Int J Speech Technol (2017) 20:261–280
DOI 10.1007/s10772-017-9403-7

TAMEEM V1.0: speakers and text independent Arabic automatic continuous speech recognizer

Mohammad A. M. Abushariah[1]

Received: 23 April 2016 / Accepted: 30 January 2017 / Published online: 24 February 2017

Abstract This research work aims to disseminate the efforts towards developing and evaluating TAMEEM V1.0, which is a state-of-the-art pure Modern Standard Arabic (MSA), automatic, continuous, speaker independent, and text independent speech recognizer using high proportion of the spoken data of the phonetically rich and balanced MSA speech corpus. The speech corpus contains speech recordings of Arabic native speakers from 11 Arab countries representing Levant, Gulf, and Africa regions of the Arabic World, which make about 45.30 h of speech data. The recordings contain about 39.28 h of 367 sentences that are considered phonetically rich and balanced, which are used for training TAMEEM V1.0 speech recognizer, and another 6.02 h of another 48 sentences that are used for testing purposes, which are mostly text independent and foreign to the training sentences. TAMEEM V1.0 speech recognizer is developed using the Carnegie Mellon University (CMU) Sphinx 3 tools in order to evaluate the speech corpus, whereby the speech engine uses three-emitting state Continuous Density Hidden Markov Model for tri-phone based acoustic models, and the language model contains uni-grams, bi-grams, and tri-grams. Using three different testing data sets, this work obtained 7.64% average Word Error Rate (WER) for speakers dependent with text independent data set. For speakers independent with text dependent data set, this work obtained 2.22% average WER, whereas 7.82% average WER is achieved for speakers independent with text independent data set.

✉ Mohammad A. M. Abushariah
m.abushariah@ju.edu.jo; m.abushariah@gmail.com

[1] Department of Computer Information Systems, King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan

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

1 Introduction

Automatic Speech Recognition (ASR) refers to the process of converting human speech into text for a particular language. This technology has been rapidly evolving and widely spreading for many languages all over the world. It is found nowadays for Arabic, English, Mandarin, Spanish, Persian, Korean, Japanese, Urdu, and many other languages. Personal computers, laptops, and mobile devices nowadays have the ASR support. In addition, there are many research efforts being conducted addressing this technology worldwide and more researchers have indulged into ASR research initiatives in the past decade, which indicates that the ASR research community is expanding and evolving.

Like many other languages, Arabic language has been considered for ASR research by researchers in the Arab World as well as the entire world. This consideration is due to the importance of Arabic language, whereby it is the largest Semitic language which is still in existence and one of the six official languages of the United Nations (UN). The number of Arabic native speakers exceeds 250 million, whereas the number of Arabic non-native speakers can reach four times the number of native 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 native speakers. In addition, there are three main forms of Arabic language namely Classical Arabic (CA), Modern Standard Arabic (MSA),

and Dialectal Arabic (DA), whereby MSA is the current formal linguistic standard of Arabic language, which is widely taught in schools and universities, and often used in the office and the media, and it is the only acceptable form of Arabic language for all native speakers (Elmahdy et al. 2009a). In spite of its importance, its research is lacking in many aspects and research effort on Arabic ASR still requires more emphasis worldwide.

Research on Arabic ASR requires written and spoken language resources and corpora that are not readily available. Various important surveys were conducted to explore the need for Arabic language resources and tools (Nikkhou and Choukri 2004, 2005). Such surveys motivated the researcher to develop new Arabic language resources that are phonetically rich and balanced, whereby 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 (Abushariah et al. 2012b). This approach is highly 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). As far as Arabic language is concerned, text and speech resources and processing are rather limited. This research work aims to deal with Arabic ASR based on phonetically rich and balanced text and speech resources. Therefore, it is important to explore this approach in order to find its efficiency for Arabic ASR research.

The word تميم/ *TAMEEM*/ is an Arabic word, which means complete, perfect, and outright, whereby TAMEEM V1.0 is the first stable version of my efforts towards developing a state-of-the-art pure Modern Standard Arabic (MSA), automatic, continuous, speaker independent, and text independent speech recognizer using spoken data of the phonetically rich and balanced MSA speech corpus, which was explained in Abushariah et al. (2012b). My previous research works and efforts were based on portions of the phonetically rich and balanced MSA speech corpus (Abushariah et al. 2012a, b, c).

The following section, Sect. 2, presents a literature investigation for Arabic language and research efforts conducted for Arabic Automatic Speech Recognition (ASR). Sect. 3 provides all design and implementation details of TAMEEM V1.0 speakers and text independent Arabic continuous speech recognition system, whereas the details about the phonetically rich and balance Modern Standard Arabic (MSA) speech corpus that is used to develop TAMEEM V1.0 are highlighted in Sect. 4. Experimental results and analysis are stated and analyzed in Sect. 5. The conclusions are finally presented in Sect. 6.

2 Arabic language and automatic speech recognition

Research interests have grown significantly in the past decade for Arabic ASR research in accordance to the increase and improvement in the performance of ASR systems for many other languages including English, Spanish, Mandarin, and many others. In addition, the availability of open source tools such as Carnegie Mellon University (CMU) Sphinx 3 and Pocketsphinx tools, the Hidden Markov Model Toolkit (HTK) produced by Cambridge University, and many others have accelerated the progress of the development of ASR systems in many languages including Arabic language.

In this section, Arabic language forms are explained and their major differences are summarized. Arabic language written and spoken resources are highlighted and their needs are also addressed. In addition, taxonomy of contributions and initiatives of the research community towards Arabic ASR research are investigated, whereas the available open source software and tools used for Arabic language ASR research are also identified.

2.1 Arabic language forms

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; Elmahdy et al. 2009a, b; Alotaibi and Meftah 2010; Abushariah 2012). Al-Sulaiti and Atwell (2006) believed that there is another form of Arabic language 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. This section discusses the characteristics of the three main forms in more details and provides key differences between them.

The first form of Arabic is Classical Arabic (CA), which is treated as the most formal and standard form of Arabic mainly because it is the language of the Qur'an, religious instructions of Islam, and classical literature. It is also referred to as the Qur'anic Arabic language and the parent language of all varieties of spoken Arabic (Elmahdy et al. 2009a, b; Al-Sulaiti and Atwell 2006; Alotaibi and Meftah 2010). CA scripts are fully vowelized and include all diacritical marks, therefore, phonetics of the word are completely represented. According to Elmahdy et al. (2009a), phonetics represented in the CA scripts include all original and basic sounds of MSA—28 original consonants and six vowels—as described in the next section, with some additional sounds that can be clearly found in the recitation of the Qur'an such as the vowel prolongation, nasalization,

shaking, merging, hiding, and many others (Abushariah 2012).

Detailed information of the rules for vowels and consonants in CA and Tajweed for the proper recitation of the Qur'an can be found in Elshafei (1991), and Harrag and Mohamadi (2010). From speech and speaker recognition perspectives, it is noticed that the Qur'anic language is very limited to applications that help learning the proper recitation of the Qur'an by correctly embedding the Tajweed rules, and also for identifying the correct reciters and recitation styles (Qiraat) (Abushariah 2012). Research efforts on Qur'anic language are further investigated in Sect. 2.3.

The second form of Arabic is Modern Standard Arabic (MSA), which is the current formal linguistic standard of Arabic language, which is widely taught in schools and universities, used in the office, the media, newspapers, formal speeches, courtrooms, and any kind of formal communication (Elmahdy et al. 2009a; Alotaibi and Meftah 2010). As classified by Elmahdy et al. (2009a), MSA is the only acceptable form of Arabic language for all native speakers, where its spoken form can be understood by all native speakers. Habash (2010) and Alotaibi (2010) agreed that there is a tight relationship between CA and MSA forms, where the latter is syntactically, morphologically, and phonologically based on the earlier. However, MSA is a lexically more modernized version of CA (Abushariah 2012).

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. 2009a; 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 (Abushariah 2012).

Similar to CA, MSA scripts contain 34 basic sounds—28 original consonants and six vowels—as agreed by most Arabic language researchers. However, Elmahdy et al. (2009a, b) have gone further to include four additional sounds, which they consider them as foreign and rare consonants. As a result, a total of 38 sounds are introduced. The four foreign and rare consonants which include /g/, /p/, /v/, and /l`/ are normally grouped together with the closest consonants and not considered as extra sounds, for instance both /f/ and /v/ are grouped together, and similarly the case for /b/ and /p/. From language resources perspective, MSA spoken and written resources are mostly available from broadcast news due to the low price and the ease for collection. In addition, since MSA is the only acceptable form of Arabic language for all native speakers, it becomes the main focus of this work and current Arabic ASR research efforts (Elmahdy et al. 2009a; Abushariah 2012).

The third form of Arabic is Dialectal Arabic (DA) or Colloquial Arabic, which 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. It is important to mention that neither CA nor MSA forms can be treated as the natural spoken language for all Arabic native speakers. From writing and publishing perspectives, DA cannot be used as a standard form of Arabic language and does not have any commonly accepted standard for the writing system, because each dialect has its own characteristics that can be different from all other dialects and even from the MSA form, which affect the pronunciation, phonology, vocabulary, morphology, and syntax of Arabic language (Newman 2002; Kirchhoff et al. 2003; Elmahdy et al. 2009a, b; Alotaibi 2010; Abushariah 2012).

Although there are many dialects for Arabic language, researchers mostly categorize them into two major categories namely (1) Western Arabic, which includes the Moroccan, Tunisian, Algerian, and Libyan dialects, and (2) Eastern Arabic, which includes the Egyptian, Gulf, and Levantine dialects (Haraty and El Ariss 2007; Elmahdy et al. 2009a, b). From language resources and speech recognition perspectives, DA language resources are mostly collected from telephone conversations and certain broadcast news such as the CallHome and the OrienTel collections. However, such efforts still suffer from high word error rates (WER), which can reach 56–61% (Canavan et al. 1997; Siemund et al. 2002; Kirchhoff et al. 2003; Cieri et al. 2006; Abushariah 2012). Table 1 summarizes the major differences between the CA, MSA, and DA forms of Arabic language.

Being the formal standard linguistic form of Arabic language, the ability to understand it, and its acceptability by all native speakers, this research work has selected MSA as the main form of Arabic language for making the written and spoken resources, which are then used for developing and evaluating the speakers and text independent TAMEEM V1.0 speech recognizer.

2.2 Arabic language written and spoken resources

Written and spoken corpora are examples of linguistic resources for a language, which normally consist of large sets of machine readable data that are used for developing, improving, and evaluating natural language, and speech algorithms and systems. Advancements in these technologies elevated the need by many communities for written and spoken resources in large volumes with relatively different types of data and variety of languages (Godfrey and Zampolli 1997; Ejerhed and Church 1997; Lamel and Cole 1997; Cieri et al. 2006).

Table 1 Summary of major differences between all Arabic language forms (Abushariah 2012)

Arabic language form	Coverage of speakers	Diacritical representation	Structures and standards for written representation	Alphabetical differences	Availability of written and spoken resources
Classical Arabic (CA)	CA is a comprehensive form of Arabic language that is usually understood by all native speakers	Arabic diacritics are fully represented in the CA scripts	CA has structures and standards for its written scripts	CA has a set of standard consonants and vowels	CA language resources (spoken and written) can be collected mostly from the Qur'an, books and media on religious instructions of Islam and classical literature
Modern Standard Arabic (MSA)	Similar to CA, MSA is also a comprehensive form of Arabic language that can be understood by all native speakers	Arabic diacritics are normally omitted or partially represented in the MSA scripts	Similar to CA, MSA has structures and standards for its written scripts	MSA also has a set of standard consonants and vowels The four foreign and rare consonants are grouped with the closest consonants	MSA language resources (spoken and written) can be collected mostly from broadcast news, textbooks, newspapers, formal speeches, and various other formal means
Dialectal Arabic (DA)	DA is very limited to a specific group, country, or region and cannot be treated as a comprehensive form of Arabic language	Arabic diacritics are mostly omitted in the DA scripts	There are no structures and standards for the written scripts of DA	There are differences in the way some consonants are spoken such as: /t/ and /s/ are used to replace /θ/ in MSA, /g/ is used to replace /ʤ/ in MSA, /ʔ/ is used to replace /q/ in MSA and various other differences	DA language resources (spoken and written) can be collected mostly from telephone conversations, some broadcast news, movies, series, and other related forms

Depending on the type of data to be collected and the application to be developed, the written corpus can be produced prior to the spoken corpus or vice-versa. However, both the written and spoken forms are closely related and very necessary to exist in order to develop any ASR system. Spoken corpora contain signals that correspond to the pronunciation of utterances by various speakers, which are used to develop the acoustic models in ASR systems. On the other hand, the written corpora contain texts that correspond to the utterances pronounced in the spoken corpora, which are used to develop the language model in ASR systems. For instance, the written corpora must be prepared prior to the spoken corpora in read speech, whereas in conversational speech the spoken corpora are normally produced first and the written corpora are transcribed either manually or using semi-automatic approaches (Mariani 1995; Ejerhed and Church 1997; Lamel and Cole 1997).

Since the written and the spoken forms are closely related and either form of the corpus can come first, ASR systems (the focus of this work) require large volumes of the spoken form. Therefore, for the purpose of this section, the spoken data type is given more emphasis, because the written corpora can be transcribed manually or using semi-automatic approaches as stated earlier. As a result, the type and contents of the written corpora are dependent on and determined by the type and contents of the spoken corpora.

Based on this assumption, Jorschick (2009) identified four main speech styles of the corpus that are determined by the task used to collect the data. These four categories of speech styles, which are (1) read speech, (2) elicited experimental speech, (3) semi-spontaneous monologue speech, and (4) conversational speech, each of which contains a range of tasks and speech styles that can often overlap. In the case of Arabic language, written and spoken resources are mostly collected from broadcast news and telephone conversations, and publically available to all communities through membership subscription to the Linguistic Data Consortium (LDC) and the European Language Resources Association (ELRA) online catalogs. Some of these resources are available in large volumes especially those collected from broadcast news (Abushariah 2012).

Based on the language archives as summarized by Open Language Archives Community (OLAC 2016a), the LDC online language resources catalog contains 701 language

resources that serve 91 distinct languages, from 1993 until December 21st, 2016. Among the total of 186 Arabic language resources, 57 are spoken corpora, and 129 are written corpora. Majority of the written corpora are in MSA form, whereas 38 and 19 spoken corpora are in DA and MSA forms, respectively. Similar to the summary of the language archives for LDC, the OLAC (2016b) provides a summary on the ELRA online language resources catalog, which contains 1062 language resources that serve 63 distinct languages, from 1995 until February 10th, 2015. Among the total of 48 Arabic language resources provided by ELRA, 19 are spoken corpora, and 29 are written corpora. The written corpora are mostly in MSA form, whereas the spoken corpora are 10 and 9 resources for DA and MSA forms, respectively. However, there are five out of the nine MSA based spoken corpora, which can be considered as DA too as they are the products of OrienTel project (Siemund et al. 2002) that seeks to collect MSA data as spoken in certain countries. These five OrienTel spoken MSA corpora are collected from Jordan, Egypt, Morocco, Tunisia, and United Arab Emirates.

Based on the language resources catalogs provided by the LDC and the ELRA as summarized above, there are about 48 (38 from LDC, and 10 from ELRA) and 28 (19 from LDC, and 9 from ELRA) spoken corpora for DA and MSA forms, respectively. There are also about 158 (129 from LDC, and 29 from ELRA) written corpora. This analysis indicates that the written corpora for Arabic language are available in large volumes. However, there is real lack of spoken corpora especially for MSA form. Therefore, this work empathizes on providing written and spoken corpora for MSA form. There is a large number of available corpora especially those for DA form, and due to the scope of this work, Table 2 is devoted for providing general information on the LDC and ELRA available spoken corpora for MSA form and accented MSA from 2002 until 2016 as claimed by their developers.

2.3 Taxonomy of Arabic ASR research efforts

Arabic ASR research has witnessed significant demands and research efforts worldwide in the last decade. Taxonomy of Arabic research efforts was developed by Abushariah (2012), which includes (1) isolated Arabic part of word (consonants, vowels, syllables, phonemes, and phones) recognition systems, (2) isolated Arabic words recognition systems, and (3) continuous Arabic speech recognition systems. The larger the unit of speech to be recognized, the harder is the ASR task. The third category that is continuous Arabic speech recognition systems—the mode of this research- includes very complex systems. As far as Arabic language is concerned, this category includes speech recognition tasks such as The Holy Qur'an, phonetically rich and balanced sentences, proverbs, questions, broadcast news, broadcast conversations, broadcast reports, and telephone conversations. Continuous Arabic speech recognition systems are more complex than other types such as isolated words speech recognition, and require large volumes of data in order to achieve excellent recognition rates. Based on my literature investigation, it is found that the broadcast news, broadcast conversations, broadcast reports, and telephone conversations dominate this category. Tables 3, 4 and 5 provide a detailed comparison for major continuous Arabic speech recognition research efforts.

For this third category, researchers including Tabbal et al., (2006), Mourtaga et al. (2007), Hyassat and Abu Zitar (2008), Abdo et al. (2010), Hafeez et al. (2014), and El Amrani et al. (2016) have contributed to The Holy Qur'an ASR systems for Classical Arabic (CA) using the widely available recordings of famous reciters as shown in Table 3. Some other research efforts including Nofal et al. (2004), Azmi and Tolba (2008), Droua-Hamdani et al. (2010, 2013), and Zarrouk et al. (2014, 2015) are directed towards recognizing Arabic sentences, proverbs, and questions as shown in Table 4. However, majority of the research efforts focused on developing ASR systems that are able to recognize the speech of broadcast news, broadcast conversations, and broadcast reports not only in the Arab world, but also worldwide including Messaoudi et al. (2006), Soltau et al. (2007, 2009), Rybach et al. (2007), Vergyri et al. (2008), Alghamdi et al. (2009), AbuZeina et al. (2011), Nahar et al. (2013, 2016), and Ali et al. (2014) as shown in Table 5. These systems are common today due to the low cost and availability of the data, which is very important to be in large volumes to help in achieving high performance for continuous ASR systems. Similar to broadcast news, broadcast conversations, and broadcast reports, telephone conversations are getting popular nowadays due to the availability of data; however, many of these efforts are focused on DA instead of MSA especially in line with the OrienTel project (Siemund et al. 2002). Important legends for Tables 3, 4 and 5 are as follows:

AM	Acoustic model
AR	Accuracy rate
BC	Broadcast conversations
BN	Broadcast news
BR	Broadcast reports
DBN	Dynamic Bayesian networks
DNN	Deep neural networks
HMM	Hidden Markov model
HTK	Hidden Markov model toolkit
LM	Language model
LVQ	Learning vector quantization
MFCC	Mel-frequency cepstral coefficient
MLP	Multilayer perceptron

Table 2 Summary of the LDC and ELRA available MSA and accented MSA spoken corpora from 2002 until 2016

Name	Catalog reference	Corpus size	Used tasks to elicit speech	Year being available
West Point Arabic Speech	LDC2002S02	11.42 h	Recordings of participants are collected by reciting one prompt from four prompt scripts using microphone	2002
TDT4 Multilingual Broadcast News Speech Corpus	LDC2005S11	N/A	Recordings of broadcast news collected from Voice of America satellite radio and Nile Television	2005
OrienTel Jordan MSA database	ELRA-S0290	N/A, but stored on 1 DVD, so must be <4 GB	Recordings of 15 items using Jordanian fixed and mobile telephone networks. These 15 items include (isolated single digit, sequences of five isolated digits, connected digits, currency money amounts, natural numbers, spelled words, directory assistance utterances such as city name and company name, yes/no questions, application keywords/keyphrases, 1 word spotting phrase, 4 phonetically rich words, 9 phonetically rich sentences, and spontaneous items)	2005
OrienTel Morocco MSA database	ELRA-S0184	N/A, but stored on 1 CD and 1 DVD	Recordings of 15 items using Moroccan fixed and mobile telephone networks. These 15 items are the same items as in the OrienTel Jordan MSA Database	2005
OrienTel Tunisia MSA database	ELRA-S0187	N/A, but stored on 1 CD and 1 DVD	Recordings of 15 items using Tunisian fixed and mobile telephone networks. These 15 items are the same items as in the OrienTel Jordan MSA Database	2005
Arabic Broadcast News Speech	LDC2006S46	10 h	Recordings of broadcast news collected from Voice of America satellite radio during transmission time	2006
NEMLAR Broadcast News Speech Corpus	ELRA-S0219	40 h	Recordings of broadcast news and interviews are collected from four different radio stations, which are (1) Medi1, (2) Radio Orient, (3) RMC – Radio Monte Carlo, and (4) RTM – Radio Television Maroc	2006
NEMLAR Speech Synthesis Corpus	ELRA-S0220	10 h	Recordings were collected from 2 native Egyptian Arabic speakers, whereby the speakers read 2032 prompted sentences covering approximately 42,000 words in three categories: transcribed speech (6600 words—20%), written text (16,500 words—50%), and constructed phrases (10,300—30%)	2006
GlobalPhone Arabic	ELRA-S0192	450 h for all the 22 Languages, about 2 GB per language	For Arabic portion of the corpus, recordings are collected by reading about 100 sentences by 78 speakers, which was produced using the Assabah newspaper	2006
OrienTel Egypt MSA database	ELRA-S0222	N/A, but stored on 1 CD and 1 DVD	Recordings of 15 items using Egyptian fixed and mobile telephone networks. These 15 items are the same items as in the OrienTel Jordan MSA Database	2006
Arabic Broadcast News Speech	LDC2006S46	10 h	Recordings from Voice of America satellite radio news broadcasts in Arabic transmitted between June 2000 and January 2001. The corresponding transcripts are available as Arabic Broadcast News Transcripts	2006
NetDC Arabic BNSC (Broadcast News Speech Corpus)	ELRA-S0157	22.5 h	Recordings of broadcast news speech collected from Radio Orient (France)	2007
2003 NIST Rich Transcription Evaluation Data	LDC2007S10	2 h	Recordings of broadcast news and telephone conversations	2007

Table 2 (continued)

Name	Catalog reference	Corpus size	Used tasks to elicit speech	Year being available
OrienTel United Arab Emirates MSA database	ELRA-S0259	N/A, but stored on 2 DVDs	Recordings of 15 items using United Arab Emirates fixed and mobile telephone networks. These 15 items are the same items as in the OrienTel Jordan MSA Database	2007
A-SpeechDB	ELRA-S0315	20 h	Recordings of continuous speech of sentences that cover all Arabic phonetics using microphone in office environment	2011
GALE Phase 2 Arabic Broadcast Conversation Speech Part 1	LDC2013S02	123 h	Recordings were collected for Arabic broadcast conversation speech in 2006 and 2007 by LDC as part of the DARPA GALE (Global Autonomous Language Exploitation) Program	2013
GALE Phase 2 Arabic Broadcast Conversation Speech Part 2	LDC2013S07	128 h	Recordings were collected for Arabic broadcast conversation speech in 2007 by LDC, MediaNet, Tunis, Tunisia and MTC, Rabat, Morocco during Phase 2 of the DARPA GALE (Global Autonomous Language Exploitation) Program	2013
King Saud University Arabic Speech Database	LDC2014S02	590 h	This speech corpus was developed by Speech Group (SG) at King Saud University, which contains recordings of Arabic speech from 269 male and female speakers. The utterances include read and spontaneous speech, which were conducted in various environments including quiet and noisy settings. The corpus is mainly designed for speaker recognition research, but other possible applications include first language recognition, mobile effect, multichannel effect, and use of different type of microphones can make use of this corpus	2014
GALE Phase 2 Arabic Broadcast News Speech Part 1	LDC2014S07	165 h	Recordings were collected for Arabic broadcast news speech in 2006 and 2007 by LDC, MediaNet, Tunis, Tunisia and MTC, Rabat, Morocco during Phase 2 of the DARPA GALE (Global Autonomous Language Exploitation) Program	2014
United Nations Proceedings Speech	LDC2014S08	N/A specifically for Arabic, but 8500 h for all six official UN languages	Recordings were collected for recorded proceedings in the six official UN languages, Arabic, Chinese, English, French, Russian and Spanish. The data was recorded in 2009–2012 from sessions 64–66 of the General Assembly (GA) and First Committee (FC) (Disarmament and International Security), and meetings 6434–6763 of the Security Council	2014
GALE Phase 2 Arabic Broadcast News Speech Part 2	LDC2015S01	170 h	Recordings were collected for Arabic broadcast news speech in 2007 by LDC, MediaNet, Tunis, Tunisia and MTC, Rabat, Morocco during Phase 2 of the DARPA GALE (Global Autonomous Language Exploitation) Program	2015
Arabic Learner Corpus	LDC2015S10	N/A, but audio files are either 44,100 Hz 2-channel or 16,000 Hz 1-channel mp3 files	Recordings of written essays by Arabic learners collected in Saudi Arabia in 2012 and 2013. The corpus includes 282,732 words in 1585 materials, produced by 942 students from 67 nationalities studying at pre-university and university levels. The average length of an essay is 178 words	2015
GALE Phase 3 Arabic Broadcast Conversation Speech Part 1	LDC2015S11	123 h	Recordings were collected for Arabic broadcast conversation speech in 2007 by LDC, MediaNet, Tunis, Tunisia and MTC, Rabat, Morocco during Phase 3 of the DARPA GALE (Global Autonomous Language Exploitation) program	2015

Table 2 (continued)

Name	Catalog reference	Corpus size	Used tasks to elicit speech	Year being available
GALE Phase 3 Arabic Broadcast Conversation Speech Part 2	LDC2016S01	129 h	Recordings were collected for Arabic broadcast conversation speech in 2007 and 2008 by LDC, MediaNet, Tunis, Tunisia and MTC, Rabat, Morocco during Phase 3 of the DARPA GALE (Global Autonomous Language Exploitation) program	2016
GALE Phase 3 Arabic Broadcast News Speech Part 1	LDC2016S07	132 h	Recordings were collected for Arabic broadcast news speech in 2007 by the LDC, MediaNet, Tunis, Tunisia and MTC, Rabat, Morocco during Phase 3 of the DARPA GALE (Global Autonomous Language Exploitation) program	2016
Arabic Speech Corpus	ELRA-S0384	3.7 h	Recordings were collected by one male speaker in South Levantine Arabic (Damascian accent) in a professional studio. The transcript was collected from "Aljazeera Learn" (Aljazeera 2015), a language learning website which was chosen because it contained fully diacritised text which makes it easier to phonetise	2016

MPE Minimum phone error
MPFE Minimum phone frame error
N/A Not available
PLP Perceptual linear prediction
SVM Support vector machine
WER Word error rate
WRCR Word recognition correctness rate

2.4 Software and tools used for Arabic language ASR research

As far as Arabic language ASR research efforts are concerned, Hidden Markov Model Toolkit (HTK) and Carnegie Mellon University (CMU) Sphinx engine are the most widely used open source ASR toolkits, and they are getting more and more popular as the ASR technology is applied into new languages. The HTK and CMU Sphinx contain ready-to-use downloadable tools, which are devoted for training the acoustic models due to their capabilities in implementing large vocabulary, speaker-independent, continuous speech recognition system in any language (Samudravijaya and Barot 2003; Kacur and Rozinaj 2008; Novak et al. 2010; Abushariah 2012).

Although both HTK and CMU Sphinx have common goal to achieve, they have various differences. Samudravijaya and Barot (2003) believed that CMU Sphinx has more advanced features and its license is meant for unrestricted use as compared to HTK. They also experimented the use of HTK and CMU Sphinx and concluded that the CMU Sphinx is able to produce better quality acoustic models than that of the HTK. Major technical differences include (1) HTK is more flexible in terms of allowing the users to specify the number of states for each unit, whereas CMU Sphinx has fixed the number of states to five-state models, (2) For language modeling, HTK supports the use of bi-gram models, whereas CMU Sphinx supports both bi-gram and tri-gram language models, (3) HTK is more user-friendly than CMU Sphinx. (4) Overall, CMU Sphinx is believed to be better than HTK especially in terms of performance and accuracy rates. Based on the above as well as the tables presented earlier, it is noticed that many researchers utilized the CMU Sphinx tools especially for large vocabulary, speaker-independent, continuous speech recognition systems (Abushariah 2012). Therefore, this research work has selected the CMU Sphinx 3 tools to be used for implementing and evaluating the ASR systems.

It is also noticed that CMU Pocketsphinx and KALDI are also used nowadays for ASR research, which contain many state-of-the-art optimization techniques. Therefore, it is assumed that many researchers will follow the trend and utilize CMU Pocketsphinx and KALDI in their research.

3 TAMEEM V1.0 Arabic automatic continuous speech recognizer

TAMEEM V1.0 is an Arabic speakers and text independent automatic continuous speech recognizer, which is the product of this research. TAMEEM V1.0 is a multi-disciplinary task, whereby Arabic phonetics, Arabic speech processing techniques and algorithms, and Natural Language Processing (NLP) are integrated, which result in improved and optimized performance of the developed recognizer. This section describes the major implementation requirements and components for developing TAMEEM V1.0, namely feature extraction, Arabic phonetic dictionary, the acoustic

Table 3 Performance comparison of the Holy Qur'an ASR research efforts

Source	Main task/vocabulary size	Techniques		Tools	Speech data		Speaker dependency	Systems' performance		
		Features extraction	Features classification		Training	Testing		WRCR (%)	AR (%)	WER (%)
El Amrani et al. (2016)	-Recognizing verses of the Qur'an AM size = N/A LM size = 65 words	N/A	N/A	CMU Sphinx-4	90%	10%	Speaker-independent	N/A	N/A	50
Hafeez et al. (2014)	-Recognizing verses of the Qur'an AM size = N/A LM size = 910 words	N/A	HMM	CMU Sphinx-4	N/A	N/A	Speaker-dependent	N/A	81	N/A
Hyassat and Abu Zitar (2008)	-Recognizing verses of the Qur'an AM size = 25,740 words LM size = 25,740 words	MFCC	HMM	CMU Sphinx-4	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 readers AM size = 2000 words LM size = 2000 words	N/A	HMM	HTK	2431 tokens	N/A	Speaker-independent	N/A	Average of 78.60	N/A
Tabbal et al. (2006)	-Recognizing verses of the Qur'an of 13 readers AM size = 16 words LM size = 16 words	MFCC	HMM	CMU Sphinx-4	1 h	1 h	Speaker-dependent	Average of 91	N/A	N/A

Table 4 Performance comparison of Arabic sentences, proverbs, and questions ASR research efforts

Source	Main task/vocabulary size	Techniques		Tools	Speech data		Speaker dependency	Systems' performance			
		Features extraction	Features classification		Training	Testing		WRCR (%)	AR (%)		WER (%)
Zarrouk et al. (2015)	-Recognizing Arabic sentences AM size = N/A LM size = 3622 words	MFCC	HMM Hybrid SVM/HMM DBN Hybrid SVM/DBN	HTK	620 statements	171 statements	N/A	N/A	N/A		Average of 13.32 Average of 11.42 Average of 10.66 Average of 8.67
Zarrouk et al. (2014)	-Recognizing Arabic sentences AM size = N/A LM size = 3622 words	MFCC	HMM Hybrid MLP/HMM Hybrid SVM/HMM	HTK	620 statements	171 statements	N/A	N/A		N/A	Average of 13.32 Average of 11.92 Average of 11.42
Droua-Hamdani et al. (2013)	-Recognizing medium sized Arabic sentences AM size = N/A LM size = 1080 sentences	MFCC	HMM	HTK	N/A	N/A	Speaker-independent	N/A	N/A		9
Droua-Hamdani et al. (2010)	-Recognizing Arabic phonetically balanced sentences AM size = 200 sentences LM size = 200 sentences	MFCC	HMM	HTK	413 tokens	157 tokens	Speaker-independent	92.11	91.65		8.35
Azmi and Tolba (2008)	-Recognizing Arabic sentences and proverbs AM size = 16 sentences LM size = 16 sentences	MFCC	HMM	HTK	528 tokens	416 tokens	Speaker-independent	70.13	N/A		N/A
Nofal et al. (2004)	-Recognizing medium sized Arabic sentences AM size = N/A LM size = 1340 words	MFCC	HMM	HTK	10 h	N/A	N/A	N/A	N/A		5.26

model training, and the statistical language model training. The used speech corpus is also described in this section.

In order to develop TAMEEM V1.0, there are four essential and required implementation components to be developed, namely, feature extraction, Arabic phonetic dictionary production, the acoustic model training, and the statistical language model training in accordance to the HMM-based architecture of the system as shown in Fig. 1, whereby the input speech uses the phonetically rich and balanced MSA speech corpus (Abushariah et al. 2012b; Abushariah 2012).

The decoder is then used when all implementation requirements are developed. It receives the new input features Y converted into a sequence of fixed size acoustic vectors at the feature extraction stage. It then attempts to identify the sequence of words W that is most likely to have generated Y. Therefore, the decoder attempts to find (Gales and Young 2008):

$$\hat{W} = \arg\max_{W} \mathrm{P(W|Y)} \tag{1}$$

The conditional probability P(W|Y) is difficult to be modeled directly, and therefore, Bayes' Rule is used in order to transform Eq. (1) to an equivalent problem resulting in Eq. (2) (Gales and Young 2008):

$$\hat{W} = \arg\max_{W} \mathrm{P(Y|W)}P(W) \tag{2}$$

The acoustic model is determined by the likelihood conditional probability P(Y|W) in order to observe a signal Y given a word sequence W was spoken, whereas the statistical language model is determined by the priori probability P(W) that word sequence W was spoken.

ASR systems are expected to serve a large number of words; and therefore, each word has to be decomposed into a subword (phone) sequence. The acoustic model that corresponds to a given W is synthesized through concatenating the phone models in order to make words according to the way they are defined by the pronunciation dictionary.

3.1 Feature extraction

Feature extraction is the initial stage of TAMEEM V1.0 recognizer 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) is used to extract features from the set of spoken utterances, which is the main feature extraction technique used in CMU Sphinx 3 tools (Chan et al. 2007). As a result, a feature vector that represents unique characteristics of each recorded utterance is produced, which is considered as an input for training and testing the acoustic model. Refer to Abushariah et al. (2012a) and Abushariah (2012) for further details about feature extraction component.

3.2 Arabic phonetic dictionary

The pronunciation or phonetic dictionary is one of the key components of the modern large vocabulary ASR systems, which serves as an intermediary link between the acoustic model and the language model in ASR systems Abushariah (2012). A rule-based approach to automatically generate the Arabic phonetic dictionary for large vocabulary ASR systems based on a given transcription is used. This tool uses the classic Arabic pronunciation rules, common pronunciation rules of MSA, and morphologically driven rules. Arabic pronunciation follows certain rules and patterns when the text is fully diacritized. According to Ali et al. (2008), this tool helps in developing the Arabic phonetic dictionary through choosing the correct phoneme combination based on the location of the letters and their neighbors, and providing multiple pronunciations for words that might be pronounced in different ways. Further details about Arabic phonetic dictionary can be found in Abushariah et al. (2012a, b). In developing TAMEEM V1.0, the transcription file contains 2110 words and the vocabulary list contains 1626 unique words. The number of pronunciations in the developed phonetic dictionary is 2482 entries.

3.3 Acoustic model training

The acoustic model in TAMEEM V1.0 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. A tri-phone not only models an individual phoneme, but it also captures distinct models from the surrounding left and right phones.

Continuous Hidden Markov Model (CHMM) technique is also supported in CMU Sphinx 3 for parametrizing the probability distributions of the state emission probabilities. Training the acoustic model using CMU Sphinx 3 tools requires successfully passing through three phases of Context-Independence (CI), Context-Dependence (CD), and Tied States, whereby each phase consists of three main steps, which are (1) model definition, (2) model initialization, and (3) model training (Rabiner 1989; Alghamdi et al. 2009).

In order to train the acoustic model in TAMEEM V1.0 recognizer, Baum-Welch re-estimation algorithm is used during the first phase in order to estimate the transition

Table 5 Performance comparison of Arabic broadcast news, broadcast conversations, and broadcast reports ASR research efforts

Source	Main task/vocabulary size	Techniques		Tools	Speech data		Speaker dependency	Systems' performance		
		Features extraction	Features classification		Training	Testing		WRCR (%)	AR (%)	WER (%)
Nahar et al. (2016)	-Arabic broadcast news (BN) AM size = N/A LM size = 15,873 words	MFCC	Hybrid LVQ/HMM	CMU Sphinx-3	50 K frames	20 K frames	N/A	N/A	89	N/A
Ali et al. (2014)	- Arabic broadcast conversations (BC) - Arabic broadcast reports (BR) AM size = N/A LM size = 1.4 M words	MFCC	DNN MPE	KALDI	194 h	9 h	N/A	N/A	N/A	BC 32. 21 BR 15.81 Combined 26.95
Nahar et al. (2013)	-Arabic T.V news AM size = N/A LM size = N/A	MFCC	HMM	CMU Sphinx-3	70% (out of 5 h)	30% (out of 5 h)	N/A	N/A	56.79	N/A
AbuZeina et al. (2011)	- Arabic Broadcast News (BN) AM size = N/A LM size = 15,873 words	N/A	HMM	CMU Sphinx-3	4.3 h	1.1 h	Speaker-independent	N/A	N/A	9.91
Alghamdi et al. (2009)	- Arabic broadcast news (BN) Test size = 3585 words AM size = 17,236 words LM size = 17,236 words	MFCC	HMM	CMU Sphinx-3	7 h	400 tokens = ½ h	Speaker-independent	90.78	N/A	10.87
Soltau et al. (2009)	- Arabic broadcast news (BN) AM size = N/A LM size = 829 M words	PLP	HMM	N/A	About 7930 h	About 6 h	Speaker-independent Speaker-dependent	N/A	N/A	30.20 20.30
Vergyri et al. (2008)	- Arabic Broadcast News (BN) - Broadcast Conversation (BC) AM size = N/A LM size = 1.1B Words	MFCC	MPFE	N/A	1120 h	5.5 h	Speaker-independent Speaker-dependent	N/A	N/A	BN 19.70 BC 29.80 BN 9.80 BC 16.50

Table 5 (continued)

Source	Main task/vocabulary size	Techniques		Tools	Speech data		Speaker dependency	Systems' performance		
		Features extraction	Features classification		Training	Testing		WRCR (%)	AR (%)	WER (%)
Rybach et al. (2007)	- Arabic broadcast news (BN) - Broadcast conversation (BC) AM size = N/A LM size = 256k Words	MFCC	HMM	N/A	About 450 h	About 6 h	Speaker-independent Speaker-dependent	N/A	N/A	BN 25.40 BC 33.90 BN 13.50 BC 19.70
Soltau et al. (2007)	- Arabic broadcast news AM size = N/A LM size = 617k words	PLP	HMM	N/A	About 2000 h	About 13 h	Speaker-independent	N/A	N/A	Average of 30.15
Messaoudi et al. (2006)	- Arabic broadcast news AM size = N/A LM size = 65k words	N/A	HMM	N/A	About 150 h	About 1.25 h	Speaker-independent Un-vowelized Speaker-independent Vowelized	N/A	N/A	16 14.80

probabilities of the Context-Independent (CI) HMMs. Arabic basic sounds are classified into phonemes or phones, whereby in this work 44 Arabic phonemes and phones are used including silence. 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 (Rabiner 1989; Rabiner and Juang 1993). Further details about acoustic model training can be found in Abushariah et al. (2012a). There are 4705 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 6.

The acoustic model training has also undergone several training attempts aiming to identify the best combination of parameters in order to optimize the performance of TAMEEM V1.0 recognizer. The acoustic model in TAMEEM V1.0 recognizer is first trained using default values of major parameters as identified in Sphinx 3 configuration file, whereby the number of Gaussian mixture distributions is 8, and the number of senones is 1000. However, these default values may not necessarily be the best. Therefore, different values must be examined in order to find the best combination that yields the best performance in terms of the WER. Therefore, different ranges for Gaussian mixture distributions and senones are tested in order to identify their best combination. In this work, Gaussian mixture distributions range from 2 to 64, whereas senones range from 300 to 2500.

3.4 Language model training

The language model component provides the grammar used in TAMEEM V1.0 recognizer. The grammar's complexity depends on the system to be developed. The language model computes the probability P(W) of a sequence of words $W = w_1, w_2, \ldots, w_L$. The probability P(W) can be expressed as shown in Eq. (3):

$$P(W) = P(w_1, w_2, \ldots w_L) = \prod_{i=1}^{L} P(w_i | w_1, \ldots w_{i-1}) \quad (3)$$

In order to develop the statistical language model for TAMEEM V1.0 recognizer, the CMU-Cambridge Statistical Language Modeling toolkit is used, which contains the uni-grams, bi-grams, and tri-grams of the language for the subject text to be recognized (Clarkson and Rosenfeld 1997). The language model is created through computing the word uni-gram counts, which are then converted into

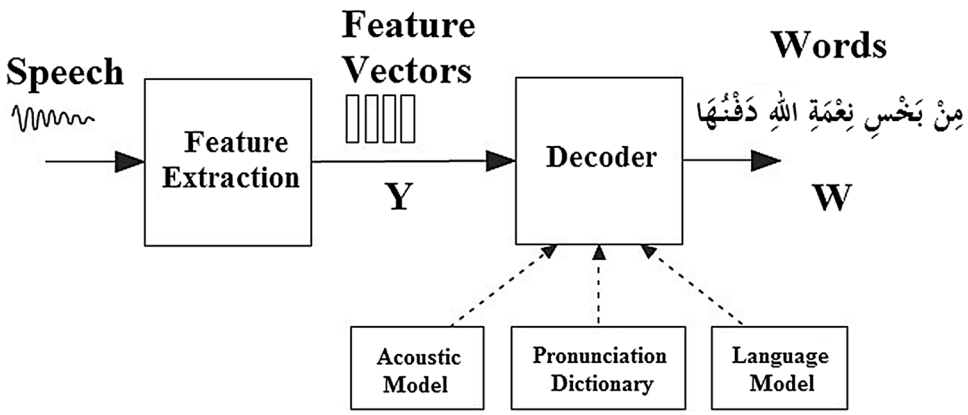

Fig. 1 Architecture of the HMM-based TAMEEM V1.0 recognizer

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; Abushariah et al. 2012b). For TAMEEM V1.0 recognizer, the number of uni-grams is 1627, whereas the number of bi-grams and tri-grams is 2083 and 2085 respectively.

3.5 The decoder

Viterbi search algorithm and beam search heuristics are used for decoding purposes that are also available within the CMU Sphinx 3 decoder, which uses a lexical-tree search structure. In order to perform decoding, the acoustic model, language model, phonetic dictionary, and feature vector of the unknown utterance are required. 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). The decoder relies on Word Insertion Penalty (WIP), Language Model Weight (LW), and Beam Pruning (BP) parameters that take place at decoding (recognition) level. CMU Sphinx 3 decoder has set the WIP to 0.7, LW to 9.5, and BP to 1.0e-35.

4 Phonetically rich and balanced MSA speech corpus

The speech corpus is an important requirement for developing and evaluating any ASR system. It is important to note that TAMEEM V1.0 recognizer as discussed in the previous section is developed and evaluated using the phonetically rich and balanced MSA speech corpus whose preparation and production are discussed and explained in sufficient detail in Abushariah et al. (2012b). This corpus contains recordings of 415 Arabic sentences. The 367 phonetically rich and balanced sentences are used for training the acoustic model. According to Alghamdi et al. (2003), although this set of 367 Arabic sentences contains only 1835 words, yet they contain all Arabic phoneme clusters that are in line with the Arabic phonotactic rules. For testing the acoustic model on the other hand, 48 additional sentences representing Arabic proverbs were created by an Arabic language specialist for the purpose of this corpus.

The phonetically rich and balanced MSA speech corpus was developed in order to provide large amounts of high quality recordings of MSA making it suitable for the design and development of any speaker-independent, continuous, and automatic Arabic ASR system. For the purpose of training and testing TAMEEM V1.0 recognizer, speech recordings of 36 speakers were randomly used from the entire corpus. Table 7 shows some statistical analysis for the selected portion of the phonetically rich and balanced

Table 6 Occurrences of tri-phones for each Arabic phoneme in TAMEEM V1.0 recognizer (Abushariah et al. 2012b)

Phone	Tri-phones	Phone	Tri-phones	Phone	Tri-phones
AA	71	F	118	R	136
AA	32	GH	61	S	98
AE	456 (Max.)	H	89	SH	77
AE	200	HH	97	SS	75
AH	44	IH	364	T	109
AH	18 (Min.)	IX	57	TH	60
AI	118	IX	18 (Min.)	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	4705
E	207	Q	97		

Table 7 Statistical analysis of the phonetically rich and balanced MSA 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	1422 words	241 words	1663 words
Number of unique words based on training and testing sentences in combined transcription file	N/A	N/A	1626 words
Total frequencies of words in the transcription file	178,704 words	28,110 words	206,814 words
No. of utterances (.wav)	36,071 utterances	4934 utterances	41,005 utterances
Average no. of (.wav) utterances/sentence	98 sound files/sentence	103 sound files/sentence	N/A
Size of utterances (.wav)	4.29 GB	0.66 GB	4.95 GB
Size of feature extracted utterances (.mfc) files	771 MB	117 MB	888 MB
Duration of utterances (.wav)	39.28 h	6.02 h	45.30 h
Average duration/sentence	6.42 min/sentence	7.53 min/sentence	N/A
Average duration/utterance (.wav)	3.92 s/utterance	4.39 s/utterance	N/A

MSA speech corpus, which is used for developing and evaluating TAMEEM V1.0 recognizer.

A total of 41,005 utterances were used resulting in about 45.30 h of speech data collected from 36 Arabic native speakers from 11 different Arab countries. The leave-one-out cross validation and testing approach was applied, where every round speech data of 35 out of 36 speakers are trained and speech data of the 36th are tested. As a result, 36 different experiments are conducted that represent different data sets.

The phonetically rich and balanced MSA speech corpus covers 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. Therefore, this corpus adds a new variety of possible speech data for Arabic language based text and speech applications besides other varieties such as broadcast news.

5 Experimental results and analysis

In order to validate the uniqueness of the phonetically rich and balanced MSA speech corpus and its positive impact, TAMEEM V1.0 recognizer is evaluated and its performance is analyzed and discussed in this section. Experimental work conducted as part of this research is evaluated using the WER, which is computed using Eq. (4) and the lower the WER the better the recognizer's performance:

$$\text{Word error rate (WER)} = 100\% - \text{percent accuracy} = \frac{\text{D} + \text{S} + \text{I}}{\text{N}} \times 100\% \quad (4)$$

where Percent Accuracy $= \frac{\text{N}-\text{D}-\text{S}-\text{I}}{\text{N}} \times 100\%$, N is the total number of words 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, which are resulted when comparing the recognized words sequence with the reference (spoken) words sequence.

As stated earlier in Sect. 3.3, the acoustic models are normally trained using default values of number of Gaussian mixture distributions (8) and number of senones (1000). Based on these default values, the WER is 12.57% for speakers dependent with text independent data set. However, the achieved results using CMU Sphinx 3 default values may not necessarily be the best, and it is always advisable to examine different values in order to find their optimal combination that leads to the best performance. Therefore, the acoustic model is trained using different combinations of number of Gaussian mixture distributions that range from 2 to 64 and number of senones that range from 300 to 2500.

Based on the range values of the number of Gaussian mixture distributions (G) and number of senones as identified in Sect. 3.3 with G = (2, 4, 8, 16, 32, and 64), and number of senones = (300, 350, 400, 450, 500, 1000, 1500, 2000, and 2500), 54 different combinations are produced, each of which corresponds to a unique experiment. Experiments conducted here are initial and meant for identifying the best combination of the two parameters, which are then applied for other experimental data sets that are resulted from the leave-one-out cross validation and testing approach. The optimal combination as resulted from the initial experimental work indicates that the number of Gaussian mixture distributions is influenced by the size and the number of speakers of the language resources used to train the ASR systems.

Table 8 TAMEEM V1.0 default and modified recognizers using gaussian mixture distributions and senones

TAMEEM V1.0	Number of gaussian mixture distributions	Number of senones	Speakers dependent with text independent WER (%)
Default recognizer	8	1000	12.57
Modified recognizer	64	350	7.42

The number of Gaussian mixture distributions is best when it is 64 in TAMEEM V1.0 recognizer. This is due to the fact that this work involves training data collected from 36 speakers with each having his/her own speaking style and characteristics. In addition, the optimal number of senones varies from one application to another depending on the amount of available training data and the number of triphones present in the task. If the number of triphones in the task is high and the training data is in high volume, the optimal number of senones are expected

Table 9 Performance in WER (%) for TAMEEM V1.0 recognizer

Experiment number	Speakers dependent with text independent WER (%)	Speakers independent with text dependent WER (%)	Speakers independent with text independent WER (%)
1	7.42	1.36	3.93
2	7.64	3.04	9.40
3	7.44	5.74	9.09
4	7.47	0.67	2.68
5	8.05	1.15	6.40
6	7.19	0.74	4.49
7	7.46	2.77	6.66
8	8.11	0.88	5.86
9	7.18	1.23	4.03
10	6.77	2.68	8.73
11	8.10	0.79	6.64
12	7.27	0.32	6.12
13	6.75	4.75	16.39
14	7.16	1.46	8.74
15	8.25	2.41	8.14
16	7.60	0.78	5.15
17	8.56	1.92	16.30
18	7.20	1.14	5.83
19	7.43	1.77	7.75
20	8.08	1.04	7.97
21	7.29	1.81	6.88
22	8.17	7.03	14.26
23	6.98	1.04	5.80
24	7.24	0.77	6.05
25	7.25	1.08	4.63
26	7.76	4.83	14.05
27	9.08	1.02	3.60
28	6.73	6.39	10.63
29	7.71	2.47	8.02
30	7.77	1.81	7.97
31	7.67	1.94	9.10
32	9.27	1.34	5.71
33	8.09	2.67	9.55
34	7.45	4.14	14.64
35	8.55	2.98	5.04
36	6.99	1.88	5.37
Average WER	7.64	2.22	7.82

Table 10 Summary of the systems' performance in WER (%)

Development phase	Speakers dependent with text independent WER (%)	Speakers independent with text dependent WER (%)	Speakers independent with text independent WER (%)
1st Phase (4 h)	14.37	16.71	25.88
2nd Phase (8 h)	11.27	5.78	15.59
3rd Phase (11 h)	9.38	3.37	10.40
4th Phase (current work)	7.64	2.22	7.82

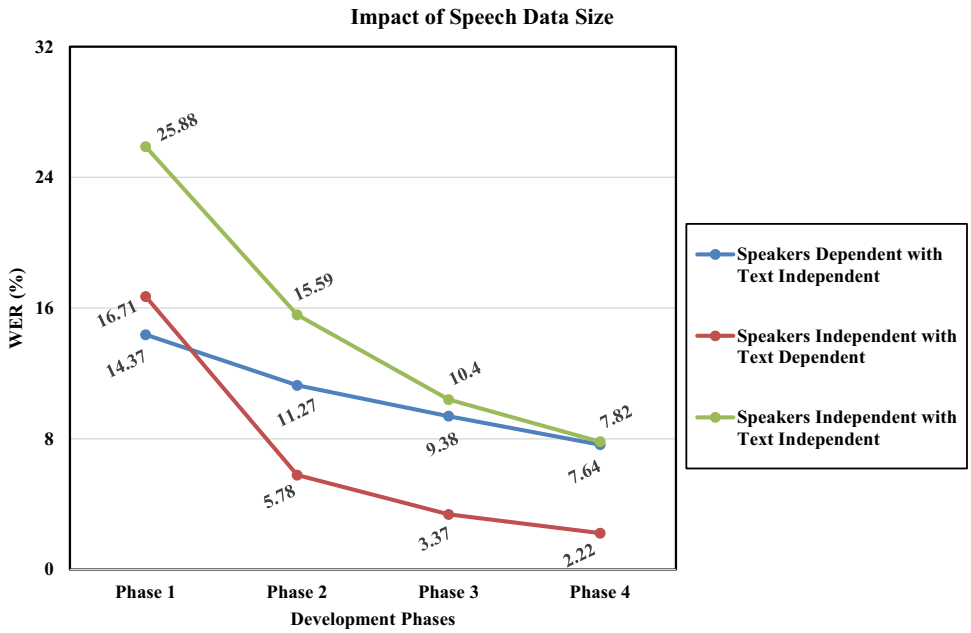

Fig. 2 Impact of speech data size on the overall performance in WERs (%)

to be high too. However, sometimes it is impractical to have a high number of senones as the accuracy of the system will degrade. In TAMEEM V1.0 recognizer, the number of senones is best when it is 350. This combination (64 Gaussian mixture distributions and 350 senones) achieved a WER of 7.42% for speakers dependent with text independent data set. Table 8 presents the TAMEEM V1.0 recognizer's performance for both default and modified parameters at training level. This best combination is also selected in training the acoustic model for Experiment 2 through Experiment 36 data sets, and the corresponding results are shown in Table 9.

Based on the default and modified recognizers' performance as shown in Table 8, it is found that CMU Sphinx 3 default values used to train the acoustic models are not the optimal combination. Therefore, it is always recommended to identify the best combination of the training parameters at this stage, and apply them to the rest of the data sets.

Based on this work, the performance of TAMEEM V1.0 recognizer as presented in Table 9 has shown marked improvements in the WER as compared to the previous development phases, which were published in Abushariah et al. (2012a, b, c). TAMEEM V1.0 recognizer is able to recognize the testing data set 1 (speakers dependent with text independent) with an average WER of 7.64%. From speaker independence perspective, data set 3 (speakers independent with text independent) achieves an average WER of 7.82%. Therefore, the gap (7.64–7.82%, is 0.18%) between the results obtained from data sets 1 and 3 is considered very minimal, which indicates that this work is the best in achieving speaker independence. Table 10 summarizes the average WERs for previously published results in Abushariah et al. (2012a, b, c) of three prior phases compared to this phase (fourth phase).

Based on Table 10 and Fig. 2, it is important to highlight the impact of speech data size on the overall performance of TAMEEM V1.0 recognizer (4th Phase) and all previous recognizers of 1st Phase through 3rd Phase. It is found that the more speech data size is used to train the recognizer, the lower the WER and the better the performance. This is logical and in line with the fact that the training data size is considered as the major contributor to lower WERs.

6 Conclusions

In conclusion, this research work has found that the modified systems perform better than the default systems using standard and default CMU Sphinx 3 setup. Therefore, it is advisable to try different combinations of parameters in order to identify the best combination that is more suitable to the data used in order to obtain better performance.

Speaker independence and text independence are highly achieved and witnessed in this research work. If we refer to Table 9, we can see that for the same speakers with different sentences (speakers dependent with text independent), the systems obtain an average WER of 7.64%, whereas for different speakers with different sentences (speakers independent with text independent) they obtain an average WER of 7.82%. This is important due to the fact that ASR 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 this research work, as more data to train the systems is added, it is realized that the systems become more speakers and text independent, and they could perform similar to those speakers used in training the systems.

During this research work using about 39.28 h of training speech data, the acoustic model is based on 64 Gaussian mixture distributions and the state distributions are tied to 350 senones. Using three different data sets, this work obtains 7.64% average WER for the same speakers with different sentences (speakers dependent with text independent). For different speakers with same sentences (speakers independent with text dependent), this work obtains 2.22% average WER, whereas for different speakers with different sentences (speakers independent with text independent) this work obtains 7.82% average WER.

It is important to highlight that the phonetically rich and balanced MSA speech corpus is able to have positive impact on the performance of TAMEEM V1.0 speaker independent, text independent, large vocabulary, automatic, and continuous speech recognizer for Arabic language. This is due to its uniqueness compared to other speech corpora such as broadcast news corpora, 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 text-to-speech synthesis, covering different research needs. Throughout this research work, the size of training data is noticed to play a crucial role in achieving better performance for TAMEEM V1.0 recognizer.

Finally, this paper reported the research work towards developing TAMEEM V1.0 recognize, which is a high performance Arabic speaker independent, text independent, large vocabulary, automatic, and continuous speech recognition system based on an in-house developed phonetically rich and balanced MSA speech corpus. Experimental results were also reported in detail showing that the developed TAMEEM V1.0 recognizer is truly speakers and text independent, and is highly comparable and better than many reported Arabic ASR research efforts as investigated in the literature review.

References

Abdo, M. S., Kandil, A. H., El-Bialy, A. M., & Fawzy, S. A. (2010). Automatic Detection for Some Common Pronunciation Mistakes Applied to Chosen Quran Sounds. *IEEE Proceedings of the 5th Cairo International Biomedical Engineering Conference*. Egypt, pp. 219–222.

Abushariah, M. A. M. (2012). Automatic Continuous Speech Recognition Based On Phonetically Rich and Balanced Arabic Speech Corpus. *Ph.D. Thesis*, University of Malaya, Malaysia.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Alqudah, A. A. M., Ahmed, E., & Khalifa, O. O. (2012a). Modern standard Arabic speech corpus for implementing and evaluating automatic continuous speech recognition systems. *Journal of the Franklin Institute, 349*(7), 2215–2242.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2012b). Phonetically rich and balanced text and speech corpora for Arabic language. *Language Resources and Evaluation Journal, 46*(4), 601–634.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2012c). Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus. *The International Arab Journal of Information Technology, 9*(1), 84–93.

AbuZeina, D., Al-Khatib, W., Elshafei, M., & Al-Muhtaseb, H. (2011). Cross-word Arabic pronunciation variation modeling for speech recognition. *International Journal of Speech Technology, 14*(3), 227–236.

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

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

Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S., & Glass, J. (2014). A Complete KALDI Recipe for Building Arabic Speech Recognition Systems. *IEEE Proceedings of Spoken Language Technology Workshop (SLT)*, USA, pp. 525–529.

Ali, M., Elshafei, M., Alghamdi, M., Almuhtaseb, H., & Al-Najjar, A. (2008). Generation of Arabic Phonetic Dictionaries for Speech Recognition. *IEEE Proceedings of the International Conference on Innovations in Information Technology*. UAE, pp. 59–63.

Alotaibi, Y. A. (2010). Is Phoneme Level Better than Word Level for HMM Models in Limited Vocabulary ASR Systems? *Proceedings of the IEEE Seventh International Conference on Information Technology—New Generations*, Las Vegas, USA, pp. 332–337.

Alotaibi, Y. A., & Meftah, A. H. (2010). Comparative Evaluation of Two Arabic Speech Corpora. *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.

Azmi, M. M., & Tolba, H. (2008). 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.

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

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

Canavan, A., Zipperlen, G., & Graff, D. (1997). *CALLHOME Egyptian Arabic Speech*. Philadelphia, PA: Linguistic Data Consortium.

Chan, A., Gouvˆea, E., Singh, R., Ravishankar, M., Rosenfeld, R., Sun, Y., Huggins-Daines, D., & Seltzer, M. (2007). *The Hieroglyphs: Building Speech Applications Using CMU Sphinx and Related Resources*. http://www-2.cs.cmu.edu/~archan/documentation/sphinxDocDraft3.pdf, Accessed 15 Sept 2010.

Chou, F. C., & Tseng, C. Y. (1999). The Design of Prosodically Oriented Mandarin Speech Database. *ICPhS'99*, San Francisco, pp. 2375–2377.

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

Clarkson, P., & Rosenfeld, R. (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, Greece, pp. 2707–2710.

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

Droua-Hamdani, G., Sellouani, S. A., & Boudraa, M. (2013). Effect of characteristics of speakers on MSA ASR performance. *IEEE Proceedings of the 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13)*, pp. 1–5.

Droua-Hamdani, G., Selouani, S. A., & Boudraa, M. (2010). Algerian Arabic speech database (ALGASD): Corpus design and automatic speech recognition application. *The Arabian Journal for Science and Engineering, 35*(2 C), 157–166.

Ejerhed, E., & Church, K. (1997). Language resources: Written language corpora. In R. Cole, J. Mariani & H. Uszkoreit. (Eds.), *Survey of the state of the art in human language technology* (pp. 359–363). Italy: Cambridge University Press and Giardin.

El Amrani, M. Y., Rahman, M. H., Wahiddin, M. R., & Shah, A. (2016). Building CMU Sphinx language model for the Holy Quran using simplified Arabic phonemes. *Egyptian Informatics Journal, 17*(3), 305–314.

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

Elmahdy, M., Gruhn, R., Minker, W., & Abdennadher, S. (2009b). Modern Standard Arabic Based Multilingual Approach for Dialectal Arabic Speech Recognition. *IEEE Proceedings of the Eighth International Symposium on Natural Language Processing*, Bangkok, Thailand, pp. 169–174.

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

Gales, M., & Young, S. (2008). *The Application of Hidden Markov Models in Speech Recognition*. Hanover: Now Publishers Inc.

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.

Godfrey, J. J., & Zampolli, A. (1997). Language resources: Overview. In R. Cole, J. Mariani & H. Uszkoreit. (Eds.), *Survey of the state of the art in human language technology* (pp. 357–359). Cambridge: Cambridge University Press and Giardin.

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

Hafeez, A. H., Mohiuddin, K., & Ahmed, S. (2014). Speaker-dependent live quranic verses recitation recognition system using Sphinx-4 framework. *IEEE Proceedings of the 17th International Conference in Multi-Topic (INMIC)*, pp. 333–337.

Haraty, R. A., & El Ariss, O. (2007). CASRA+: A colloquial Arabic speech recognition application. *American Journal of Applied Sciences, 4*(1), 23–32.

Harrag, A., & Mohamadi, T. (2010). QSDAS: New Quranic speech database for Arabic speaker recognition. *The Arabian Journal for Science and Engineering, 35*(2C), 7–19.

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

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

Jorschick, A. (2009). Sound to Sense Corpus Manual. *Technical Report*, Faculty for Linguistics and Literary Sciences, University of Bielefeld, Germany.

Kacur, J., & Rozinaj, G. (2008). Practical issues of building robust HMM models using HTK and SPHINX systems. In F. Mihelič & J. Žibert. (Eds.), *Speech recognition, technologies and applications* (pp. 171–192). Vienna: I-Tech Education and Publishing.

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

Lamel, L., & Cole, R. (1997). Language resources: Spoken language corpora. In R. Cole, J. Mariani & H. Uszkoreit. (Eds.), *Survey of the state of the art in human language technology* (pp. 363–367). Cambridge: Cambridge University Press and Giardin.

Liang, M. S., Lyu, R. Y., & Chiang, Y. C. (2003). An Efficient Algorithm to Select Phonetically Balanced Scripts for Constructing a Speech Corpus. *IEEE Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, China, pp. 433–437.

Mariani, J. (1995). Tasks of a European Center for Spoken Language Resources (ECSLR). *Technical Report*, Mlap SPEECHDAT Project, Computer Sciences Laboratory for Mechanics and Engineering Sciences, National Center for Scientific Research, France.

Messaoudi, A., Gauvain, J. L., & Lamel, L. (2006). Arabic Broadcast News Transcription Using a One Millionword Vocalized Vocabulary. *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, pp. 1093–1096.

Mourtaga, E., Sharieh, A., & Abdallah, M. (2007). Speaker Independent Quranic Recognizer Based on Maximum Likelihood Linear

Regression. *Proceedings of World Academy of Science, Engineering and Technology*. BrazilVol. 36, pp. 61–67.

Nahar, K. M., Al-Khatib, W. G., Elshafei, M., Al-Muhtaseb, H., & Alghamdi, M. M. (2013). Data-driven Arabic phoneme recognition using varying number of HMM states. *Proceedings of the IEEE 1st International Conference in Communications, Signal Processing, and their Applications (ICCSPA'2013)*, pp. 1–6.

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

Newman, D. L., (2002). The phonetic status of Arabic within the world's languages: the uniqueness of *the lu"At Al-d∓AAd. Antwerp Papers in Linguistics*, No. 100, pp. 65–75.

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.

Nofal, M., Abdel-Raheem, E., El Henawy, H., & Abdel Kader, N. (2004). Acoustic Training System for Speaker Independent Continuous Arabic Speech Recognition System. *IEEE Proceedings of the Fourth International Symposium on Signal Processing and Information Technology*. Italy, pp. 200–203.

Novak, J. R., Dixon, P. R., & Furui, S. (2010). An Empirical Comparison of the T^3, Juicer, HDecode and Sphinx3 Decoders. *INTERSPEECH'10*. Japan, pp. 1890–1893.

Open Language Archives Community (OLAC). (2016a). Retrieved Dec 25, 2016 from http://www.language-archives.org/metrics/www.ldc.upenn.edu.

Open Language Archives Community (OLAC). (2016b). Retrieved Dec 25, 2016 from http://www.language-archives.org/metrics/catalogue.elra.info.

Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Ravishankar, M., Rosenfeld, R., Seymore, K., Siegler, M., Stern, R., & Thayer, E. (1997). The 1996 Hub-4 Sphinx-3 System. *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.

Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Rybach, D., Hahn, S., Gollan, C., Schluter, R., & Ney, H. (2007). Advances in Arabic broadcast news transcription at RWTH. *IEEE Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU'07)*. Japan, pp. 449–454.

Samudravijaya, K., & Barot, M. (2003). A Comparison of Public Domain Software Tools for Speech Recognition. *Workshop on Spoken Language Processing*. India.

Siemund, R., Heuft, B., Choukri, K., Emam, O., Maragoudakis, E., Tropf, H., Gedge, O., Shammass, S., Moreno, A., Rodriguez, A. N., Zitouni, I., & Iskra, D. (2002) OrienTel—Arabic speech resources for the IT market. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, Spain.

Soltau, H., Saon, G., Kingsbury, B., Kuo, J., Mangu, L., Povey, D., & Zweig, G. (2007). The IBM 2006 Gale Arabic ASR System. *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, Hawaii, USA, pp. 349–352.

Soltau, H., Saon, G., Kingsbury, B., Kuo, H. K. J., Mangu, L., Povey, D., & Emami, A. (2009). Advances in Arabic speech transcription at IBM under the DARPA GALE program. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(5), 884–894.

Tabbal, H., El Falou, W., & Monla, B. (2006). Analysis and Implementation of a "Quranic" Verses Delimitation System in Audio Files Using Speech Recognition Techniques. *IEEE Proceedings of the 2nd International Conference on Information and Communication Technologies (ICTTA'06). Syria, 2*, 2979–2984.

Uraga, E., & Gamboa, C. (2004). VOXMEX Speech Database: Design of a Phonetically Balanced Corpus. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Portugal, pp. 1471–1474.

Vergyri, D., & Kirchhoff, K. (2004). Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland, pp. 66–73.

Vergyri, D., Mandal, A., Wang, W., Stolcke, A., Zheng, J., Graciarena, M., Rybach, D., Gollan, C., Schluter, R., Kirchhoff, K., Faria, A., & Morgan, N. (2008). Development of the SRI/Nightingale Arabic ASR System. *INTERSPEECH'08*. Australia vol. 1, pp. 1437–1440.

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

Zarrouk, E., Benayed, Y., & Gargouri, F. (2015). Graphical Models for the Recognition of Arabic Continuous Speech Based Triphones Modeling. *IEEE/ACIS Proceedings of the 16th International Conference on Software Engineering*, Artificial Intelligence, Networking and Parallel/*Distributed Computing (SNPD)*, Japan, pp. 603–609.