Visually Impaired Assistance with Arabic Speech Recognition on GPS

Siwar Rekik
College of Computer and Information Sciences
Imam Mohammad Bin Saud Islamic University
Riyadh, Saudi Arabia
sarekik@imamu.edu.sa

Lamya Alaqeel
College of Computer and Information Sciences
Imam Mohammad Bin Saud Islamic University
Riyadh, Saudi Arabia

Roaa Kordi
College of Computer and Information Sciences
Imam Mohammad Bin Saud Islamic University
Riyadh, Saudi Arabia

Sara Al-Rashoud
College of Computer and Information Sciences
Imam Mohammad Bin Saud Islamic University
Riyadh, Saudi Arabia

***Abstract*— People who have impaired vision regularly need a guide to assist in obstacle avoidance. Several electronic devices are currently used to provide guidance for a remote location. One of the latest trends in technology is Automatic Speech Recognition (ASR) which has become a primary communication tool for special needs people such as visually impaired and blind people. Nowadays, these people in Saudi Arabia could not find public places offering services such as braille readings on menus and flyers, sound facilities, ease of movements, health care and so on. Our application (Ayn) provides a database of several locations and all the information needed. This project investigated the suitability of a user-centered and client-server approach for the development of a talking GPS planned to fill a niche for outdoor wayfinding. We highlight the importance of having more places serving blind people in public places such as restaurants, centers, hospitals, parks … etc. The developed application used a speech-recognition speech-synthesis interface. The prototype solution incorporates a custom web application that accesses the Google Maps API. The system is intended to be scalable and extensible with additional features. The quality of Arabic speech recognition is improved over Google Speech Recognition API for Arabic using one of the machine learning algorithms: Artificial Neural Network (ANN).**

Keywords— Automatic speech recognition, speech recognition, Global Positioning System, impaired vision, GPS, Navigation.

I. Introduction

In a human interaction with the software, there are multiple needs and different priorities from one user to another. The recognition technology could be based on fingerprint, voice, or face. Speech Recognition (SR) is the process that converting a speech signal to a sequence of words [1] using a special algorithm model to implement speech recognition software. There are four types of SR systems based on the speech utterance and they are isolated words, connected words, continuous speech, and spontaneous speech. Isolated word speech recognition recognizes one single word at a time and it is suitable when the system is requiring a word or command from the user. While connected words and continuous SR take multiple words at a time without silence between the different words. [2].

SR system is based on two phases, the first one is the training phase and second one testing phase. In the training phase, the speech will be recorded then parameters of the classification model will be extracted and stored in the database. In the testing phase, the feature of the test pattern is extracted and compared with the trained model. The tested data decide which class of data is belonging to depending on the matching pattern [3].

Recent developments in Automatic Arabic Speech Recognition (AASR) research exists, but it still suffers from limited sources in terms of the difficulty of the Arabic language. Moreover, the difficulty of building AASR arises because it is involving more than one field at the same time such as natural language processing, Arabic speech processing techniques, and Arabic phonetic [4]. Usually, building such a phonetic AASR requires plenty of time even years to achieve a complete corpus of phones that consist of the probability of all different pronunciation to form a specific word from a listed dictionary. Thus, this research will focus on the learning algorithm Neural Network (NN) for the speech training and the voice processing in the ASR using Application Programming Interfaces (APIs) that support the Arabic language such as Microsoft Speech API or Google Voice API.

Nowadays, ASR is used in the research field for different purposes like dictation speech to text, searching by voice, etc. But there is a lack of research that deploys ASR and Global Positioning System (GPS). Therefore, combining these two technologies is quite new and not widely experienced before especially for the targeted user who is visually Impaired. Therefore, the proposed application aims to serve the visually impaired in one of their most important needs, which is moving between places through an Arabic SR application depends on GPS that guide those people to the needed places in order to reach their desired locations with the electronic maps on mobile phones.

978-1-7281-5467-1/20/$31.00©2020 IEEE

II. BACKGROUND

Nowadays people with disabilities and special needs require some special services. Generally, there is a lack of applications that assists the disabled in getting enough information about the desired places. As the awareness of disabled needs increases and several types of facilities and accessibilities are provided for them, it becomes a necessity to have an application that enables the disabled from searching and locating such places on Global Positioning System (GPS). The potential users of this project system are the people with visual disabilities or visually impaired people. Through this system, they will be able to search a location that provides a specific type of accessibilities using voice-in technique via ASR which will support the Arabic language. The system then will locate the spoken word -location name- to a matching name from the database that stores the information of the places, that supports visually impaired needs and then spell the location name to the user using the voice-out technique. This research will assist in understanding these concepts, which are associated with speech recognition. Amid all of these approaches, artificial intelligence is found to be the most successful and incorporated methodology, which has strengthened and improved speech recognition practices [5].

The proceeding manuscript will commendably help in illustrating the core concept of artificial intelligence, as well as the technological advances, which have been occurred in artificial intelligence. In addition, this paper will also provide an enhanced mobile application that supports the Arabic language supplied with an SR training algorithm for better recognition to find the desired locations that support visually impaired needs. Providing such an application will encourage more places to offer accessibilities for the visually disabled and will allow them to find the place that best suits their needs.

Automatic Speech recognition (ASR) is the ability of a machine or program to identify words and phrases in that a person speaks through a microphone and converts them to written text. ASR has a wide area of applications, that mostly recognizes an input, such as command and control, data entry, or interactive conversation that needs to be understood by an intelligent agent. There are some challenges in SR like different speakers' pronunciation and accent, disfluencies, size of vocabulary words, language modeling, and Noisy environment [5]. Basically, there are four main tasks of automatic speech recognition as described below:

A. *Analysis*

Applied to the data type or word type, the entered voice will be either isolated words or continuous speech. Also, the analysis considers the design of the SR based on the following factors: the size of vocabulary, speech style, and environment (noise /signal).

B. *Feature extraction*

Consists of three stages: speech analysis, vector compilation of static or dynamic features, then transformation of vectors to the recognizer.

C. *Modeling*

There are so many modeling techniques used in SR, but the most common are: hidden Markov modeling (HMM), Neural Networks (NN), and Gaussian Mixture Model (GMM)[6]

D. *Testing*

In this stage of evaluation, a machine learning mechanism could be used to achieve the best possible results [7,8,9]. Some speech languages are characterized by challenges more than others, such as the Arabic language. So, conducting an Arabic speech recognition system need lots of effort and studying. The most difficult problems in developing highly accurate ASRs for Arabic are the morphological complexity, natural language, the enormous dialectal variety, etc. Therefore, we decided to start where the others ended up and decided to use an Arabic speech Application Programming Interface (API) that will support the Arabic language in the first place and allow learning algorithms to be running on the voice parameters.

III. RELATED WORK

Nowadays, Arabic Speech Recognition researchers are interested in the extent of voice accuracy using different techniques and specific algorithms to extract the correct word. In order to use these algorithms for recognition and training with a selected language, there are specific phases that must be taken into consideration. The speech recognition system is composed of three components and they are an: acoustic model which contains sound and acoustic data, a dictionary for words pronunciation, and a Language Model (LM) that consists of the rules and pattern of the word which will be followed in order to get the resulted words [10]. Through searching Arabic papers in speech recognition, we found that the Arabic ASR mostly was built for classification purposes and identification of speakers. Most of those papers are old and there is a lack of research in the field of AASR. So, the following paragraphs will talk about some Arabic papers that took the lead of building AASR for the recognition purpose whether it is a phone or word recognition.

Satori et. Al in [10, 11] has developed an Arabic speech recognition application using CMUSphinx4 that supports java and using the SphinxTrain as a training model. In the training phase, a corpus was generated of the 10 Arabic digits with the aid of Moroccans speakers asked to pronounce each digit 5 times. In the first paper, they (Satori and the authors) asked 6 speakers to do the test and in the second they had investigated more and ask 60 speakers to do the test. The results were determined by taking the mean recognition ratio of correct words for each speaker. In [10], the mean recognition ratio result for three male speakers was 86.66 %, 86.66%, and 83.33%. In [11], the mean recognition ratio for 35 men speakers was 96.67 %, 93.33%, and 93.33% while for females it was 86.66%, 83.33% and 90,00% respectively.

Al-Qatab et. Al [12] developed an Arabic speech recognition system using Hidden Markov Model Toolkit (HTK) which is a tool created by Steve Young at Cambridge University Engineering Department (CUED) in 1989. This AASR system works for both continuous and isolated words speech recognition. It takes the speech as an input and compares it with the HMM transition model (also called training model) to find any match. In the study, the authors had built their Arabic dictionary by a specifying list of the word with its meaning in English. For training and testing, they used 13 Arabic native speakers from 8 countries. They were asked to speak 33 words and repeat each one 4 times. The results show that the system achieved 98.01 % of word correction and 97.99%-word accuracy.

Authors in [13], proposed a phonetically rich and balanced speech corpus for an Arabic speaker-independent continuous ASR. They used a database of Arabic sounds developed by King Abdulaziz City for Science and Technology (KACST). The database consists of 663 phonetically rich words and 367 sentences. The sound corpus has a collection of characteristics because it contains sound files from different ages, regions, nationalities, and professions. For the training phase 367 sentences are used and an extra 48 sentences for testing. A MATLAB program was used in the data preparation process for performing the segmentation of sentences on the recording for every single speaker. It considers silence as the main factor of segmentation. CMU Sphinx 3 which is HMM-based is used for the acoustic model.

The experiments tested different speakers from both genders with similar sentences resulting 95.92% and 96.29% of accuracy rate, and a Word Error Rate (WER) of 5.78% and 5.45% while different speakers with different sentences got 89.08% and 90.23%, and a WER of 15.59% and 14.44%.

Neural Network is proven to be a powerful algorithm technique in AI in various fields especially speech recognition. Usually, it is used as a complementary component of the ASR HMM system for the sake of recognition or classification tasks. Each of the following papers applies the Neural Network(NN) for implementing their system:

Zhou et al. [14], proposed multi-stream features using Deep Neural Network (DNN) which is a neural network but with more than one hidden layer instead of using the normal Artificial Neural Network (ANN) for speech recognition. Their hybrid system (HMM-DNN) does not randomly generate weights, instead, they used a first fitting generative model called Deep Belief Network (DBN). They have also used TIMIT database that contains 6300 sentences spoken by 630 speakers from 8 different dialect regions. The results showed a fact that multi-stream ASR systems usually beat the best single-stream system, especially in noisy environments. The proposed method, which has 4 hidden layers reached a percentage of 6.1% (PER)Phone Error Rate reduction.

Also, for the hybrid NN-HMM model, authors in [15] conducted a fast speaker adaptation method that was based on the adaption methodology. It consists of giving each speaker a unique code, then in a later step; a weight is given in the NN model. They had used TIMIT corpus in their experiment and they achieved 21.34% PER passing the whole NN layer. Also, the results in the activation function give PER of 21.65% for one adaptation utterances. and 20.91% for 7 utterances.

According to [19], authors have proposed a Multi Based Depth Neural Network (MBDNN) which is a multi-layer perception neural network but with more than one network running in parallel to the output. Each sub-network is called a base. They used the TIMIT dataset and they partition it into two parts: the first for training and the second for testing. For the learning process, they use a smaller learning rate with range values (0.01- 0.1). Two comparisons were tested in order to perform an experiment. First, they compare the DNN-HMM model with the DBN-HMM model and the result shows that DNN-HMM has the shortest recognition time. The second was the proposed system MBDNN-HMM and DNN-HMM which resulted that MBDNN-HMM had a higher recognition rate by 4% than DNN-HMM.

Finally, a related paper to this project idea had discussed the google Arabic voice search recognition in terms of recognition goal with Arabic using google speech API. Biadsy et. Al [21] present a huge effort in building a commercial automatic speech recognition (ASR) for the Arabic language they use it to recognize five different Arabic dialects: Saudi Arabia(SA), United Arab Emirates(AE), Egypt(EG), Lebanon(LB) and Jordan(JO). they collect their spoken dictionary in environments (noisy, silent, car … etc.)

Also, there are various types of neural network or Artificial Neural Network models to use in Speech recognition such as Recurrent Neural Networks (RNN) as in [16,17] and Convolutional Neural Network (CNN) as in [18].

To summarize, there are different approaches obtained not just in ASR but also in the use of the NN algorithm based on the purpose of the study[24,25]. Therefore, this paper will be focusing on the core structure of Artificial Neural Network (ANN) and its effect on voice recognition and then compare the testing results with existing systems such as: google as used in [22].

The research group considered mobile applications with a similar idea but different features and operating system such as BlindSquare [23] which is an iOS application use Siri voice recognition. It is a paid software that uses GPS for blinds. It has no android version. Also, we had found another similar application called “voice navigation" in the android app store, its basic idea to dictate the voice to text using an android speech API, and then shows the user the navigation root to his spoken location (if it matches) from the database. Also, in a graduation project provided by authors of [22] in the university Al-Imam Muhammad Bin Saud Islamic University, they developed a CMUsphnix recognition to train the user voice with a list of previously defined words to increase system performance.

Therefore, and after collecting the key advantages of the missing points of previous applications and combine the pros and cons of them all, and noticing the lack of Arabic applications that supports voice recognition, the idea of visually impaired assistance came on board to apply the AI algorithm neural network to enhance the voice recognition for the visually impaired people to guide them to places on GPS that support their needs.

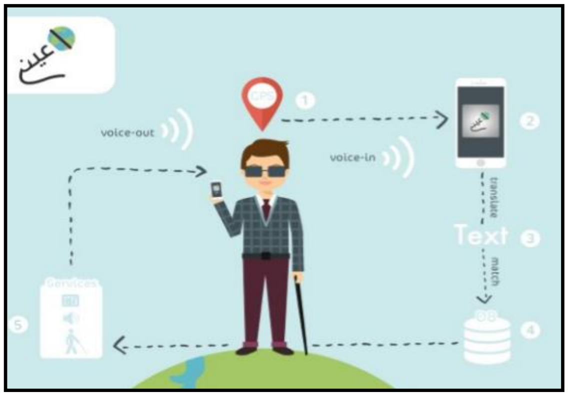

Fig. 1. Application sample scenario.

IV. METHODOLOGY

The system was designed using object orientated methods. Figure 1 gives a simple scenario of our proposed application. Our proposed technique needs a flexible software process model; therefore, the incremental model approach is used. The incremental model is a method where developers add piece by piece and test each unit by performing unit testing before testing the system in general. Also, some of the advantages of this model that it is less costly in case of scope or requirements modification and is easier to test and debug during a smaller iteration. The development of our application is based on the following steps:

- The application will get user location GPs.
- The user speaks Arabic location name to the app.
- The application will translate the spoken words from Speech to Text(STT).
- The system will match the text with the location name stored in the location services database.
- The system will read the location information containing services such as Braille language, Sound facilities, accessibility… etc.

A. System Architecture

The proposed application is called "Ayn" in Arabic "عين" and it consists of four major building blocks as is shown in Figure 2**.** The system components below show the different components system and the relation between them.

1) Speech recognition tool: Android Speech API (google dictionary).

2) Speech recognition algorithm: Neural Network (NN).

3) Database of locations and services: using the SQLite android tool.

4) Location services information: that had been added by the developer will be retrieved in the case the search matches the information in the database.

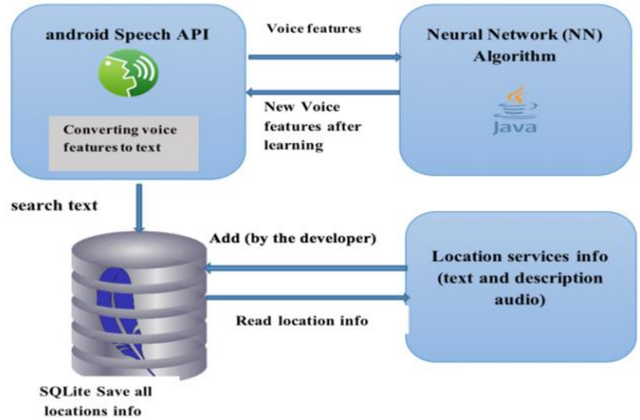

Fig. 2. System components.

B. Modular Decomposition

The methodology employed the following tools and technologies we will specify major sub-systems component and the relationships between these components.

1) Neural network classes structure

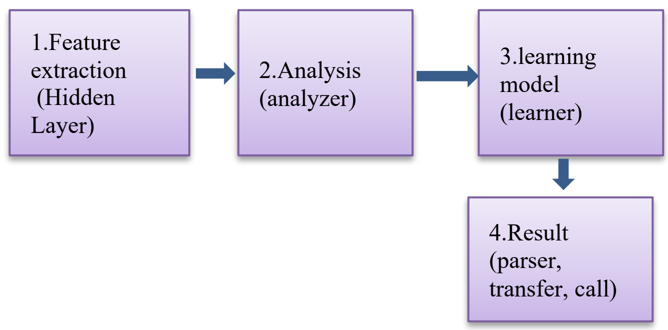

Fig. 3. NN algorithm steps

Neural Network (NN) is our chosen algorithm to implement. This choice has many reasons the main reasons are: it is mostly used for isolated words and second its results are so effective on voice accuracy upon other algorithms. However, we will review the NN components and its classes here from Object-Oriented Design (OOD) view.

Every step represents a specific class of the algorithm as shown in Figure 3. These steps are the steps that every speech recognition follows and they are:

a) Feature extraction: or data preparation that was held in the Hidden Layer class that does the initialization process for the weights and bias for each neuron.

b) Analysis: that represents in the analyzer class who perform the calculations of the neural network equation of weights and bias for each neuron and output the array of voice with a new value called "Fout".

c) *Learning model:* represented in the Learner class that adjusts the learning rate (0.05 f) and calculates the error of the original voice array and the output voice from the analyzer they apply some equations to modify the weight and bias.

d) *Result:* the resulting class will use a combination of packages. Each package has one main function to do in the NN algorithm.

V. Results And Discussion

The development of the raw user requirements included the following considerations: the software and tools used in our system and discuss in detail the steps to build our application of speech recognition with the neural network using android tools, libraries, databases, and Google Maps Application Programming Interface (API). The implementation of the proposed system is based on three phases: starting from building the android speech recognizer with neural network analysis algorithm ending up with connecting to the google map.

A. *NN algorithm in android code*

The neural network concepts and the logical procedure were explained in detail. However, this paper will highlight the core java methods in the android code of the algorithm and represent its parameter inputs and return values.

The voice features were stored in arrays of bytes into two copies: one for input and the other one for testing in the learner method. Although we did not use a neural network library, we have implemented four methods that simulate the 4 steps of the NN model as mentioned in Figure 1 using Java code in the android studio application. At the end of the program execution, the speech will be displayed as text results of the probability and likelihood of the same pronunciation in the dictionary:

Fig. 4. NN Results.

The text result is up to 5 results as shown in Figure 4. These are the results of "معهد النور" word.

B. *Database attributes*

Database attributes were derived from the locations table our team collected from interviews and social network about places that provide services for disabled people. Each location has some attributes such as name, description, neighborhood, X, and Y coordination. The database, to be used in the location services database in the searching phase and google map as well.

C. *Audio Database and storage*

As the visually impaired user need to hear the information of the location to know its services, locations descriptions were recorded in audio files by human voice to be added to the database information. The audio files must be in .mp4/.mp3/.m4a/.wav format. The file should not exceed 2 MB otherwise it will not be added to the database. To solve the storage problem of the audio files in the database, we use the base64AudioData encoding method in android to store the description audio files as a string in the database. Then, after the application finds the target location, it retrieves all the data associated with the location name including the description audio which was encoded previously to strings to decode it for the sake of reading the results for the user by voice.

The searching function is an integral part of the first phase and the second. First, we construct the function in the Database Handler class. The function query took the "location name" attribute to go through the database record and find all location information associated with the specific input name of the location. After that, we must locate the dictated text obtained from phase 1 goes to the searching query. Since we have 5 possible text results we took the first result as a default to use in the query by stating the index 0 of the results arrays in the search function placed in the main activity class. If the database records are empty or there is no location name with the same string match an audio error message will appear to the user. Otherwise, the result will be equal to the location name attributes and the data of that location will be retrieved by *getServices()* function.

VI. System Testing

In order to test our system, we had met the visually impaired personally and test the application on them. The user test involves 10 females of different places, different ages, specialty, and the type of visual impairment. Google voice search is the voice recognition tool in all android devices that use Google applications. thus, to have a reliable testing experience we test our application Ayn with google voice search on four different types of places categories, basically 7 places on each category listed on the following: restaurants, centers/institutions, pharmacies, and universities.

After taking the results on each category we sum up the total and divide it into the total number of words in all categories to obtain the final WER for Ayn which was 96%. Table 1 shows a comparison done with google voice recognition to compare the accuracy rate in both systems.

Table1. Results Comparison

App	Accuracy
Ayn (NN with google API)	96%
Google voice recognition	98%

VII. Conclusion

The proposed system "Ayn application" is a system that links between visually impaired people and locations are used to help Arabic people with visual disabilities to communicate with the environment around them in the easiest way. The system depends mainly on Speech Recognition "SR" technology which is the process of converting a speech signal to a sequence of words as a part of machine learning. The client-server model enabled the system to make use of existing tools suitable for dealing with location and places information we used the local SQL database within the application as well as Global Positioning System (GPS) the Google Maps API.

The method we used to implement is the neural network Algorithm. This algorithm focuses on the analysis of the voice features database on two phases in such recognition algorithms: training and testing. Our system is intended to add additional services such as the ability to calculate the distance and showing the route path using push-to-talk to reduce the effect of background noise. Further, more cities can be added for visually impaired services.

References

[1] M. Young, *The technical writer's handbook: writing with style and clarity.* University Science Books, 2002.

[2] S. K. Saksamudre, P. P. Shrishrimal, and R. R. Deshmukh, "A review of different approaches for speech recognition system" in *International Journal of Computer Applications,* 2015, vol. 115, no.22.

[3] M. R Gamit, K. Dhameliya, and N. S. Bhatt, "Classification techniques for speech recognition: a review," in *International Journal of Emerging Technology and Advanced Engineering,* 2015, vol. 5, no.2, 58-63.

[4] M. A. Abushariah, R. N. Ainon, R. Zainuddin, M. Elshafei, and O. O. Khalifa, "Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools," in *IEEE International Conference on Computer and Communication Engineering (ICCCE'10)*, May. 2010, pp. 1-6.

[5] W. Ghai, and N. Singh, "Literature review on automatic speech recognition," in *International Journal of Computer Applications*, , 2012, vol. 41, no. 8.

[6] M. Gales, and S. Young, "The application of hidden Markov models in speech recognition," in *Foundations and trends in signal processing*, 2008, vol.1, no. 3,195-304.

[7] M. A. Anusuya, and S. K. Katti, "Speech recognition by machine, a review," in *International Journal of Computer Science and Information Security*, IJCSIS, Dec. 2009, ISSN 1947 5500.

[8] P. Saini, and P. Kaur, "Automatic speech recognition: A review," in *International Journal of Engineering Trends and Technology*, 2013, vol. 4, no. 2, 132-136.

[9] Y. Zhang, "Speech Recognition Using Deep Learning Algorithms," Stanford University, 2017

[10] H. Satori, M. Harti, and N. Chenfour, "Introduction to Arabic speech recognition using CMUSphinx system," 2007.

[11] H. Satori, H. Hiyassat, M. Haiti, and N. Chenfour, "Investigation Arabic Speech Recognition Using CMU Sphinx System," in *International Arab Journal of Information Technology (IAJIT),* 2009, vol. 6, no. 2.

[12] B. A. Al-Qatab, and R. N. Ainon, "Arabic speech recognition using hidden Markov model toolkit (HTK)," in *Proc. 2010 IEEE Int. Symposium on Information Technology*, June. 2010, vol. 2, pp. 557-562.

[13] M. A. Abushariah, R. Ainon, R. Zainuddin, M. Elshafei, and O. O. Khalifa, "Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus," in *International Arab Journal of Information Technology (IAJIT),* 2012, vol. 9, no. 184-93.

[14] P. Zhou, L. Dai, Q. Liu, and H. Jiang, "Combining information from multi-stream features using deep neural network in speech recognition," in *Proc. IEEE 11th Int. Conf. on Signal Processing (ICSP)*, Oct. 2012, vol. 1, pp. 557-561.

[15] O. Abdel-Hamid, and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* May. 2013, pp. 7942-794.

[16] A.Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. on Acoustics, speech and signal processing (ICASSP)*, May. 2013, pp. 6645-6649.

[17] A. Graves, and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. 31st Int. Conf. on Machine Learning (ICML-14)*, 2014, pp. 1764-1772.

[18] Y. Qian, and P. C. Woodland, "Very deep convolutional neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. on Spoken Language Technology Workshop (SLT),* Dec. 2016, pp. 481-488.

[19] C. Jun, L. Fei, Z. Yi, and L. Yu, "Research on multi-base depth neural network speech recognition," in *Proc. IEEE 2nd Advanced Information Technology, Electronic, and Automation Control Conference (IAEAC),* Chongqing, 2017, pp. 1540-1544, doi: 10.1109/IAEAC.2017.8054271.

[20] V. Nigam, "Understanding Neural Networks. From neuron to RNN, CNN, and Deep Learning" Medium, Sep. 11, 2018 [Online]. Available: https://towardsdatascience.com/understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90 [accessed February 13, 2020]

[21] F. Biadsy, P. J. Moreno, and M. Jansche, "Google's cross-dialect Arabic voice search" in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4441-4444.

[22] M. Al-Bogami, M. AL-Otaibi, and F. Al-Mutairi, Speech recognition application, Imam Muhammad Ibn Saud Islamic University, Riyadh, KSA, 2014.

[23] Blind Square, 2016 [Online]. Available: http://blindsquare.com/about/ [accessed June 10, 2016]

[24] Welch Labs, "Neural Networks Demystified, Part 1: Data and Architecture" Nov. 4, 2014 [Online]. Available: http://www.welchlabs.com/blog/2015/1/16/neural-networks-demystified-part-1-data-and-architecture [accessed August 4, 2017]

[25] M. A. Nielsen, *Neural networks and deep learning*, 2015. [E-books]