A Review: Isolated Arabic Words Recognition Using Artificial Intelligent Techniques

S R Shareef1,*, Y F Irhayim2
1Department of Computer science, College of computer science and mathematics, University of Mosul, Iraq
2Department of Computer science, College of computer science and mathematics, University of Mosul, Iraq
*Email: sura.ramzishareef@uomosul.edu.iq
Yusrafaisalc@uomosul.edu.iq

Abstract. In recent few years, deep learning has fast growing in many fields as natural language processing, image recognition, handwriting recognition, computer vision, and speech recognition. Automatic speech recognition (ASR) is a technique that refers to translating spoken words from an acoustic waveform into a text equivalent to what the speaker says. More recently, the advances in deep learning can support ASR in improving the performance of systems accuracies. Arabic is a Semitic language, one of the oldest used and most communicated languages in the world. But, it least concentrated in the case of Arabic speech recognition and under-resourced languages. This paper presents a survey that focuses on an automatic speech recognition system based on isolating words technique for Arabic speech. It also highlights the facilities and tools for developing speech recognition systems. This work is intended to be a useful starting point for those who are interested in ASR.

1. Introduction

Recently, deep learning's have made big developments in many machine learning areas as speech recognition, machine translation and natural language processing [1]. Arabic is an official language for more than 22 countries. The Arabic language is one of the most widely native languages for more than 313 million people around the world [2]. It is written from the right direction to the left direction and it consists of 28 letters. The forms of letters' are changed according to their locations within a word. In fact, any Arabic word may have several meanings. Languages are communication systems that enable speakers to make more effective utilize of usually learned words. The characteristics of a speech sound is based on the certain human language. The speech natural way, a large information can be transferred to the listener in a short time. Speech is particularly the main method of communication among people, it is important for understanding, learning reading or writing. Today speech technologies are enable machines to respond fast, correct and reliable by using humans’ voices rather keyboards [2][3].

Speech recognition technology has begun in 1950. Then, in 1952 it was improved for isolating basic words by the digital identification system. Speech recognition system enables isolating words of users by a machine [4]. Speech recognition identifies specific spoken words depending on collected features from the speech signals. The purpose of speech recognition is to find the variability of features for the applied words. Change it to "Through last sixty and seventy years, the automatic speech recognition system was able to
process isolated words to perform templates for pattern recognitions [5]. Automatic Speech Recognition (ASR) is exploited to identify the used words by a person in order to authenticate his/her identity. It uses hardware and software techniques of computers to identify and process the human’s voice. ASR is widely used in many areas of our lives such as education, health, security and communication [5].

In 1980, the artificial neural networks, which are the predecessors of deep neural networks, change the directions of speech recognitions. Then, the deep learning appears as a modern field of machine learning. It utilizes in different areas such as handwriting recognition, natural language processing, speech recognition and computer vision as shown in Figure 1. The deep neural network as a type of machine learning attempt to learn by extracting specific features and information. It can hierarchy learn to extract features from multiple layers at a time [6] [7].

The purpose of this paper is to review previous work by highlighting the available studies on isolating Arabic words by exploiting deep learning techniques in terms of speech recognition. The remaining of this paper is distributed as follows: Section 2 provides related work, Section 3 illustrates the background of this study, Section 4 describes the challenges and Section 5 presents the conclusion.

![Figure 1. Statistical demonstrations for the deep learning applications of computer vision, speech recognitions and natural language processing [8]](image-url)
2. Related work

In this section, previous work in the area of speech recognition for isolating Arabic phrases will be clarified. In addition of providing a summarized table that reports the speech representations, feature extraction techniques, databases and performances.

ASR study was illustrated for the field of isolating words from small to medium predefined vocabularies. Phonetics were used to perform the pattern recognition template. ASR, which is the process of deriving the transcription (word sequence) of an utterance, was given for the speech waveform.[5]

Dua M. et al. [9]. In 2012 a system approach was implemented for ASR isolating words. Indian regional language of Punjabi was utilized. Hidden Markov Model Toolkit (HTK) was exploited based on Hidden Markov Model (HMM). The dataset here was reported for 115 various words from speakers, each word was spoken 3 times for every speaker. System result was 95.63% [9].

Rehmam B.et al. [11]. In 2015 a system of speaker independent on speech recognition was presented for isolating words from oriental languages. This paper proposed a combined work between the Feed Forward Artificial Neural Network (FFANN) and Digital Wavelet (DWT) to recognize isolated words. The dataset was created by capturing speech signals and then employing pre-processing operations. The DWT was consisted of 5 levels and an analysis of db-8/ Haar for the features extractions. This system obtained high accuracies for two and five classes. For db-8, level-5 of the DWT filter attained 95.20%, 95.73% and 98.40% accuracy rates with 20, 15, and 10 classes, respectively. For Haar, level-5 of the DWT filter achieved 91%, 94.40% and 97.20% accuracy rates with 20, 15, and 10 classes, respectively [11].

Almisreb A. A .et al. [11] presented the training and testing over a corpus that consisted of recorded consonant Arabic phonemes from 20 Malay speakers. The data are divided into five waveforms for training the proposed Maxout network and fifteen waveforms for testing. Test performance of Deep Neural Networks (DNN) based on Maxout used the Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction and Maxout algorithm with dropout function to improve the DNN efficiency. Experimentally, the proposed dropout function for training has shown considerable performance over the sigmoid and Rectified Linear Unit (ReLU) functions. The result yielded that the Maxout based deep structure gave better performance with lowest error rate than other deep networks such as the Restricted Boltzmann Machine (RBM), Deep Belief Network (DBN), Convolutional Neural Network (CNN), Conventional Feedforward Neural Network (CFNN) and Convolutional Auto-Encoder (CAE) [11].

EL-Mashed S.Y., Sharway M.I.and Zayed H.[12]. In 2017, recognizing a speaker of Arabic speech by using the Support Vector Machine (SVM) was studied. In this work a model was applied for the connected Arabic digits (numbers) by exploiting the neural networks. The input information to the neural networks was numbers for the recognition stage. A corpus of 1000 values was built in the system composing 10000 numbers reported by 20 speakers of various genders, ages, physical conditions…, and in an unquiet environment. Secondly, each reported value was digitized into 10 separate numbers. The features of these numbers were extracted by exploiting the MFCC. The system performance was reached 94% by using the SVM [12].

Boussaid L a, Hassine M. [13]. In 2017, a speaker-dependent speech recognition system was implemented for 11 standard Arabic isolated words. Feature extraction of MFCC was utilized. In addition to perceptual linear prediction and relative perceptual linear prediction. This system used the learning algorithm of the Feed Forward Back-Propagation Neural Network (FFBPNN). Hybrid recognition rates have been produced a tested error rate reached to 0.26% when using a combination between the Relative Spectral-Perceptual Linear Prediction (RASTA-PLP), Principal Component Analysis (PCA) and FFBPNN techniques [13].
Saleh, A., Wazir M. Ba. and Chuah J. H. [14]. In 2018, a study was presented for developed deep learning in the case of speech recognition. 1040 samples of Arabic language dataset were exploited, 840 for the training phase and 200 for the testing phase. In this work, feature was extracted by using the MFCC and Long Short-Term Memory (LSTM) techniques. The obtained result here was recorded as 94% of accuracy [14].

Zerari N.et al. [15]. In 2019, using the neural network and Long Short-Term Memory (LSTM) was suggested for the framework of Arabic speech recognition. The feature extraction techniques were the Mel Frequency (MF) and Filter Banks (FB) coefficients. The sequences of MFCC/FB values were encoded as a specific size vector. Then, this vector presented to the Multi- Layer Perceptron (MLP). This work represented a deep architecture by a Gated Recurrent Unit (GRU) or recurrent LSTM to perform classifications. This system used two different datasets. Firstly, the spoken digit recognition. Secondly, the spoken TV commands. All experiments are carried out for the two datasets. The accuracy of 95% was just satisfied to symbolize the speech signal for both feature extraction techniques the FBs and MFCCs. whilst, accuracy over 96% was obtained for the delta features task [15].

Zada B., Ullah R.[16]. In 2020, isolating digits for Pashto language recognition was developed using the Convolutional Neural Network (CNN). A dataset of digits from 0 to 9 was exploited with 50 utterances for each digit. The MFCC was used as a feature extraction to isolate digits and fed inputs to the CNN. The network consisted of 4 convolution layers, tailed by ReLU and max-pooling layers. It was trained and tested. A total of average results was benchmarked for 84.17% accuracy [16].

Khudeyer R S, Alabbas M. and Radif M.[17]. In 2020, combining multiple machine learning classifiers for recognizing multi-fonts separated printed Arabic letters was considered. This technique was depended on fusing three classifiers of machine learning. These are the generalized Regression Neural Network (GRNN), k-Nearest Neighbors (kNN), and SVM. Then, the majority vote and Structural Similarity Index (SSIM) were utilized to generate the final output. Experimental performances show that combined techniques are always outperformed individual methods in terms of in isolation recognition. The total recognition rates of each system were 63.8%, 64.5%, 58.4%, and 76.7% for System 1(uses GRNN), System 2 for uses KNN, System 3 for uses SVM, and System4 is a combination of the three previous systems. [17].

Table 1 shows a summary of feature extraction methods, architectures of deep learning techniques, types of employed datasets and the accuracy rates that was utilized in previous work.

| Reference | Feature extraction | Classification Method | Dataset | Accuracy |
|-----------|--------------------|-----------------------|---------|----------|
| Moner N. M.et al. [18] | MFCC | HMM | KNN | 63 children (ages 5-11), requested to produce 28 Arabic phonemes for 10 times. | 83.75% |
| El Kourd1 A., El Kourd K. [19] | 1-MFCC, 2-LPC | DTW | Gaussian mixture model | 600 Samples | 94.56% |
| Authors | Feature Extraction Method | Model | Recording Details | Results |
|---------|---------------------------|-------|-------------------|---------|
| Wahyuni E. S. [20] | MFCC | Neural networks | Recording voices for human with the pronounced letters of sa (س), sya (ش) and (ث) | 92.42% |
| Emami A., Mangu, L. [21] | N-gram model N-gram model. | Neural network language models for arabic speech recognition | Samples of words acquired from broadcast conversations and Arabic broadcast news | 3.8% WER |
| Sadeghian R. [22] | MFCC | HMM | DNN | 206 children (ages 5-11), each child has spoken 100 individual words that were selected from a dictionary | 72.38% (age 10) 68% (age 9) 65% (age 8) |
| Mahfoudh A. S. [14] | 1-MFCC 2-LSTM | HMM | RNN | 1040 samples of data (840 samples for training & 200 samples for testing) | 94% |
| Bourouba, H.et al. [23] | MFCC log pitch of energy | HMM | KNN SVM | 920 samples (92 speakers × 10 digits) 92 participants (46 males and 46 females) | 89.79% for KNN 87.48% for SVM |
| Hachkar Z et al. [24] | 1-MFCC 2-Energy | DTW | DHMM | 500 samples (10 digits × 5 speakers × 10) | 92% |
| Zada B., Ullah R. [16] | MFCC | HMM | CNN | 50 utterance digits from 0 to 9 | 84.1% |

From the feature extraction method perspective, features extracted from the spoken utterance is the acoustic front-end task. Here, can discuss some properties of these methods. The PLP and MFCC parameters take into account the nature of speech when extracting the features. Therefore, can see these methods (PLP, MFCC) most frequently with ASR systems. PLP is one of perceptual prediction methods, it can suitable for noisy conditions. While the LPC parameter relies on previous features to predicts the other future features. But, because of its nature-based on linear computation, make it is not so acceptable. The MFCC and PLP is a good choice comparing to LPC for ASR,
due to the non-linear nature of the voice of the speaking person. From the perspective of pattern recognition techniques or classification model, there are many models that were being used. From table 1 can see the classifiers based on Neural Network and HMM, it had better results compared to other classifiers. And can be observed it clearly the poor results of models based on N-gram. To summarize Table 1, it can be noted that the MFCC as a feature extraction was adopted in most studies and works presented by this review. The works that have adopted a combined MFCC with other feature extraction methods such as [11] [14] [19] [24] are achieved better accuracy results than others. And more specifically, among these works can also turn out that the adoption of the RNN as a classification algorithm yielded more accurate results between it.

3. Background

3.1 Automatic Speech Recognition (ASR)

Speech is a natural way of communication among people. Automatic Speech recognition is the process by which a computer can identify spoken words and efficaciously recognize what are saying. The ASR is the first part of an intelligent system. It is a technology of mapping an acoustic signal into a string of words, these words can be the final outputs or inputs to natural language processing. The goal of ASR systems is to recognize natural languages that are spoken by humans [25]. ASR technology has widely used in computers that have the user voice interface, foreign languages' application, dictation, hands free operations and controls which make interactions between machines and humans more faster and easier than utilizing keyboards, where people can interact with society [25][6].

Speech recognition systems can be separated into several different categories such as speaker dependent systems, which work only for the speech of a certain speaker, speaker independent systems, which can be work for the speech of any participant, and isolated continuous speech recognitions for small/large vocabularies, where this is based on the fact that an ASR has the ability to determine when a speaker starts and finishes an utterance. Some interesting categories can be illustrated as follows:

- Isolated words: isolated words strategy is used in the wide variety of recorded speech archives. Its recognizers normally require every utterance to have equal size of a window pattern. It accepts a single phrase or single utterance at a time. These structures have "Listen/Not-Listen" states that represent the required places for the speaker to wait between utterances.
- Connected words: connected word strategies (or more correctly 'connected utterances') are close to isolated words systems, but allow a minimal pause between the separated utterances to be used together.
- Continuous speech: continuous speech recognizing strategies permit participants to often speak naturally, during determining the contents. Non-stop speech abilities are considered as some of the most tough challenges as they use exclusive techniques to decide utterance borders.
- Spontaneous speech: spontaneous speech refers to the normal sounding and not rehearsed speech. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters [19][26].

A speech recognition system can be seen as mathematical model as Hidden Markov Model (HMM) /GMM of supervised, unsupervised and semi-supervised models. It generally composes of four components. These are the pre-processing, feature extraction, classification and data as shown in Figure 2.
When an input signal acquired, it is converted into a fixed-size acoustic vector. The pre-processing prepares the speech input to the feature extraction. In this component, a set of common functions can be implemented as pre-emphasis, noise removal, framing, endpoint detection, and normalization [5]. Consequently, the feature extraction component is used to extract a group of information from the pre-processed signal. Extracted features should be capable to differentiate between classes, there are various methods of feature extractions such as the Linear Predictive Coding (LPC), MFCC and DWT [27]. The next component is classification, it utilizes to efficiently classify the input speech signals depending on these extracted features. Classification stage can be done by assigning the joint probability distributions over the given observations to the class labels. Examples of methods in this stage are HMM and Gaussian Mixture Model (GMM) [28]. To illustrate the data component, it firstly refers to the represented information in overall system such as the language model and acoustic model. The language model maps the personal fluent sequence of information to final transcription [29]. It is necessary to produce the meaningful representation for a speech signal according to the following equation:

\[
\text{argmax}_y p(y \mid x) = \text{argmax}_y p(x \mid y) p(y)
\]

Where: \(\text{argmax}_y p(y \mid x)\) represents the translation model, \(p(x \mid y)\) is called acoustic model and \(p(y)\) represents the language model. This equation describes the probability associated with a postulated sequence of words. The acoustic model is a file that contains statistical representations of sequences of feature vectors which are computed from the speech signal [30].
3.2 Deep learning algorithms

Deep learning algorithms are part of machine learning algorithms that inspired by the act of personal brain [19]. The architecture of deep learning is consisted of input layer, output layer and many hidden layers. Hidden layers are responsible to do complex computations especially for extracting features in order to obtain better representations. There are many different deep learning algorithms such as the Recurrent Neural Network (RNN) and CNN [11] [31]. Deep learning technologies are available in reinforcement learning, unsupervised learning and supervised learning. The reinforcement learning is for methods based on Markov Decision Process (MDP). The unsupervised learning is for methods that are capable to extract meaningful representations to unlabeled data. The supervised learning is for methods which associate inputs to determined outputs.

Deep learning has be synthetic talent to classify data. It plays roles in the ASR after the large growth in computer hardware and machine learning algorithms. Generally, the DNN refers to a feed forward neural network with more than one hidden layer [32]. It hierarchically representing features in its multiple hidden layers of neurons [33]. Inputs are applied to the DNN by multiplying them with a weight vector and summing the result, then, passing them to an activation function. Figure 3 shows a DNN that has been one input layer, three hidden layers and one output layer.

![DNN diagram](image)

**Figure 3.** DNN that has one input layer, one output layer and three hidden layers [22]

Usually, a deep learning has the following properties:

- Extracting robust and significant features from the input data
- Efficiently dealing with merging multiple feature extraction vectors.
- Using dropout technique to prevent over fitting problem.

3.2.1 Convolutional Neural Networks

CNN is a type of the DNN. It also a developed type of the feed-forward artificial neural network. The term ‘convolutional’ is derived from the mathematical operation of convolution. It can be efficiently employed for image recognition. Generally, input data in CNNs are processed in a range of connected layers by three bases: local connectivity, shared weights and pooling the sequence. Figure 4 shows a CNN architecture of
input layer, feature-extraction layers and classification layers. It transfers data from input layer to all connected layers in order to give the output.

Figure 4. CNN architecture of input layer, feature-extraction layers and classification layers [8]

Generally, the input layer accepts three-dimensional input in the form of size (width × height) and a depth to represent the color channel. Then, feature-extraction layers have a repeated patterns of the sequence: convolution layer and pooling layer. The Rectified Linear Unit (ReLU) refers to an activation function, its layer is usually directly employed after the convolution layer. A pooling layer is commonly located between successive convolutional layers. Its aim is to reduce the data representation size in the network. It includes the same feature values that are calculated in different locations, where a single value is extracted by using the maximum operation. This is referred to as max pooling type to spatially resize the information (width and height) this leads to minimize the size of extracted features. Classification layers are completely linked to all of the neurons in the last feature extraction layer as the identification implies. Usually, the output of these layers produce two-dimensional outputs \( [b \times N] \), where \( b \) is the variety of examples in the mini-batch and \( N \) is the wide variety of lessons that are fascinated in scoring [34].

3.2.2 Recurrent Neural Networks

RNN is a type of artificial neural network. It involves a sequential data connection with the hidden neurons. It can be applied for the applications of text, audio and video [35]. It deals with sequential data from the analyzed the sequence at each time depending on the previous time in a directed cycle [36]. LSTM units and Gated Recurrent Units (GRUs) are variations type of RNN. Figure 5. shows a model of the RNN which provides input samples and contains more interdependencies by offering the output to be as an input to the hidden layer [37][38].
Figure 5. A model of the RNN which provides input samples and contains more interdependencies by offering the output to be as an input to the hidden layer [38]

The RNN also known as Auto-associative (feedback network). It consists of a single layer of neurons with each neuron feeding its output signal back to the inputs of all neurons. The RNN is also developed for the DNN. For example, deep network with one input layer, three hidden layers and one output layer. Each hidden layer has its own information group of biases and weights, let’s assume: the information group for the 1st hidden layer is \((w_1, b_1)\), the information group for the 2nd hidden layer is \((w_2, b_2)\) and the information group for the 3rd hidden layer is \((w_3, b_3)\). Thus, it appears that each hidden layer is independent to others.

The RNN has many characteristics, as follow [39]:

- It converts the independent activations to the dependent activations by using in all layers the same biases and weights, where each output is being as input to the next hidden layer. This leads to reduce the complication of increasing memories and parameters of prior outputs.
- Multiple layers can be connection altogether with the same biases and weights.
- Calculating the current state as follows:

\[
\hat{h}_t = f(h_{t-1}, x_t)
\]  

Where: \(\hat{h}_t\) represents the current state, \(h_{t-1}\) represents the previous state and \(x_t\) represents the the input state.

- Applying the tanh activation function as follows:

\[
\hat{h}_t = \tanh(w_{hh}\hat{h}_{t-1} + w_{xh}x_t)
\]

Where: \(w_{hh}\) represents the weight of a recurrent neuron and \(w_{xh}\) represents the weight of an input neuron.

- Calculating the output as follows:
\[ y_t = w_{hy}h_t \]  \hspace{1cm} (4)

Where: \( y_t \) represents the output and \( w_{hy} \) represents the weight of an output neuron.

4. Challenges

Speech recognition is one of the difficult areas in a computer science. Lots of contributed methodologies tried to unravel proper ways and better shares of recognition.

The main challenges for the speech recognition in deep learning are:

- Collecting a big number of signals from participants.
- Producing a deep learning network that has the capability to recognize spoken words.
- Improving the speech recognition accuracy.

5. Conclusion

This paper introduced a review of automatic Arabic speech recognition systems that are based on isolated words technique utilizing deep learning networks. The current survey has focused on isolated words techniques for speech recognition. However, it should be clear that many stages i.e. feature extraction and classification and many steps in the pre-processing stage, it is adopted from other speech recognition techniques such as connected words and continuous speech. The paper presents previous studies that are associated with the isolating words techniques of ASR to recognition words are spoken in Arabic. And, it presents a review of the speech recognition types, classification methods, feature extraction methods, and accuracies were also provided. This survey has also discussed a variety of deep learning methods, which commonly used in ASR systems designing and can contribute to improve recognition accuracy. Also, Deep learning strategies of CNN and RNN are described as types of DNN.

References

[1] Du Guiming, Wang Xia, Wang Guangyan, Zhang Yan, and Li Dan, 2016, “Speech recognition based on convolution neural network , IEEE international conference on signal and image processing (ICSIP), No. 16776620, Beijing, China, DOI: 10.1109/SIPROCESS.2016.7888355.

[2] M. M. El Choubassi, H. E. El Khoury, C. E. Jabra Alagha, J. A. Skaf and M. A. AlAlaoui, 2004, “Arabic speech recognition using recurrent neural networks”, IEEE International Symposium on Signal Processing and Information Technology IEEE Cat. No.03EX795, DarmstadtGermany, pp543–547,ISBN:0-7803-8292-7 ,DOI: 10.1109/ISSPIT.2003.1341178.

[3] Deepankar Dayal, Firoz Alam, Himanshu Varun, Neha Singh, 2020, “Review on Speech Recognition using Deep Learning”, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 8, Issue V, ISSN: 2321-9653, pp 1-7.

[4] Lars Schillingmann, Jessica Ernst, Verena Keite, Britta Wrede, Antje S. Meyer, Eva Belke,2018, “ Align tool : the automatic temporal alignment of spoken utterances in german, dutch, and british english for psycholinguistic purposes”,50:466–489, https://doi.org/10.3758/s13428-017-1002-7.

[5] J. McKechnie, B. Ahmed, R. Gutierrez-Osuna, P. Monroe, P. McCabe , K. J. Ballard, 2018, “Automated speech analysis tools for children’s speech production: A systematic literature review”, International Journal of Speech-Language Pathology, DOI: 10.1080/17549507.2018.1477991 https://doi.org/10.1080/17549507.2018.1477991.

[6] Rubi, Chhavi Rana,2015, “A Review Speech Recognition with Deep Learning Methods”, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.5, pg. 1017-1024 IJCSCMC, http://www.ijcsmc.com.

[7] Venkat N. Gudivada, C.R. Rao, 2018, “Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications”, Vol.38, Pages 317-328 , https://doi.org/10.1016/bshost.2018.05.001.
[8] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon and K. M. Iftekharuddin, 2017, “Survey on Deep Neural Networks in Speech and Vision Systems”.

[9] Mohit Dua, R.K. Aggarwal, Virender Kadyan, Shelza Dua, 2012, “Punjabi Automatic Speech Recognition Using HTK”, *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 4, No. 1, ISSN: 1694-0814, www.IJCSI.org.

[10] Bacha Rehmam, Zahid Halim, Ghulam Abbas, Tufail Muhammad, 2015, “Artificial neural network based speech recognition using dwt analysis applied on isolated words from oriental languages”, *Malaysian Journal of Computer Science*. Vol. 28(3), pp 242–262.

[11] Ali AbdAlmisreb, Ahmad Farid Abidin, Nooritawati Md Tahir, 2015, “Maxout based deep neural networks for Arabic phonemes recognition”, *IEEE 11th International Colloquium on Signal Processing & its Applications (CSPA2015)*, 6-8, Kuala Lumpur, Malaysia.

[12] Shady Y. EL-Mashed, Mohammed I. Sharway, Hala H. Zayed, 2017, “speaker independent arabic speech recognition using support vector machine”, Cairo, Egypt.

[13] Lottfi Boussaid, Mohamed Hassine, 2017, “Arabic isolated word recognition system using hybrid feature extraction techniques and neural network”, *International Journal of Speech Technology, Springer Science +Business Media, LLC, part of Springer Nature*, 21, pp 29–37.

[14] Abdulaziz Saleh Mahfoudh Ba Wazir, Joon Huang Chuah, 2019, “Spoken Arabic digits recognition using deep learning”, *IEEE International Conference on Automatic Control and Intelligent Systems (ICACIS 2019)*, Selangor, Malaysia.

[15] Moner N. M. Arafa, Reda Elbarougy, A. A. Ewees, G. M. Behery, 2018, “A dataset for speech recognition to support arabic phoneme pronunciation”, *International Journal of Image, Graphics & Signal Processing*, Vol. 10, No. 4, pp. 31-38. http://www.mecs-press.org DOI:10.5815/ijigsp.2018.04.04

[16] Elvira Sukma Wahyuni, 2017.”Arabic speech recognition using MFCC feature extraction and ANN classification”, *International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITI))*, 1-2 Nov., IEEE Xplore, INSPEC Accession Number: 17595168, pp. 22–25, Publisher: IEEE, Yogyakarta, Indonesia DOI: 10.1109/ICITI))2017.8285499

[17] Amer El Kourd1, Kaouther El Kourd, 2016, “Arabic isolated word speaker dependent recognition system”, *British Journal of Mathematics & Computer Science*.14, pp 1-15, ISSN: 2231-0851 www.sciencedomain.org

[18] Ahmad Emami , Lidia Mangu, 2007, “Empirical study of neural network language models for Arabic speech recognition”. *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)* Date of Conference: 9-13 Dec. 2007, INSPEC Accession Number: 9821959, Publisher: IEEE Kyoto, Japan, DOI: 10.1109/ASRU.2007.4430100
[22] Roozbeh Sadeghian, 2017, “automatic speech recognition techniques for diagnostic predictions of human health disorders”, University of New York

[23] Bourouba, H., Djemili, R. and et al., 2006, “New hybrid system (supervised classifier/HMM) for isolated Arabic speech recognition”, International Conference on Information & Communication Technologies IEEE Xplore, Date :24-28 April 2006, , INSPEC Accession Number: 9061662, ISBN:0-7803-9521-2 ,Publisher: IEEE, Damascus, Syria DOI: 10.1109/ICICTTA.2006.1684560

[24] Z.Hachkar, A. Farchi , B.Mounir , J. EL Abbadi, 2011, “A comparison of DHMM and DTW for isolated digits recognition system of Arabic language”, International Journal on Computer Science and Engineering (IJCSE) Vol. 3 No. 3, ISSN : 0975-3397.

[25] Moyen Mohammad Mustaquim, 2011, “Automatic speech recognition- an approach for designing inclusive games”. 66:131–146, Springer Science+Business Media, LLC.

[26] M.A. Ammaya, S.K.Katti , 2009, “Speech Recognition by Machine, A Review”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, ISSN 1947-5500.

[27] Mumtaz Begum Mustafa, , Fadhliah Rosdi, Siti Salwah Salim, Muhammad Umair Mughal, 2015, “Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker”, Expert system with application, Vol. 42, Issue 8, Pages 3924-3932.

[28] Michelle Cutajar, Edward Gatt , Ivan Grech, Owen Casha, Joseph Micallef, 2013, “Comparative study of automatic speech recognition techniques”, IET Signal Processing, vol. 7, issue: 1, page(s): 25 - 46 Publisher: IET DOI: 10.1049/iet-spr.2012.0151

[29] Yu, D., Deng, L., 2015, “Automatic speech recognition: A deep learning approach,”, Signals and Communication Technology, Springer, London, pp. 13–21. https://doi.org/10.1007/978-1-4471-5779-3

[30] Karpagavalli S, Chandra E, 2016, “A Review on Automatic Speech Recognition Architecture and Approaches”, International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.9, No.4, pp.393-404, http://dx.doi.org/10.14257/ijisp.2016.9.4.34

[31] I. Gavat, O. Dumitrd, C. Iancd, G. Costache,2005. “Learning strategies in speech recognition”, International Symposium ELMAR-2005, 08-10 , pp.237-240

[32] Ossama Abdel-Hamid, Abdelrahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, Dong Yu,2014, “Convolutional Neural Networks for Speech Recognition”, IEEE/ACM transactions on audio, speech, and language processing, vol. 22, No. 10, PP. 1533-1545.

[33] Muhammad Imran Razzaq, Saeeda Naz , Ahmad Zaib , 2017, “Deep Learning for Medical Image Processing: Overview, Challenges and Future”,arXiv.org, vol. 1.No.4, PP. 1-30.

[34] John Murphy,2016,“An Overview of Convolutional Neural Network Architectures for Deep Learning”, Microway, Inc. PP 1-22 https://www.semanticscholar.org

[35] Yong Yu, Xiaosheng Si, Changhua Hu and Jianxun Zhang, 2019, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures”, Vol. 31, Issue 7, p.1235-1270, China.

[36] Mike Schuster and Kuldip K. Paliwal, Member, 1997, “Bidirectional Recurrent Neural Networks”. IEEE transactions on signal processing, vol.45, NO. 11, PP.2673-2681. https://doi.org/10.1109/78.650093

[37] Filippo Maria Bianchi, Enrico Maiorinob , Michael C. Kampffmeyera , Antonello Rizzib , Robert Jensena , 2017, “ Overview and comparative analysis of recurrent neural networks for short term load forecasting”, arXiv:1705.04378, vol 1, pp.1-41
[38] Skevin zhou, Daniel Rueckert, Gabor Fichtinger, 2020, “Handbook of Medical Image Computing and Computer Assisted Intervention”, ISBN 978-0-12-816176-0, No. of pages 1072. DOI: https://doi.org/10.1016/C2017-0-04608-6

[39] https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network, 2018.