymbol{W_{i,j}^{(k)}}$ is the weight matrix, and $\boldsymbol{b^{(k)}}$ is the bias. *f*, k, and (l, m) are the activation function, the current layer, and the two dimensions of the featured graph of the previous layer, respectively.

The activation function for CNNs is commonly ReLUs. The output dimensions and the input dimensions are the same after passing through the ReLU activation layers. The ReLU layer adds non-linearity in the network and provides non-saturation of gradients for positive inputs [31].

The pooling process involves a two-dimensional filter sliding over the feature map that decreases the dimensions of each feature map but keeps the most powerful features (rotation and position invariants) are unchanged. Thus, it provides effective training for the model and decreases the computing power. Also, pooling help to make the representation

invariant to small translations of the input. In other words, the values of the outputs from the pooling layer did not change [32]. Furthermore, pooling has many approaches; the most commonly used are max-pooling [33], [34], and average pooling [35]. In each pooling region, the max-pooling extracts the max activations while the average pooling considered all activations averaged [36]. It was found that the fully connected layers of CNN are similar to those layers in classical neural networks [37]. A fully connected layer is usually used in the last layers with a softmax function to perform the classification tasks [38]. The output is converted in the last layer into a decimal probability distribution over all the classes. The output of this layer will be composed as follow [30]:

$$\boldsymbol{F_i^{(k)} = X_i^{(k-1)} W^k + b^k} \qquad (2)$$

where both $\boldsymbol{X_i^{(k-1)}}$ **and** $\boldsymbol{F_i^{(k)}}$ refer to the (k-1)-th layer featured map, and fully-connected layer featured map, respectively. $\boldsymbol{b^k}$ is the offset term. The output layer is depicted by a one-hot vector, which has the dimension of the number of classes. The probability of class for the output vector x can be written as [30]:

$$\boldsymbol{P(\hat{y}) = \max\big(P(y_i)\big) \quad 0 < i < k} \qquad (3)$$

where k is the number of classes, $P(y_i)$ is the probability predicted as y_i class and $P(\hat{y})is$ the probability for prediction $\hat{y}$. The correct prediction result is obtained when $\hat{y} = y_i$.

CNNs have four attractive preferences: weight sharing, pooling, local connections, as well as the use of several layers [39]. Therefore, CNNs have achieved a remarkable performance in ASR. Weights sharing refers to using the same weights within a kernel (filter) for a certain receptive field (a small section of the image) and move it (the same filter) through the whole image. In a conventional neural network, each weight element is used once in which it is multiplied by one element of the input when computing the output. In a CNN, each component of the kernel is used at every position of the input. The weights sharing used by the convolution operation means that we use one set of weights for every location [32]. The idea behind sharing the same weights for different locations of the image is to detect the same pattern in different locations of the image [39]. This reduces the memory requirements for the model without affecting the runtime of forwarding propagation. Thus convolution is more efficient and suitable than dense matrix multiplication in terms of memory requirements [32]. For these reasons, CNNs are convenient for images, video, and audio processing.

Figure 1 depicts the architecture of CNN for speech recognition [36]. By turning the speech signals frequencies into images in some manner and then directed to eyes to differentiate. It might be able to understand a larger range of frequencies, and therefore we can use CNNs. Speech can be represented as an image presented as frequency vs. time in the spectrogram. Spectrogram can be considered as an image, and we can apply CNN on it.

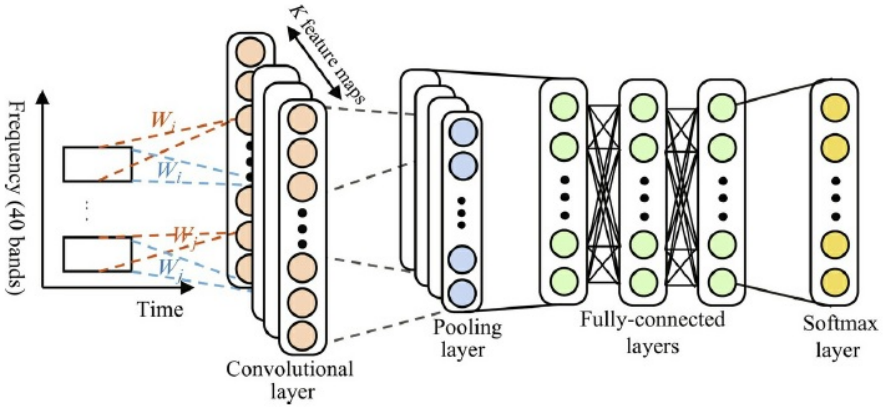


Figure 1: CNN architecture for speech recognition [36]

Speech signals have certain variations caused by the disparity in vocal tract length amongst different speakers. This makes the same speech patterns that result from the same speech units to appear in slightly different frequencies according to the speaker's vocal tract shape. By applying convolution and pooling along the frequency axis, the resulting neural network enjoys more stability against these variations, leading to better performance. Moreover, the convolutional layer neurons receive input from local frequency regions. This results in better stability against band-limited noise.

C. Research Motivation

The Arabic language is one of the most fifth languages in the world because there are about 295 million native Arabic speakers [40]. Besides, it is the official language of the Arab world that consists of 22 countries. It is morphologically

rich and highly ambiguous. The are many colloquial dialects for the Arabic language that differ from region to region used in daily communications. These different dialects are different from Modern Standard Arabic (MSA), used in newspapers and formal communication. Hence, we chose the Arabic language for this research. The motivation for using CNN is due to its achievement in the area of ASR. The use of CNNs is good at handling individual variations in the speech signal and improving the speaker invariance of the acoustic model [5]. This work aimed to examine the use of CNN for automatic Arabic speech recognition using various feature extraction techniques to improve the performance of isolated word AASR.

3 EXPERIMENT

Feature extraction phase and CNN model were implemented in python 3.7 using a laptop with NVIDIA GeForce GTX 1050 GPU in window 10 64 operating system. Librosa version 0.6.3 library [41] and spafe library [42] were used for feature extraction and Keras [43] version 2.3.1 with Tensorflow [44] version 2.1.0 backend for CNN implementation. The proposed model architecture is illustrated in Fig. 2. First, the speech signal is read and normalized. Then relevant features are extracted. The extracted features are passed as inputs to the CNN, which reduces the input vector dimension then predicts the word class.

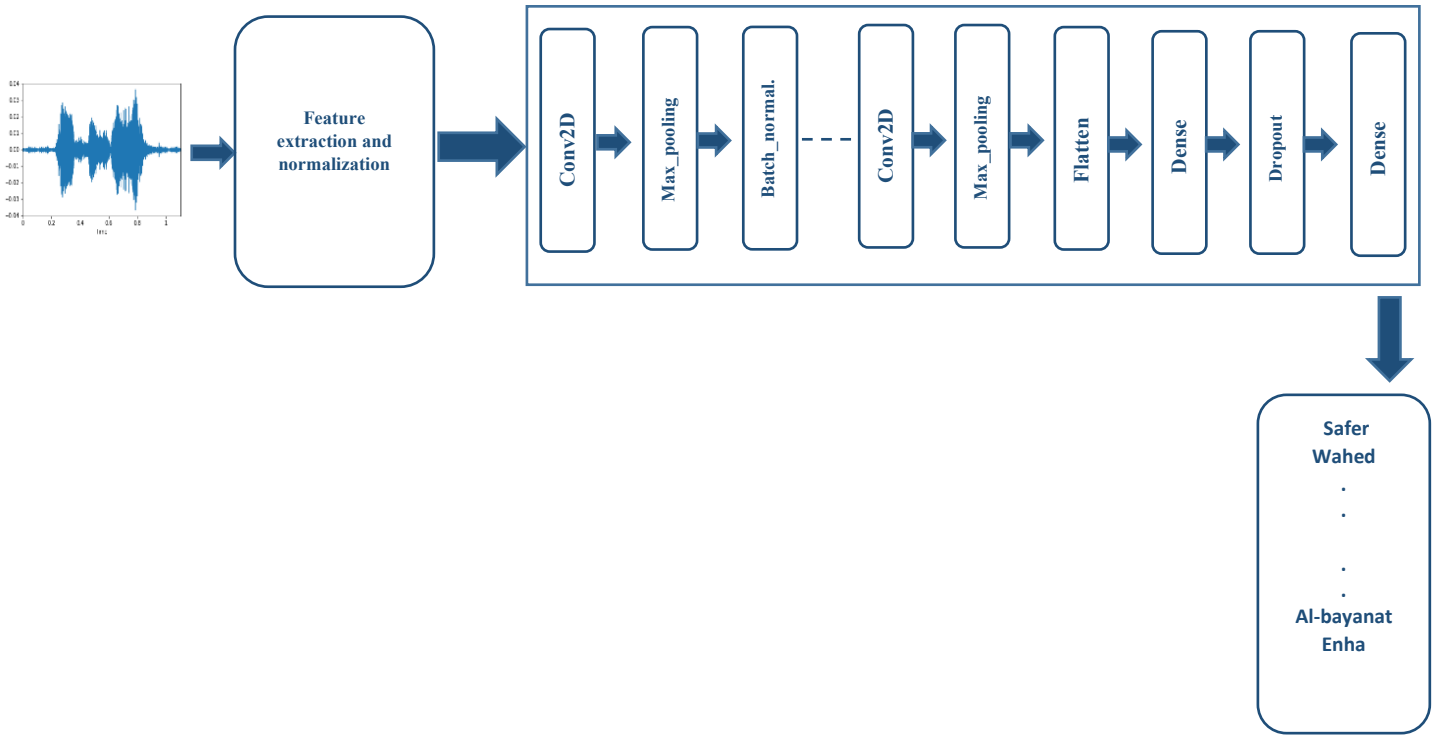


Figure 2: Block diagram of the proposed model

A. Corpus

We evaluate our model on the Arabic speech corpus for isolated words, which was conducted at the Department of Management Information Systems, King Faisal University. It contains 9,992 utterances of 20 words spoken by 50 native male Arabic speakers, as shown in Table 1. The corpus was recorded with a 44100 Hz sampling rate and 16-bit resolution [43]. A method of enlarging the dataset is data augmentation [32]. Using data augmentation, meaning artificially, creates extra speech signals. We create an additional dataset that contains 9,992 utterances by changing pitch, speed, dynamic range, adding noise, and forward and backward shift in time. The new dataset (original & augmented) contains 29,972 utterances is divided into two parts: a training set (training and validation) with 80% of the samples (23,977 tokens) and the test set with the remaining 20% samples (5,995 tokens).

TABLE 1
THE USED CORPUS WORDS, THE NUMBER OF UTTERANCES FOR EACH WORD, ITS ENGLISH APPROXIMATION, AND ITS TRANSLATION [45].

Arabic	Translation	English Approximation	IPA	No. of Utterance
صَفَّرَ	Zero	Safer	sˤfr	493
واحد	One	Wahed	wa:hid	500
اثنَان	Two	Ethnan	PTna:n	500
ثَلاثة	Three	Thlatha	θala:θh	500
أربعَة	Four	Arbah	ʔrbaʕh	500
خَمسة	Five	Khamsah	xmsat	500
سِتّة	Six	Setah	sitat	500
سَبعة	Seven	Sabah	sabʕah	500
ثَمَانِيَة	Eight	Thamanah	θma:njh	500
تِسعة	Nine	Tesah	tisʕah	500
التنشيط	Activation	Al-tansheet	a:tanʃytˤ	500
التحويل	Transfer	Al-tahweel	a:taħwyl	499
الرصيد	Balance	Al-raseed	a:rasyd	500
التسديد	Payment	Al-tasdeed	a:tasdyd	500
نَعِمَ	Yes	Naam	nʕm	500
لا	No	Laa	la:	500
التمويل	Funding	Al-tamueel	a:tamwyl	500
البيانات	Data	Al-bayanat	a:lbayana:t	500
الحساب	Account	Al-hesab	a:lħisa:b	500
انهاء	End	Enha	ʔinha:ʔ	500

B. Feature Extraction

One of the problems of ASR is the high variance of the speech signal due to many parameters: speaking rates, different speakers, contents, and acoustic settings. So, it is better to extract the deterministic features from the speech to reduce variability [46]. In this work, we used Log frequency spectral coefficients (MFSC) and gammatone-frequency cepstral coefficients (GFCC) feature extraction techniques.

1) Log Frequency Spectral Coefficients (MFSC)

The advantage of MFSCs is that they reduce the dimensionality of the STFT spectra and provide a compact set of features of speech [5]. The MFSC is represented by log-mel energy spectrum as an input image for the CNN. The N MFSC coefficients can be determined by using the formula [45]:

$$\boldsymbol{MFSC(n) = \log\left(\sum_{k=0}^{K} H_n(K) * |F(K)|^2\right), n = 1 \ldots . N} \qquad \boldsymbol{(4)}$$

where the $\boldsymbol{|F(K)|^2}$ describes the energy spectrum in the points of kth energy, n is the number of the filter banks, k is the point of the FFTs, and $\boldsymbol{H_n(K)}$ is the mel filter bank. Figure 3 represents the steps involved in MFSC feature extraction.

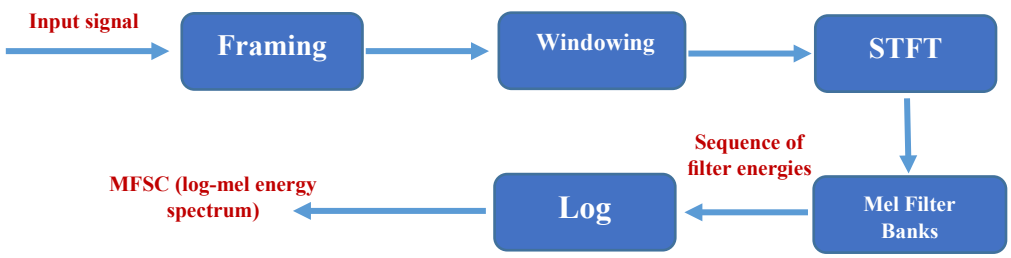


Figure 3: MFSC feature extraction process

For the CNN model, it is useful and important to recognize the features of the speech waveform and its log-mel energy spectra, as displayed in Fig. 4.

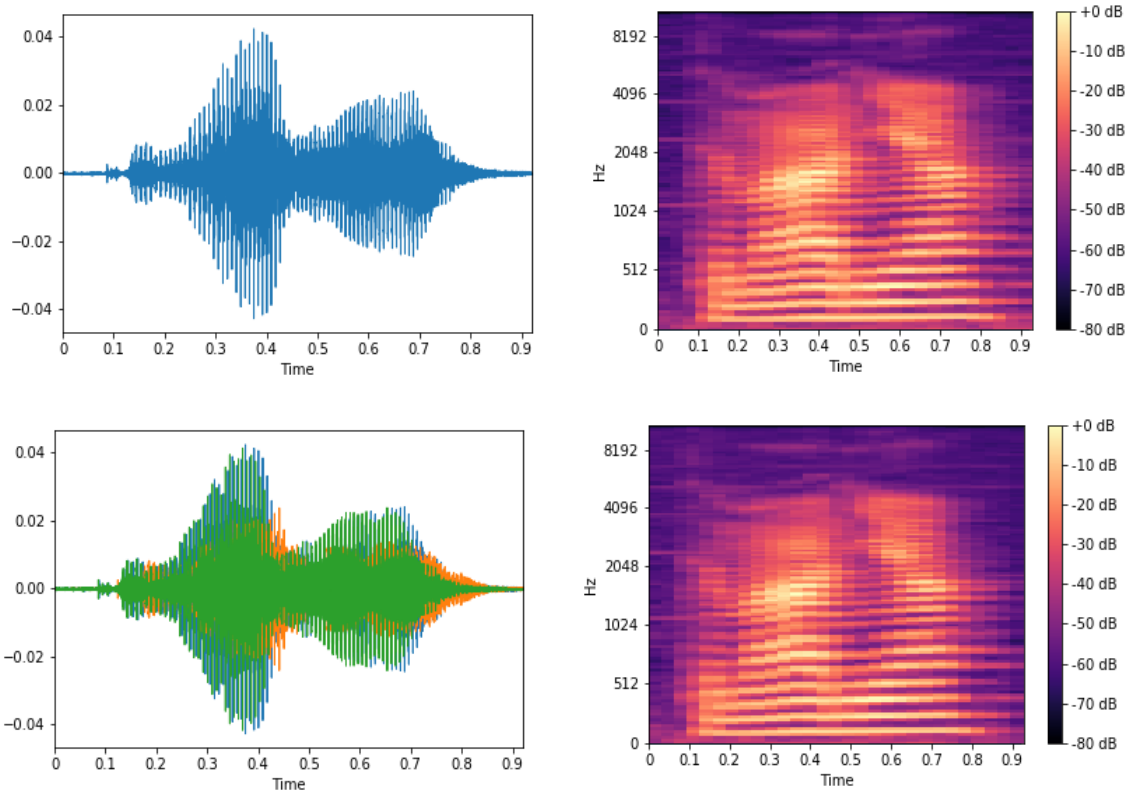


Figure 4: (a) speech waveform (b) log-mel energy spectrogram of speech waveform (c) augmented speech waveform (d) log-mel spectrogram of the augmented speech waveform

2) Gammatone-frequency Cepstral Coefficients (GFCC)

The gammatone filter is motivated to mimic the structure of the peripheral auditory processing stage. The impulse response of a gammatone filter is given in eq. 5, which is similar to the magnitude characteristics of a human auditory filter [47].

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi ft + \varphi) \qquad (5)$$

where *f* is the central frequency, φ is the phase of the carrier, a is the amplitude, n is the filter's order, b is the filter's bandwidth, and t is time. The GFCC computation, shown in Fig 5, starts using a bank of gammatone filters (GT) to decompose the input speech signal into the time-frequency domain, followed by a down-sampling operation of the filter-bank responses along the time dimension. Then a cubic root operation is used to compress the magnitudes of the down-sampled filter-bank responses. Finally, decorrelation is performed using the discrete cosine transform. The signal and its GFCC coefficients are shown in Fig. 6.

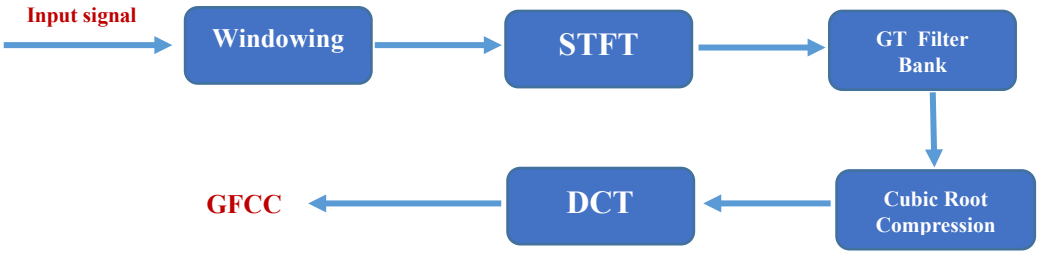


Figure 5 GFCC feature extraction process

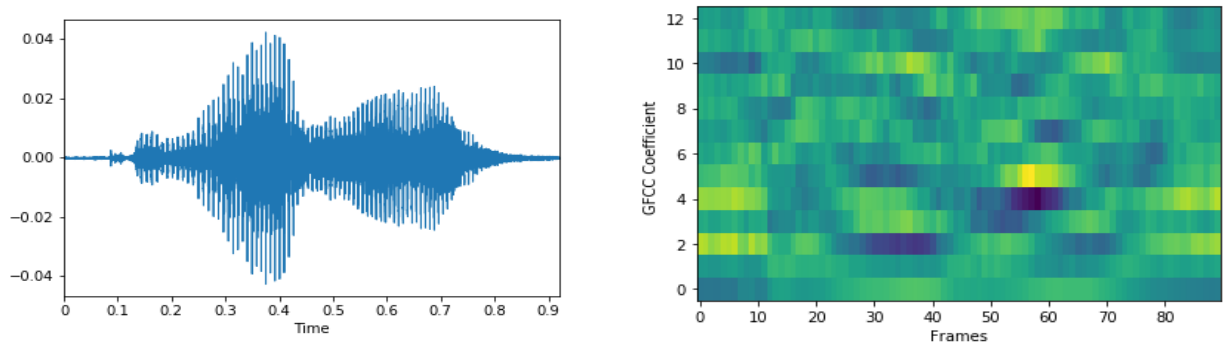


Figure 6: (a) speech waveform (b) GFCCs

C. CNN Implementation

Our CNN model consists of an input, an output layer, and fourteen hidden layers. Hidden layers include five convolutional layers, five pooling layers, one batch normalization, one dropout layer, one dense layer, and a flattened layer. The convolutional layers and pooling layers are filter stages that are used to learn features, and it is followed by a fully connected layer and a softmax layer for the classification stage. ReLU activation function was used for all convolutional and pooling layers except for the fully-connected layer. For the fully connected layer, we used the softmax activation function to output probability. The CNN parameters and details of the layers are shown in Table 2. We trained our network with Adam as an optimizer and categorical cross-entropy loss function with a batch size of 40 for 200 epochs.

TABLE 2
PARAMETERS USED FOR ALL LAYERS IN THE CNN MODEL (CNN ARCHITECTURE)

Layer	No. of filters	Filter size	No. of Nodes	Dropout	Activation
Conv2D x 5	32	(3, 3)			
Max_pooling x 5		(2, 2)			
Activation x 5					relu
Batch_normalization x 1					
Flatten x 1					
Dense x 1			128		
Dropout x 1				0.25	
Output			20		softmax

4 RESULTS AND DISCUSSION

First, all speech signals are normalized, and therefore, each vector dimension has a zero mean and unit variance. Speech signals are padded with silence to fit the length of the longest speech signal then generated the features from speech. The features coefficients vector was fed as an input to CNN. Features are extracted using the following settings 16000 sample rate, the length of FFT and hop size of 512, and 160 samples, respectively, and Hamming window function. The number of cepstral coefficients for GFCC was set to 13. For each frame 13 GTCC was extracted with their first and second-order derivatives. A vector of size (39,187) is obtained. For MFSC, 128-bands log-mel-spectrograms are computed, generating a vector of size (128,187).

The first CNN layer performs a convolutional over the input feature vector with 32 ReLU kernels of 3x3 receptive fields and stride of 1 in both dimensions. The obtained feature maps are then downsampled with a 2 x 2 max-pooling layer and followed by a batch normalization layer, which reduces the covariant shift by normalizing the input of each intermediate activation layer. Then the other five convolutional and max-pooling layers are used, followed by a flatten layer, dense layer with 128 nodes, and dropout layer with a probability of retention p = 0.25. Finally, the classification involves 20 classes. The last is a softmax layer composed of 20 fully-connected neurons. Table 3 shows the results for a series of recognition experiments to determine the effect of different feature extraction techniques. As observed from the results that the two feature extraction techniques give good results with CNN, but as shown GFCC gives the highest accuracy.

For the test set data, the classes are compared to each other using the confusion matrix in Fig 9. The number of correct and incorrect predictions are collected with count values. Besides, Table 4 presented the precision, recall, and F1-score, which showed the quality of predictions of the proposed model. F1 is given as:

$$F_1 = 2x\frac{Precision\ x\ recall}{Precision + Recall} \quad (6)$$

Precision is defined as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (7)$$

And recall is defined as:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (8)$$

It would be better to compare the results of the model that we used with those of previous reports [18], [22], [23] (see Table 5).

TABLE 3
DATA FOR TRAINING AND TESTING ACCURACY.

Features	Training accuracy	Training loss	Validation accuracy	Validation loss	Test accuracy	Test loss
MFSE	99.91	0.0653	99.95	0.0552	99.22	0.076
GFCC	99.94	0.0509	99.78	0.0547	**99.77**	0.0540

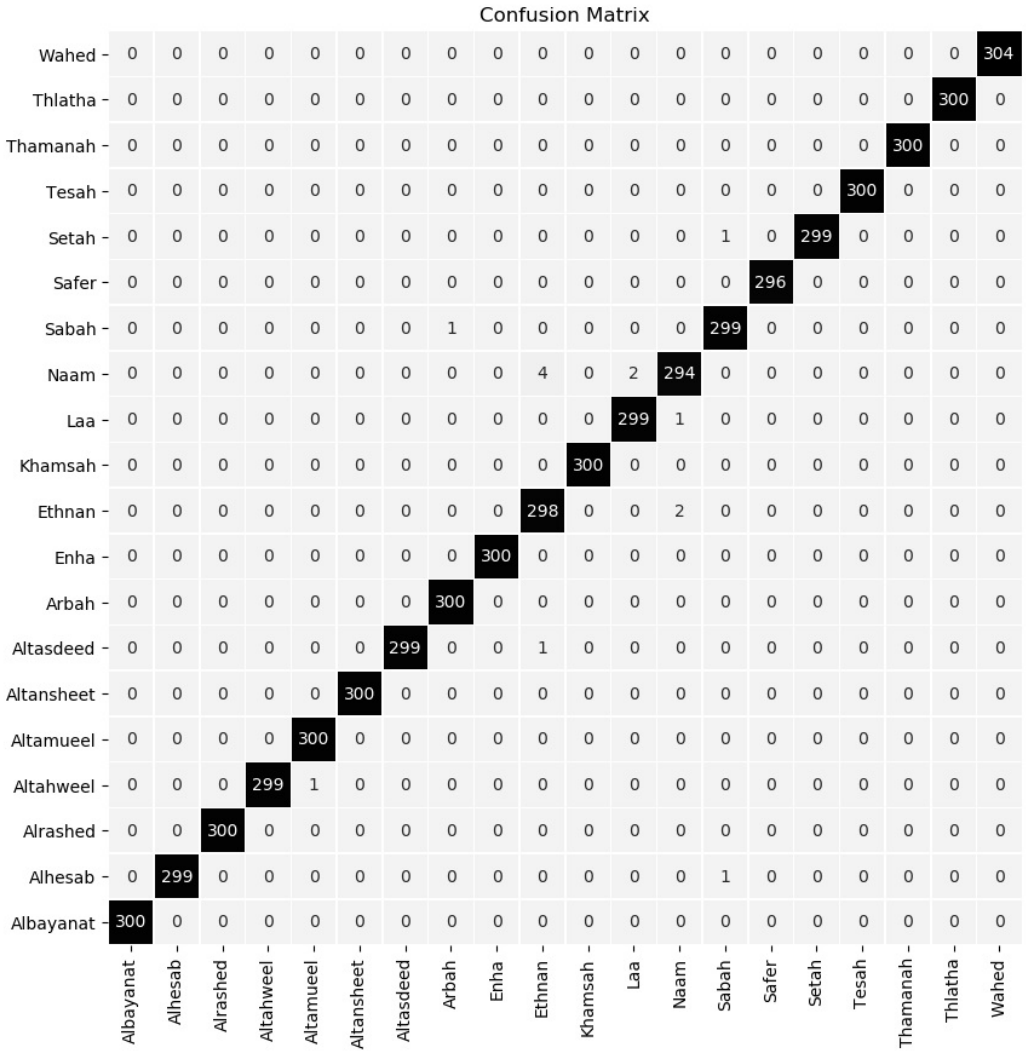


Figure 9: Confusion matrix of the proposed model using GFCC features

TABLE 4
CLASSIFICATION REPORT SHOWING CLASS-WISE PRECISION, RECALL, AND F1-SCORE OF THE PROPOSED MODEL USING GFCC FEATURES.

	Precision	**Recall**	**F1-score**
Al-baynat	1.0000	1.0000	1.0000
Al-hesab	1.0000	0.9967	0.9983
Al-raseed	1.0000	1.0000	1.0000
Al-tahweel	1.0000	0.9967	0.9983
Al-tamueel	0.9967	1.0000	0.9983
Al-tansheet	1.0000	1.0000	1.0000
Al-tasdeed	1.0000	0.9967	0.9983
Arbah	0.9967	1.0000	0.9983
Enha	1.0000	1.0000	1.0000
Ethnan	0.9835	0.9933	0.9884
Khamsah	1.0000	1.0000	1.0000
Laa	0.9934	0.9967	0.9950
Naam	0.9899	0.9800	0.9849
Sabah	0.9934	0.9967	0.9950
Safer	1.0000	1.0000	1.0000
Setah	1.0000	0.9967	0.9983
Tesah	1.0000	1.0000	1.0000
Thamanah	1.0000	1.0000	1.0000
Thlatha	1.0000	1.0000	1.0000
Wahed	1.0000	1.0000	1.0000
All	0.9977	0.9977	0.9977

TABLE 5
COMPARISON BETWEEN THE RECOGNITION ACCURACY OBTAINED BY OUR MODEL AND THE RESULTS OBTAINED RESULTS IN [18], [22], [23]

Models	**Dataset**	**Features**	**Recognizer**	**Accuracy**
Model in [18]	Ten Arabic letters (400 tokens)	MFSC+delta+double delta	CNN	80.75 %
Model in [22]	Arabic digits 0 through 9 (1040 tokens)	MFCCs	LSTM	69 %
Model in [23]	Arabic Speech Corpus for Isolated Words (9,992 tokens)	RASTA-PLP	ESNEKM	99.69 %
Proposed model	Arabic Speech Corpus for Isolated Words (29,972 tokens)	GFCC	CNN	**99.77 %**

5 CONCLUSION

In this work, we implemented a robust speaker-independent automatic Arabic speech recognition model based on a convolutional neural network to recognize Arabic words. CNNs have considerable advantages in acoustic modeling. The problem of small dataset size was solved by using data augmentation. We compared several feature extraction techniques. The suggested model was evaluated on Arabic speech corpus for isolated words. It was shown that GFCC performed the best recognition accuracy compared to other feature extraction techniques. The proposed model showed efficient results and improved performance compared to the results obtained in previous works.

REFERENCES

[1] L. D. Dong Yu, Automatic speech recognition - A Deep Learning Approach. *Springer*, 2015.
[2] E. Elmaghraby, A. Gody, and M. Farouk, “Speech Recognition Using Historian Multimodal Approach,” Egypt. *Journal of Language Eng*ineering, vol. 6, no. 2, pp. 44–58, Sep. 2019.
[3] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion, “Seeing it all: Convolutional network layers map the

function of the human visual system," *Neuroimage*, vol. 152, pp. 184–194, May 2017.

[4] F. Sultana, A. Sufian, and P. Dutta, "Advancements in Image Classification using Convolutional Neural Network," in *Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN),* pp. 122–129, Kolkata; India, Nov. 2018.

[5] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[6] D. Palaz, R. Collobert, and M. M. -Doss, "End-to-end Phoneme Sequence Recognition using Convolutional Neural Networks," *Proc. NIPS Deep Learn. Work.*, pp. 1–8, Dec. 2013.

[7] D. Palaz, M. Magimai.-Doss, and R. Collobert, "Convolutional Neural Networks-based continuous speech recognition using raw speech signal," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2015-August, pp. 4295–4299, 2015.

[8] D. Palaz, M. Magimai-Doss, and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2015-January, pp. 11–15, Dresden, Germany, 2015.

[9] E. ElMaghraby, A. Gody, and M. Farouk, "Noise-Robust Speech Recognition System based on Multimodal Audio-Visual Approach Using Different Deep Learning Classification Techniques," *Journal of Language Engineering*, vol. 7, no. 1, pp. 27–42, Egypt, Apr. 2020.

[10] D. Scherer, H. Schulz, and S. Behnke, "Accelerating large-scale convolutional neural networks with parallel graphics multiprocessors," in *Diamantaras K., Duch W., Iliadis L.S. (eds) Artificial Neural Networks – ICANN 2010. Lecture Notes in Computer Science, vol. 6354. Springer*, pp. 82–91, Berlin, Heidelberg, 2010.

[11] M. Bojarski et al., "VisualBackProp: Efficient visualization of CNNs for autonomous driving," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4701–4708, Brisbane, QLD, Australia, 2018.

[12] G. K. Sheela and Mt. Student, "A Survey on Different Algorithms for Automatic Speaker Recognition Systems," *Int. J. Eng. Res. Gen. Sci.*, vol. 4, no. 1, 2016.

[13] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, vol. 2017-January, pp. 936–944, Honolulu, HI, USA, 2017.

[14] M. A. Haque, A. Verma, J. S. R. Alex, and N. Venkatesan, "Experimental evaluation of cnn architecture for speech recognition," in *Luhach A., Kosa J., Poonia R., Gao XZ., Singh D. (eds) First International Conference on Sustainable Technologies for Computational Intelligence. Advances in Intelligent Systems and Computing*, vol. 1045, Springer, pp. 507–514, Singapore, 2020.

[15] J. Li et al., "Jasper: An End-to-End Convolutional Neural Acoustic Model," *INTERSPEECH,* Apr. 2019.

[16] D. Nagajyothi and P. Siddaiah, "Speech recognition using convolutional neural networks," *Int. J. Eng. Technol.*, vol. 7, no. 4.6 Special Issue 6, pp. 133–137, 2018.

[17] B. Zada and R. Ullah, "Pashto isolated digits recognition using deep convolutional neural network," *Heliyon*, vol. 6, no. 2, Feb. 2020.

[18] R. A. Rajagede, C. K. Dewa, and Afiahayati, "Recognizing Arabic letter utterance using convolutional neural network," in *Proceedings - 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 181–186, Kanazawa, Japan, 2017.

[19] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, "Bidirectional deep architecture for Arabic speech recognition," *Open Comput. Sci.*, vol. 9, no. 1, pp. 99–102, Jan. 2019.

[20] L. Boussaid and M. Hassine, "Arabic isolated word recognition system using hybrid feature extraction techniques and neural network," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 29–37, 2018.

[21] E. S. Wahyuni, "Arabic speech recognition using MFCC feature extraction and ANN classification," in *Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, vol. 2018-January, pp. 22–25, Yogyakarta, Indonesia, 2018.

[22] A. S. Mahfoudh Ba Wazir and J. Huang Chuah, "Spoken Arabic Digits Recognition Using Deep Learning," in *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS) - Proceedings,* pp. 339–344, Selangor, Malaysia, 2019.

[23] A. Alalshekmubarak and L. S. Smith, "On improving the classification capability of reservoir computing for Arabic speech recognition," in *Wermter S. et al. (eds) Artificial Neural Networks and Machine Learning – ICANN 2014. ICANN 2014. Lecture Notes in Computer Science*, vol. 8681. Springer, pp. 225–232, 2014.

[24] D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," *Speech Commun.*, vol. 108, pp. 15–32, Apr. 2019.

[25] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, "Bi-directional recurrent end-to-end neural network classifier for spoken Arab digit recognition," in *2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pp. 1–6, Algiers, Algeria, 2018.

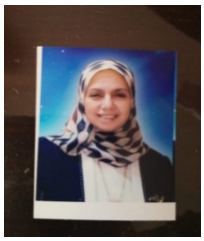
[26] "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way." [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. [Accessed: 30-Jan-2020].

[27] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-September-2016, pp. 410–414.
[28] S. Thomas, G. Saon, H. Kuo, L. Mangu, I. B. M. T. J. Watson, and Y. Heights, "The IBM BOLT Speech Transcription System.", in *Sixteenth Annual Conference of the International Speech Communication Association*, pp. 3150–3153, Dresden, Germany, 2015.
[29] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.,* vol. 160, no. 1, pp. 106–154, Jan. 1962.
[30] H. Fan, S. Gao, X. Zhang, X. Cao, H. Ma, and Q. Liu, "Intelligent Recognition of Ferrographic Images Combining Optimal CNN with Transfer Learning Introducing Virtual Images," *IEEE Access*, vol. 8, pp. 137074-137093, Jul. 2020.
[31] S. Pattanayak, Pro Deep Learning with TensorFlow. *Apress Berkely*, CA, USA, 2017.
[32] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, *Cambridge: MIT press*, Vol. 1, No. 2, 2016.
[33] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2519–2523, Florence, Italy, 2014.
[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks." *Communications of the ACM,* Vol. 60, no. 6, pp. 84–90, 2017.
[35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2323, 1998.
[36] N. Dey, Intelligent Speech Signal Processing, *Academic Press*, 2019.
[37] D. Sarkar, R. Bali, and T. Sharma, Practical machine learning with Python : a problem-solver's guide to building real-world intelligent systems, *Springer*, New York, NY, 2018.
[38] W. Di, A. Bhardwaj, and J. Wei, Deep learning essentials : your hands-on guide to the fundamentals of deep learning and neural network modeling, *Packt Publishing*, New York, NY, 2018.
[39] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436–444, 2015.
[40] "Most Spoken Languages In The World (native & non - native)." [Online]. Available: https://www.saigonizers.com/the-most-spoken-languages-in-the-world/. [Accessed: 04-Feb-2020].
[41] "librosa/librosa: Python library for audio and music analysis." [Online]. Available: https://github.com/librosa/librosa. [Accessed: 13-Aug-2020].
[42] "SuperKogito/spafe: spafe: Simplified Python Audio-Features Extraction." [Online]. Available: https://github.com/SuperKogito/spafe. [Accessed: 13-Aug-2020].
[43] F. Chollet, "Keras," GitHub repository, 2015. [Online]. Available: https://github.com/fchollet/keras.
[44] "TensorFlow." [Online]. Available: https://www.tensorflow.org/. [Accessed: 03-Feb-2020].
[45] "The Arabic Speech Corpus for Isolated Words." [Online]. Available: http://www.cs.stir.ac.uk/~lss/arabic/. [Accessed: 04-Apr-2020].
[46] U. Shrawankar and V. M. Thakare, "Techniques for Feature Extraction In Speech Recognition System : A Comparative Study," *Computer and Information Sciences*, vol. 6, no. 1, pp. 58-69, 2013.
[47] J. M. Liu et al., "Cough signal recognition with gammatone cepstral coefficients," in *IEEE China Summit and International Conference on Signal and Information Processing*, pp. 160–164, Beijing, China, 2013.

BIOGRAPHY

Engy R. Rady received the B.Sc. (Honor's) degree in Physics from the Faculty of Science, Fayoum University in 2006. She received the M.Sc. degree in speech recognition systems from Faculty of Science, Ain shams University in 2013. She is currently a Ph.D. student at the Faculty of Science, Fayoum University. She is working as Assistant Lecturer at the Basic Science Department, Faculty of Computers and Information, Fayoum University. Her research interest is in signal processing and deep learning.

Hassen received Ph.D. in Natural Science from the Faculty of Science and Mathematics, University of Augsburg, Germany, in 2003. He was awarded postdoctoral scholarships from Indian National Science Academy (INSA) in 2005 and South Korea (BK21 project), Pusan National University, in 2007. He is a Professor of Materials Science, Department of Physics, Faculty of Science, Fayoum University. He was the Dean (from Aplil 2017 to April 2020) of the Faculty of Science, Fayoum University, Egypt. He published many papers in international peer-reviewed journals and conferences. He is an active reviewer for different high-ranking international journals. He becomes Associate Editor for two international Journals.



N. M. Hassan
Dean Faculty of Computers and Information
Prof. of Theoretical Physics- High Energy.
Previewer in National Authority for Quality Assurance and Accreditation of Education

Mohamed H. Farouk received the B.Sc. in Electronics Engineering from the Faculty of Engineering, Cairo University. Egypt, in 1982. He received the M.Sc. and PhD. of Engineering Physics from the Faculty of Engineering, Cairo University, Egypt, in1988 and 1994 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Cairo University, Egypt in 1984. His Current Position is full Professor, Engineering Math. & Physics Dept., Faculty of Engineering, Cairo Univ from 2007-Till Now. He is the author and co-author of about 40 papers in national and international conference proceedings and journals.

إستخدام الشبكة العصبية التلافيفية لتصميم نظام للتعرف علي الأصوات العربي

انجي رجائي راضي*[1]، محمد هشام فاروق**[2]، نبيلة محمد حسن *[3] ، عرفة صبري حسن***[4]

*قسم العلوم الاساسية, كلية الحاسبات والمعلومات, جامعة الفيوم, الفيوم, مصر
[1]era00@fayoum.edu.eg
[3]nmh00@fayoum.edu.eg

**قسم هندسة الرياضيات والفيزيقا, كلية الهندسة, جامعة القاهرة, جيزة, مصر
[2]mhesham@eng.cu.edu.eg

***قسم الفيزياء, كلية العلوم, جامعة الفيوم, الفيوم, مصر
[4]ash02@fayoum.edu.eg

ملخص

ركز هذا العمل على التعرف علي الأصوات العربي باستخدام كلمات منفصلة. تم إستخدام تقنيات مختلفة أثناء إستخراج خصائص الصوت مثل MFSC ، GFCC بمشتقاتها من الدرجة الأولى والثانية. تُستخدم الشبكة العصبية التلافيفية (CNN) لأداء تعلم الميزات والتصنيف. حققت CNN أداءً جيد في التعرف التلقائي على الكلام (ASR). يعد الإتصال المحلي ومشاركة الوزن والتجميع من الخصائص الرئيسية لشبكات CNN التي لديها القدرة على تحسين ASR . لقد تم اختبار نموذج CNN علي قاعدة البيانات لبعض مقاطع اللغة العربية. تم تعزيز قاعدة البيانات المستخدمة من خلال تطبيق بعض التحويلات مثل تغيير درجة الصوت والسرعة والنطاق الديناميكي وإضافة الضوضاء علي المقاطع. وجد أن أقصى دقة تم الحصول عليها عند إستخدام GFCC مع CNN هي 99.77٪. تم مقارنة نتائج هذا العمل بنتائج الأبحاث السابقة ولوحظ أن شبكة CNN حققت أداءً أفضل في ASR.

الكلمات المفتاحية

التعرف التلقائي علي الكلام العربي، MFSC، GFCC، الشبكة العصبية التلافيفية، كلمات منفصلة