Lang Resources & Evaluation (2012) 46:601–634
DOI 10.1007/s10579-011-9166-8

ORIGINAL PAPER

Phonetically rich and balanced text and speech corpora for Arabic language

Mohammad A. M. Abushariah · Raja N. Ainon · Roziati Zainuddin · Moustafa Elshafei · Othman O. Khalifa

Published online: 5 November 2011

Abstract This paper describes the preparation, recording, analyzing, and evaluation of a new speech corpus for Modern Standard Arabic (MSA). 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 three major regions (Levant, Gulf, and Africa). Three hundred and sixty seven sentences are considered as phonetically rich and balanced, which are used for training Arabic Automatic Speech Recognition (ASR) systems. The rich characteristic is in the sense that it must contain all phonemes of Arabic language, whereas the balanced characteristic is in the sense that it must preserve the phonetic distribution of Arabic language. The remaining 48 sentences are created for testing purposes, which are mostly foreign to the training

M. A. M. Abushariah (✉) · R. N. Ainon · R. Zainuddin
Faculty of Computer Science and Information Technology, University of Malaya,
50603 Kuala Lumpur, Malaysia
e-mail: shariah@siswa.um.edu.my; shariah_um@yahoo.com

R. N. Ainon ·
e-mail: ainon@um.edu.my

R. Zainuddin ·
e-mail: roziati@um.edu.my

M. A. M. Abushariah
King Abdullah II School for Information Technology, University of Jordan,
Amman 11942, Jordan

M. Elshafei
Department of Systems Engineering, King Fahd University of Petroleum and Minerals,
KFUPM Box 405, Dhahran 31261, Saudi Arabia
e-mail: elshafei@kfupm.edu.sa

O. O. Khalifa
Electrical and Computer Engineering Department, Faculty of Engineering, International Islamic
University Malaysia, 53100 Gombak, Kuala Lumpur, Malaysia
e-mail: khalifa@iium.edu.my

sentences and there are hardly any similarities in words. In order to evaluate the speech corpus, Arabic ASR systems were developed using the Carnegie Mellon University (CMU) Sphinx 3 tools at both training and testing/decoding levels. The speech engine uses 3-emitting state Hidden Markov Models (HMM) for tri-phone based acoustic models. Based on experimental analysis of about 8 h 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 uni-grams, bi-grams, and tri-grams. For same speakers with different sentences, Arabic ASR systems obtained average Word Error Rate (WER) of 9.70%. For different speakers with same sentences, Arabic ASR systems obtained average WER of 4.58%, whereas for different speakers with different sentences, Arabic ASR systems obtained average WER of 12.39%.

Keywords Modern Standard Arabic · Speech corpus · Text corpus · Phonetically rich · Phonetically balanced · Automatic continuous speech recognition

1 Introduction

Arabic language is the largest Semitic language still in existence and one of the six official languages of the United Nations (UN). The number of first language speakers of Arabic exceeds 250 million, whereas the number of second language speakers can reach four times the number of first language speakers. It is the official language in 21 countries situated in Levant, Gulf, and Africa. Arabic language is ranked as fourth after Mandarin, Spanish and English in terms of the number of first language speakers (Elmahdy et al. 2009).

According to Elmahdy et al. (2009), Arabic language consists of three main forms, each of which has distinct characteristics. These forms are (1) Classical Arabic (CA), (2) Modern Standard Arabic (MSA), and (3) Colloquial or Dialectal Arabic (DA). Al-Sulaiti and Atwell (2006) believe that there is another form of Arabic language they referred to as Educated Spoken Arabic (ESA), which is considered as a hybrid form that derives its features from both the standard and dialectal forms, and is mainly used by educated speakers.

Being the most formal and standard form of Arabic, CA can be found in the Qur'an, religious instructions of Islam, and classical literature. These scripts have full diacritical marks, therefore, Arabic phonetics are completely represented (Elmahdy et al. 2009).

MSA is the current formal linguistic standard of Arabic language, which is widely taught in schools and universities, often used in the office, the media, newspapers, formal speeches, courtrooms, and any kind of formal communication (Elmahdy et al. 2009; Alotaibi and Meftah 2010). As classified by Elmahdy et al. (2009), MSA is the only acceptable form of Arabic language for all native speakers, where its spoken form can be understood by all native speakers.

According to Habash (2010), there is a tight relationship between CA and MSA, where the latter is syntactically, morphologically, and phonologically based on the earlier. However, MSA is lexically more modernized version of CA.

Although almost all written Arabic resources use MSA, diacritical marks are mostly omitted and readers must infer missing diacritical marks from the context (Elmahdy et al. 2009; Alotaibi and Meftah 2010). However, the issue of diacritization has been studied, where diacritics are derived automatically when they are manually unavailable (Vergyri and Kirchhoff 2004). Many software companies such as Sakhr, Apptek, and others also provide commercial software products for automatic diacritization of Arabic scripts.

Similar to CA, MSA scripts contain 34 basic sounds (28 original consonants and 6 vowels) as agreed by most Arabic language researchers. However, Elmahdy et al. (2009) have gone further to include 4 additional sounds, which they consider as foreign and rare consonants. As a result, a total of 38 sounds are introduced.

Since MSA is the only acceptable form of Arabic language for all native speakers (Elmahdy et al. 2009), it became the main focus of current Arabic ASR research efforts. However, previous Arabic ASR research efforts were directed towards DA serving a specific cluster of the Arabic native speakers (Kirchhoff et al. 2003).

DA is the natural spoken language in everyday life. It varies from one country to another and includes the daily spoken Arabic, which deviates from the standard Arabic and sometimes more than one dialect can be found within a country. From writing and publishing perspectives, DA is not used as a standard form of Arabic language (Elmahdy et al. 2009).

Lack of spoken and written resources is one of the main issues encountered by Arabic ASR researchers. A list of most popular (from 1986 through 2005) corpora is provided by Al-Sulaiti and Atwell (2006) showing only 19 corpora (14 written, 2 spoken, 1 written and spoken, and 2 conversational). However, Nikkhou and Choukri (2005) identified over 100 language resources including 25 speech corpora, 45 lexicons and dictionaries, 29 text corpora, and 1 multimodal corpus. A majority of the available spoken and written resources are not readily available to the public and many of them can only be obtained by purchasing from the Linguistic Data Consortium (LDC), the European Language Resource Association (ELRA), or other external vendors.

The need for Arabic spoken resources was surveyed by Nikkhou and Choukri (2004). This survey examined the industrial needs for Arabic language resources, where 20 companies situated in Lebanon, Palestine, Egypt, France, and US responded to the survey expressing the need for prepared and read Arabic spoken resources. Some responding companies have not purchased any data claiming that the suitable language resources were either not available, or the available resources were too expensive and did not meet standard quality requirements. They also reported that the available resources were lacking in various aspects covering adaptability, reusability, quality, coverage, and adequate information types.

Nikkhou and Choukri (2005) conducted a complementary survey on Arabic language resources and tools in the Mediterranean countries. This survey targeted players of Arabic language technologies in academia and industry, where a total of 55 respondents were received (36 institutions and 19 individual experts) representing 15 countries located in North Africa, Near and Middle East, Europe, and North America. The respondents insisted on the need for Arabic language resources for both MSA and DA. They also emphasized on the importance of automatic Arabic large-vocabulary (dictation) speech recognition systems for office environment, and Arabic speech understanding and synthesis.

The two surveys conducted by Nikkhou and Choukri (2004 and 2005) not only showed the need for language resources for MSA within the Arab world, but also beyond that covering many western countries.

The available spoken corpora for Arabic language such as OrienTel (Siemund et al. 2002), NEMLAR broadcast news speech corpus (ELRA 2005), and many others were mainly collected from broadcast news (radios and televisions), and telephone conversations. Broadcast news corpora are widely used in many recent ASR research efforts not only for its central interest and broad vocabulary coverage, but also for its abundant availability. However, according to Cieri et al. (2006), systems developed using broadcast news corpora may lack generality, because this kind of data may not provide adequate variability among speakers and broadcast conditions since they are collected from a single source or small number of sources. On the other hand, with the spread of telephones, conversational corpora collection from samples (not necessarily local) in the population is now possible. Therefore, variability among speakers is somewhat improved. However, the telephone-based collection of data is a limited solution, because of its quality and variation characteristics of telephone networks and handsets.

Cieri et al. (2006) stated that sampling of subjects and the loss of their anonymity are the two major risks for linguistic data collection. They also asserted that language resources need to cover important categories related to gender, age, region, class, education, occupation, and others in order to provide an adequate representation of the subjects.

The relationship between the written and spoken forms of the language resources is essential to be addressed since both forms are required for various applications especially ASR research. Many of the available Arabic spoken resources are collected prior to having the written form. In such resources, the written form is produced as a result to what has been collected in the spoken form. According to Alansary et al. (2007), the coverage of any corpora cannot contain complete information about all aspects of language lexicon and grammar due to the limited written training data and therefore inadequate spoken training data.

From the investigation of linguistic characterization of speech and writing (Parkinson and Farwaneh 2003), writing is more structurally complex and elaborate, more explicit, and more organized and planned than speech. These differences generally lead to the approach that the written form of the corpora needs to be created prior to producing and recording the spoken form. Therefore, linguists and phoneticians carefully produce written corpora before handing them to speech recording specialists for recording purposes.

In the past few years, a lot of effort has been devoted to the design and development of speech corpora for different languages. These efforts have addressed the relationship between the written and spoken forms of the corpora, and gave more emphasis to designing quality written form that embeds the language's phonetic knowledge prior to collecting the spoken form. According to Uraga and Gamboa (2004), speakers would have their own speaking style; however, their speech of the same language has the same phonological structure. Therefore, the phonological level of the language is selected to design phonetically rich and balanced text and speech corpora for many languages.

Creating phonetically rich and balanced text corpora 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 (Pineda et al. 2004). Such text corpora are then recorded in order to produce phonetically rich and balanced speech corpora.

This approach has been adopted in languages such as English (Garofolo et al. 1993; Black and Tokuda 2005; D'Arcy and Russell 2008), Mandarin (Chou and Tseng 1999; Liang et al. 2003), Spanish (Uraga and Gamboa 2004), and Korean (Hong et al. 2008).

Based on literature investigation, our research work provides Arabic language resources that meet academia and industrial expectations and recommendations. The phonetically rich and balanced Arabic speech corpus is developed in order to provide a state-of-the-art spoken corpus that bridges the gap between currently available Arabic spoken resources and the research community expectations and recommendations. The following motivational factors and speech corpus characteristics were considered for developing our spoken corpus:

1. MSA is the only acceptable form of Arabic language for all native speakers and is highly demanded for Arabic language research; therefore, our speech corpus is based on MSA form.
2. The newly developed Arabic speech corpus is prepared in a high quality and specialized sound-attenuated studio, which suits a wide horizon of systems especially for office environment as recommended by Nikkhou and Choukri (2005).
3. The speech corpus is designed in a way that would serve any Arabic ASR system regardless of its domain. It focuses on the presence of Arabic phonemes as much as possible using the least possible Arabic words and sentences based on phonetically rich and balanced speech corpus approach.
4. The availability of a phonetically rich and balanced text corpus developed in (Alghamdi et al. 1997, 2003). Further details are provided in Sect. 3.
5. The opportunity to explore differences of speech patterns between Arabic native speakers from 11 different countries representing the three major regions (Levant, Gulf, and Africa) in the Arab world.
6. The need for prepared and read Arabic spoken resources as illustrated in Nikkhou and Choukri (2004) is also considered. Companies did not show interest in Arabic telephone and broadcast news spoken data. Therefore, this phonetically rich and balanced Arabic speech corpus provides neither telephone nor broadcast news spoken resources. It is a prepared and read Arabic spoken corpus.

The following section, Sect. 2, provides a statistical analysis and description of the text and speech corpora. Implementation requirements for developing the Arabic automatic continuous speech recognition system are presented in Sect. 3. Section 4 emphasizes on the testing and evaluation of the text and speech corpora using the developed Arabic automatic continuous speech recognition system. Conclusions are finally presented in Sect. 5.

2 Statistical analysis and description of the text and speech corpora

In order to produce a robust speaker-independent, continuous, and 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 phonemes of Arabic language, whereas the balanced characteristic is in the sense that it must preserve the phonetic distribution of Arabic language. 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

As stated earlier, creating phonetically rich and balanced text corpus requires the presence of phonetically rich words that are used to form sentences and phrases, which are verified for balanced phonetic distribution.

King Abdulaziz City of Science and Technology (KACST) created a database for Arabic language phonemes. The purpose of this work was to create a list of the least number of phonetically rich Arabic words. As a result, a list of 663 phonetically rich words containing all Arabic phonemes based on 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 (TTS) synthesis applications. The list of 663 phonetically rich words was created based on the following characteristics and guidelines (Alghamdi et al. 1997):

- Cover all Arabic phonemes which must be balanced so as to be close in frequency as possible.
- Contain all phonotactic rules of Arabic, which means coverage of all Arabic phoneme clusters.
- 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.
- To be of words in circulation and use as far as possible.

Based on the above characteristics and guidelines, two specialized linguists manually prepared a list of about 7,000 words. It was difficult to know all covered Arabic phoneme clusters while writing the list; therefore, the list had to be this huge. At this stage, a linguist might have written a word in the list in order to achieve a certain phonotactic rule of Arabic, and also have written another word to achieve another phonotactic rule of Arabic, while a single word could have achieved both phonotactic rules of Arabic. For example, the linguist could have written the word (معلومٌ) in order to cover the phonotactic rule (Presence of Two Consonants) in this case the two consonants are /ع/ and /ل/, and also have written the word (مسلولٌ) in order to cover another phonotactic rule (Presence of a Consonant followed by a Vowel) in this case the consonant /ل/ and the vowel /ـٌ/. It is noticed that the word (معلومٌ) could cover the said two phonotactic rules.

In order to reduce such redundancies, a computer program was developed and applied on the initial list of about 7,000 words. As a result, a list of 663 phonetically rich words was produced, which covers all possible Arabic phonotactic rules (Alghamdi et al. 1997).

Statistical analysis of the 663 phonetically rich words show that all Arabic phonemes are covered in this list as illustrated 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. Each Arabic phoneme is also represented in the International Phonetic Alphabet (IPA) symbols (Wikipedia 2011; Habash 2010). The number of repetitions is classified further to include repetitions of the each Arabic phoneme in three places (Front, Inside, End) of the 663 phonetically rich words.

From Table 1, it is noticed that the Arabic vowel (ـَ) was repeated 609 times, which is considered very high compared to other vowels and consonants. According to Alghamdi et al. (1997), the Arabic vowel (ـَ) has a high repetition in Arabic words that exceeds all other Arabic phonemes, which might even reach 43% from the total repetition of Arabic phonemes. However, the average repetition for each Arabic phoneme for the list of 663 phonetically rich words was 82 times, if excluding the Arabic vowel (ـَ).

As an extension to Alghamdi et al. (1997) work, KACST produced a technical report of the project "Database for Arabic Phonemes: Sentences" in Alghamdi et al. (2003). 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:

- To have the minimum word repetitions as far as possible.
- To have an average of 2–9 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.

As a result, a list of fully diacritical 367 phonetically rich and balanced sentences was produced using 1,835 Arabic words. An average of 2 phonetically rich words and 5 other words were used in each single sentence. Statistical analysis shows that 1,333 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.

Table 1 Arabic phoneme repetitions for the 663 phonetically rich words

Arabic alphabets and vowels	IPA symbols	Repetitions			Total	Percentage (100%)
		Front	Inside	End		
ء	ʔ	76	38	18	132	3.95
ب	b	27	45	32	104	3.11
ت	t	21	30	19	70	2.09
ث	θ	4	30	13	47	1.40
ج	dʒ	9	39	12	60	1.79
ح	ħ	29	39	16	84	2.51
خ	x	16	36	14	66	1.97
د	d	6	38	19	63	1.88
ذ	ð	5	37	6	48	1.43
ر	r	36	46	53	135	4.04
ز	z	4	34	10	48	1.43
س	s	28	29	17	74	2.21
ش	ʃ	11	32	18	61	1.82
ص	sˤ	10	27	13	50	1.49
ض	dˤ	11	31	10	52	1.55
ط	tˤ	11	28	18	57	1.70
ظ	zˤ	6	25	5	36	1.07
ع	ʕ	35	34	20	89	2.66
غ	ɣ	11	34	6	51	1.52
ف	f	27	46	24	97	2.90
ق	q	25	36	18	79	2.36
ك	k	14	41	12	67	2.00
ل	l	25	37	48	110	3.29
م	m	77	36	53	166	4.97
ن	n	40	41	39	120	3.59
هـ	h	3	44	50	97	2.90
و	w	21	47	14	82	2.45
ي	j	74	45	17	136	4.07
ـَ	a	0	597	12	609	18.26
ـَـَ	a:	0	74	21	95	2.84
ـُ	u	0	124	11	135	4.04
ـُـُ	u:	0	46	5	51	1.52
ـِ	i	0	115	0	115	3.44
ـِـِ	i:	0	29	19	48	1.43
Total repetitions for all Arabic phonemes and vowels					3,334	100

The main aim of this work was to produce a set of Arabic sentences that are phonetically rich and also balanced. According to Alghamdi et al. (2003), although this set of 367 Arabic sentences contains only 1,835 words, yet they contain all Arabic phoneme clusters that are in line with the Arabic phonotactic rules.

Table 2 Samples of the phonetically rich and balanced sentences

Sample 1	مِنْ بَخْسِ نِعْمَةِ اللّٰهِ دَفْنُهَا
Sample 2	أَبْجَلَنِي هَذَا الطَّعَامُ
Sample 3	الْأَعْشَى الشَّاعِرُ مِنْ أَدْهَى الشُّعَرَاء

Table 3 Samples of the testing sentences

Sample 1	أَصَابَنَا مَطَرٌ سَالَ مِنْهُ الغُبَّانُ
Sample 2	زُرْ غِبّاً تَزْدَدْ حُبّاً
Sample 3	لَسْتُ بِالْخِبِّ وَلَا الْخِبُّ يَخْدَعُنِي

This set of phonetically rich and balanced sentences can be used for training and testing Arabic ASR engines, Arabic TTS synthesis, and many others. Any Arabic ASR system that is based on this set of phonetically rich and balanced sentences is expected to perform successfully against any other Arabic sentences (Alghamdi et al. 2003). Table 2 shows three sample sentences of the 367 phonetically rich and balanced sentences.

KACST 367 phonetically rich and balanced sentences are used for training purposes in our experimental work, 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. Table 3 shows three sample sentences of the 48 testing sentences.

Table 4 shows the number of repetitions as well as the percentage for each Arabic phoneme and grapheme in both the KACST 367 phonetically rich and balanced training sentences and the 48 testing sentences sorted in an ascending order. It is found that the Arabic vowel (ـَ) is still maintained as the highest in repetition compared to the rest of Arabic phonemes and graphemes shown in Table 4. In this list the Arabic vowel (ـَ) repetition was 18.46%, whereas it was roughly the same percentage in the list of 663 phonetically rich words as shown in Table 1. This indicates that almost all properties found in the list of 663 phonetically rich words are also reflected on the list of phonetically rich and balanced training sentences and testing sentences. This also shows that there is a high possibility that the recorded speech corpus of the 415 sentences will maintain such properties.

After finalizing the text corpus and identifying the training and testing texts, it is important to record this corpus and produce its spoken version. Details on our phonetically rich and balanced speech corpus are discussed in the following Sect. 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 and were recorded and used for training the acoustic models. For testing the acoustic models, 48 additional sentences representing Arabic proverbs were created by an Arabic language specialist for the purpose of our corpus.

Table 4 Arabic phonemes and graphemes repetitions for the 367 phonetically rich and balanced training sentences and the 48 testing sentences

Arabic phonemes and graphemes	IPA symbols	Repetitions			Percentage (%)
		Training sentences	Testing sentences	Total	
ؤ	ʔ	8	2	10	0.07
آ	ʔ	11	1	12	0.08
ئ	ʔ	19	0	19	0.13
ـَـَ	a:	34	14	48	0.32
ء	ʔ	45	10	55	0.37
ظ	zˤ	54	6	60	0.40
ـِـِ	i	59	6	65	0.44
إ	ʔ	57	16	73	0.49
ى	a	79	5	84	0.57
غ	ɣ	75	11	86	0.58
ث	θ	80	6	86	0.58
ز	z	78	10	88	0.59
ض	dˤ	96	7	103	0.69
ـُـُ	u:	93	19	112	0.75
ذ	ð	108	7	115	0.77
خ	x	104	19	123	0.83
ص	sˤ	108	17	125	0.84
ط	tˤ	113	22	135	0.91
ة	t	118	23	141	0.95
ش	ʃ	133	12	145	0.98
ج	dʒ	162	17	179	1.21
ك	k	165	26	191	1.29
ق	q	173	19	192	1.29
ح	ħ	195	24	219	1.47
س	s	186	39	225	1.51
ت	t	205	31	236	1.59
أ	ʔ	203	44	247	1.66
د	d	219	41	260	1.75
هـ	h	246	27	273	1.84
ف	f	251	31	282	1.90
ع	ʕ	280	33	313	2.11
و	w	295	45	340	2.29
ب	b	370	58	428	2.88
ر	r	393	63	456	3.07
ن	n	421	72	493	3.32
م	m	479	75	554	3.73
ي	j	552	56	608	4.09
ـُ	u	910	119	1,029	6.93
ل	l	1,037	149	1,186	7.98

Table 4 continued

Arabic phonemes and graphemes	IPA symbols	Repetitions			Percentage (%)
		Training sentences	Testing sentences	Total	
ا	a:	1,023	181	1,204	8.11
ـِ	i	1,333	178	1,511	10.17
ـَ	a	2,337	405	2,742	18.46
Total repetitions		12,907	1,946	14,853	100

The motivation behind the creation of our phonetically rich and balanced speech corpus was to provide large amounts of high quality recordings of MSA suitable for the design and development of any speaker-independent, continuous, and automatic Arabic ASR system. The uniqueness of our speech corpus can be characterized as follows:

- It contains large amounts of MSA speech.
- It contains the phonetic transcription of all recorded speech.
- It contains high quality recordings captured using specialized equipments located in a sound-attenuated studio.
- It contains speech recordings that can be used for training as well as testing any Arabic speech based system.
- It contains speech from 40 (20 male and 20 female) native speakers having different characteristics and variabilities.
- It contains speech from native speakers from 11 Arab countries representing the three major regions (Levant, Gulf and Africa). The minimum number of speakers was 11 for Gulf, followed by 14 and 15 speakers for Africa and Levant, respectively. This allows researchers to study within country and region variability.
- It contains large amounts of data for every speaker. An average of 1 h ready to use speech recordings was captured in order to allow researchers to study within speaker variability.

2.2.1 Speech corpus participants

The phonetically rich and balanced Arabic speech corpus was initiated in March 2009. Although participants were selected based on their interest to join this work, speakers were indirectly selected based on predetermined characteristics as follows:

- They have a fair distribution of gender and age.
- Their current professions vary.
- They have a mixture of educational backgrounds with a minimum of high school certification. This is important to secure an efficient reading ability of the participants.
- They belong to various native Arabic speaking countries.

Table 5 Speakers' region, country, and gender distribution

Region	Country	Gender		Total	Total/region	Ratio/region (%)
		Male	Female			
Levant	Jordan	8	4	12	15	37.50
	Palestine	2	–	2		
	Syria	1	–	1		
Gulf	Iraq	–	4	4	11	27.50
	Saudi Arabia	–	3	3		
	Yemen	–	3	3		
	Oman	–	1	1		
Africa	Sudan	3	3	6	14	35.00
	Algeria	3	2	5		
	Egypt	2	–	2		
	Morocco	1	–	1		
Total		20	20	40	40	100
Total (%)		50	50	100	100	

Table 6 Speakers' age and gender distribution

No	Age category	Gender		Total
		Male	Female	
1	Less than 30 years	6	14	20
2	30 Years and above	14	6	20
Total		20	20	40

- They belong to any of the three major regions where Arabic native speakers mostly live (Levant, Gulf, and Africa). This is important to produce a comprehensive speech corpus that can be used by all Arabic language research community.

As a result, speech recordings of 40 speakers were collected. Table 5 shows the distribution of the 40 speakers according to region, country, and gender, whereas Table 6 shows that speakers are divided into two main age groups.

A complete list of the selected speakers is summarized in Table 7, which shows the assigned Speaker ID, and the corresponding gender, age, age category, current profession, educational background, country, and region of each participant. The abbreviations (U.), (S.), and (M.) in the current profession column refer to (University), (School), and (Medical), respectively.

2.2.2 Speech corpus recording set-up

Recording sessions were conducted in a sound-attenuated studio shown in Fig. 1. Participants were asked to complete their recordings in one session. However, some participants exceeded one session and completed their recordings in 2–3 sessions

Table 7 Summary of the selected speakers' details

Speaker ID	Gender	Age	Age category	Current profession	Educational background	Country	Region
SP01	Female	19	<30	Student	High school	Yemen	Gulf
SP02	Male	53	≥30	U. Lecturer	Master	Sudan	Africa
SP03	Male	32	≥30	Student	Master	Algeria	Africa
SP04	Female	28	<30	Student	Bachelor	Algeria	Africa
SP05	Female	23	<30	Student	Bachelor	Yemen	Gulf
SP06	Female	20	<30	Student	High school	Yemen	Gulf
SP07	Male	35	≥30	Student	Master	Jordan	Levant
SP08	Female	25	<30	Student	Bachelor	Jordan	Levant
SP09	Female	18	<30	Student	High school	Sudan	Africa
SP10	Male	57	≥30	U. Lecturer	PhD	Egypt	Africa
SP11	Female	24	<30	Student	Bachelor	Saudi	Gulf
SP12	Female	25	<30	Student	Bachelor	Saudi	Gulf
SP13	Male	53	≥30	U. Lecturer	Master	Sudan	Africa
SP14	Male	30	≥30	Student	Bachelor	Jordan	Levant
SP15	Male	27	<30	Student	Bachelor	Jordan	Levant
SP16	Male	20	<30	Student	High school	Palestine	Levant
SP17	Male	38	≥30	Student	Master	Jordan	Levant
SP18	Female	30	≥30	S. Teacher	Master	Jordan	Levant
SP19	Male	29	<30	Student	Bachelor	Algeria	Africa
SP20	Female	27	<30	Student	Master	Jordan	Levant
SP21	Male	58	≥30	U. Lecturer	PhD	Jordan	Levant
SP22	Male	49	≥30	U. Lecturer	Master	Egypt	Africa
SP23	Male	58	≥30	U. Lecturer	Master	Sudan	Africa
SP24	Male	27	<30	Student	Bachelor	Syria	Levant
SP25	Male	21	<30	Student	High school	Palestine	Levant
SP26	Female	25	<30	Student	Bachelor	Saudi	Gulf
SP27	Male	35	≥30	Student	Master	Jordan	Levant
SP28	Male	30	≥30	Student	Master	Algeria	Africa
SP29	Female	21	<30	Student	High school	Sudan	Africa
SP30	Female	66	≥30	S. Teacher	Bachelor	Iraq	Gulf
SP31	Female	46	≥30	Student	Master	Iraq	Gulf
SP32	Female	34	≥30	Student	Bachelor	Iraq	Gulf
SP33	Female	42	≥30	U. Lecturer	PhD	Algeria	Africa
SP34	Female	25	<30	M. Doctor	Bachelor	Iraq	Gulf
SP35	Male	35	≥30	U. Lecturer	Master	Morocco	Africa
SP36	Male	61	≥30	U. Lecturer	PhD	Jordan	Levant
SP37	Male	29	<30	Student	Master	Jordan	Levant
SP38	Female	28	<30	Student	Bachelor	Oman	Gulf
SP39	Female	19	<30	Student	High school	Jordan	Levant
SP40	Female	30	≥30	Student	Bachelor	Sudan	Africa

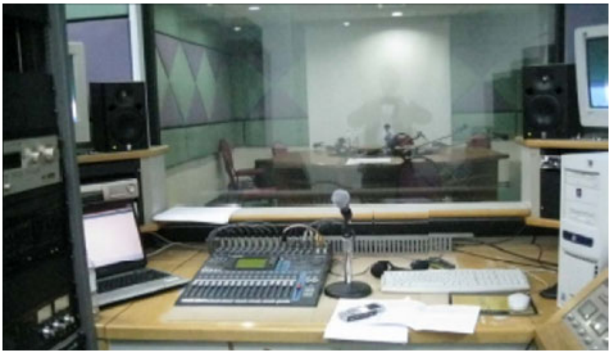

Fig. 1 Sound-attenuated studio

due to scheduling reasons. Participants were asked to read the 415 sentences prepared for this task. Each sentence was recorded at least twice depending on the participant's reading ability and quality. Some participants had to utter sentences for 10 times due to pronunciation deficiencies and mistakes.

Participants had to use headsets in order to listen to the instructor's comments, announcements, and directions. The instructor is located in a different control room separated by glassed window. However, the instructor and participants can see each other.

Participants were allowed to stop at any point of time for short rests. They were also allowed to ask or talk to the instructor for relevant and irrelevant topics. At times they used to laugh, cough, and sneeze.

Recording sessions were conducted in a sound-attenuated studio room. Speakers used the SHURE SM58 wired unidirectional dynamic microphone to utter the recordings. They also used the Beyerdynamic DT 231 Headphone in order to listen to instructions from the recording specialist. In addition, the YAMAHA 01V 96 Version 2 (Digital Audio Mixer) was used. In terms of software, Sony Sound Forge 8 was used on a normal Personal Computer (PC) located in the studio with Windows XP in order to record the utterances from the speakers. Default recording attributes were initially used as shown in Table 8.

These recording attributes were then converted at a later stage to be used for developing ASR applications as shown in Table 9 using features provided by Sony Sound Forge 8. A Matlab program was also developed in order to assure the converted attributes are achieved. It is important to highlight that converting from 2 channels (Stereo) to 1 channel (Mono) does not affect the utterances since the second channel does not have any important information. Therefore, this conversion does not make the utterances lose anything and it is meant to meet standards of ASR applications.

2.2.3 Speech corpus preparation and pre-processing

In order to use our phonetically rich and balanced speech corpus for training and testing Arabic ASR systems, a number of Matlab programs were developed in order

Table 8 Initial recording attributes

Recording attribute	Value
Sampling rate (Hz)	44,100
Bit-depth (bits)	16
Channels	2 (Stereo)

Table 9 Converted recording attributes for speech recognition tasks

Recording attribute	Value
Sampling rate (Hz)	16,000
Bit-depth (bits)	16
Channels	1 (Mono)

to produce a ready to use speech corpus. These Matlab programs were developed for the purpose of (1) Automatic Arabic speech segmentation, (2) Parameters conversion of speech data, (3) Directory structure and sound filenames convention, and (4) Automatic generation of training and testing transcription files. Manual classification and validation of the correct speech data were conducted with great care and precision. This process was very crucial in order to ensure and validate the pronunciation correctness of the speech utterances before using them in training the system's acoustic model (Abushariah et al. 2010a).

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 three ".wav" files depending on the 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.

The second Matlab program was developed re-sample the sampling rate from 44,100 into 16,000 Hz and to convert number of channels from 2 into 1, which are used in most ASR research.

In addition, 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. Therefore, 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”.

After finalizing the ready-to-use speech and text corpora, an open source concordance tool (aConCorde) developed by the School of Computing at University of Leeds for analyzing Arabic text corpora was deployed (Roberts et al. 2006). Statistical analysis of the transcription file associated with the final ready-to-use speech corpus shows that the minimum repetition of words is 87 in examples like the words (رَأَى) and (أَحْلامٍ) which mean (saw) and (dreams), respectively. The word (في) which means (in) in English language is still considered as the maximum repeated word and is repeated 7,310 times. In addition, only 12 out of the 1,626 unique words are repeated between 1,001 and 7,500 times, whereas 111 words are repeated between 200 and 1,000 times. Therefore, 1,503 unique words are repeated between 87 and 199 times, which indicates that the word has been recorded in an average of 2–5 times from each of the 40 speakers.

After finalizing the ready to use speech corpus, statistical analysis was conducted and summarized in Table 10. It is important to highlight that the number of unique words if both training and testing sentences are combined together in one transcription file is 1,626 words. However, if training and testing sentences have been divided into two transcription files, it is found that the number of unique words are 1,422 and 241 for training and testing sentences, respectively, which sum to 1,663 words. The difference between 1,663 and 1,626 is 37 words, which comprise the similar words between training and testing sentences. As a result, testing sentences are mostly foreign to the training sentences and there are hardly any similarities in words.

Table 11 shows the number of repetitions as well as the percentage for each Arabic phoneme and grapheme in the final transcription file of the speech corpus sorted in an

Table 10 Statistical analysis of the phonetically rich and balanced speech corpus

Criteria	Training sentences	Testing sentences	Total
No. of sentences	367 sentences	48 sentences	415 sentences
Number of unique words based on training and testing sentences in isolated transcription files	1,422 words	241 words	1,663 words
Number of unique words based on training and testing sentences in combined transcription file	–	–	1,626 Words
Total frequencies of words in the transcription file	198,426 words	31,172 words	229,598 words
No. of utterances (.wav) sound files	40,025 utterances	5,469 utterances	45,494 utterances
Average no. of (.wav) sound files/sentence	109 sound files/sentence	114 sound files/sentence	–
Size of utterances (.wav) sound files (GB)	4.72	0.72	5.44
Size of feature extracted utterances (.mfc) files (MB)	849	129	978
Duration of utterances (.wav) sound files (h)	43.20	6.63	49.83
Average duration/sentence	7.06 minutes/sentence	8.29 minutes/sentence	–
Average duration/utterance (.wav) sound file	3.89 seconds/utterance	4.36 seconds/utterance	–

ascending order. It is found that the Arabic vowel (ـَ) is still maintained as the highest in repetition compared to the rest of Arabic phonemes and graphemes similar to what was shown earlier in Table 4.

It is vital to emphasize the concept of phonetically rich and balanced approach. The rich characteristic is in the sense that it must contain all phonemes of Arabic language, whereas the balanced characteristic is in the sense that it must preserve the phonetic distribution of Arabic language. It is noticed that all Arabic phonemes are covered in the 367 phonetically rich and balanced sentences; and therefore, it is a phonetically rich corpus. On the other hand, the phonetically balanced aspect does not mean that the Arabic phonemes must have equal number of occurrences in the corpus. Instead, it must preserve the phonetic distribution of the language. To validate this characteristic, Meeralam (2007) stated in his report that various old Arabic linguists such as Alkindi, Ibn Dunaineer, and Ibn Adlan have classified the occurrences of the Arabic alphabets into high, average, or low repeated. The high repeated alphabets are seven, which make up the word (الموهين). The average repeated alphabets are eleven, which make up the following three words (رعفت بكدس قحج). Finally the low repeated alphabets are ten, which are the first alphabet of each word of the following poetry (ظلم غزا طاب زورا ثاويا خوف ضنى شبت صبا ذاويا). Meeralam (2007) also stated that the Arabic alphabets /ا/, and /ل/ are the most frequently used alphabets, whereas the Arabic alphabets /ظ/, and /غ/ are the least frequently used

Table 11 Arabic phonemes and graphemes repetitions for the final transcription (training and testing) files associated with the speech corpus

Arabic phonemes and graphemes	IPA symbols	Repetitions			Percentage (%)
		Training sentences	Testing sentences	Total	
ؤ	ʔ	895	258	1,153	0.07
آ	ʔ	1,201	95	1,296	0.08
ئ	ʔ	2,011	0	2,011	0.12
ـَـَ	a:	3,811	1,596	5,407	0.33
ء	ʔ	4,965	1,166	6,131	0.38
ظ	zˤ	5,835	664	6,499	0.40
ـِـِ	i	6,351	743	7,094	0.43
إ	ʔ	6,155	1,798	7,953	0.49
ى	a:	8,627	593	9,220	0.57
غ	ɣ	8,094	1,246	9,340	0.57
ث	θ	8,784	644	9,428	0.58
ز	z	8,655	1,102	9,757	0.60
ض	dˤ	10,439	751	11,190	0.69
ـُـُ	u:	10,076	2,133	12,209	0.75
ذ	ð	11,627	805	12,432	0.76
ص	sˤ	11,780	1,861	13,641	0.84
خ	x	11,581	2,220	13,801	0.85
ط	tˤ	12,500	2,484	14,984	0.92
ة	t	12,699	2,700	15,399	0.94
ش	ʃ	14,485	1,340	15,825	0.97
ج	dʒ	17,859	1,866	19,725	1.21
ك	k	18,017	2,904	20,921	1.28
ق	q	19,170	2,065	21,235	1.30
ح	ħ	21,303	2,681	23,984	1.47
س	s	20,549	4,466	25,015	1.53
ت	t	22,627	3,710	26,337	1.61
أ	ʔ	21,730	4,962	26,692	1.64
د	d	24,007	4,514	28,521	1.75
هـ	h	26,918	3,006	29,924	1.83
ف	f	27,293	3,540	30,833	1.89
ع	ʕ	30,765	3,752	34,517	2.12
و	w	32,182	5,317	37,499	2.30
ب	b	40,482	6,527	47,009	2.88
ر	r	43,274	7,143	50,417	3.09
ن	n	45,993	8,226	54,219	3.32
م	m	52,364	8,563	60,927	3.73
ي	j	60,753	6,337	67,090	4.11
ـُ	u	100,118	13,578	113,696	6.97
ل	l	113,242	16,783	130,025	7.97

Table 11 continued

Arabic phonemes and graphemes	IPA symbols	Repetitions			Percentage (%)
		Training sentences	Testing sentences	Total	
ا	a:	111,233	20,305	131,538	8.06
ـِ	i	145,523	20,248	165,771	10.16
ـَ	a	255,130	45,770	300,900	18.44
Total repetitions		1,411,103	220,462	1,631,565	100

alphabets. Another study of Arabic letter frequency analysis conducted by Madi (2010) using an Arabic letter and word frequency analyzer known as 'Intellyze' is referred. This study used sources adding up to 3,378 pages, generating 1,297,259 words, or, 5,122,132 letters. The letter frequency distribution for this data shows that the Arabic letters / ا /, / ل /, / ن /, / م /, / ي /, / و /, / هـ /, / ب /, / ر /, and / ع / have the most frequency among all, whereas the Arabic letters / ؤ /, / آ /, / ظ /, / ئ /, / ء /, / غ /, / ض /, / ط /, / ز /, and / ش / are least frequent letters. In other studies, this analysis is roughly the same. Therefore, the corpus is considered phonetically balanced when it meets and preserves such phonetic distribution. In the analysis of the phonetically rich and balanced text and speech corpora as illustrated in Tables 4 and 11, this phonetic distribution is preserved and establishes the phonetically balanced corpora.

3 Arabic automatic continuous speech recognition system

This section describes the major implementation requirements and components for developing the Arabic automatic speech recognition system namely feature extraction, Arabic phonetic dictionary, the acoustic model training, and the statistical language model training, which are clearly shown in Fig. 2 (Abushariah et al. 2010b, c, d, 2011).

The decoder is then used once all implementation requirements are achieved. It takes the new input features converted at the feature extraction stage, the search graph, the trained acoustic model, the trained language model, and the phonetic dictionary in order to recognize the speech in the features. A brief description of each component is discussed in the following sub-sections.

3.1 Feature extraction

Feature extraction, also referred to as front end component, is the initial stage of any ASR system that converts speech inputs into feature vectors in order to be used for training and testing the speech recognizer. The dominating feature extraction technique known as Mel-Frequency Cepstral Coefficients (MFCC) was applied to extract features from the set of spoken utterances. The MFCC is also used in CMU Sphinx 3 tools (Chan et al. 2007) as the main feature extraction technique. As a result, a feature vector that represents unique characteristics of each recorded

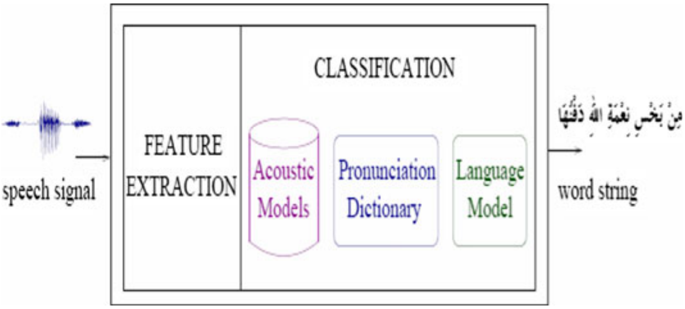

Fig. 2 Components of Arabic automatic continuous speech recognition system

آلامُ 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

أبِي E AE B IY

أبْجَلَنِي E AE B JH AE L AE N IY

Fig. 3 Sample of the rule-based phonetic dictionary

utterance is produced, which is considered as an input to the classification component (Abu Shariah et al. 2007).

3.2 Arabic phonetic dictionary

The phoneme pronunciation dictionary serves as an intermediary link between the acoustic model and the language model in all ASR systems. A rule-based approach to automatically generate a phonetic dictionary for a given transcription was used. A detailed description of the development of this Arabic phonetic dictionary can be found in the work of Ali et al. (2008). Arabic pronunciation follows certain rules and patterns when the text is fully diacritized. A detailed description of these rules and patterns can be found in the work of Elshafei (1991).

In this work, the transcription file contains 2,110 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 pure MSA-based phonetic dictionary, which is based on the transcription file that combines the training and testing sentences.

3.3 Acoustic model training

The acoustic model component provides the Hidden Markov Models (HMMs) of the Arabic tri-phones to be used in order to recognize speech. The basic HMM structure known as Bakis model, has a fixed topology consisting of five states with three emitting states for tri-phone acoustic modeling (Rabiner 1989; Bakis 1976).

In order to build a better acoustic model, CMU Sphinx 3 (Placeway et al. 1997) uses tri-phone based acoustic modeling. Continuous Hidden Markov Models (CHMM) technique is also supported in CMU Sphinx 3 for parametrizing the probability distributions of the state emission probabilities. A tri-phone not only models an individual phoneme, but it also captures distinct models from the surrounding left and right phones.

Training the acoustic model using CMU Sphinx 3 tools requires successfully passing through three phases. 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 6. 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 (Alghamdi et al. 2009).

There are 4,705 unique tri-phones extracted from the training transcripts. The minimum occurrence of tri-phones is 18 times for (AH: and IX:) whereas the maximum is 456 times for (AE) as shown in Table 12.

During the development phase, a small portion of the entire speech corpus is experimented. A total of 8,043 utterances are used resulting in about 8 h of speech data collected from 8 (5 male and 3 female) Arabic native speakers from 6 different Arab countries namely Jordan, Palestine, Egypt, Sudan, Algeria, and Morocco.

In order to show a fair testing and evaluation of the Arabic ASR performance, the leave-one-out cross validation and testing approach was applied, where every round speech data of 7 out of 8 speakers were trained and speech data of the 8th were tested. This is also important to examine the speaker-independence of the developed systems.

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 300 to 2,500. A total of 54 experiments were conducted at this stage producing different results as shown in Sect. 4. During the second stage, the best combination of Gaussian mixture distributions and number of senones was used to train the remaining seven out of eight training data sets (Abushariah et al. 2010b, d).

3.4 Language model training

The language model component provides the grammar used in the system. The grammar's complexity depends on the system to be developed. In this work, the language model is built statistically using the CMU-Cambridge Statistical Language

Table 12 Occurrences of tri-phones for each phoneme

Phone	Tri-phones	Phone	Tri-phones	Phone	Tri-phones
AA	71	F	118	R	136
AA	32	GH	61	S	98
AE	456	H	89	SH	77
AE	200	HH	97	SS	75
AH	44	IH	364	T	109
AH	18	IX	57	TH	60
AI	118	IX:	18	TT	70
AW	31	IY	103	UH	342
AY	39	JH	89	UW	77
B	148	K	96	UX	57
D	104	KH	74	W	70
DD	66	L	178	Y	70
DH	58	M	137	Z	59
DH2	40	N	195	Total	4,705
E	207	Q	97		

Modeling toolkit, which is based on modeling the uni-grams, bi-grams, and tri-grams of the language for the subject text to be recognized (Clarkson and Rosenfeld 1997).

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 (Alghamdi et al. 2009). For this work, the number of uni-grams is 1,627, whereas the number of bi-grams and tri-grams is 2,083 and 2,085, respectively (Abushariah et al. 2010b, c, d).

3.5 The decoder

This work used CMU Sphinx 3 decoder, which is based on the conventional Viterbi search algorithm and beam search heuristics. It uses a lexical-tree search structure. The decoder requires certain inputs and resources such as the acoustic model, language model, phonetic dictionary, and feature vector of the unknown utterance. The result is a recognition hypothesis, which is a single best recognition result for each utterance processed. It is a linear word sequence, with additional attributes such as their time segmentation and scores (Chan et al. 2007).

4 Testing and evaluation

This section presents the testing and evaluation of the Arabic automatic continuous speech recognition system based on a small portion of our phonetically rich and balanced speech corpus.

It is important to highlight that each speaker has two kinds of recordings, training recordings and testing recordings. Therefore, although the leave-one-out cross validation and testing approach was adopted, those speakers used in training the system still have their testing recordings. In other words, it is expected to see results of three different data sets, we refer to them as (1) Data of same speakers with different sentences, (2) Data of different speakers with same sentences, and (3) Data of different speakers with different sentences. Data sets 1 and 3 refer to our 48 testing sentences, whereas data set 2 refers to the 367 phonetically rich and balanced training sentences. As a result, testing sentences are totally foreign to the training sentences and there are hardly any similarities in words. In addition, data set 1 comprises of the testing utterances collected from the same speakers used in training. However, data sets 2 and 3 belong to the speaker who is left out in order to examine the speaker-independence of the systems.

As stated earlier, a small portion of the newly developed speech corpus is used for the development and evaluation of Arabic ASR systems. As a result, 8 different data sets were used as shown in Table 13. During the first stage of training the acoustic model, the first data set (Experiment 1) was used to identify the best combination of Gaussian mixture distributions and number of senones.

This is important in order to examine the possibilities to utilize this corpus in tasks such as ASR systems. The overall performance of the developed Arabic ASR systems based on our corpus should reflect the quality of this corpus compared to the available speech corpora especially those of broadcast news and telephone conversation speech corpora.

4.1 Performance measures

Experimental works are evaluated using two main performance metrics known as word recognition correctness rate and the WER. Corresponding formulas are as follows:

$$\text{Word Recognition Correctness Rate} = \frac{\mathrm{N} - \mathrm{D} - \mathrm{S}}{\mathrm{N}} \times 100\% \qquad (1)$$

$$\text{Percent Accuracy} = \frac{\mathrm{N} - \mathrm{D} - \mathrm{S} - 1}{\mathrm{N}} \times 100\% \qquad (2)$$

$$\text{WER} = 100\% - \text{Percent Accuracy Or WER} = \frac{\mathrm{D} + \mathrm{S} + \mathrm{I}}{\mathrm{N}} \times 100\% \qquad (3)$$

where (N) is the total number of labels in the reference transcriptions, (D) is the number of deletion errors, (I) is the number of insertion errors, and (S) is the number of substitution errors.

4.2 Testing and evaluation of Arabic ASR systems

Arabic ASR systems have undergone several modifications and enhancement approaches at both training and testing/decoding levels in order to optimize their

Table 13 Training and testing data sets

Experiment_ID	Training data size (Utterances)	Testing data size			Total testing data size (Utterances)
		Same speakers	Different speakers		
		Different sentences	Same sentences	Different sentences	
Experiment 1	6,379	906	678	80	1,664
Experiment 2	6,288	871	769	115	1,755
Experiment 3	5,569	755	1,488	231	2,474
Experiment 4	6,308	888	749	98	1,735
Experiment 5	6,296	889	761	97	1,747
Experiment 6	6,331	891	726	95	1,712
Experiment 7	6,219	861	838	125	1,824
Experiment 8	6,009	841	1,048	145	2,034

performance. This section highlights our work towards producing Arabic ASR systems with better performance based on parameters modification and enhancement at the training and testing/decoding level.

4.2.1 Modifications using basic parameters at training level

For the development of Arabic ASR systems, the first data set of Experiment 1 was used to identify the best combination of values at training level. Such values are then applied for the rest of the experiments. As stated earlier, in order to identify the best combination of Gaussian mixture distributions and senones at training level, 54 experiments were conducted. Each experiment has its own combination of the two parameters.

Gaussian mixture distributions ranged from 2 to 64, whereas senones ranged from 300 to 2,500. It is found that 16 Gaussians with 500 senones obtained the best performance as shown in Table 14 and Fig. 4.

Figure 5 shows all combinations with their corresponding word recognition correctness rates (%). Therefore, this work used the combination of 16 Gaussians with 500 senones for training the acoustic model in Experiment 2 through Experiment 8 data sets.

Based on this combination, results of the data sets identified in Table 13 are presented in Table 15.

Table 14 Performance of Arabic ASR systems at training level

Experiment_ID	Same speakers with different sentences WER (%)	Different speakers with same sentences WER (%)	Different speakers with different sentences WER (%)
(Experiment 1)	10.73	6.28	13.48

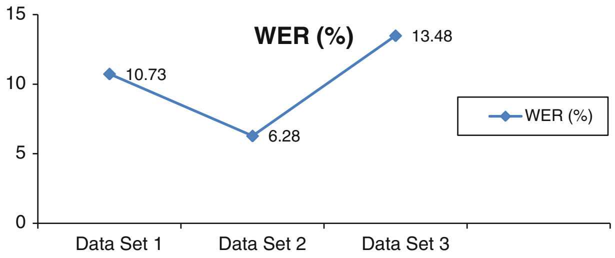

Fig. 4 Performance of Arabic ASR systems at training level

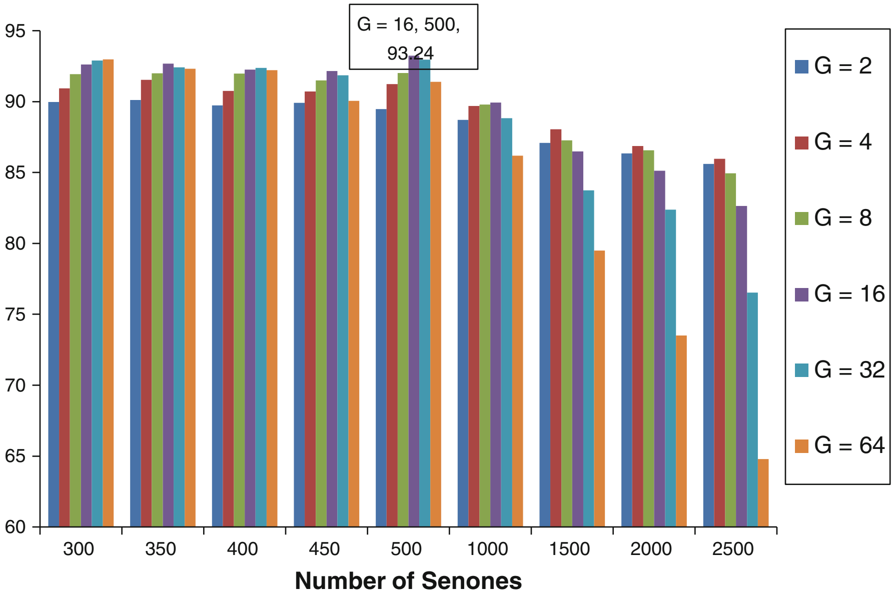

Fig. 5 Word recognition correctness rates (%) in reference to number of senones and Gaussians

4.2.2 Modifications using basic parameters at testing/decoding level

For performance optimization, a modified version of the decoder is used in order to identify possible combinations of Word Insertion Penalty (WIP) ranging between 0.2 and 0.7, Language Model Weight (LW) ranging between 8 and 11, and Beam Pruning (Beam) ranging between 1.e-40 and 1.e-85 that yield a higher word recognition correctness rate and lower WER compared to what the standard decoder could achieve. As a result, 160 iterations of the decoder were required at this initial stage. However, it is found that the ranges are too broad and some results are even worse than what the standard decoder used to achieve. Therefore, the Word Insertion Penalty (WIP) is now ranged between 0.4 and 0.7, Language Model Weight (LW) remained the same ranging between 8 and 11, and Beam Pruning (Beam) is fixed to be 1.e-85.

Table 15 Performance of Arabic ASR systems at training level for all data sets

Experiment_ID	Same speakers with different sentences WER (%)	Different speakers with same sentences WER (%)	Different speakers with different sentences WER (%)
Experiment 1	10.73	6.28	13.48
Experiment 2	11.96	10.62	27.87
Experiment 3	10.53	3.66	14.94
Experiment 4	11.42	4.16	11.76
Experiment 5	10.09	7.13	14.86
Experiment 6	11.56	7.37	14.23
Experiment 7	11.15	4.51	14.25
Experiment 8	12.75	2.51	13.31
Average results	11.27	5.78	15.59

Table 16 Systems' performance at testing/decoding level after performance optimization

Experiment_ID	Same speakers with different sentences WER (%)	Different speakers with same sentences WER (%)	Different speakers with different sentences WER (%)
Experiment 1	9.59	5.14	9.44
Experiment 2	10.50	8.65	23.43
Experiment 3	8.96	3.00	10.88
Experiment 4	9.65	2.81	9.27
Experiment 5	8.54	5.82	11.96
Experiment 6	9.77	5.36	11.42
Experiment 7	9.37	3.89	11.85
Experiment 8	11.22	1.94	10.87
Average results	9.70	4.58	12.39

New optimized results of the WER are presented in Table 16. It clearly shows that lower WER are achieved by the new modified decoder. The impact of the modified decoder on the WER is clearly shown in Fig. 6.

4.3 Overall experimental results analysis

Based on the experimental work, it is advisable to try different combination of parameters in order to identify the best combination that is more suitable to the data used in order to optimize the performance.

The modified decoder used at testing level using different combination of Word Insertion Penalty (WIP), Language Model Weight (LW), and Beam Pruning (Beam), obtained better performance than the standard CMU Sphinx 3 decoder. Therefore, it is important to look for the best combination of such key parameters in

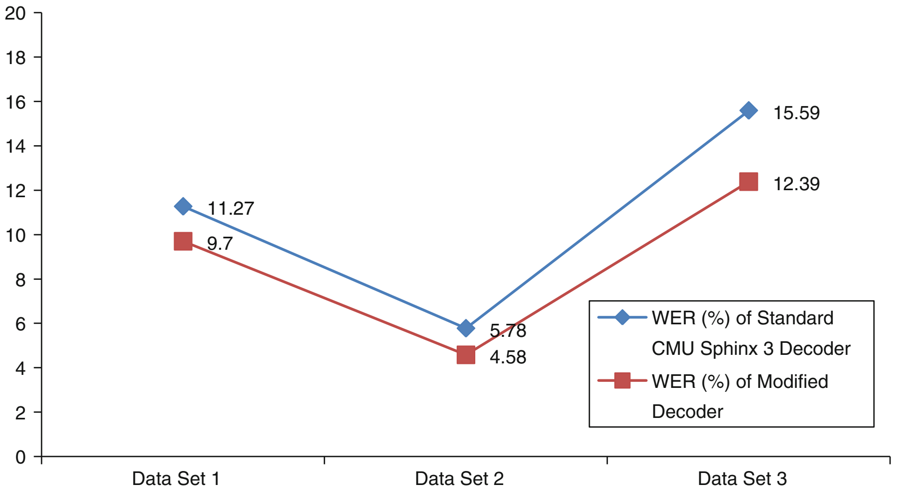

Fig. 6 Comparisons of systems' performance in terms of the WER (%) using standard CMU Sphinx 3 decoder and a modified decoder at testing/decoding level

order to enhance the performance of the decoder and obtain better performance compared with the standard version based on default values of the parameters.

Speaker-independence is highly realized. If we refer to Table 16, we can see that for same speakers with different sentences, the systems obtained an average WER of 9.70%, whereas for different speakers with different sentences they obtained an average WER of 12.39%. This is important due to the fact that speech recognition systems must adhere to the differences between speakers. Obviously not all potential users can be used in training, therefore, the systems must be able to adapt to users who are not being used in training the systems. In our work, as we added more data to train the systems, it is realized that the systems become more speaker-independent and they could perform similar to those speakers used in training the systems.

The systems performance is expected to improve further once our speech corpus is fully utilized due to the fact that training data play very crucial role in enhancing and improving the performance of speech recognition systems as they are considered as the major contributor to better systems' performance.

It is important to highlight that our phonetically rich and balanced speech corpus is able to have positive impact on the performance of our automatic continuous speech recognition systems for Arabic language. It is believed that this corpus will have more impact when fully used in our research. This is due to its uniqueness compared to other speech corpora such as broadcast news corpus, since participating speakers have fair distribution of age and gender, vary in terms of educational backgrounds, belong to various native Arabic speaking countries, and belong to the three major regions where Arabic native speakers are situated. This speech corpus can be used for Arabic speech based applications including speaker recognition and TTS synthesis, covering different research needs. Table 17 shows a brief comparison between our Arabic ASR systems' performance and the state-of-the-art research efforts on Arabic ASR systems.

Table 17 Performance comparison of state-of-the-art Arabic ASR research efforts

General domains	Source	Main task	Algorithms	Tools	Speech data		Speaker dependency	Correctness rate (%)	Accuracy rate (%)	WER (%)
					Training	Testing				
Arabic digits recognition	Hyassat and Abu Zitar (2008)	Arabic digits (0–9)	Hidden Markov Models (HMM)	CMU Sphinx-IV	1,213 tokens	143 tokens	Speaker-Independent	N/A	99.21	0.78
	Satori et al. (2007)	Arabic digits (0–9)	Hidden Markov Models (HMM)	CMU Sphinx-IV	300 tokens	N/A	Speaker-Dependent	85.56% (Males) 83.34% (Females)	N/A	N/A
	Alotaibi (2008)	Arabic digits (0–9)	Artificial Neural Networks (ANN)	N/A	340 tokens	1,700 tokens	Speaker-Dependent	99.50%	N/A	N/A
					400 tokens	1,300 tokens	Speaker-Independent	94.50%		
	Alotaibi (2008)	Arabic digits (0–9)	Hidden Markov Models (HMM)	HTK	340 tokens	1,700 tokens	Speaker-Dependent	98.10%	N/A	N/A
					400 tokens	1,300 tokens	Speaker-Independent	94.80%		
	Alotaibi et al. (2008)	Arabic digits (0–9)	Hidden Markov Models (HMM)	HTK	11,610 tokens	3,870 tokens	Speaker-Independent	93.72%	N/A	N/A

Table 17 continued

General domains	Source	Main task	Algorithms	Tools	Speech data		Speaker dependency	Correctness rate (%)	Accuracy rate (%)	WER (%)
					Training	Testing				
Broadcast News	Alghamdi et al. (2009)	Arabic broadcast news	Hidden Markov Models (HMM)	CMU Sphinx-III and HTK	7.0 h	½ h = 400 tokens	Speaker-Independent	90.78%	N/A	10.87
	Solatu et al. (2007)	Arabic Broadcast News	Hidden Markov Models (HMM)	N/A	About 2,000 h from BBN and LDC	About 13 h from BBN and LDC and others	Speaker-Independent	N/A	N/A	Average of 30.15
	Solatu et al. (2007)	Arabic Broadcast News	Hidden Markov Models (HMM)	N/A	About 2,000 h from BBN and LDC	About 13 h from BBN and LDC and others	Speaker-Dependent Un-Vowelized	N/A	N/A	Average of 20.38
							Speaker-Dependent Vowelized			Average of 18.73
	Messaoudi et al. (2006)	Arabic Broadcast News	Hidden Markov Models (HMM)	N/A	About 150 h	About 75 min	Speaker-Independent Un-Vowelized	N/A	N/A	16.00%
							Speaker-Independent Vowelized			14.80

Table 17 continued

General domains	Source	Main task	Algorithms	Tools	Speech data		Speaker dependency	Correctness rate (%)	Accuracy rate (%)	WER (%)
					Training	Testing				
Command and control	Hyassat and Abu Zitar (2008)	Recognizing Arabic computer based command and control terms	Hidden Markov Models (HMM)	CMU Sphinx-IV	5,628 tokens	372 tokens	Speaker-Independent	N/A	98.18	1.81
Qur'an	Hyassat and Abu Zitar (2008)	Recognizing verses of the Qur'an	Hidden Markov Models (HMM)	CMU Sphinx-IV	18.35 h	N/A	N/A	N/A	70.81	N/A
	Mourtaga et al. (2007)	Recognizing verses of the Qur'an from 5 famous readers	Hidden Markov Models (HMM)	HTK	2,431 tokens	N/A	Speaker-Independent	N/A	68–85% Average of 78.60	N/A

5 Conclusions

This paper reports our work towards building a phonetically rich and balanced MSA speech corpus, which is necessary for developing speaker-independent, automatic, and continuous Arabic ASR systems. This work includes creating the phonetically rich and balanced speech corpus with full diacritical marks transcription from different speakers with different varieties of attributes, and making all preparation and pre-processing steps in order to produce a ready-to-use speech data for further training and testing purposes of Arabic ASR systems.

Based on our literature investigation, majority of Arabic spoken resources are collected from broadcast news or telephone conversations. However, they lack generality, variability among speakers, and quality. From industrial and academia perspectives, available spoken corpora are also lacking in various aspects covering adaptability, reusability, quality, coverage, and adequate information types.

Language resources need to cover important categories related to gender, age, region, class, education, occupation, and others in order to provide an adequate representation of the subjects, which are not considered in many available Arabic spoken resources.

This work adds a new variety of possible speech data for Arabic language based text and speech applications besides other varieties such as broadcast news and telephone conversations.

The newly developed phonetically rich and balanced MSA speech corpus has a total of about 50 h of high quality speech, which are collected from 40 native speakers differing in gender, age, country, geographical region, profession, educational background, and mastery of Arabic language. Based on our experience with this corpus, it really bridges the gap between the available spoken resources and the industrial and academia expectations as depicted from our literature investigation.

This speech corpus is not publically available yet and will hopefully be distributed through proper language resources providers such as the ELRA and LDC. However, interested researchers can contact the corresponding author for distribution details and probably ask for an evaluation portion of the corpus.

Since this phonetically rich and balanced speech corpus contains training and testing written and spoken data of 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, TTS synthesis, speaker recognition, and many others.

Experimental recognition results presented in this paper show that the developed systems are speaker-independent and are highly comparable and better than many reported Arabic ASR research efforts. The systems performance is also expected to improve further once our speech corpus is fully utilized.

In conclusion, the introduction section of this paper clearly states the advantages and disadvantages of the broadcast news and telephone conversation speech corpora. Therefore, our speech corpus is meant to overcome such disadvantages by producing a new variety of speech corpus with high quality having in mind that the training data is considered as the major contributor to highly performing systems. In addition, the experimental section was meant to evaluate part of the corpus. This evaluation reflects on the quality of the speech corpus and promotes this speech corpus as a potential substitute to the available Arabic speech corpora. Although the size of the corpus is about 50 h, which is far less than many Arabic broadcast news corpora, we believe that our corpus managed to perform better than such corpora because it is properly prepared and recorded with clear goals. In addition, even though 40 speakers maybe considered small to achieve speaker-independent systems, our study shows it can be achieved largely because the training data is phonetically rich and balanced. Our study also shows that using only 8 h of our speech corpus can produce speaker-independent systems. When using the entire corpus, speaker-independence will certainly improve further. This work also emphasizes on the relationship between the written and spoken corpora. In many cases, the available corpora are reverse engineered. In other words, in the case of broadcast news in many cases the speech corpus is collected then only transcribed and produced in its written form. This shows that such corpora are not properly prepared and recorded. Therefore, we produced a new corpus (although small as some might argue), but the corpus (even small portion of it) is able to produce ASR systems with highly impressive and competitive performance compared with the available corpora. This in summary forms the hypothesis of our work.

Acknowledgments We would like to extend our appreciation to University of Malaya and University of Jordan for funding this research work.

References

Abu Shariah, M. A. M., Ainon, R. N., Zainuddin, R., & Khalifa, O. O. (2007). Human computer interaction using isolated-words speech recognition technology. In: *Proceedings of the IEEE international conference on intelligent and advanced systems (ICIAS'07)* (pp. 1173–1178). Kuala Lumpur, Malaysia.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Al-Qatab, B. A., & Alqudah, A. A. M. (2010d). Impact of a newly developed modern standard Arabic speech corpus on implementing and evaluating automatic continuous speech recognition systems. In *Proceedings of the second international workshop on spoken dialogue systems technology (IWSDS'10)* (Lecture Notes in Computer Science (LNCS)) (Vol. 6392, pp. 1–12). Springer.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Alqudah, A. A. M., Elshafei, M. A., & Khalifa, O. O. (2011). Modern standard Arabic speech corpus for implementing and evaluating automatic continuous speech recognition systems. *Journal of the Franklin Institute*. Elsevier. doi: 10.1016/j.jfranklin.2011.04.011.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2010b). Natural speaker-independent Arabic speech recognition system based on Hidden Markov models using Sphinx tools. In *Proceedings of the IEEE international conference on computer and communication engineering (ICCCE'10)*. Kuala Lumpur, Malaysia.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2010c). Phonetically rich and balanced speech corpus for Arabic speaker-independent continuous automatic speech recognition systems. In *Proceedings of the IEEE 10th international conference on information science, signal processing and their applications (ISSPA 2010)* (pp. 65–68). Kuala Lumpur, Malaysia.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Khalifa, O. O., & Elshafei, M. (2010a). Phonetically rich and balanced Arabic speech corpus: An overview. In *Proceedings of the IEEE international conference on computer and communication engineering (ICCCE'10)*. Kuala Lumpur, Malaysia.

Alansary, S., Nagi, M., & Adly, N. (2007). Building an international corpus of Arabic (ICA): Progress of compilation stage. *8th international conference on language engineering*, Egypt.

Alghamdi, M., Alhamid, A. H., & Aldasuqi, M. M. (2003). Database of Arabic sounds: sentences. *Technical Report*, Saudi Arabia: King Abdulaziz City of Science and Technology (in Arabic).

Alghamdi, M., Basalamah, M., Seeni, M., & Husain, A. (1997). Database of Arabic sounds: words. In *Proceedings of the 15th National computer conference* (pp. 797–815). Saudi Arabia (in Arabic).

Alghamdi, M., Elshafei, M., & Al-Muhtaseb, H. (2009). Arabic broadcast news transcription system. *International Journal of Speech Technology*. Springer, 183–195.

Ali, M., Elshafei, M., Alghamdi, M., Almuhtaseb, H., & Al-Najjar, A. (2008). Generation of Arabic phonetic dictionaries for speech recognition. In *IEEE proceedings of the international conference on innovations in information technology* (pp. 59–63). UAE.

Alotaibi, Y. A. (2008). Comparative study of ANN and HMM to Arabic digits recognition systems. *Journal of King Abdulaziz University: Engineering Sciences, 19*(1), 43–59.

Alotaibi, Y. A., Alghamdi, M., & Alotaiby, F. (2008). Using a telephony Saudi accented Arabic corpus in automatic recognition of spoken Arabic digits. *4th international symposium on image/video communications over fixed and mobile networks (ISIVC08)*, Bilbao, Spain.

Alotaibi, Y. A., & Meftah, A. H. (2010). Comparative evaluation of two Arabic speech corpora. In *IEEE proceedings of the international conference on natural language processing and knowledge engineering*, Beijing, China.

Al-Sulaiti, L., & Atwell, E. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*. John Benjamins Publishing Company, pp. 1–36.

Bakis, R. (1976). Continuous speech recognition via centisecond acoustic states. *The Journal of the Acoustical Society of America, 59*(S1), S97.

Black, A. W., & Tokuda, K. (2005). The Blizzard Challenge—2005: Evaluating corpus-based speech synthesis on common datasets. *INTERSPEECH'05* (pp. 77–80). Portugal.

Chan, A., Gouv^ea, E., Singh, R., Ravishankar, M., Rosenfeld, R., Sun, Y. et al. (2007). The Hieroglyphs: building speech applications using CMU Sphinx and related resources. http://www-2.cs.cmu.edu/~archan/documentation/sphinxDocDraft3.pdf. Accessed on 15 September 2010.

Chou, F. C., & Tseng, C. Y. (1999). The design of prosodically oriented Mandarin speech database. *ICPhS'99* (pp. 2375–2377), San Francisco.

Cieri, C., Liberman, M., Arranz, V., & Choukri, K. (2006). Linguistic data resources. In T. Schultz & K. Kirchhoff (Eds.), *Multilingual speech processing* (pp. 33–70). USA: Academic Press, Elsevier.

Clarkson, P., & Rosenfeld, R. (1997). Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of the 5th European conference on speech communication and technology* (pp. 2707–2710), Rhodes, Greece.

D'Arcy, S., & Russell, M. (2008). Experiments with the ABI (Accents of the British Isles) Speech Corpus. *INTERSPEECH'08* (pp. 293–296), Australia.

Elmahdy, M., Gruhn, R., Minker, W., & Abdennadher, S. (2009). Survey on common Arabic language forms from a speech recognition point of view. *International conference on acoustics (NAG-DAGA)* (pp. 63–66), Rotterdam, Netherlands.

ELRA. (2005). *NEMLAR broadcast news speech corpus*. catalogue Reference S0219. http://catalog.elra.info/product_info.php?products_id=874. Accessed on 10 May 2011.

Elshafei, A. M. (1991). Toward an Arabic text-to-speech system. *The Arabian Journal of Science and Engineering, 16*(4B), 565–583.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). *DARPA TIMIT acoustic-phonetic continuous speech corpus*. University Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*. USA: Morgan and Claypool Publishers.

Hong, H., Kim, S., & Chung, M. (2008). Effects of Allophones on the performance of Korean speech recognition. *INTERSPEECH'08* (pp. 2410–2413), Australia.

Hyassat, H., & Abu Zitar, R. (2008). Arabic speech recognition using SPHINX engine. *International Journal of Speech Technology*. Springer, 133–150.

Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Ji, G. et al. (2003). Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop. *ICASSP'03* (Vol. 1, pp. 344–347), Hong Kong.

Liang, M. S., Lyu, R. Y., & Chiang, Y. C. (2003). An efficient algorithm to select phonetically balanced scripts for constructing a speech corpus. In *IEEE Proceedings of the international conference on natural language processing and knowledge engineering* (pp. 433–437), China.

Madi, M., (2010). A study of Arabic letter frequency analysis. http://www.intellaren.com/articles/en/a-study-of-arabic-letter-frequency-analysis. Accessed on 6 June 2011.

Meeralam, Y. (2007). *Contributions of cryptography scholars in Arabic linguistics*. Diwan al Arab. http://www.diwanalarab.com/IMG/pdf/Is_hamaatUolamaaAltaumieat1-1.pdf. Accessed on 10 May 2011 (in Arabic).

Messaoudi, A., Gauvain, J. L., & Lamel, L. (2006). Arabic broadcast news transcription using a one million word vocalized vocabulary. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP'06)* (pp. 1093–1096), Toulouse, France.

Mourtaga, E., Sharieh, A., & Abdallah, M. (2007). Speaker independent Quranic recognizer based on maximum likelihood linear regression. In *Proceedings of world academy of science, engineering and technology* (Vol. 36, pp. 61–67), Brazil.

Nikkhou, M., & Choukri, K. (2004). *Survey on industrial needs for language resources*. Technical Report, NEMLAR—Network for Euro-Mediterranean Language Resources.

Nikkhou, M., & Choukri, K. (2005). *Survey on Arabic language resources and tools in the mediterranean countries*. Technical Report, NEMLAR—Network for Euro-Mediterranean Language Resources.

Parkinson, D. B., & Farwaneh, S. (Eds.). (2003). *Perspectives on Arabic linguistics XV* (pp. 149–180). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Pineda, L. V., Montes-y-Gómez, M., Vaufreydaz, D., & Serignat, J. -F. (2004). Experiments on the construction of a phonetically balanced corpus from the web. In *5th international conference on computational linguistics and intelligent text processing* (Lecture Notes in Computer Science, Springer) (Vol. 2945/2004, pp. 416–419) Korea.

Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B. et al. (1997). The 1996 Hub-4 Sphinx-3 System. In *Proceedings of the 1997 ARPA speech recognition workshop* (pp. 85–89).

Rabiner, L. R. (1989). A tutorial on Hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Roberts, A., Al-Sulaiti, L., & Atwell, E. (2006). aConCorde: Towards an open-source, extendable concordancer for Arabic. *Corpora Journal, 1*(1), 39–60.

Satori, H., Harti, M., & Chenfour, N. (2007). Arabic speech recognition system based on CMUSphinx. In *IEEE proceedings of ISCIII'07* (pp. 31–35) Morocco.

Siemund, R., Heuft, B., Choukri, K., Emam, O., Maragoudakis, E., Tropf, H. et al. (2002). OrienTel—Arabic speech resources for the IT market. In *Proceedings of the 3rd international conference on language resources and evaluation (LREC'02)*, Spain.

Solatu, H., Saon, G., Kingsbury, B., Kuo, J., Mangu, L., Povey, D. et al. (2007). The IBM 2006 GALE Arabic ASR system. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP'07)* (pp. 349–352), Hawaii, USA.

Uraga, E., & Gamboa, C. (2004). VOXMEX speech database: Design of a phonetically balanced corpus. In *Proceedings of the 4th international conference on language resources and evaluation* (pp. 1471–1474), Portugal.

Vergyri, D., & Kirchhoff, K. (2004). Automatic Diacritization of Arabic for acoustic modeling in speech recognition. In *Proceedings of the workshop on computational approaches to Arabic script-based languages* (pp. 66–73) Geneva, Switzerland.

Wikipedia. (2011). *IPA for Arabic*. http://en.wikipedia.org/wiki/Wikipedia:IPA_for_Arabic. Accessed on 10 May 2011.