Handicapped Wheelchair Movements Using Discrete Arabic Command Recognition

Khalid M.O. Nahar(1), Moyawiah al-shannaq(1), Rafat Alshorman(1), Ra’ed M. Al-Khatib(1), Mohammad Ashraf Ot.tom(2)

(1) Computer Science Department, Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid, Jordan

(2) Computer Information Systems Department, Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid, Jordan

Received 20 December 2018 - Accepted 4 September 2019

https://doi.org/10.37575/b/com/2034

ABSTRACT

Automatic Speech Recognition (ASR) is an effective and widespread method used to interpret the human voice into commands. Over the past few decades and due to the great evolution in data science, the processing of speech has been embedded in various new technologies. The benefits of such evolution motivated us to develop a discrete voice system for controlling the wheelchair movements for Arab people with physical impairment (handicapped people). The wheelchair *via* voice system was able to recognize seven isolated words in the Arabic language. The approach uses the CMU-Sphinx4 ASR for evaluating the effectiveness and the accuracy of the recognition system. A corpus was built, and then the system was trained on 70% of this corpus and tested using the rest of 30%. The experiments reveal that the recognition rate of the proposed approach is 96%. The average level of accuracy of the real experiment and corpus-based testing experiment reached 94%.

Keywords: Acoustic Model, Arabic Speech Recognition, Handicapped, Hidden Markov Model, Language Model, Wheelchair.

INTRODUCTION

Speech is the main method of communication for information exchange among humans. Therefore, mutual understanding among humans takes place through speech. Difficulties in communication using speech do arise, due to the variety of spoken languages. Recently, many computer applications in the field of computational linguistics have been used intensively to research the problem of recognizing and translating spoken language (Jurafsky, 2000). The productivity of such software applications statistically enhances and enriches the disciplines within the natural language processing field (Jurafsky, 2000).

Physically handicapped persons need a wheelchair for their daily activities inside or outside home. Using human hands in moving the chair is exhausting; hence, it is preferable to move it using some automotive manner such as speech control. This could be achieved by translating vocal Arabic commands into electrical and mechanical actions to initiate a motor for moving the wheelchair. Based on these facts, Automatic Speech Recognizer (ASR) is to be produced to recognize discrete pronounced commands. Fig. 1 shows the general framework of an ASR system.

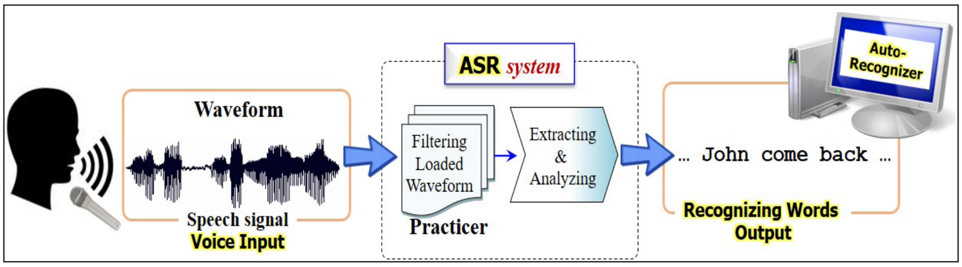


Fig. 1. General framework of the Automatic Speech Recognition (ASR) system for recognizing spoken words.

As shown in Fig. 1, the spoken words are processed by the ASR to convert them into written commands that are interfaced with the wheelchair through the Programmable

Integrated Circuit (PIC), and then it will be reflected as a movement on the wheelchair. The proposed ASR system recognizes seven isolated Arabic words. Using the speech recognizer (SMU-Sphinx4), which support Arabic language speech recognition, a working prototype of the proposed approach was implemented. SMU-Sphinx4 recognizer is an open source speech recognizer that is based on Hidden Markov Model (HMM) of three states.

In this research, the words used to represent a set of commands to control the wheelchair of the disabled user. However, the idea could be extended to cover other equipment such as Unmanned Arial Vehicles (UAV) or for controlling games. This research focuses on serving those physically impaired persons with complete disability whose mother tongue is Arabic. Any application that can be adapted in the same way can be implemented via voice systems.

This paper is organized as follows: The literature review section provides a comprehensive overview of previous works. The next section is a background and some information about ASR system and Mel-Frequency Cepstrum Coefficients (MFCC) algorithm. After that, the proposed approach is introduced followed by the experimental results and discussion with précised calculations are shown. Finally, the conclusion of the work and possible future direction are given. References are provided at the end of the research paper.

LITERATURE REVIEW

This section discusses speech recognition in general and the previous speech recognition systems used for Arabic language. In our world, many people are completely disabled and that prevents them from carrying out everyday activities. Because of this, the objective of this paper is to propose, design, and develop an Arabic voice- recognition system suitable for operating a motorized wheelchair. The proposed Automatic Arabic Speech Recognition System (AASRS) could replace traditional wheelchairs that require a human interaction (via a joystick) or human assistance. The ASR systems are defined as a set of methods for converting voice signals into a series of words or linguistic units (Al-Qatab and Ainon, 2010) an Arabic dictionary was built by composing the words to its phones. Next, Mel Frequency Cepstral Coefficients (MFCC. ASR has become widely adopted in different types of applications such as those that targeted disabilities in order to facilitate their interactions with the environment easily (Satori *et al*., 2007). In the following subsections, the concepts of SRS and ASR are briefly presented.

Speech Recognition System (SRS):

Speech recognition refers to applications that identify spoken words or sentences through a computer. The goal is to convert the recognized spoken words into a format readable by a machine or a human. Speech recognition systems (SRSs) may be phoneme-based or syllable-based so that the recognizer can identify the closest spoken words or phrases. The SRS usually depends on a language model to enhance recognition accuracy (Allen *et al*., 2000) and (Al-Barhamtoshy *et al*., 2014)

Arabic ASR System:

One of the oldest Semitic languages in the world is the Arabic language, which is officially the sixth most spoken languages and one of the formal United Nations languages. There is an official Arabic linguistic form known as Modern Standard Arabic (MSA), which is used in formal media, courtrooms, offices, schools, and universities (Abdul-Mageed *et al*., 2014) and (Kirchhoff *et al*., 2002). Recently, many Automatic Speech Recognition (ASR) systems have been proposed in the field of Arabic Speech Recognition. The earliest Arabic ASR systems were developed to recognize the MSA, because there are many challenges such as lack of lexical variety and data sparseness in Arabic in general, and it is also one of most complex languages due to the morphological variations of the letters in its alphabet (Diehl *et al.*, 2012) and (Ali *et al*., 2015a). Arabic language ASR research is still in its infancy compared to research in other languages (Zarrouk *et al*., 2014). Many Arabic ASR researchers use different learning algorithms such as Hidden Markov Models (HMM), SVM and hybrid from multi-system (Zarrouk *et al*., 2014) and (Ali *et al.*, 2015b).

For effective natural speaker-independent Arabic continuous speech recognition, (Abushariah *et al.*, 2010)implementation, and evaluation of a research work for developing a high performance natural speaker-independent Arabic continuous speech recognition system. It aims to explore the usefulness and success of a newly developed speech corpus, which is phonetically rich and balanced, presenting a competitive approach towards the development of an Arabic ASR system as compared to the state-of-the-art Arabic ASR researches. The developed Arabic AS R mainly used the Carnegie Mellon University (CMU) and Abdou *et al.* (2014) developed a system that aims to promote a phonetically rich and balanced speech corpus for the Arabic ASR system. The authors used the CMU-Sphinx4 tools with the Cambridge HTK tool. For the tri-phone acoustic modeling, the authors used a five- stage HMM with three pronouncing states. The model includes uni-gram, bi-gram, and tri-gram words. For evaluating the proposed system, the authors trained the system with 7 hours of phonetically rich recordings. They also use another one hour of the recorded corpus for testing purposes. The results obtained reveal 92.67% accuracy for word recognition with 11.27% Word Error Rate (WER) for the same speaker using different utterances. In general, the proposed system yields an accuracy of 95.92% for word recognition with 5.78% Word Error Rate over different speakers reading the same sentences. On the other hand, the proposed system achieves 89.08% word recognition accuracy with 15.59% word Error Rate if different speakers read different sentences. Recently, researchers worldwide have tended to employ various techniques for Arabic speech recognition systems. These techniques include segmentation, vector quantization and hybrid techniques. For instance, (Frihia and Bahi, 2016) present a system for speech recognition by using automatic Arabic speech segmentation The system depends on HMM. However, such approaches require a special corpus of transcribed, recorded, and categorized data. Abdou *et al*. (2012*)* and Nahar *et al.* (2016) offer a hybrid algorithm for recognition. Mainly the algorithm is a combination between linear vector quantization (LVQ) with a hidden Markov model (HMM). The hybrid algorithm was able to recognize and identify Arabic sounds in continuous speech recognition. An Arabic corpus of multiple TV news recordings was used for the training and testing along with a data driven approach to extract the training characteristic vectors for neighboring correlation information. Their approach generates the phoneme by a splitting algorithm. The generated phonemes were modeled using a learning vector quantization algorithm to achieve 89% recognition accuracy of the Arabic phonemes.

BACKGROUND

General ASR System:

The process of speech recognition has two components: Front-End and Back-End. The Front-End receives the speech signals from an outer source (such as the live speech from a microphone or an audio file as streamed speech). In the pre-processing, some signal characteristics were enhanced in order to achieve accurate results. The feature extraction process is done using the MFCC (Mel Frequency Cepstral Coefficients) technique. The Back-End is responsible for pattern matching, decoding and mapping between the extracted features and the phonetic dictionary, and modeling the language. The components of a general ASR system are illustrated in Fig. 2.

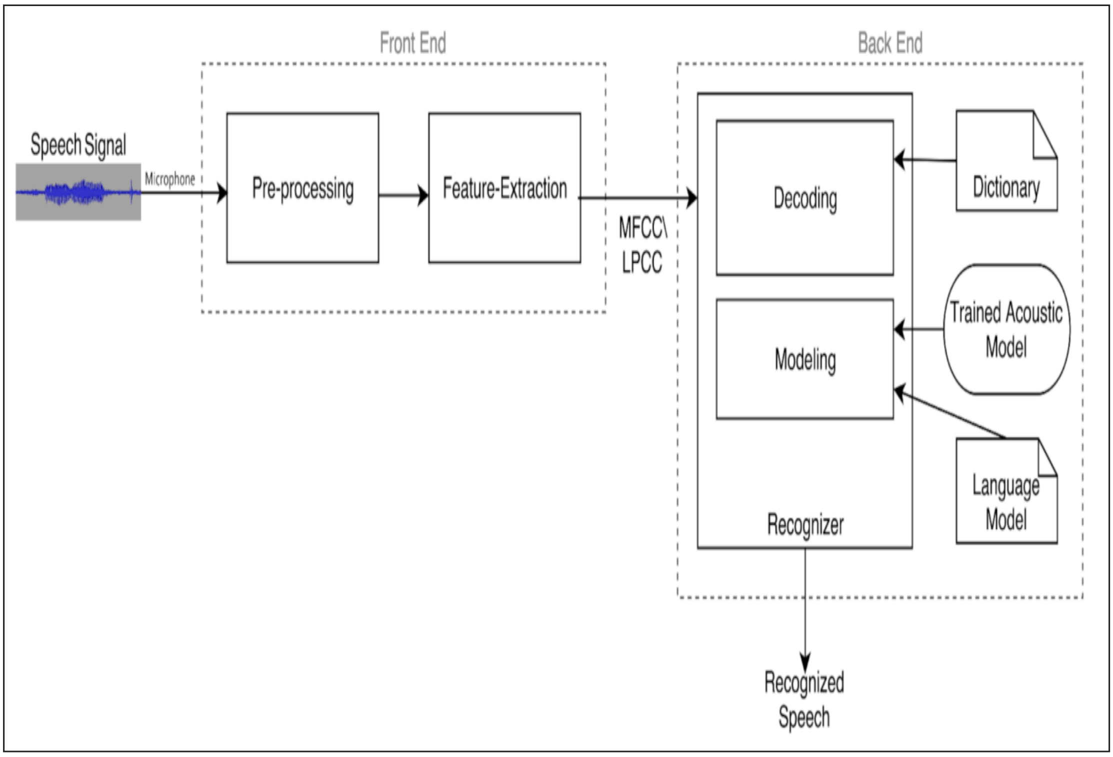


Fig. 2. Components of Automatic Speech Recognition (ASR).

Mel-Frequency Cepstrum Coefficients (MFCC):

MFCC is a short-period power spectrum of sound wave representation. Mel frequencies are based on the human ear's respond to bandwidth variation below 1 kHz frequencies and logarithmically at higher ones to capture the phonetically important characteristics of speech (Majeed et al., 2015). In speech recognition, the correct form of pronunciation depends on the context and controlled by the speaker's voice. Moreover, the more closeness of the pronunciation to modern standard Arabic the more its correctness. The steps involved in MFCC extraction are illustrated in Fig. 3. The MFCC is calculated using- Equation 1 and its implementation was carried out based on MFCC Algorithm in (Nahar et al., 2016).

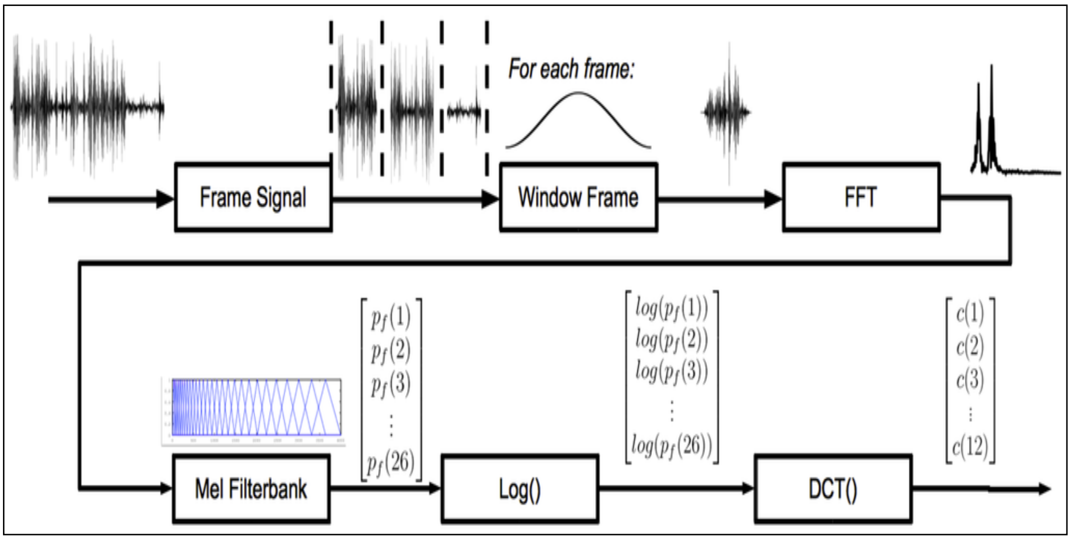

For each frame:
Frame Signal
Window Frame
FFT
Mel Filterbank
Log()
DCT()
$p_f(1)$ $p_f(2)$ $p_f(3)$ ⋮ $p_f(26)$
$log(p_f(1))$ $log(p_f(2))$ $log(p_f(3))$ ⋮ $log(p_f(26))$
$c(1)$ $c(2)$ $c(3)$ ⋮ $c(12)$

Fig. 3. MFCC Feature Extraction Steps (Nahar *et al.*, 2016).

$$C_i = \sum_{k=1}^{N} X_k cos\left(\frac{[\pi_i(k-0.5)]}{N}\right), \quad for\ i = 1,2,\dots,p \qquad (1)$$

where C_i are the cepstral coefficients, p is the order, k is the number of discrete Fourier transforms magnitude coefficients, X_k is the k^{th} order log-energy output from the filter bank, and is the number of filters (usually 13 in our case). Thus, 12 coefficients and an energy feature were extracted, forming 13 MFCC features.

Since the commands are of different lengths, a 13-feature representation of each Arabic word (Command) were extracted for all the acoustic waves as MFCC files corresponding to the acoustic waves.

Hidden Markov Model (HMM):

Training the speech recognizer either continuous or discrete is usually done by HMM model. Hidden Markov Model (HMM) is established by Markov (Baum and Petrie, 1966) and (Rabiner and Juang, 1986), which is a model statistically based on Markov chain for Markov processes. It is used to model time wrapping systems with unobserved (hidden) states (Das and Nahar, 2016). HMM consists of a set of states with transition probabilities from one state to another. There is a hidden state, which depends on the observed state and formulates a stochastic process in order to predict or classify unobserved state. Fig. 4 shows three states of HMM model of an acoustic wave (adapted from[1]), which can be used to represent the Arabic phoneme. This phoneme is composed of three states (start state (S1), suspending state (S2), and exit state (S3)), respectively. The P11 is the probability to be in the same state, while P12 is the probability to move from S1 to S2; and so on until S3 is performed at the end of the phoneme. The HMMs of Arabic phonemes that represent and utterance (for example, a spoken sentence) make a chain called Markov chain. The HMM is a time series learning algorithm, which is usually used in speech recognition with Gaussian Mixture distribution for accurate learning.

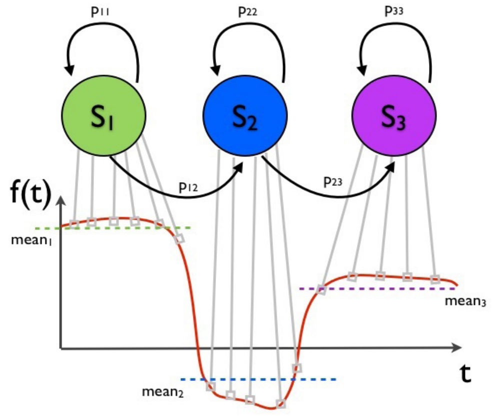


Fig. 4. General framework of Three States HMM algorithm.

PROPOSED APPROACH

In this research, the Arabic command recognition wheelchair was considered in order to ease the life of those with complete physical disability by data collection and features extraction. Fig. 5 illustrates an overview of the proposed recognition system. The novelty of the proposed approach includes the following:

1. A general framework for Arabic commands recognition using Sphinx4 ASR was produced.
2. A corpus recorded and for completely physically handicapped persons was built by a handicapped society in northern Jordan. Seven Arabic words (Commands) were uttered specifically for the chair movement.
3. The Via-Voice control model for Arabic handicapped persons is more realistic and can be extended to include more commands in the future, unlike other similar models which are hardware-dependent like the one in Nishimori *et al.* (2007)
4. All possible pronunciations of the commands are taken into considerations which have made the model more accurate.
5. The support and dedication to Arabic language were not fully supported before.
6. Results in terms of the accuracy of the model shows its superiority over other similar works.

[1] https://cutt.us/Prh06.

The proposed system has been trained and tested using the sphinx-4 ASR system. The proposed recognizer is developed and run on several phases, which include: i) Data collection and building the corpus, ii) Data set pre-processing, including the use of Mel-Frequency Cepstrum Coefficients (MFCC) to extract the set of informative features from the corpus, iii) Speech processing (Training), and iv) Speech processing testing and accuracy calculations. Fig. 5 illustrates the phases of the proposed discrete recognition system.

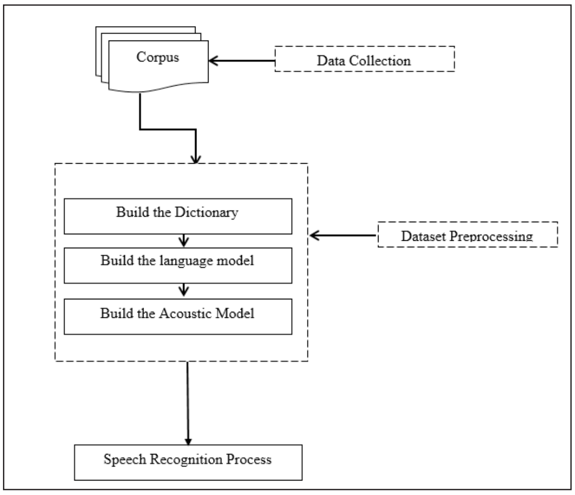


Fig. 5. Main Steps of Proposed Recognition System.

Data Collection (Building the Corpus):
The proposed discrete Arabic command recognizer starts with preparing the corpus where Arabic words and utterances were collected from different speakers. The corpus consists of 700 discrete audio files (15 males and 5 female speakers). The 20 speakers were asked to utter seven different predetermined commands 5 times each. Table 1 summaries the description of the corpus used. Speakers were selected from the handicapped society in North Jordan. The handicapped speakers came from different areas of Northern Jordan and all had complete physical disability except speech. The corpus was limited, since the work targeted handicapped people resident in the north of Jordan only. Naturally, the corpus can be extended to include a wider range of speakers, and not necessarily handicapped speakers. However, since this research is oriented towards handicapped, it was decided to record utterances from only those completely physically impaired subjects. The northern society for the handicapped has only 5 females and 15 males suffering from such a disability.

Table 1. The Corpus Summary

Gender	Count	# of Commands & # of Utterance	Total
Male (M)	15	Uttered 5 7	Audio-Wave 525
Female (F)	5	Uttered 5 7	Audio-Wave 175
Total # of Audio-Wave Files:			700

Using the Audacity software tool for sound manipulation, the files with the specifications showed in Table 2 were generated. The corpus was divided into 70% for training (490 Audio files) and 30% for testing (210 Audio files)

Table 2. Audio File Specifications

Parameter	Values
Sampling Rate	Hz 16000
Output	bit 16
Type	Little Indian
Format	WAV, Mono

Dataset Pre-processing:
This phase of the recognition system has three sub-phases, which included building the phonetic dictionary, building the language model, and building the acoustic model. Each of these sub-phases had its own requirements and processing.

Building the Phonetic Dictionary:
The collected wave files had to be converted to their corresponding MFCC files. It is preferable to keep the same file name. In fact, the extensions of these files are "*.mfc*". The

phonetic dictionary provides a system to map vocabulary words to sequence of phonemes. Fig. 6 shows part of our phonetic dictionary.

أمام	A A M A M
أَمَامْ	E A E M A E:M
خلف	KH AA L F

Fig. 6. Phonetic Dictionary Sample Lines.

The dictionary can contain alternative pronunciations, as seen from Fig. 6, where the word “أمـام” is completely different from the diacritic version of the same word “أَمَـامْ” since the pronunciation is very different. For accurate results, the dictionary has to contain all pronunciation variations of all the expected commands.

Different phone sets template exists in CMU-Sphinx4 for phonemes representation like IPA and SAMPA. CMU-Sphinx4 provide enough flexibility to deal with phones and their processing. Table 3 shows the lists of Arabic phoneme sets that are used for transcription and their representation in English symbols.

Table 3. Arabic phoneme list used in training with their transcription

Phoneme	Arabic Letter	Phoneme	Arabic Letter
SIL	–	TT	ط
A	ـَ	AI	ع
AE	ـَ	F	ف
:AE	ـاَ	Q	ق
AA	ـَ	K	ك
IY	ـِي	L	ل
IX	ـِ	M	م
:IX	ـِ	N	ن
E	ءَ	Y	ي
B	ب	TT	ط
HH	ح	AI	ع
KH	خ	F	ف
R	ر	Q	ق
S	س	K	ك

To identify words by the recognizer, they must be in both dictionary, and the language model, otherwise it will not be recognized. Moreover, the dictionary may contain some extra unused- words, they have no effect and they could be removed or left. The final phonetic transcriptions of the Arabic commands with and without discretization are shown in Table 4. The Arabic words in Table 4 are based on the phonetic unit that will be used by the proposed system. The proposed ASR systems were designed to use Arabic phonemes as the phonetic unit.

Table 4. Phonetic dictionary of 14 words and their transcription.

Words	Phonetic Representation	Words	Phonetic
أمام	AA M A M	بَطِيءْ	B AE TT IY E IX
أَمَامْ	E AE M AE: M	يسار	Y S AE: R
خلف	KH AA L F	يَسَارْ	Y AE S AE AE: R
خَلْف	K HH AE L IX F	يمين	Y AE M IY N
سريع	S AE R IX: AI	يَمِينْ	Y AE M IY N
سَرِيعْ	S R IX: AI	قف	Q IX F
بطيء	B TT IY E		

Building the Language Model:

The language model is an important component of the configuration, which tells the decoder which sequence of words it is possible to recognize. Generally, there are two types of models to describe a language: the grammar and the statistical language models. The grammar model describes simple types of language for command and control. The grammar model is usually handwritten or automatically generated. In this paper, the statistical language model is used and automatically generated using an online tool from a CMU-Cambridge statistical language model toolkit called SIRLM to extract the n-gram for each word. SRILM is a toolkit for creating language models (LMs), which are frequently used in speech recognition, and machine translation. The statistical language model describes more complex language. The statistical model also includes the probabilities of the words and their combinations. These probabilities are estimated from sample data with some flexibility. In this work, the number of uni-grams is 11, whereas the number of bi-grams and tri-grams are 18 and 9 respectively. In general, given a word sequence w_1, w_2, w_3, and w_n, the model will assign the probability to the sequence in the ratio of the frequency of occurrence of the sequence in training. During training, the model will construct the distribution of unique n-grams (Rosenfeld, 2000) , which is approximated using equation 2:

$$P(W_1^N) \approx \prod_{1}^{N} P(W_i | W_{i-n+1}^{i-1}) \qquad (2)$$

Building Acoustic Model:
The acoustic model is used in the Automatic Speech Recognition to illustrate the relationship between a speech audio signal and its corresponding phonemes or its other linguistic units. The model is trained with a set of audio recordings and their corresponding transcripts. The model is created by taking audio speech recordings, and their text transcriptions. Since the speech sound consists of consecutive words, the statistical representations of the sounds are created using Audacity software. A HMM of three states is provided by the sphinx tool and the acoustic model is trained using the sphinx-train tool. The process of creating the Arabic Acoustic Model using sphinx-train is shown in Fig. 7.

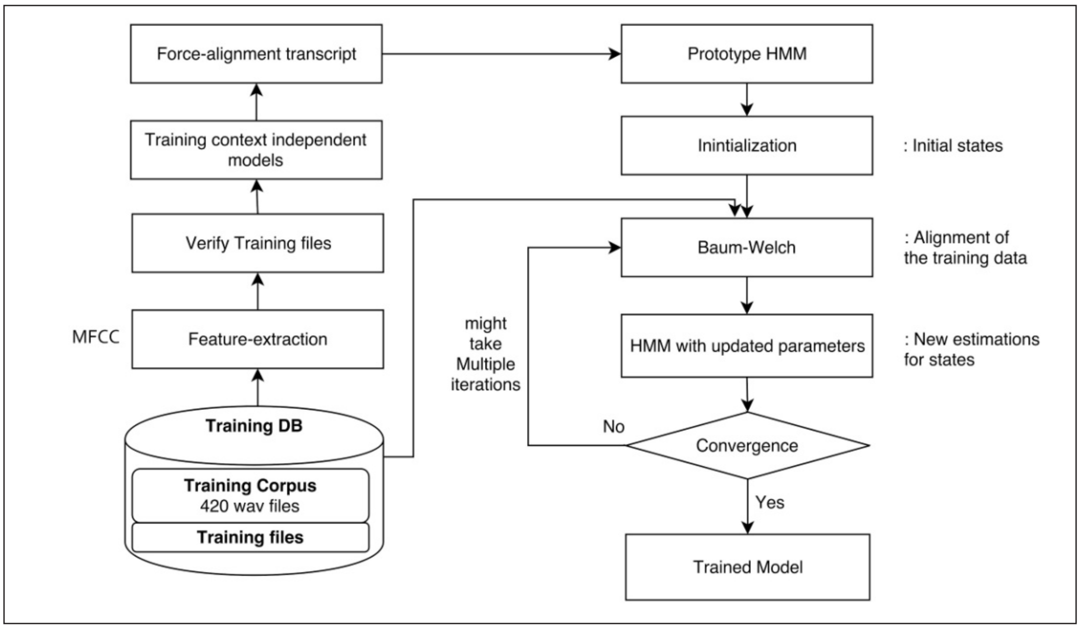
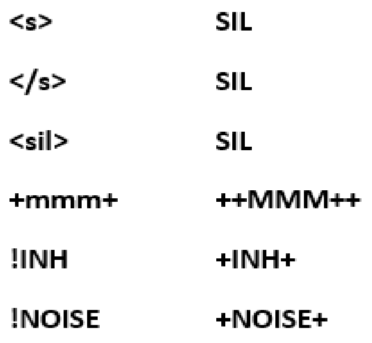


Fig. 7. Training Acoustic Model.

The sphinx tools help matching the live voice speech stream (audio) to its corresponding command in the database. In the training phase, the phones in the phone file are stored one per line with number of phones that matches the phones in the dictionary as well as the *silence* phone. Based on the Arabic phoneme set, a predetermined rule were used to create the phonetic pronunciations of the seven Arabic words. The sphinx-train tool, throughout its training phase, yields statistically a range of possible audio representations for the phonemes. The sphinx-train tool employs the Baum-Welch algorithm iteratively until no improvement is gained in the training phase.

Speech Recognition Process:
As mentioned earlier, the corpus was divided into 70% for training the system and 30% for testing. For now, the training directory has been formulated to include the following: the phonetic dictionary, phone set files, language model, list of fillers, list of wave files for

training (490 files), transcriptions for training files, a list of wave files for testing (210 files) and transcriptions for the testing files. The acoustic model had been trained and ready for use, and the remaining 30% of the recordings were fed to the recognizer in order to measure its accuracy. Moreover, some extra recordings from outside the collected corpus were also fed to the recognizer. The accuracy was calculated. The list of fillers or the filler dictionary is something different from a standard dictionary. It contains the filler phones that are phones that are not covered by a language model and include non-linguistic sounds like breath, inhalation, amah (mmm) or noise just as they may contain silences. Fig. 8 shows part of the filler dictionary.



<s>	SIL
</s>	SIL
<sil>	SIL
+mmm+	++MMM++
!INH	+INH+
!NOISE	+NOISE+

Fig. 8. Filler Phones List.

RESULTS AND DISCUSSION

For the purpose of evaluating the effectiveness of the proposed system, the system acoustic model was trained on 70% of the corpus using sphinx train utility, and the system was tested on the remaining 30%. The recognition rate (RR) was calculated based on Equation 3.

$$\mathrm{RR} = \frac{\text{Number of Correctly Recognized words}}{\text{Number of Tested Words}} \qquad (3)$$

In sentence-wise recognition, another equation was used to measure the Word Error Rate (WER). It compares a Reference (REF) to a Hypothesis (HYP) and is defined in equation 4 (Ali *et al.*, 2009).

$$WER=S+D+I/N \qquad (4)$$

Where,

- S denotes to number of substitutions,
- D denotes to number of deletions,
- I denotes to number of insertions and
- N denotes to number of words in the reference

In the training phase the Baum Welch algorithm iterates several times to find the best values for a HMM which represent the phoneme statistics. For each phoneme, there will be a HMM which has to be trained. As an outlet test, 6 volunteers were permitted to utter commands to the system 30 times with random selection between commands. The test results given by volunteers are illustrated in Table 5.

Table 5. Results of Voiced Commands

Speaker	Gender	Environment	Random Spoken Words	Number of Hits (Correctly (Recognized words	Accuracy
1	Male	Noisy	30	26	87%
2	Female	Noisy	30	29	97%
3	Male	Silent	30	27	90%
4	Female	Silent	30	29	97%
5	Male	Normal	30	29	97%
6	Female	Normal	30	28	93%
Average Accuracy Among All Speakers					**92%**
Average Error Rate Among All Speakers					**8%**

Since this experiment represents a real test, based on outside data, the result gained is considered very good in the field of ASR. In this experiment, the number of hits or correctly recognized words were considered. Moreover, though the error rate is generally more than 15% in such experiments it was less than 10% in our case. A graph to reflect the relation between factors of gender, environment, and accuracy is shown in Fig. 9. It indicates that this system could recognize female voices much better than male ones, especially in normal environment.

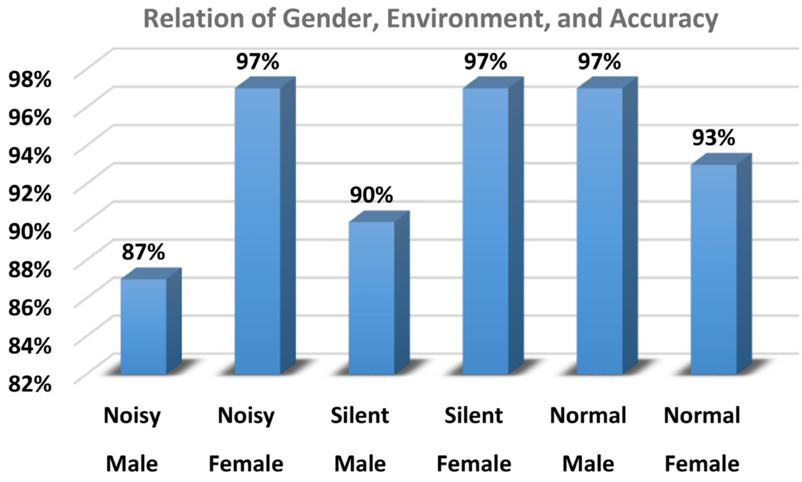


Fig. 9. Relationship of Gender, Environment, and Accuracy.

From Fig. 9, the effect of the surrounding environment on the accuracy is clearly shown. Naturally, the noisy environment produces less accuracy though it is still within an acceptable rate. Normal and silent environments have almost the same level of accuracy. The previous experiment was considered as a test for the recognizer. Since the corpus is divided into 70% for training and 30% for testing, respectively, the 30% of the corpus was fed to the recognizer. The word error rate and the accuracy obtained (i.e. percentage of correctness) are shown in Table 6. The percent of precision is calculated using equation 5. (Al-Qatab and Ainon, 2010)an Arabic dictionary was built by composing the words to its phones. Next, Mel Frequency Cepstral Coefficients (MFCC.

Table 6. Recognition accuracy from the Corpus

Total number of words in the 30%	Correctly recognized words	Correct percentage	WER (Phoneme Level)	Accuracy
126	120	96%	4%	96%

Percent of Correctly recognized words = (N-D-S)/N×100%, (5)
where N denotes to number of labels in the reference transcriptions, and D denotes to number of deletions which may occur when the recognition take place. The deleted items are a set of phonemes that are omitted, and S is the number of substitutes. The substituted items are the set of phonemes that are replaced based on the language model, when more than one phoneme is placed in the same position. However, this measure ignores insertion errors and therefore, the total percentage of accuracy as defined by equation 5 is used for better understanding of the results (Al-Qatab and Ainon, 2010) Percent of Accuracy
(N-D-S-I)/N×100%, (6)
Where *I*, is the number of insertions. The insertions occur when a new phoneme is inserted to complete the most probable word. It is similar to word insertion in equation 4, but at the phoneme level.

For the sake of fact and consistency, no recognizer can give you 100% of accuracy due to several factors such as noise, quality of the training recordings, pronunciation variations between speakers, and the quality of outlet data. In our case, WER was about 4%, which is acceptable and relatively low compared to others Table 6 where the WER exceeds 4%. Moreover, the average accuracy of the corpus-based experiment and the real test experiment reached 94%, which compares favorably to others. In addition, our system built upon a famous recognizer, the CMU-Sphinx4 recognizer and it can be adapted to work with any microcontrollers. It is fully independent of any hardware, while most of the work previously done was mainly built upon a targeted special hardware. Qidwai and Ibrahim, (2010) and Wang *et al.* (2015) had a limited set of commands with a maximum of 5, while the current work allow for 7 commands and these can be extended according to our needs.

CONCLUSION AND FUTURE WORK

In this paper, an automatic Arabic speech recognition system for a handicapped people to support and facilitate their daily life is proposed, designed, and developed. First, a corpus of 700 recording waves for 7 words from 20 different male and female speakers was created. Second, from the phone set, about 23 phones were used in this system. A phonetic dictionary which consists of 14 words (seven with discretization or "*tashkeel*" and seven without) were built for the system. Third, the files and transcription for the training using a CMU-Sphinx4and testing (490 files for training and 210 files for testing) were determined. Finally, the proposed system was developed using java programming language using the Sphinx-4 libraries to receive the spoken voice commands and convert the recognized speech to a text. The proposed system achieves an accuracy of 96% where most of the words of the different voices were recognized correctly. The average accuracy of the real experiment and corpus test is 94%.

One of the main limitations of our system, and most of the AASR systems, is its dependence on Modern Standard Arabic (MSA). MSA is usually used for broadcast news, TV reports, and other functions. Accents are not covered in ASR systems, as it is a sophisticated issue in speech recognition. However, some issues in this area can be resolved by building a corpus for each accent though this would be hard to achieve. In the future, it is planned to improve the accuracy of the proposed system as well as allowing it to support more Arabic words from different accents close to MSA to strengthen the corpus. A cost-effective solution will be explored to build a working prototype of a voice controlled wheelchair that will move according to its user's commands and which can be easily connected to our AASR. In fact, some visual simulations were initiated through MATLAB for this purpose.

REFERENCES

Abdou, S., Rashwan, M., Al-Barhamtoshy, H., Jambi, K., and Al-Jedaibi, W. 2012. Enhancing the confidence measure for an Arabic pronunciation verification system. Stockholm, Sweden, 6-8 June 2012, In Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (pp. 85-90).

Abdou, S., Rashwan, M., Al-Barhamtoshy, H., Jambi, K., and Al-Jedaibi, W. 2014. Speak correct: A computer aided pronunciation training system for native Arabic learners of English. Life Science Journal. 11(10): 370-380.

Abdul-Mageed, M., Diab, M., and Kübler, S. 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer Speech & Language. 28(1): 20-37.

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

Al-Barhamtoshy, H., Abdou, S., and Jambi, K. 2014. Pronunciation evaluation model for none native English speakers. Life Science Journal. 11(9): 216-226.

Ali, A., Dehak, N., Cardinal, P. Khurana, S., Yella, S.H., Glass, J., Bell, P., and Renals, S., 2015a. Automatic dialect detection in arabic broadcast speech. arXiv preprint arXiv:1509.06928.

Ali, A., Magdy, W., and Renals, S. 2015b. Multi-reference evaluation for dialectal speech recognition system: A study for Egyptian ASR. Beijing, China, 30 July 2015, *In* Proceedings of the Second Workshop on Arabic Natural Language Processing (pp. 118-126)

Ali, M., Elshafei, M., Al-Ghamdi, M., and Al-Muhtaseb, H. 2009. Arabic phonetic dictionaries for speech recognition. Journal of Information Technology Research (JITR). 2(4): 67-80.

Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. 2000. An architecture for a generic dialogue shell. Natural Language Engineering. 6(3-4): 213-228.

Al-Qatab, B.A., and Ainon, R.N. 2010. Arabic speech recognition using hidden Markov model toolkit (HTK). Kuala Lumpur, Malaysia, *In* 2010 International Symposium on Information Technology. May 2010,Kuala Lumpur, Malaysia. 2: 557-562.

Baum, L.E., and Petrie, T. 1966. Statistical inference for probabilistic functions of finite state Markov chains. The Annals of Mathematical Statistics37(6): 1554-1563.

Das, T.K., and Nahar, K.M. 2016. A voice identification system using hidden Markov model. Indian Journal of Science and Technology. 9(4): 1-6.

Diehl, F., Gales, M.J., Tomalin, M., and Woodland, P.C. 2012. Morphological decomposition in Arabic ASR systems. Computer Speech & Language. 26(4): 229-243.

Frihia, H., and Bahi, H., 2016. Embedded learning segmentation approach for Arabic speech recognition. Brno, Czech Republic, 12-16 Septemebr 2016, *In* International Conference on Text, Speech, and Dialogue (pp. 383-390). Springer, Cham.

Jurafsky, D. 2000. Speech & Language Processing. Pearson Education, India.

Kirchhoff, K., Bilmes, J., Henderson, J., Schwartz, R., Noamany, M., Schone, P., Ji, G., Das, S., Egan, M., He, F., and Vergyri, D. 2002. Novel speech recognition models for Arabic. Maryland, USA, 1 June 2002, *In* Johns-Hopkins University summer research workshop.

Majeed, S.A., Husain, H., Samad, S.A., and Idbeaa, T.F. 2015. Mel frequency cepstral coefficients (Mfcc) feature extraction enhancement in the application of speech recognition: A comparison study. Journal of Theoretical & Applied Information Technology. 79(1): 38-56.

Nahar, K.M., Shquier, M.A., Al-Khatib, W.G., Al-Muhtaseb, H., and Elshafei, M. 2016. Arabic phonemes recognition using hybrid LVQ/HMM model for continuous speech recognition. International Journal of Speech Technology. 19(3): 495-508.

Nishimori, M., Saitoh, T., and Konishi, R., 2007. Voice controlled intelligent wheelchair. Kagawa, Japan, 17-20 September 2007, *In* SICE Annual Conference 2007 (pp. 336-340).

Qidwai, U., and Ibrahim, F. 2010. Arabic speech-controlled wheelchair: A fuzzy scenario. Kuala Lumpur, Malaysia, 10-13 May 2010, *In* 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010) (pp. 153-156).

Rabiner, L.R., and Juang, B.H. 1986. An introduction to hidden Markov models. IEEE ASSP Magazine. 3(1): 4-16.

Rosenfeld, R. 2000. Two decades of statistical language modeling: Where do we go from here? Proceedings of the IEEE. 88(8): 1270-1278.

Satori, H., Harti, M., and Chenfour, N. 2007. Arabic speech recognition system based on CMUSphinx. Agadir, Morocco, 28-30 March 2007 *In* International Symposium on Computational Intelligence and Intelligent Informatics (pp. 31-35).

Wang, H., Li, T., Zheng, F., and Yan, Y. 2015. A wheelchair platform controlled by a multimodal interface. Shanghai, China, 24-26 April 2015 *In* 2nd International Conference on Information Science and Control Engineering (pp. 587-590).

Zarrouk, E., Ayed, Y.B., and Gargouri, F. 2014. Hybrid continuous speech recognition systems by HMM, MLP and SVM: A comparative study. International Journal of Speech Technology. 17(3): 223-233.

تحريك كرسيّ المُقعدين من خلال تمييز الأوامر العربية المتقطعة

خالد نهار[1] و معاوية الشناق[1] و رأفت الشرمان[1] و رعد الخطيب[1] و ”محمد أشرف“ العتوم[2]

(1) قسم علوم الحاسوب، كلية تكنولوجيا المعلومات وعلوم الحاسوب، جامعة اليرموك، المملكة الأردنية الهاشمية
(2) قسم نظم المعلومات الحاسوبية، كلية تكنولوجيا المعلومات وعلوم الحاسوب، جامعة اليرموك، المملكة الأردنية الهاشمية

استلام 20 ديسمبر 2018م - قبول 4 سبتمبر 2019م

https://doi.org/10.37575/b/com/2034

المستخلص

يعد تعرف الكلام آليا وسيلة فعالة وواسعة الانتشار في تحويل الصوت إلى أوامر، وفي العقود القليلة الماضية، ونتيجة لثورة المعالجة الرقمية للبيانات، أصبحت نظم معالجة الكلام مضمنة في معظم التكنولوجيا الحديثة. إن من فوائد هذه الثورة أنها دفعتنا باتجاه تطوير نظام تمييز الصوت المتقطع للتحكم في حركة كرسي المعوقين الناطقين بالعربية. سيكون بمقدور «نظام عبر الصوت» الخاص بكرسي المعوقين تمييز سبعة أوامر منفصلة في اللغة العربية. لقد تم استخدام نظام «سفينكس 4 للتمييز الأوتوماتيكي للكلام من جامعة كارغيني ميلون الأمريكية» لبناء النظام والتحقق من فعاليته ودقته. لقد تم بناء مستوعب صوتي خصيصا لهذا الغرض وتم تقسيمه إلى جزأين: الجزء الأول خاص بالتدريب وبلغ 70 % من المستوعب، والجزء الثاني لأغراض التحقق من صحة النظام وحساب الدقة وبلغ 30 %. بعد تنفيذ الاختبارات اللازمة تم التوصل إلى دقة تمييز للأوامر تصل لغاية 96 %. إن مستوى معدل الدقة في تمييز الأوامر بين ما تم إجراؤه من اختبارات على 30 % من المستوعب الصوتي وبين بيانات حقيقيه أخذت من خارج المستوعب بلغ 94 %.

الكلمات المفتاحية: تمييز الكلام العربي، كرسي متحرك، معوق، النموذج الصوتي، نموذج ماركوف الخفي، نموذج اللغة.