Arabic Speech Recognition with Deep Learning: A Review

Wajdan Algihab, Noura Alawwad, Anfal Aldawish,
and Sarah AlHumoud(✉)

College of Computer and Information Science,
Al-Imam Mohammad Ibn Saud Islamic University (IMSIU),
Riyadh, Saudi Arabia
{Waghaihb, Naawad, Anaduweish}@sm.imamu.edu.sa,
Sohumoud@imamu.edu.sa

Abstract. Automatic speech recognition is the area of research concerning the enablement of machines to accept vocal input from humans and interpreting it with the highest probability of correctness. There are several techniques to implement speech recognition models. One of the emerging techniques is using neural networks with deep learning for speech recognition. Arabic is one of the most spoken languages and least highlighted in terms of speech recognition. This paper serves as a brief review on the available studies on Arabic speech recognition. In addition, it sheds some light on the services and toolkits available for Arabic speech recognition systems' development.

Keywords: Automatic speech recognition (ASR) · Arabic Automatic Speech Recognition (AASR) · Deep learning · Artificial neural networks (ANN) · Deep neural network (DNN) · Recurrent neural network (RNN)

1 Introduction

Arabic is one of the most widely spoken languages around the world with an estimated number of over 313 million speakers with 270 million as a second language speaker of Arabic ranked as the forth after Mandarin, Spanish and English [1]. Moreover, it is the language of the Islamic holy book "Quran" with 1.8 billion Muslims around the world in 2015 and projected to increase to 3 billion in 2060 [2]. There have been relatively little speech recognition researches on Arabic compared to other languages [3].

The Arabic language has three types: classical, modern, and dialectal. Classical Arabic is the language Quran. Modern Standard Arabic (MSA) is based on classical Arabic but with dropping some aspects like diacritics. It is mainly used in modern books, education, and news. Dialectal Arabic has multiple regional forms and is used for daily spoken communication in non-formal settings. With the advent of social media, dialectal Arabic is also written. Those forms of the language result in lexical, morphological and grammatical differences resulting in the hardness of developing one Arabic NLP application to process data from different varieties.

G. Meiselwitz (Ed.): HCII 2019, LNCS 11578, pp. 15–31, 2019.
https://doi.org/10.1007/978-3-030-21902-4_2

Al-Anzi and AbuZeina in [4] addressed challenges in speech recognition such as different acoustic conditions, different accents, and the variety of expressing words. Meanwhile, they introduce Arabic speech recognition challenges such as the Arabic script discretization. The authors claimed that the Arabic language is in the early stages compared to English.

Deep learning is a branch of machine learning that inspired by the act of the human brain in processing data based on learning data by using multiple processing layers that has a complex structure or otherwise, composed of multiple non-linear transformations that is capable of unsupervised learning from unstructured or unlabeled data. Deep learning research has been successful in the last few years and it is used in various fields such as in computer vision, speech recognition, natural language processing, handwriting recognition. Deep learning is one of the promising areas in machine learning for the future tasks involved in machine learning especially in the area of the neural network [5].

The published research on models, techniques and applications on English speech recognition based on deep learning is comparably higher than that of Arabic. For Arabic, the literature is limited and scattered. This review serves as pivot point aiming at shedding the light on the available literature on Arabic speech recognition using deep learning.

In later sections, we will introduce the following: Sect. '2' Review methodology. Then, in Sect. '3', the background and related work presented and discussed. After that, in Sect. '4' Application of Arabic Automatic Speech Recognition (AASR) is presented. Further, Sect. '5', discusses the techniques used for deep learning with AASR. Finally, Sect. '7' is the conclusion.

2 Review Methodology

In developing this review, we are inspired with the methodology described by [6, 7]. Additionally, the focus of this review is depicted in the following research questions:

RQ1: what are the techniques for ASR
RQ2: what are the studies on Arabic ASR using deep learning
RQ3: what are the available services and frameworks for developing Arabic ASR.

The databases we did search in are ACM, IEEE, Springer, Sage journals and Science Direct. The keywords used are: "Deep learning", "Arabic automatic speech recognition", "Speech recognition", "Arabic speech recognition", "Neural Networks", "Deep neural networks", "Recurrent neural networks", "Voice recognition". Moreover, timeframe of the review focused on published papers from year 2000 until now. After eliminating papers that does not answer the research questions, we are let with 17 papers. Figure 1 shows the distribution of the papers across the different databases.

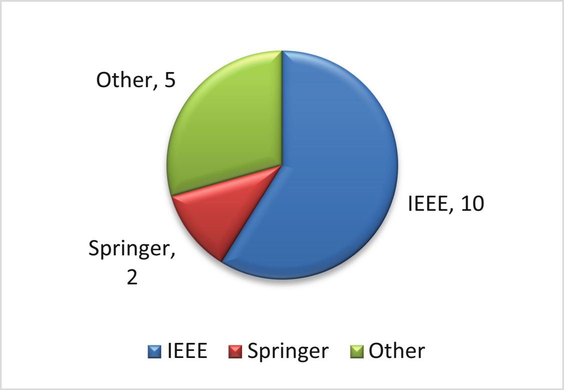

Fig. 1. Research paper results based on the publisher

3 Background and Related Work

The aim of speech recognition is to enable machines to accept sounds and act based on it. Automatic speech recognition is the ability for a machine to recognize "receive and interpret" the speech and convert it into readable form or text and performing an action based on the instructions defined by the human [8].

Speech analysis is the process of analyzing voice and different speech patterns. Speech analysis techniques are divided into segmentation analysis, sub-segment analysis, and surpa segmental analysis [9]. Meanwhile, speech feature extractions are done using Linear predictive coding (LPC) and Mel frequency cepstral coefficients (MFCC).

In addition, the approaches of speech recognition are the Template-Based approach, Knowledge-based approaches, Neural network based approaches, Dynamic time warping (DTW) based approaches and Hidden Markov model (HMM) based speech recognition [10].

Template based approach is the process of matching unknown spoken word and then comparing them with speech patterns templates (pre-recorded words) to find the best match. While the Knowledge-based approach deals with acoustic aspects of spoken words. It analyses sound wave properties based on observed features and then incorporate them with the knowledge of the relationship between the acoustic feature and phonetic symbol. The spoken words are decoded to obtain a sequence of phonemes and other linguistic units. Moreover, Dynamic Type Warping approach is based on an algorithm for measuring the similarity between two sequences, to find the optimal match between a given sequence that may be varied in time or speed. Finally, Hidden Markov Model which is widely used in the stochastic approach, where the Hidden Markov model is characterized using a set of distribution outputs and finite state Markov model. Word and phone boundaries are automatically determined in the training process.

In this paper, we will focus on neural network-based approaches that are represented as an important class of discriminative techniques and as it inspired the biological neural networks. In the following paragraph, we will discuss the literature based in terms of AASR classification, stages, and techniques.

Authors Yu and Deng [11] dissected the AASR system into four stages. First, preprocessing stage. Second, Feature extraction stage. Third, decoding using Acoustic model, Language model, and pronunciation dictionary. Fourth, Post-processing results were the best hypothesis is produced. The stages work as following: First, speech waveform used as input in the preprocessing stage. Then, the output is processed speech waveform and this is used as input in feature extraction stages where we have the feature vector as output and use it as input in the next stage, the decoding stage. In this stage, the Acoustic model, is employed along with a pronunciation dictionary. After that, the n-best hypothesis - the output of the pronunciation dictionary stage is used in post-processing as input. As a result, the best hypothesis is produced from this work operation.

Turab, Khatatneh, and Odeh in [12] discussed the phoneme recognition as it is related to speech recognition. The techniques used are as follows: Gaussian Low Pass filtering algorithm along with the neural network in the pre-processing stage to have an improvement on the results. Furthermore, the stages of phoneme recognition are: catching a signal, sampling, quantization and setting energy. After that, a neural network is used to enhance the results. Moreover, this paper shows the enhanced impact in results after applying the Gaussian Low Pass filter in voice signals hence, the noise was reduced. After that, in the training phase, the neural network has been used to train the system in order to recognize the speech signals.

Ahmed and Ghabayen in [3] proposed three approached to enhance the AASR. The paper started with the first approach which is the punctuation modeling, in this approach Ahmed and Ghabyaen proposed a decision tree with variant pronunciation generation. After that, a hybrid approach proposed and used to adapt the native acoustic model with another native acoustic model. Finally, the language model is enhanced and improved using a processed text. The model efficiency was measured by Word Error Rate (WER) which is a metric to measure the performance of speech recognition and calculates misrecognitions at the word level. Consequently, the pronunciation model reduced WER by 1%, The acoustic modeling reduced the WER by 1.2% and the language model reduced WER by 1.9%.

Emami and Mangu [13], examine the neural network usage for Arabic speech recognition using a distributed word representation. Furthermore, the model of the neural network allows robust generalization and enhance the ability to fight the data sparseness problem. Also, the investigation process includes different configuration neural probabilistic model, n-gram order parameter experiment, output vocabulary, the method of normalization, model size and parameters. The experiment has been done on the Arabic news broadcast, and conversation broadcast. As a result, some improvement has been achieved using the optimized neural network model over the 4-gram baseline model resulting in up to 0.8% absolute reductions and 3.8% relative WER. However, different parameters do not have a significant impact on model performance. The paper was based on analyzing first. Then, feature extraction. After that, modeling and finally, testing.

Based on Desai, Dhameliya, and Desai [14], the proposed speech recognition system contains four stages. First, feature extraction. Second, database. Third, network training. Fourth, testing or decoding.

In [4] Al-Anzi and AbuZeina used WER metric to evaluate the performance of isolated-word recognition and continuous speech recognition. The evaluation of continuous speech was presented for seven papers based on the improvement of WER. The results were as follow: Kirchhof, Bilmes and Stolcke in [15] performed performance evaluation using a language model for morphology and the improvement of WER for two different test sets is 1.8% and 1.5% respectively. In [16], Emami, Ahmad and Lidia use two different configurations of neural probabilistic models and the improvement of WER is 0.8% and 3.8% respectively. The authors in [17] used broadcast news corpus and improve the WER by 13.66%. Hyassat and Abu Zitar in [18] used the holy Quran corpus and WER improved by 46.182%. In [19] Elmahdy and Mohamed used Egyptian Colloquial Arabic and reached 99.34% of recognition accuracy. Selouani, Sid Ahmed and Malika Boudraa in [20] used MSA continues speech corpus and reach an accuracy rate of 91.65%. In [21] the authors Jurafsky and Martin used MSA continues speech corpus and the improvement of WER using diacritical marks and without resulting 11.27% and 10.07% respectively.

4 Arabic Automatic Speech Recognition with Deep Learning

Some popular techniques used in ASR and AASR are artificial neural networks, dynamic time warping and Hidden Markov modeling. In this review, we are going to focus on artificial neural network techniques. Moreover, speech recognition systems can be classified in different classes based on what type of utterances they can recognize. Those are of four types: isolated words, connected words, connected speech, and spontaneous speech. Those are discussed in more detail in the following subsections.

4.1 Isolated Words

Isolated word recognizers require to have Listen/Not-Listen states between each utterance, it processes the words during the “not listen” state [22].

The authors in [23] show a comparison between general regression neural network (GRNN) algorithm and the traditional multi-layer perceptron in the recognition of a large set of Arabic words. The results show that the GRNN gives better results than those based on the feedforward backpropagation in the recognition rate. The proposed architecture consists of two parts: pre-processing phase which consists of segmental normalization and feature extraction and a classification phase which uses neural networks based on nonparametric density estimation. Using MLP the error rate was respectively 8%, 8% and 12% for the digit “2”, “3” and “8” pronounced by male speakers. It is less significant when they used the non-parametric regression (respectively 2%, 6% and 6%). The GRNN gives better recognition rate and it was the faster algorithm when having a large dimension of input vectors.

The authors in [24] designed a speech recognition system that investigates Arabic digits based on a recurrent neural network. They implement it as a multi-speaker mode and a speaker-independent mode. The system in the case of a multi-speaker mode achieved 99.5% correct digit recognition, and in the case of the speaker-independent mode, the system achieved 94.5%.

A novel approach was presented in [1] describing the implementation of Arabic isolated speech recognition system by modular recurrent Elman neural networks (MRENN). The authors claimed that the results have shown that this new neural network approach can compete with the traditional HMM-based speech recognition approaches. They show a table with the obtained results of 6 speakers some of them with a noise background and other with clean background. The recognition rate for the different speakers was around 85% and 100%.

4.2 Connected Words

Connected word systems are similar to isolated words but allow separate utterance to be run together with “minimum pause between them” [22].

In [25] the author introduces a “simple and effective time alignment” for spoken Arabic digit recognition systems. The algorithms are simple and low in computational power, and in the understanding of the algorithm also. The speech recognition system designed based on an artificial neural network tested with automatic Arabic digit recognition and implemented in a multi-speaker mode. The authors used the time alignment algorithm to compensate for the differences in the utterance and the misalignment of the phoneme the time alignment algorithm. The algorithm was tested on a MLP neural network based recognizer; the overall system performance for Arabic digit recognition was 99.49%.

The authors aimed in [26] to observe the differences in the 29 letters of the Arabic alphabet. They proposed a system based on a fully-connected recurrent neural network with a backpropagation through time learning algorithm. The purpose was to improve the knowledge of the Arabic alphabet. They compared the LPCCC and MFCC performance with different hidden node (40, 50 and 60) for different 4 speakers, overall the LPCCC outperform the MFCC performance by 0.7%.

An approach in learning to deal with a non-uniform sequence length of the speech utterances have been proposed in [27] based on Long Short-Term Memory (LSTM). The system consists of two phases: feature extraction with the Mel Frequency Cepstral Coefficients algorithm (MFCC), and then process the features with a deep neural network. They used a recurrent LSTM or GRU architecture to encode sequences of MFCC features like a fixed size vector to feed a multilayer perceptron network to perform the classification.

4.3 Continuous Speech

Continuous speech recognizers allow the user to speak almost naturally. Due to the utterance boundaries, it uses a special method, which is why it considered as one of the most difficult systems to create [22].

The authors in [28] present three different system structures. They manually constructed an Arabic phoneme database. The Mel Frequency Cepstral Coefficients algorithm (MFCC) was used to extract the features from the input signal. The normalized dataset was used to train and test the three different systems. The performances of these systems were 47.52%, 44.58% and 46.63% frame recognition for single MLP identification system, category-based phonemes recognition system and individual Phoneme classifier system respectively.

Also, an argument in the improvement of the performance of speech recognition in mobile communication system has been shown in [29], the authors used in the feature extraction phase the Multitaper Frequency Cepstral Coefficients features and the Gabor features, and in the processing phase they have investigated three different systems: Continues Hidden Markov Models (CHMM), Deep Neural Network (DNN) and HMMDNN hybrid. They focused on HMMDNN and claimed that it can get consistently almost 8% of clean speech, 13% of AMR-NB coder and 8.5% of DSR coders.

A novel approach [30] where the authors combines the benefits of the morpheme-based LMs and feature-rich modeling with the DNN-LMs for the Egyptian Arabic. A result have been shown when a single hidden layer, 2 hidden layers, 3 hidden layers and 4 hidden layers. The most improvement was obtained in the single hidden layer. Incorporating the conventional n-gram LM, the DNN-LM and the feature-rich DNN-LM achieve the best performance.

An AASR system was developed in [31] with a 1,200-h speech corpus. The authors modeled a different DNN topologies including: Feed-forward, Convolutional, Time-Delay, Recurrent Long Short-Term Memory (LSTM), Highway LSTM (H-LSTM) and Grid LSTM (GLSTM). A table with all the models and its result has been shown. The best performance was from a combination of the top two hypotheses from the sequence trained GLSTM models with 18.3% WER.

A comparison for some of the state-of-the-art speech recognition techniques was shown in [32]. The authors applied those techniques only to a limited Arabic broadcast news dataset. The different approaches were all trained with a 50-h of transcription audio from a news channel “Al-jazirah”. The best performance obtained was the hybrid DNN/HMM approach with the MPE (Minimum Phone Error) criterion used in training the DNN sequentially, and achieved 25.78% WER.

An Arabic broadcast news speech recognition system was built using the KALDI toolkit in [33]. The system was trained with 200 h broadcast news database. They build a broadcast news system with 15.81% WER on Broadcast Report (BR) and 32.21% WER on Broadcast Conversation (BC) with a combined WER of 26.95%.

A LIUM ASR system win the second position in the 2016 Multi-Genre Broadcast (MGB-2) Challenge in the Arabic language [34]. Their main idea was to combine the GMM derived features for training a DNN with the use of time-delay neural networks for acoustic models for automatically phonetic the Arabic words. The key features was the training data selection approach, where a five neural network AMs of different types with a various acoustic features and also a different techniques for speaker adaptation and two types of phonetization. The final system was a combination of a five systems where the result obtained succeeded the best single LIUM ASR system with a 9% of WER reduction and also succeeded the baseline MGB system that was provided by the organizers with a 43% WER reduction.

Also in the same 2016 Multi-Genre Broadcast (MGB-2) Challenge in the Arabic language, The lowest WER was achieved among the nine participating teams by [35] with 14.2%. They built a system that is a combination of three LF-MMI trained models; TDNN, LSTM and BLSTM. Before combinations, The models were rescored using a four-gram and RNNME LM. The system was trained using 1,200 h audio with lightly supervised transcription.

4.4 Spontaneous Speech

It is a speech that is natural sounding and not rehearsed. An ASRS should be able to handle a variety of natural speech features like words being run together [22].

An approach that integrates into adverse acoustic conditions multiple components to improved speaker identification in spontaneous Arabic speech has been presented in [36]. They used two acoustic speakers models the maximum likelihood linear regression support vector machine (MLLR-SVM) and the Cepstral Gaussian Mixture Models (GMM) models and a neural network combiner. A result of the Arabic portion of the NIST (National Institute of Standards and Technology) mixer data is shown. The authors apply noises like babble and city traffic, in both they found an equal error rate reductions over the no-compensation condition. Which gave a complementary gain for both acoustic models. The authors show different tables with the result, Surprisingly, they found the combiner that trained in clean conditions gives a similar performance to the one that trained in matched conditions.

The authors in [37] presented a comparative study between two identification engines to identify speakers automatically from their voices when speaking spontaneously in Arabic. The continuous hidden Markov models (CHMMs) was used in the first engine, and in the second engine, they used artificial neural networks (ANNs). In the feature extraction phase of the signal, the Mel frequency cepstral coefficients (MFCCs) were used. They used the general Gaussian density distribution HMM, as for the ANN-based engine they used the Elman network. The identification rate was found to be 100% for both engines during text dependent experiments. However, for text-independent experiments, the performance for the CHMM-based engine outperformed that of the ANN-based engine. The identification rates for the CHMM- and the ANN-based engines were found to be 80% and 50% respectively.

5 AASR Deep Learning Techniques

Deep learning has different techniques which can be applied on AASR. In this paper we focus on the artificial neural network technique. We cover the main types of ANN. Table 1 cover the summary of the main types of ANN.

5.1 Neural Networks

It is more convenient to use NN for speech recognition than serial programming which execute only one operation at a time. In this section, we review the three available papers for AASR using NNs.

Emami and Mangu [20] showing the use of distributed representations and neural network for AASR. They used the AASR decoder to generate a set of lattices with an average link density. The training samples were 7 M words collected from Arabic broadcast news and broadcast conversations. This paper used a baseline 4-gram model which helped in improving the NN by reducing up to 0.8% absolute and 3.8% relative in WER. Experimented the parameters of NN language models (LMs) with different configurations of NN LMs which concluded that the performance of NN LMs was not affected by parameters. The size of the NN has no effect on the performance.

Ettaouil et al. [19] used a hybrid model ANN/HMM for AASR to determine the optimal codebook generated by Kohonen network Self Organizing Maps (SOM). The Optimal codebook used to the classification of the Arabic digits this leads to optimization of Kohonen approach. The numerical results are satisfactory showing that the classification was affected by the size of the dictionary. The codebook vectors are with size 34, 36, and 48 they had a recognition rate 84%, 85%, and 86% respectively.

Wahyuni in [20] used Mel-Frequency Cepstral Coefficients (MFCC) based on feature extraction and ANN to distinguish between the pronounce of three different letters (sa, sya, and tsa) by Indonesian speakers which have the same sound for those different letters which is (sa) according to their usual using Bahasa. The result showing that the usage of

MFCC with ANN gave the better recognize the three letters average accuracy of 92.42%.

5.2 Recurrent Neural Networks

Recurrent Neural Networks (RNN) is one of the best models applied for sequential data [21]. It allows for both feedforward and feedback paths. For AASR only two papers used the Elman RNN which is the type of RNN. Elman has an advantage when compared with fully RNN [4]. It can use backpropagation to train the network.

Alotaibi in [5] show the usage of recurrent ANN namely recurrent Elman network for AASR to the recognition of ten Arabic digits (from Zero to nine). Asking 17 male Arabic native speakers to repeat the digits ten times. This created the database with 1700 token, that is, 170 samples for every digit. Operating the system in two different modes. The first mode is a multi-speaker mode which used the same speakers sound for both training and testing phases. The training tokens set has 340 tokens. That is 17 speakers, 2 repetitions and 10 digits. Where the test set used 1700 tokens. The system performance was 99.47% in this mode. The second mode is speaker-independent mode which used the different speaker sounds for training and testing phases. The training tokens set has 400 tokens, that is 4 speakers, 10 repetitions, and 10 digits. Where the testing set using 1,300 tokens that is 13 speakers, 10 digits, and 10 repetitions. The system performance was 96.46% in this mode. The system in both modes cannot recognize the digit 9 that according to the dissimilarity of this digit and the other digits. Digit 1, 4 and 8 have high error rates, particularly in the second mode.

Choubassi et al. [4] used small RNN and their recognizer has recognized a limited set of isolated words there were: “manzel” (house), “hirra” (cur), “chajara” (tree), “tariq” (road), “ghinaa” (singing), “zeina” (zeina). All those words have individual RNN to detected only the specific word. Training has two phases first is consistent of consistent training then discriminative training. Consistent training used the different utterances of the dedicated word. In discriminative training use utterances of other words not only the dedicated word. This paper used 4 female speakers in a clean environment without any noise for training. For testing they used one women speaker and one-male speaker both in a clean environment. They used back-propagation with momentum and variable learning rate as a training algorithm. This paper used MATLAB to simulate the result. They took a slot of output curves to determine the classification of an utterance over other dedicated words by comparing its result slope s with minimum slope sm. The result of the paper indicated that the usage of RNN gives the same recognition rate matching as the HMM-based approach as mention in Sect. '4'.

5.3 Deep Neural Networks

Deep Neural Networks (DNNs) have reached to suitable performance. DNNs have three advantage when used over other NN [22]. First, DNNs can extract robust and significant features of the input data via several non-linear hidden layers. Second, DNNs can merge multiple extracted feature vectors efficiently. Third, DNNs can prevent overfitting problem by using dropout technique. We find only one paper use this method for AASR.

AbdAlmisreb et al. [22] presented the DNN with three hidden layers, 500 Maxout units with 2 neurons for the unit and used Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction. This approach was trained and tested over a corpus which consisted of 20 Malay speakers of consonant Arabic phonemes recorded. The training set consisted of 5 waveforms and the tested set contained 15 waveforms. The result show that the Maxout based deep structure gave better performance with lowest error rate than other deep networks such as Restricted Boltzmann Machine (RBM), Deep Belief Network (DBN), Convolutional Neural Network (CNN), the conventional feedforward neural network (NN) and Convolutional Auto-Encoder (CAE) which had error rate between 2800 and 3000 (numbers).

Table 1. Summary of the main types of ANN

Approach	Aims	Preprocessing	Type of NNs	Used datasets	Result
Emami and Mangu [20]	Showing the use of distributed representations and neural network for AASR	Used the AASR decoder to generate a set of lattices with an average link density Use different order of neural probabilistic model by taking those parameters:	Neural Networks (NN)	The training samples were 7M words collected from Arabic broadcast news and broadcast conversations	This paper used a baseline 4-gram model which helped in improving the NN by reducing up to 0.8% absolute and 3.8% relative in WER. Experimented the parameters of NN language models (LMs) with different

(*continued*)

Table 1. (*continued*)

Approach	Aims	Preprocessing	Type of NNs	Used datasets	Result
		- N-gram order - Output vocabulary - Normalization method - Model size			configurations of NN LMs which concluded that the performance of NN LMs was not affected by parameters. The size of the NN has no effect on the performance
Ettaouil et al. [19]	Determine the optimal codebook generated by Kohonen network Self Organizing Maps (SOM)	Used a hybrid model ANN/HMM for AASR. By generate three dictionaries with three neural networks the first with 34 neurons, the second with 36 and the third with 48 neurons	Neural Networks (NN)	Consists of 8800 tokens for Arabic digits Dataset divide to: 1–75% of the samples for training set 2–25% of the samples for test set	The classification was affected by the size of the dictionary. The codebook vectors are with size 34, 36, and 48 they had a recognition rate 84%, 85%, and 86% respectively
Wahyuni in [20]	Distinguish between the pronounce of three different letters (sa, sya, and tsa) by Indonesian speakers which have the same sound for those different letters which is (sa) according to their usual using Bahasa	They extract feature by using Mel-Frequency Cepstral Coefficients (MFCC) then use ANN for classification	Neural Networks (NN)	738 data of three letters as: 248 data of sa (س), 254 data of sya (ش), 236 data of tsa (ث) Collect them by recording human voice with pronounces letters sa (س), sya (ش), tsa (ث), by depending on the *makhraj* pronunciation of hijaiyah	The usage of MFCC with ANN gave the better recognize the three letters average accuracy of 92.42%
Alotaibi in [5]	Show the usage of recurrent ANN namely recurrent Elman network for AASR to the recognition of ten Arabic	- Extract feature by using (MFCC) - Used VECTOR QUANTIZATION technique to compression data	Recurrent Neural Networks (RNN)	Asking 17 male Arabic native speakers to repeat the digits ten times. this created the database with	The system performance was 99.47% The system performance was 96.46% The system in both modes cannot

(*continued*)

Table 1. (*continued*)

Approach	Aims	Preprocessing	Type of NNs	Used datasets	Result
	digits (from Zero to nine)			1700 token, that is, 170 samples for every digit Operating the system in two different modes 1-multi-speaker mode which used the same speakers sound for both training and testing phases The training set has 340 tokens. And the test set used 1700 tokens 2-Speaker-independent mode which used the different speaker sounds for training and testing phases The training set has 400 tokens. And the testing set using 1,300 tokens	recognize the digit 9 that according to the dissimilarity of this digit and the other digits. Digit 1, 4 and 8 have high error rates, particularly in the second mode
Choubassi et al. [4]	To recognize a limited set of isolated words there were: “manzel” (house), “hirra” (cur), “chajara” (tree), “tariq” (road), “ghinaa” (singing), “zeina” (zeina)	- Used small RNN and their recognizer - All those words have individual RNN to detected only the specific word - used back-propagation with momentum and variable learning rate as a training algorithm	Recurrent Neural Networks (RNN)	Two phases for Training: 1st phase consistent training which used the different utterances of the dedicated word 2nd phase discriminative training which use utterances of other words not only the dedicated word. This paper used	This paper used MATLAB to simulate the result. They took a slot of output curves to determine the classification of an utterance over other dedicated words by comparing its result slope s with minimum slope sm. The result of the paper indicated that the usage of RNN gives the same recognition rate

(*continued*)

Table 1. (*continued*)

Approach	Aims	Preprocessing	Type of NNs	Used datasets	Result
				4 female speakers in a clean environment without any noise for training. For testing they used one women speaker and one-male speaker both in a clean environment	matching as the HMM-based approach as mention in Sect. 4
AbdAlmisreb et al. [22]	Test performance of DNN based on Maxout	Use: 1-Mel-Frequency Cepstral Coefficients for feature extraction 2-Maxout Deep Neural Network by using Maxout algorithm with dropout function to improve the efficiency	Deep Neural Networks (DNN)	The training set consisted of 5 waveforms and the tested set contained 15 waveforms	The result show that the Maxout based deep structure gave better performance with lowest error rate than other deep networks such as Restricted Boltzmann Machine (RBM), Deep Belief Network (DBN), Convolutional Neural Network (CNN), the conventional feedforward neural network (NN) and Convolutional Auto-Encoder (CAE) which had error rate between 0. 2800 and 0.3000

6 AASR with Deep Learning Services

Speech recognition using deep-learning is a huge task that its success depends on the availability of a large repository of a training dataset. The availability of open-source deep-learning enabled frameworks and Application Programming Interfaces (API) would boost the development and research of AASR. There are multiple services and frameworks that provide developers with powerful deep-learning abilities for speech recognition.

6.1 API Services

One of the marked applications is Cloud Speech-to-Text service from Google [35] which uses a deep-learning neural network algorithm to convert Arabic speech or audio file to text. Cloud Speech-to-Text service allows for its translator system to directly accept the spoken word to be converted to text then translated. The service offers an API for developers with multiple recognition features.

Another service is Microsoft Speech API [36] from Microsoft. This service help developers to create speech recognition systems using deep neural networks.

IBM cloud provide Watson service API for speech to text recognition [37] support modern standard Arabic language until now there is not any work use this API with Arabic.

6.2 Toolkits

The Kaldi [38]. It is a toolkit for speech recognition using deep neural network and support Arabic language as Ali et al. in [37] showing the usage of Kaldi to build Arabic broadcast news speech recognition system. They use all Kaldi conventional models. The result showing that the building of broadcast news system on broadcast report take 15.81% WER and 32.21% WER on broadcast conversation.

Manohar et al. in [40] use Kaldi toolkit for Arabic Multi-Genre Broadcast (MGB-3) challenge which deal with dialectal Arabic of Egyptian. For their study, they take 80 different video from YouTube for seven genres. Their comparative study for the efficiency of using Kaldi by take multi-reference word error rate (MR-WER) to measure the efficiency. The system first, built with Minimum Bayes Risk (MBR) system combination of sMBR and nonsMBR system and produce the MR-WER of 32.78% on the MGB-3 test set. They conclude that the Kaldi improve the efficiency of MR-WER.

Additionally, is the Microsoft Cognitive Toolkit (Microsoft's CNTK) [41] which is an open-source toolkit that trains the deep learning algorithm. This toolkit enables using more than one model like DNN, CNN, and RNN. There is no paper that developed AASR with this tool kit.

6.3 Frameworks

one of the main frameworks offering deep-learning capabilities for developers is Tensorflow [42] which is a library that provides accuracy when used with other models to produce speech recognition. Sim et al. [17] used Tensorflow to improve the efficiency of Forward-backward algorithm with English speech recognition.

7 Conclusion

This paper presented a review on Arabic speech recognition using deep-learning Neural Networks. The literature covered seventeen papers and presented according to the recognized entity and according to the learning technique. Recognized entities are of four types: isolated word, connected word, continues word and Spontaneous speech.

Furthermore, Deep learning techniques have three main types, Neural networks, recurrent neural networks, and deep learning networks. This paper presented the state of the art frameworks and services that aid in the ASR system development.

References

1. El Choubassi, M.M., El Khoury, H.E., Alagha, C.E.J., Skaf, J.A., Al-Alaoui, M.A.: Arabic speech recognition using recurrent neural networks. In: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 03EX795), Darmstadt, Germany, pp. 543–547 (2004)
2. Lipka, M., Hackett, C.: Why Muslims are the world's fastest-growing religious group. Pew Research Center (2017). http://www.pewresearch.org/fact-tank/2017/04/06/why-muslims-are-the-worlds-fastest-growing-religious-group/. Accessed 14 Nov 2018
3. Ahmed, B.H.A., Ghabayen, A.S.: Arabic automatic speech recognition enhancement. In: 2017 Palestinian International Conference on Information and Communication Technology (PICICT), Gaza, Palestine, pp. 98–102 (2017)
4. Al-Anzi, F., AbuZeina, D.: Literature survey of Arabic speech recognition. In: International Conference on Computing Sciences and Engineering (ICCSE) (2018)
5. Rana, C.: A review: speech recognition with deep learning methods, p. 8 (2015)
6. Kitchenham, B.: Procedures for performing systematic reviews. Joint Technical report, Keele University Technical report (TR/SE-0401) and NICTA Technical report (0400011T.1), July 2004 (2004)
7. Heckman, S., Williams, L.: A systematic literature review of actionable alert identification techniques for automated static code analysis
8. Nasereddin, H.H.O., Omari, A.A.R.: Classification techniques for automatic speech recognition (ASR) algorithms used with real time speech translation. In: 2017 Computing Conference, London, pp. 200–207 (2017)
9. Shanbhogue, M., Kulkarni, S., Suprith, R.: A study on speech recognition, vol. 4, p. 6 (2016)
10. Pdfs.semanticscholar.org (2012). https://pdfs.semanticscholar.org/04c8/b7668bc09eebcb56d54ba221a26d8fd174d7.pdf. Accessed 14 Nov 2018
11. Yu, D., Deng, L.: Automatic Speech Recognition: A Deep Learning Approach, pp. 13–21. Springer, London (2015). https://doi.org/10.1007/978-1-4471-5779-3
12. Turab, N., Khatatneh, K., Odeh, A.: A novel Arabic Speech Recognition method using neural networks and Gaussian Filtering. (IJEECS) Int. J. Electr. Electron. Comput. Syst. **19** (01) (2014)
13. Emami, A., Mangu, L.: Empirical study of neural network language models for Arabic speech recognition. In: 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), The Westin Miyako Kyoto, pp. 147–152 (2007)
14. Desai, N., Dhameliya, K., Desai, V.: Feature extraction and classification techniques for speech recognition: a review, **3**(12), 5 (2013)
15. Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K., Stolcke, A.: Morphology-based language modeling for conversational Arabic speech recognition. Comput. Speech Lang. **20**(4), 589–608 (2006)
16. Emami, A., Mangu, L.: Empirical study of neural network language models for Arabic speech recognition. In: IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU. IEEE (2007)
17. Alghamdi, M., Elshafei, M., Al-Muhtaseb, H.: Arabic broadcast news transcription system. Int. J. Speech Technol. **10**(4), 183–195 (2007)

18. Hyassat, H., Abu Zitar, R.: Arabic speech recognition using SPHINX engine. Int. J. Speech Technol. **9**(3–4), 133–150 (2006)
19. Elmahdy, M., et al.: Modern standard Arabic based multilingual approach for dialectal Arabic speech recognition. In: Eighth International Symposium on Natural Language Processing, SNLP 2009. IEEE (2009)
20. Selouani, S.A., Boudraa, M.: Algerian Arabic speech database (ALGASD): corpus design and automatic speech recognition application. Arab. J. Sci. Eng. **35**(2C), 15 (2010)
21. Jurafsky, D., Martin, J.: Speech and Language Processing. Prentice Hall, Upper Saddle River (2000)
22. AbdAlmisreb, A., Abidin, A.F., Tahir, N.: Maxout based deep neural networks for Arabic phonemes recognition, p. 6 (2015)
23. Amrouche, A., Rouvaen, J.M.: Arabic isolated word recognition using general regression neural network. In: 2003 46th Midwest Symposium on Circuits and Systems, Cairo, Egypt, vol. 2, pp. 689–692 (2003)
24. Alotaibi, Y.A.: Spoken Arabic digits recognizer using recurrent neural networks. In: Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, Rome, Italy, pp. 195–199 (2004)
25. Alotaibi, Y.: A simple time alignment algorithm for spoken Arabic digit recognition. J. King Abdulaziz Univ.-Eng. Sci. **20**(1), 29–43 (2009)
26. Ahmad, A.M., Ismail, S., Samaon, D.F.: Recurrent neural network with backpropagation through time for speech recognition. In: IEEE International Symposium on Communications and Information Technology, ISCIT 2004, Sapporo, Japan, vol. 1, pp. 98–102 (2004)
27. Zerari, N., Abdelhamid, S., Bouzgou, H., Raymond, C.: Bi-directional recurrent end-to-end neural network classifier for spoken Arab digit recognition. In: 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), Algiers, pp. 1–6 (2018)
28. Hmad, N., Allen, T.: Biologically inspired continuous Arabic speech recognition. In: Bramer, M., Petridis, M. (eds.) SGAI 2012, pp. 245–258. Springer, London (2012). https://doi.org/10.1007/978-1-4471-4739-8_20
29. Bouchakour, L., Debyeche, M.: Improving continuous Arabic speech recognition over mobile networks DSR and NSR using MFCCs features transformed, **12**, 8 (2018)
30. El-Desoky Mousa, A., Kuo, H.-K.J., Mangu, L., Soltau, H.: Morpheme-based feature-rich language models using deep neural networks for LVCSR of Egyptian Arabic. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, pp. 8435–8439 (2013)
31. AlHanai, T., Hsu, W.-N., Glass, J.: Development of the MIT ASR system for the 2016 Arabic multi-genre broadcast challenge. In: 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, pp. 299–304 (2016)
32. Cardinal, P., et al.: Recent advances in ASR applied to an Arabic transcription system for Al-Jazeera, p. 5
33. Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S., Glass, J.: A complete KALDI recipe for building Arabic speech recognition systems. In: 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, pp. 525–529 (2014)
34. Tomashenko, N., Vythelingum, K., Rousseau, A., Esteve, Y.: LIUM ASR systems for the 2016 multi-genre broadcast Arabic challenge. In: 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, pp. 285–291 (2016)
35. Khurana, S., Ali, A.: QCRI advanced transcription system (QATS) for the Arabic multi-dialect broadcast media recognition: MGB-2 challenge. In: 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, pp. 292–298 (2016)

36. Graciarena, M., Kajarekar, S., Stolcke, A., Shriberg, E.: Noise robust speaker identification for spontaneous Arabic speech. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, Honolulu, HI, pp. IV-245–IV-248 (2007)
37. Tolba, H.: Comparative experiments to evaluate the use of a CHMM-based speaker identification engine for Arabic spontaneous speech. In: 2009 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, China, pp. 241–245 (2009)
38. Ettaouil, M., Lazaar, M., En-Naimani, Z.: A hybrid ANN/HMM models for arabic speech recognition using optimal codebook. In: 2013 8th International Conference on Intelligent Systems: Theories and Applications (SITA), Rabat, Morocco, pp. 1–5 (2013)
39. Wahyuni, E.S.: Arabic speech recognition using MFCC feature extraction and ANN classification. In: 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, pp. 22–25 (2017)
40. Venkateswarlu, R., Kumari, R., JayaSri, G.: Speech_recognition_by_using_recurrent_neural_networks, **2**(6), 7 (2011)
41. Cloud Speech-to-Text. https://cloud.google.com/speech-to-text/. Accessed 18 Feb 2019
42. Speech-to-Text. https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/. Accessed 18 Feb 2019
43. IBMWatsonSpeech-to-Text. https://www.ibm.com/watson/services/speech-to-text/. Accessed 18 Feb 2019
44. KALDI. http://kaldi-asr.org/. Accessed 18 Feb 2019
45. Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S.: A complete KALDI recipe for building Arabic speech recognition systems. In: Presented at the 2014 IEEE Spoken Language Technology Workshop (SLT), pp. 225–229 (2014)
46. Manohar, V., Povey, D., Khudanpur, S.: JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, pp. 346–352 (2017)
47. The Microsoft cognitive toolkit. https://www.microsoft.com/en-us/cognitive-toolkit/. Accessed 18 Feb 2019
48. An open source machine learning framework for everyone. https://www.tensorflow.org/. Accessed 18 Feb 2019
49. Sim, K.C., Narayanan, A., Bagby, T., Sainath, T.N., Bacchiani, M.: Improving the efficiency of forward-backward algorithm using batched computation in TensorFlow. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan (2017)