Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus

Mohammad Abushariah[1,2], Raja Ainon[1], Roziati Zainuddin[1], Moustafa Elshafei[3], and Othman Khalifa[4]
[1]Faculty of Computer Science and Information of Technology, University of Malaya, Malaysia
[2]King Abdullah II School for Information Technology, University of Jordan, Jordan
[3]Department of Systems Engineering, King Fahd University of Petroleum and Minerals, Saudi Arabia
[4]Faculty of Engineering, International Islamic University Malaysia, Malaysia

Abstract: *This paper describes and proposes an efficient and effective framework for the design and development of a speaker-independent continuous automatic Arabic speech recognition system based on a phonetically rich and balanced speech corpus. The speech corpus contains a total of 415 sentences recorded by 40 (20 male and 20 female) Arabic native speakers from 11 different Arab countries representing the three major regions (Levant, Gulf, and Africa) in the Arab world. The proposed Arabic speech recognition system is based on the Carnegie Mellon University (CMU) Sphinx tools, and the Cambridge HTK tools were also used at some testing stages. The speech engine uses 3-emitting state Hidden Markov Models (HMM) for tri-phone based acoustic models. Based on experimental analysis of about 7 hours of training speech data, the acoustic model is best using continuous observation's probability model of 16 Gaussian mixture distributions and the state distributions were tied to 500 senones. The language model contains both bi-grams and tri-grams. For similar speakers with different sentences, the system obtained a word recognition accuracy of 92.67% and 93.88% and a Word Error Rate (WER) of 11.27% and 10.07% with and without diacritical marks, respectively. For different speakers with similar sentences, the system obtained a word recognition accuracy of 95.92% and 96.29%, and a WER of 5.78%, and 5.45% with and without diacritical marks, respectively. Whereas different speakers and different sentences, the system obtained a word recognition accuracy of 89.08% and 90.23%, and a WER of 15.59% and 14.44% with and without diacritical marks, respectively.*

Keywords: *Arabic automatic speech recognition, arabic speech corpus, phonetically rich and balanced, acoustic model, statistical language model.*

Received December 22, 2009; accepted May 20, 2010

1. Introduction

Arabic, a Semitic language and one of the six official languages of the United Nations (UN), is one of the most widely spoken languages in the world. Statistics show that it is the first language (mother-tongue) of 206 million native speakers ranked as fourth after Mandarin, Spanish and English [15]. In spite of its importance, research effort on Arabic Automatic Speech Recognition (ASR) is unfortunately still inadequate.

Modern Standard Arabic (MSA) is the formal linguistic standard of Arabic language, which is widely taught in schools and universities, and used in the office and the media. It has been the focus and the core interest of many previous and recent researches compared to dialectal Arabic [4, 17, 18, 28, 29, 30].

Lack of spoken and written training data is one of the main issues encountered by Arabic ASR researchers. A list of most popular (from 1986 through 2005) corpora is provided [7] showing only 19 corpora (14 written, 2 spoken, 1 written and spoken, and 2 conversational). These corpora are not readily available to the public and many of them can only be obtained by purchasing from the Linguistic Data Consortium (LDC) or the European Language Resource Association (ELRA). It is clear that there is a shortage of spoken data as compared to written data resulting in a great need for more speech corpora in order to serve different domains of Arabic ASR. The available spoken corpora were mainly collected from broadcast news (radios and televisions), and telephone conversations. This kind of spoken data may not necessarily serve quality Arabic ASR research, because of the quality of the spoken data itself in terms of recording attributes and parameters used (e.g., sampling rate). They are also limited to certain applications and domains. The coverage of any corpora cannot contain complete information about all aspects of language lexicon and grammar [1], due to the limited written training data and therefore inadequate spoken training data. In addition, a clear strong relationship between written and spoken forms needs to be clarified.

Writing is claimed to be more structurally complex and elaborate, more explicit, more organized and planned than speech [23]. These differences generally lead to the approach that the written form of the corpora needs to be created carefully before producing and recording the spoken form. Therefore, linguists and phoneticians carefully produce written corpora before handing them to speech recording specialists. This can also be seen throughout the past few years, where a number of phonetically rich and/or balanced corpora for many languages have been produced. Many ASR researches are now based on phonetically rich and/or balanced corpora, e.g., English [9, 12, 14], Mandarin [10], Japanese [26], Indonesian [20], Korean [16], Cantonese [19], Hindi [11], Turkish [27], Urdu [25] and many others obtaining comparatively competitive results.

As far as Arabic language is concerned, automatic speech recognition tasks mainly addressed Arabic digits, broadcast news, command and control, The Holy Qur'an, and Arabic proverbs researches. They explored various state-of-the-art techniques and tools for Arabic speech recognition. Arabic digits recognition systems were implemented in [5, 6, 17, 28]. The system in [17] used Carnegie Mellon University's (CMU) Sphinx-IV engine based on Hidden Markov Models (HMM), which obtained a word recognition rate of 99.21% for about 35 minutes of training speech data and 7 minutes of testing speech data. The system in [28] was also using CMU Sphinx-IV engine based on HMM for the same task and obtained a word recognition rate of 85.56% for male speakers and 83.34% for female speakers. The comparative study in [5] shows two versions of Arabic digits recognition system based on Artificial Neural Networks (ANN) and HMM. For speaker-dependent mode, ANN version of the system achieved correct digit recognition rate of 99.5% whereas HMM version of the system achieved 98.1% correct digit recognition rate. However, the same versions of the system were tested for speaker-independency showing correct digit recognition rates of 94.5% and 94.8% based on ANN and HMM, respectively.

In [6], a different kind of speech data is presented for Arabic digits recognition system using telephony Saudi accented Arabic corpus. The system used Cambridge HTK tools based on HMMs and reported correct digit recognition rate of 93.67%. In addition, The Holy Qur'an was also considered for Arabic speech recognition in [17, 21]. The system in [17] used Sphinx-IV engine based on HMMs and obtained a word recognition rate of 70.81% and a WER of 40.18% for corpus of 18.35 hours. However, the system in [21] used HTK tools based on HMMs and achieved a word recognition rate ranging from 68% to 85% with an average word recognition rate of 78.6%. On the other hand, Arabic speech recognition system using broadcast news corpus was developed in [4]. The system was trained using about 7 hours of speech using Sphinx 3 tools based on HMMs and tested using 400 utterances adding to about half an hour of speech. The system obtained a correct word recognition rate of 90.78% and a WER of 10.87% with full diacritical marks, whereas it obtained a correct word recognition rate of 93.04% and a WER of 8.61% without diacritical marks.

Other Arabic automatic speech recognition systems were developed for different tasks such as in [8, 13, 17, 22]. A command and control system covering approximately 30 words was developed in [17] using Sphinx-IV engine based on HMMs and obtained a word recognition rate of 98.18%, whereas an Arabic speech recognition system using Recurrent Neural Networks (RNN) [13] was developed for recognizing six isolated words obtaining an average recognition rate of 95.58%, and an Arabic ASR system to recognize 16 sentences of Egyptian proverbs was developed [8] based on HMMs using HTK and obtained a word recognition rates of 56.8%, 66.65%, and 81.79% for Mono-phone, Tri-phone, and Syllable based recognition respectively. In addition, an acoustic training system for speaker independent continuous Arabic speech recognition system based on HMMs and different language models (bigram and context free grammar) using HTK was developed [22]. The system was designed to investigate the WER as a function of vocabulary size for different types of language models. For bigram based language model, the system obtained a WER of 5.26% and 2.72% for 1340 and 306 words respectively. Whereas for context free grammar, the system obtained a WER of 0.19% and 0.99% for 1340 and 306 words, respectively

From literature investigation, there is no published Arabic ASR research work yet based on phonetically rich and/or balanced training data and it is important to explore this approach in order to find out its efficiency and effectiveness for Arabic ASR research. The technical report in [2] is one of the earliest works on producing written Arabic training data based on phonetically rich and balanced sentences. This technical report was submitted to King Abdulaziz City of Science and Technology (KACST) in Saudi Arabia as the final deliverable of the project "Database of Arabic Sounds: Sentences". This written training data was created by experts from KACST and consists of 367 sentences written using 663 phonetically rich words.

KACST written (text) training data was used as the baseline for creating our phonetically rich and balanced speech corpus. Another 48 written sentences were created for testing purposes, which include common Arabic proverbs. Therefore, this work is the first for Arabic ASR based on high quality phonetically rich and balanced written and spoken corpora.

The paper is organized as follows. Section 2, describes KACST text corpus and our phonetically rich and balanced speech corpus and their statistical analysis. Speech data preparation and analysis is presented in section 3. Section 4 describes the speech corpus testing and evaluation for Arabic ASR systems. Conclusions are finally presented in section 5.

2. Statistical Analysis and Description of the Text and Speech Corpora

In order to produce a robust speaker-independent continuous automatic Arabic speech recognizer, a set of speech recordings that are rich and balanced is required. The rich characteristic is in the sense that it must contain all the phonemes of Arabic language. It must be balanced in preserving the phonetics distribution of Arabic language too. This set of speech recordings must be based on a proper written set of sentences and phrases created by experts. Therefore, it is crucial to create a high quality written (text) set of the sentences and phrases before recording them.

2.1. Phonetically Rich and Balanced Text Corpus

Creating phonetically rich and balanced text corpus requires selecting a set of phonetically rich words, which are combined together to produce sentences and phrases. These sentences and phrases are verified and checked for balanced phonetic distribution. Some of these sentences and phrases might be deleted and/or replaced by others in order to achieve an adequate phonetic distribution [24].

In 1997, KACST created a database for Arabic language sounds. The purpose of this work was to create the least number of phonetically rich Arabic words. As a result, a list of 663 phonetically rich words containing all Arabic phonemes, which are subject to all Arabic phonotactic rules was produced. This work is the backbone for creating individual sentences and phrases, which can be used for Arabic ASR and text-to-speech synthesis applications. The list of 663 phonetically rich words was created based on the following characteristics and guidelines [3]:

1. Cover all Arabic phonemes which must be balanced so as to be close in frequency as possible.
2. Contain all phonotactic rules of Arabic. This means coverage of all Arabic phoneme clusters.
3. The presence of the least possible number of words so that the list does not contain a single word whose goal of existence is achieved by another word in the same list.
4. To be of words in circulation and use as far as possible.

Statistical analysis of the 663 phonetically rich words show that all Arabic phonemes exist in this list as indicated in Table 1, which also shows the number of repetitions as well as the percentage for each Arabic phoneme in the KACST phonetically rich words database in an alphabetical order.

Table 1. Arabic phonemes repetitions for the 663 phonetically rich words.

Arabic Alphabets & Vowels	Phoneme Repetitions			Total	Percentage (100%)
	Front	Inside	End		
ء	76	38	18	132	3.95
ب	27	45	32	104	3.11
ت	21	30	19	70	2.09
ث	4	30	13	47	1.40
ج	9	39	12	60	1.79
ح	29	39	16	84	2.51
خ	16	36	14	66	1.97
د	6	38	19	63	1.88
ذ	5	37	6	48	1.43
ر	36	46	53	135	4.04
ز	4	34	10	48	1.43
س	28	29	17	74	2.21
ش	11	32	18	61	1.82
ص	10	27	13	50	1.49
ض	11	31	10	52	1.55
ط	11	28	18	57	1.70
ظ	6	25	5	36	1.07
ع	35	34	20	89	2.66
غ	11	34	6	51	1.52
ف	27	46	24	97	2.90
ق	25	36	18	79	2.36
ك	14	41	12	67	2.00
ل	25	37	48	110	3.29
م	77	36	53	166	4.97
ن	40	41	39	120	3.59
هـ	3	44	50	97	2.90
و	21	47	14	82	2.45
ي	74	45	17	136	4.07
ـَ	0	597	12	609	18.26
ـَـَ	0	74	21	95	2.84
ـُ	0	124	11	135	4.04
ـُـُ	0	46	5	51	1.52
ـِ	0	115	0	115	3.44
ـِـِ	0	29	19	48	1.43
Total Repetitions for all Arabic Phonemes				3,334	100%

In 2003, KACST produced a technical report of the project “Database of Arabic Sounds: Sentences”. Arabic independent sentences have been written using the said 663 phonetically rich words. The database consists of 367 sentences; 2 to 9 words per sentence. Therefore, this work aims to produce Arabic phrases and sentences that are phonetically rich and balanced based on the previously created list of 663 phonetically rich words, which were put in phrases and sentences while taking into consideration the following goals [2]:

- To have the minimum word repetitions as far as possible.

- To have an average of 2 to 9 words in a single sentence.
- To have structurally simple sentences in order to ease readability and pronunciation.
- To have as far as possible maximum number of rich and balanced words in a single sentence.
- To have the minimum number of sentences.

Table 2 shows the number of repetitions as well as the percentage for each Arabic phoneme in the KACST sentences database in an ascending order.

Table 2. Arabic phoneme repetitions for the 367 sentences.

Phonemes & Graphemes	**Phoneme Repetitions**	**Percentage (100%)**
ؤ	8	0.06
آ	11	0.09
ئ	19	0.15
ـً	35	0.27
ء	45	0.35
ظ	54	0.42
إ	58	0.45
ـٍ	59	0.46
غ	75	0.58
ز	78	0.60
ث	80	0.62
ى	81	0.63
ـٌ	92	0.71
ض	100	0.77
خ	103	0.80
ذ	107	0.83
ط	110	0.85
ص	112	0.87
ة	120	0.93
ش	132	1.02
ج	164	1.27
ك	164	1.27
ق	176	1.36
س	186	1.44
ح	194	1.50
أ	201	1.55
ت	205	1.58
د	220	1.70
ه	247	1.91
ف	256	1.98
ع	280	2.16
و	294	2.27
ب	368	2.84
ر	391	3.02
ن	426	3.29
م	478	3.69
ي	555	4.29
ـُ	909	7.02
ا	1031	7.97
ل	1035	8.00
ـِ	1344	10.39
ـَ	2337	18.06
Total Repetitions	12,940	100%

As a result, a list of fully diacritical 367 phonetically rich and balanced sentences was produced using 1835 Arabic words. An average of 2 phonetically rich words and 5 other words were used in each single sentence. Statistical analysis shows that 1333 words were repeated once only and 99 words were repeated more than once in the entire 367 sentences, whereas 17 words were repeated 5 times and more only. The word (فِي) which means (*IN*) in English language was repeated 65 times and that is the maximum repetition of words.

KACST 367 phonetically rich and balanced sentences are used for training purposes in our system, whereas a set of 48 additional sentences is created for testing purposes. Therefore, our text corpus contains two subsets of text data, the first is used for training purposes and the second is used for testing purposes. This is also reflected on the speech corpus, which is explained further in subsection (2.2).

2.2. Phonetically Rich and Balanced Speech Corpus

Speech corpus is an important requirement for developing any ASR system. The developed corpus contains recordings of 415 Arabic sentences. 367 written phonetically rich and balanced sentences were developed by KACST [2], and were recorded and used for training the acoustic model. For testing the acoustic model, 48 additional sentences representing Arabic proverbs were created by an Arabic language specialist. The speech corpus was recorded by 40 (20 male and 20 female) Arabic native speakers from 11 different Arab countries representing three major regions (Levant, Gulf, and Africa) in the Arab world. This speech corpus was recorded in a sound-proof studio using Sound Forge 8.0 software and took nearly three months to complete starting from March 2009 until June 2009.

This speech corpus was enriched with varieties of Arabic native speakers taking into consideration the following characteristics representing:

- Different age categories.
- Different nationalities.
- Different Arab regions.
- Different professions.
- Different academic qualifications.
- Different mastery of Arabic.

Since this speech corpus contains training and testing written and spoken data of a variety of Arabic native speakers who represent different genders, age categories, nationalities, regions, and professions, and is also based on phonetically rich and balanced sentences, it is expected to be used for development of many Arabic speech and text based applications, such as speaker dependent and independent ASR, Text-To-Speech (TTS) synthesis, speaker recognition, and many others. Table 3 shows important speakers' details, whereas Table 4 shows more technical details of our speech corpus.

This work adds a new kind of possible speech data for Arabic language based text and speech applications besides other kinds such as broadcast news and telephone conversations. Therefore, this work is an invitation to all Arabic ASR developers and research groups to utilize and capitalize.

Table 3. Speakers' genders, nationalities, and regions.

Region	Country	Gender		Total	Total/ Region
		M	F		
Levant	Jordan	8	4	12	15
	Palestine	2	-	2	
	Syria	1	-	1	
Gulf	Iraq	-	4	4	11
	Saudi Arabia	-	3	3	
	Yemen	-	3	3	
	Oman	-	1	1	
Africa	Sudan	3	3	6	14
	Algeria	3	2	5	
	Egypt	2	-	2	
	Morocco	1	-	1	
Total		20	20	40	40
Total (%)		50%	50%	100%	100%

Table 4. Speech corpus technical details.

Criteria	Training Sentences	Testing Sentences
No. of Sentences	367 sentences	48 sentences
Number of Words	1835 words	275 words
Average No. of Words/Sentence	5 words	5 words
Min. and Max. No. of Words/Sentence	Min. of 2 and Max. of 9	Min. of 3 and Max. of 10
No. of Phonetically Rich Words	663 words	-
Average No. of Phonetically Rich Words/Sentence	2 words	-
No. of Speakers	40 Speakers	40 Speakers
Speakers' Age	18 to 66 years	18 to 66 years
Speakers' Gender	20 Males and 20 Females	20 Males and 20 Females
Average Recording Time/Speaker	2 Hours	45 Minutes
Average No. of Sound Files/ Sentence	100 sound files/sentence	100 sound files/sentence
Sampling Rate (Hz)	44100Hz	44100Hz
No. of Bits	16 bits	16 bits
No. of Channels	2 channels	2 channels
Size of Raw Speech	65 GB	5 GB

3. Speech Data Preparation and Analysis

This section covers all preparation and pre-processing steps we developed in order to produce ready to use speech data, which are used later for training the acoustic model.

3.1. Automatic Arabic Speech Segmentation

During the recording sessions, speakers were asked to utter the 415 sentences sequentially starting with training sentences followed by testing sentences. Recordings for a single speaker were saved into one ".wav" file and sometimes up to four ".wav" files depending on number of sessions the speaker spent to finish recording the 415 sentences. It is time consuming to save every single recording once uttered. Therefore, there was a need to segment these bigger ".wav" files into smaller ones each having a single recording of a single sentence.

We developed a Matlab program that has two functions. The first function "read.m" reads the original bigger ".wav" files, identifies the starting and ending points for each sentence utterance, generates a text "segments.txt" file that automatically assigns a name for each utterance and concatenates the name with the corresponding starting and ending points. Whereas the second function "segment.m" reads the automatically generated text file "segments.txt" and compares it with the original bigger ".wav" file, it then segments the bigger ".wav" file based on starting and ending points read from "segments.txt" into smaller ".wav" files carrying the same name as identified in "segments.txt". All those smaller ".wav" files are then saved into a single directory.

It is worth mentioning that the developed Matlab program considers silence as the main factor for segmentation. Some speakers used to record slower than others; therefore, the silence allowed variable was fixed on an individual basis. However, the silence allowed variable for a majority of speakers was fixed to half a second.

3.2. Manual Classification and Validation of Correct Speech Data

The automatic Arabic speech segmentation explained in subsection (3.1) outputs all possible ".wav" files in a single directory. Therefore, a human manual classification of those ".wav" files into the corresponding sentence directories was done.

Wrongly pronounced utterances were ignored at this stage. As a result, only correct utterances are considered for further pre-processing steps.

3.3. Parameters Conversion of Speech Data

Initial recording parameters shown in Figure 1 were used to fit different research needs. However, a Matlab program was developed to re-sample the sampling rate from 44100Hz into 16000Hz and to convert number of channels from 2 into 1 as shown in Figure 2, which are used in most ASR researches.

Sampling Rate = 16000Hz
No. of Channels = 1 (Mono)
No. of Bits = 16 bits
File Format = ".wav"

Figure 1. Original parameters.

Sampling Rate = 44100Hz
No. of Channels = 2 (Stereo)
No. of Bits = 16 bits
File Format = ".wav"

Figure 2. Modified parameters.

3.4. Directory Structure, Sound Filenames Convention and Automatic Generation of Training and Testing Transcription Files

Each speaker has a single folder that contains three sub-folders namely, “Training Sentences”, “Testing Sentences”, and “Others”. “Training Sentences” sub-folder contains 367 sub-folders representing the 367 training sentences, whereas “Testing Sentences” sub-folder contains 48 sub-folders representing the 48 testing sentences. The sub-folder “Others” contains out of content utterances for each speaker. Each sentence sub-folder contains two other sub-folders namely “Correct” and “Wrong”. Utterances classified under the sub-folder “Correct” are the ones used for further pre-processing steps.

A Matlab program was developed in order to read the correctly classified utterances from all speakers and assigns them unique filenames. It also separates training utterances from testing utterances by producing two main folders namely “Training” and “Testing”. The “Training” folder contains all correctly classified utterances for the 367 training sentences for all speakers, whereas the “Testing” folder contains all correctly classified utterances for the 48 testing sentences for all speakers with unique filenames. Filenames follow the following formats: SpeakerID_SentenceType_SentenceNo_SequenceNo.

This Matlab progam also produces two corresponding transcription files associated with the utterance file_ID namely “Training.transcription” and “Testing.transcription” for all utterances produced in the two output folders. It also outputs two file_IDs files namely “Training.fileids” and “Testing.fileids”.

4. Speech Corpus Testing and Evaluation for Arabic ASR Systems

This section covers three major requirements for Arabic ASR namely, Arabic phonetic dictionary, the acoustic model training, the language model training. An evaluation of the Arabic ASR is finally presented.

4.1. Arabic Phonetic Dictionary

A rule-based approach to automatically generate a phonetic dictionary for a given transcription was used [4]. The transcription file contains 2,088 words and the vocabulary list contains 1,626 unique words. The number of pronunciations in the developed phonetic dictionary is 2,482 entries. Figure 3 shows a sample of the generated phonetic dictionary.

آلَامُ E AE: L AE: M UH
آمِنِ E AE: M IH N IH N
آيَاتُ E AE: Y AE: T UH
أَبَدَ E AE B AE D AE
أَبِي E AE B IY
أَبْجَلَنِي E AE B JH AE L AE N IY
أَبْطَأَ E AE B TT AH E AE
أَبْلَجُ E AE B L AE JH UH

Figure 3. Sample of the phonetic dictionary.

4.2. Acoustic Model Training

Training the acoustic model using CMU Sphinx 3 tools requires successfully passing through three phases shown in Figure 4. The engine uses 3-emitting state HMM for tri-phone acoustic models.

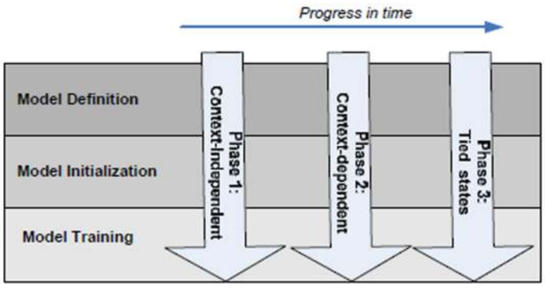


Figure 4. Acoustic model training phases for Sphinx 3 [4].

Baum-Welch re-estimation algorithm is used during the first phase in order to estimate the transition probabilities of the Context-Independent (CI) HMMs. Arabic basic sounds are classified into phonemes or phones as shown in Table 5. In this work, 44 (including silence) Arabic phonemes and phones are used. During the second phase, Arabic phonemes and phones are further refined into Context-Dependent (CD) tri-phones. The HMM model is now built for each tri-phone, where it has a separate model for each left and right context for each phoneme and phone. As a result of the second phase, tri-phones are added to the HMM set. In the Tied-States phase, the number of distributions is reduced through combining similar state distributions [4].

There are 4,705 unique tri-phones extracted from the training transcripts. The minimum occurrence of tri-phones was 18 times for (AH: and IX:) whereas the maximum was 456 times for (AE).

Acoustic model training was divided into two stages. During the first stage, one of the eight training data sets was used in order to identify the best combination of Gaussian mixture distributions and number of senones. The acoustic model is trained using continuous state probability density ranging from 2 to 64 Gaussian mixture distributions. In addition, the state distributions were tied to different number of senones ranging from 350 to 2500. A total of 54 experiments were done at this stage producing different results as shown in sub-section (4.4). During the second stage, the best combination of Gaussian mixture distributions and number of senones was used to train the other seven out of eight training data sets.

4.3. Language Model Training

Language model is another important requirement for any ASR system. Creation of a language model consists of computing the word uni-gram counts, which are then converted into a task vocabulary with word frequencies, generating the bi-grams and tri-grams from the training text based on this vocabulary, and finally converting the n-grams into a binary format language model and standard ARPA format [4]. The CMU-Cambridge statistical language modeling toolkit is used. The number of uni-grams is 1,627, whereas the number of bi-grams and tri-grams are 2,083 and 2,085 respectively.

4.4. Testing the Arabic ASR System

This work is based on 8,043 utterances gathered from 8 (5 male and 3 female) Arabic native speakers resulting about 8 hours of speech data. In order to show a fair testing and evaluation of the Arabic ASR performance, the round robin testing approach was applied, where every round speech data of 7 out of 8 speakers are trained and speech data of the 8th are tested. This is also important to show how speaker-independent the system. As a result, 8 different data sets were used as shown in Table 5.

Table 5. Training and testing data sets for each experiment.

Exp. ID	Train. Data	Test. Data	Test. Data		Total Test. Data	Ratio of Test. Data (%)
		Same Speakers	Diff. Speakers			
		Diff. Sent.	Same Sent.	Diff. Sent.		
Exp.1	6379	906	678	80	1664	20.69
Exp.2	6288	871	769	115	1755	21.82
Exp.3	5569	755	1488	231	2474	30.76
Exp.4	6308	888	749	98	1735	21.57
Exp.5	6296	889	761	97	1747	21.72
Exp.6	6331	891	726	95	1712	21.29
Exp.7	6219	861	838	125	1824	22.68
Exp.8	6009	841	1048	145	2034	25.29

During the first stage of training the acoustic model, the first data set in Exp.1 was used to identify best combination of Gaussian mixture distributions and number of senones. It is found that 16 Gaussians with 500 senones obtained the best word recognition rate of 93.24%. Therefore, this combination was used for training the acoustic model in Exp.2 through Exp.8 data sets.

Tables 6 and 7 show the word recognition rates (%) and the WER with and without diacritical marks, respectively.

Table 6. Arabic ASR performance with full diacritical marks.

Exp. ID	Same Speakers with Diff. Sent.		Diff. Speaker with Same Sent.		Diff. Speaker with Diff. Sent.	
	Rec. Rate %)	WER	Rec. Rate (%)	WER	Rec. Rate (%)	WER
Exp.1	93.24	10.73	94.98	6.28	90.11	13.48
Exp.2	91.80	11.96	93.30	10.62	83.00	27.87
Exp.3	93.07	10.53	97.22	3.66	89.81	14.94
Exp.4	92.72	11.42	96.89	4.16	91.44	11.76
Exp.5	93.43	10.09	94.92	7.13	89.49	14.86
Exp.6	92.61	11.56	95.55	7.37	90.64	14.23
Exp.7	92.65	11.15	96.37	4.51	88.15	14.25
Exp.8	91.85	12.75	98.10	2.51	89.99	13.31
Average Results	92.67	11.27	95.92	5.78	89.08	15.59

Table 7. Arabic ASR performance without diacritical marks.

Exp. ID	Same Speakers with Diff. Sent.		Diff. Speaker with Same Sent.		Diff. Speaker with Diff. Sent.	
	Rec. Rate %)	WER	Rec. Rate (%)	WER	Rec. Rate (%)	WER
Exp.1	94.41	9.57	95.22	6.04	90.79	12.81
Exp.2	93.02	10.74	93.95	10.33	84.38	26.49
Exp.3	94.29	9.31	97.42	3.46	90.88	13.87
Exp.4	93.86	10.29	97.33	3.73	92.87	10.34
Exp.5	94.57	8.95	95.32	6.73	90.76	13.59
Exp.6	93.75	10.41	95.91	7.00	91.39	13.48
Exp.7	94.06	9.74	96.68	4.20	89.42	12.98
Exp.8	93.04	11.56	98.50	2.11	91.33	11.97
Average Results	93.88	10.07	96.29	5.45	90.23	14.44

4.5. Evaluation of the Arabic ASR System

Based on our previous experiments of 4.07 hours of training speech data, the acoustic model was based on 16 Gaussian mixture distributions and the state distributions were tied to 400 senones. The system obtained 91.23% and 92.54% word recognition accuracy with and without diacritical marks respectively. However, in this work the training speech data was about 7 hours and the best combination was 16 Gaussian mixture distributions with 500 senones obtaining 93.43% and 94.57% word recognition accuracy with and without diacritical marks respectively. Therefore, the number of senones increases when there is an increase in training speech data, and it is expected to increase further when our speech corpus is fully utilized.

Speaker independency is clearly realized in this work as testing was conducted to assure this aspect. For different speakers with similar sentences, the system obtained a word recognition accuracy of 95.92% and 96.29%, and a Word Error Rate (WER) of 5.78% and 5.45% with and without diacritical marks respectively. On the other hand, for different speakers with different sentences, the system obtained a word recognition accuracy of 89.08% and 90.23%, and a WER of 15.59% and 14.44% with and without diacritical marks, respectively.

5. Conclusions

This paper reports our work towards developing a high performance Arabic ASR system based on phonetically rich and balanced speech corpus. This work includes creating the phonetically rich and balanced speech corpus with full diacritical marks transcription, build an Arabic phonetic dictionary, and an Arabic statistical language model. Recognition results show that this Arabic ASR system is speaker-independent and is highly comparable or better than many reported Arabic recognition results.

References

[1] Alansary S., Nagi M., and Adly N., "Building an International Corpus of Arabic Progress of Compilation Stage," *in Proceedings of 8th International Conference on Language Engineering*, Egypt, pp. 337-344, 2007.

[2] Alghamdi M., Alhamid A., and Aldasuqi M., "Database of Arabic Sounds: Sentences," *Technical Report*, Saudi Arabia, 2003.

[3] Alghamdi M., Basalamah M., Seeni M., and Husain A., "Database of Arabic Sounds: Words," *in Proceedings of the 15th National Computer Conference*, Saudi Arabia, pp. 797-815, 1997.

[4] Alghamdi M., Elshafei M., and Al-Muhtaseb H., "Arabic Broadcast News Transcription System," *International Computer Journal of Speech Technology*, vol. 10, no. 4, pp. 183-195, 2009.

[5] Alotaibi Y., "Comparative Study of ANN and HMM to Arabic Digits Recognition Systems," *Journal of King Abdulaziz University: Engineering Sciences*, vol. 19, no. 1, pp. 43-59, 2008.

[6] Alotaibi Y., Alghamdi M., and Alotaiby F., "Using a Telephony Saudi Accented Arabic Corpus in Automatic Recognition of Spoken Arabic Digits," *in Proceedings of 4th International Symposium on Image/Video Communications over Fixed and Mobile Networks*, Spain, pp. 43-60, 2008.

[7] Alsulaiti L. and Atwell E., "The Design of a Corpus of Contemporary Arabic," *International Computer Journal of Corpus Linguistics*, John Benjamins Publishing Company, vol. 11, no. 2, pp. 135-171, 2006.

[8] Azmi M. and Tolba H., "Syllable-Based Automatic Arabic Speech Recognition in Different Conditions of Noise," *IEEE Proceedings of the 9th International Conference on Signal Processing*, China, pp. 601-604, 2008.

[9] Black A. and Tokuda K., "The Blizzard Challenge Evaluating Corpus-Based Speech Synthesis on Common Datasets," *in Proceeding of Interspeech*, Portugal, pp. 77-80, 2005.

[10] Chou F. and Tseng C., "The Design of Prosodically Oriented Mandarin Speech Database," *in Proceedings of International Congress of Phonetics Sciences*, San Francisco, pp. 2375-2377, 1999.

[11] Chourasia V., Samudravijaya K., and Chandwani M., "Phonetically Rich Hindi Sentence Corpus for Creation of Speech Database," *in Proceedings of International Symposium on Speech Technology and Processing Systems and Oriental*, Indonesia, pp. 132-137, 2005.

[12] D'Arcy S. and Russell M., "Experiments with the ABI (Accents of the British Isles) Speech Corpus," *in Proceedings of Interspeech 08*, Australia, pp. 293-296, 2008.

[13] El Choubassi M., El Khoury H., Alagha J., Skaf J., and Al-Alaoui M., "Arabic Speech Recognition Using Recurrent Neural Networks," *in Proceedings of 3rd IEEE International Symposium on Signal Processing and Information Technology*, Germany, pp. 543-547, 2003.

[14] Garofolo J., Lamel L., Fisher W., Fiscus J., Pallett D., Dahlgren N., and Zue V., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *Technical Document*, Trustees of the University of Pennsylvania, Philadelphia, 1993.

[15] Gordon R., *Ethnologue: Languages of the World*, Texas: Dallas, SIL International, 2005.

[16] Hong H., Kim S., and Chung M., "Effects of Allophones on the Performance of Korean Speech Recognition," *in Proceedings of Interspeech*, Australia, pp. 2410-2413, 2008.

[17] Hyassat H. and Abu Zitar R., "Arabic Speech Recognition Using SPHINX Engine," *International Computer Journal of Speech Technology*, vol. 9, no. 3-4, pp. 133-150, 2008.

[18] Kirchhoff K., Bilmes J., Das S., Duta N., Egan M., Ji G., He F., Henderson J., Liu D., Noamany M., Schone P., Schwartz R., and Vergyri D., "Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns-Hopkins Summer Workshop," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal processing*, Hong Kong, vol. 1, pp. 344-347, 2003.

[19] Lee T., Lo W., Ching P., and Meng H., "Spoken Language Resources for Cantonese Speech Processing," *Computer Journal of Speech Communication*, vol. 36, no. 3-4, 327-342, 2002.

[20] Lestari D., Iwano K., and Furui S., "A Large Vocabulary Continuous Speech Recognition System for Indonesian Language," *in Proceedings of 15th Indonesian Scientific Conference*, Japan, pp. 17-22, 2006.

[21] Mourtaga E., Sharieh A., and Abdallah M., "Speaker Independent Quranic Recognizer Based on Maximum Likelihood Linear Regression," *in Proceedings of World Academy of Science, Engineering and Technology*, Brazil, pp. 61-67, 2007.

[22] Nofal M., Abdel-Raheem E., El Henawy H., and Abdel Kader N., "Acoustic Training System for

Speaker Independent Continuous Arabic Speech Recognition System," *in Proceedings of the 4th IEEE International Symposium on Signal Processing and Information Technology*, Italy, pp. 200-203, 2004.

[23] Parkinson D. and Farwaneh S., *Perspectives on Arabic Linguistics XV*, John Benjamins Publishing Company, Philadelphia, 2003.

[24] Pineda L., Gómez M., Vaufreydaz D., and Serignat J., "Experiments on the Construction of a Phonetically Balanced Corpus from the Web," *in Proceedings of 5th International Conference on Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, Korea, pp. 416-419, 2004.

[25] Raza A., Hussain S., Sarfraz H., Ullah I., and Sarfraz Z., "Design and Development of Phonetically Rich Urdu Speech Corpus," *in Proceedings of IEEE Oriental COCOSDA International Conference on Speech Database and Assessments*, Urumqi, pp. 38-43, 2009.

[26] Sagisaka Y., Takeda K., Abel M., Katagiri S., Umeda T., and Kuwabara H., "A Large-Scale Japanese Speech Database," *in Proceedings of International Conference on Spoken Language Processing*, Japan, pp. 1089-1092, 1990.

[27] Salor Ö., Pellom B., Ciloglu T., and Demirekler M., "Turkish Speech Corpora and Recognition Tools Developed by Porting SONIC: Towards Multilingual Speech Recognition," *Computer Journal of Speech and Language*, vol. 21, no. 4, pp. 580-593, 2007.

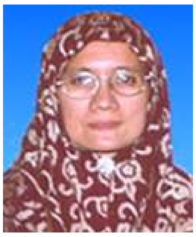
[28] Satori H., Harti M., and Chenfour N., "Arabic Speech Recognition System Based on CMUSphinx," *in Proceedings of IEEE International Symposium on Computational Intelligence and Intelligent Informatics*, Morocco, pp. 31-35, 2007.

[29] Satori H., Hiyassat H., Harti M., and Chenfour N., "Investigation Arabic Speech Recognition Using CMU Sphinx System," *International Arab Journal of Information Technology*, vol. 6, no. 2, pp. 186-190, 2009.

[30] Soltau H., Saon G., Kingsbury B., Kuo J., Mangu L., Povey D., and Zweig G., "The Ibm 2006 Gale Arabic Asr System," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Single*, USA, vol. 4, pp. 349-352, 2007.

Mohammad Abushariah received two bachelor degrees in management information systems and information technology from the International Islamic University Malaysia in 2005 and 2006, respectively. He obtained his Master degree in software engineering from University of Malaya in 2007. Currently, he is working towards his PhD in computer science and information technology in University of Malaya, specialized in arabic automatic continuous speech recognition. He has over 10 publications in IEEE international conferences, and technical reports. His research interests include: Arabic speech processing, text and speech corpora, and language resources production. He is a member of IEEE and IACSIT.

Raja Ainon is an associate professor in the Department of Software Engineering at University of Malaya. Her current research areas include HMM-based speech synthesis and recognition for malay and arabic languages, and fuzzy-genetic algorithms. She is the author of more than 30 scholarly articles in automatic timetabling, text compression, expert systems, computational linguistics, fuzzy-genetic algorithms, emotional text-to-speech synthesis, and speech recognition. Currently, she is heading the computational research group at University of Malaya.



Roziati Zainuddin is working at the Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya. Her areas of interest are intelligent multimedia, image and speech processing, computational fluid dynamics, bio-medical informatics, computer vision and visualisation, and e-learning. Her research work has been published in several international journal and conference publications. Several awards have been won for her research outcome at software exhibitions. Her professional duties include reviewing articles, editing journals, supervising research students, and appointed as an external examiner.

Moustafa Elshafei received his PhD (with Dean List) from McGill University, Canada, in electrical engineering in 1982. Since then, he has accumulated a unique blend of nine years of industrial experience and over 17 years of academic experience. He is co-inventor/sole inventor of several US patents and international patents. He has over 120 publications in international journals, conferences, and technical reports. He was the PI/CoI of many funded projects and was also involved in many internally funded or industry funded projects. His research interests include: Arabic speech processing, digital signal processing, and intelligent instrumentation. He is a member of IEEE, ISA, and SPE.

Othman Khalifa received his Bachelor's degree in electronic engineering from the Garyounis University, Libya in 1986. He obtained his Master degree in electronics science engineering and his PhD from Newcastle University, UK in 1996 and 2000, respectively. He worked in industry for eight years. Currently, he is a professor and head of the Department of Electrical and Computer Engineering, International Islamic University Malaysia. His area of research interest is communication systems, digital image/ video processing, coding and compression, wavelets, fractal and pattern recognition. He published more than 150 papers in international journals and conferences. He awarded more than 30 medals in different exhibition, and secured more than 9 research Grants. He is SIEEE member.