KSU Rich Arabic Speech Database

Mansour Alsulaiman, Ghulam Muhammad, Mohamed A. Bencherif, Awais Mahmood and Zulfiqar Ali

Speech Processing Lab, College of Computer and Information Sciences, King Saud University,
Riyadh 11543, Saudi Arabia
Email: {msuliman, ghulam, mbencherif, awais, zuali} @ksu.edu.sa

Abstract

Arabic is one of the major languages in the world. Unfortunately not so much research in Arabic speaker recognition has been done. One main reason for this lack of research is the unavailability of rich Arabic speech databases. In this paper, we present a rich and comprehensive Arabic speech database that we developed for the Arabic speaker / speech recognition research and/or applications. The database is rich in different aspects: (a) it has 257 speakers; (b) the speakers are from different ethnic groups: Saudis, Arabs, and non-Arabs; (c) utterances are both read text and spontaneous; (d) scripts are of different dimensions, such as, isolated words, digits, phonetically rich words, sentences, phonetically balanced sentences, paragraphs, etc.; (e) different sets of microphones with medium and high quality; (f) telephony and non-telephony speech; (g) three different recording environments: office, sound proof room, and cafeteria; (h) three different sessions, where the recording sessions are scheduled at least with 2 weeks interval. Because of the richness of this database, it can be used in many Arabic, and non-Arabic, speech processing researches, such as speaker / speech recognition, speech analysis, accent identification, ethnic groups / nationality recognition, etc. The richness of the database makes it a valuable resource for research in Arabic speech processing in particular and for research in speech processing in general. The database was carefully manually verified. The manual verification was complemented with automatic verification. Validation was performed on a subset of the database where the recognition rate reached 100% for Saudi speakers and 96% for non-Saudi speakers by using a system with 12 Mel frequency Cepstral coefficients, and 32 Gaussian mixtures.

Key Words: Speaker Recognition, Speech corpus, Arabic speech database, Rich database, Phonetically, Rich Database

1. Introduction

Arabic is one of the oldest and widely spoken Semitic languages. Some of its differences from other languages are unique phonemes and phonetic features, and a complicated morphological word structure. It has been reported in the literature that major difficulties in automatic speech processing of Modern Standard Arabic (MSA) are due to distinctive characteristics of the Arabic sound system, namely, emphatic, uvular, and pharyngeal consonants, and short and long vowels [1].

A speech database is an essential component in speech processing research and in developing speech processing systems. An automatic speech/speaker recognition system can be deployed successfully in real life only if it is developed using a versatile and relevant database. Without a proper speech database, speech processing related research cannot be progressed. There are many databases in major languages, like English, Spanish, German, Japanese, Chinese, etc. These databases are rich in terms of the number of speakers, amount of speech, variability of speakers and texts, environments, and transmission channels. However, Arabic speech databases are few in numbers and most of them are private. Therefore, there is a need for a publicly available comprehensive Arabic speech database. A rich and a publicly available database is an important and essential resource for research in the Arabic speech.

While developing a speech corpus, the following consideration may be taken into account: Scope of the corpora, Content, Phonological distribution, Number of speakers, Gender, Accents and/or Regional dialects, Speaking style, Environment, Recording materials, Sessions, Partition into training and testing data sets.

Our database was designed by taking care of all these considerations. We highlight these considerations in section 1.2. In section 2, we present the different richness dimensions of the database and give justification for each dimension. We also present the recording team, the volunteers, the text verification, and the pilot recording. Section 3 gives some details of the hardware and software of the system. In section 4, the main statistics of the database are given. In section 5, we proceed with the database verification methodology and we present the results of this verification. The validation of the database is discussed in section 6, and finally in section 7 we conclude the article.

1.1. Literature review

In [2], we presented some major databases in languages other than Arabic and we also did a survey on many of Arabic speech databases. Table 1 constitutes a summary of our survey of the Arabic speech databases. We also give a short description of some non-Arabic databases for many languages including English [2]. This description will be helpful in recognizing the richness of our database, which we perceive as richer than other databases in many aspects. TIMIT is one of the mostly used English databases with large number of speakers (630) with eight different dialects of American English [19]. The speakers read ten phonetically rich sentences. The text material in the TIMIT consists of 2 dialect "shibboleth" sentences (SA), 450 phonetically-compact sentences (SX) and 1890 phonetically-diverse sentences. The SA sentences were read by all 630 speakers. Each speaker read 5 of the SX sentences and each sentence was uttered by 7 different speakers. For the SI sentences, each speaker read 3 of these sentences, with each sentence being read only by a single speaker [20]. Word and phone level labeling are provided with the database. The database can be obtained from the Linguistic Data Consortium (LDC).

Table 1. Comparison of available Arabic speech databases

Database	speakers	Dialect	Prompts	Channel	Sampling rate	Environment
SAAVB [3]	1033	Saudi	Numbers, words, sentences, alphabets	Telephone (fixed and mobile)	8 KHz	Indoor, outdoor, car
BBL [4]	164	Levantine	Spontaneous	Microphone	16 KHz	-
QSDAS [5]	77	Quran recitation	Quranic verses	Microphone 1-channel	16 KHz	-
MSA Speech Corpus [6]	40	Levant, Gulf, Africa	Sentences	SHURE microphone, 2 channels are converted to 1 channel	44.1 KHz is converted to 16 KHz	Studio
ALGASD [7]	300	Algerian Arab	Sentences	Microphone, 1-channel	16 KHz	-
West Point [8]	110	Native, non-native	Sentences	SHURE microphone	22.05 KHz	-
NetDC Arabic BNSC [9]	-	-	News	Radio receiver	22.05 KHz	-
Global Phone Arabic [10]	78	Tunisia, Palestine, Jordan	Sentences from newspaper	Microphone	16 KHz	-
Egyptian Arabic Speecon [11]	550 (adults) 50 (child)	Egyptian	Spontaneous + Read (words, sentences)	Microphone, 4-channel	16 KHz	Office, entertainment , car, public place
A-Speech DB [12]	205	-	Continuous speech	Microphone	16 KHz	Office
OrienTel Morocco MCA [13]	772	Moroccan	Digits, words, sentence + spontaneous	Fixed & mobile phones	8 KHz	-
OrienTel Tunisia MCA [14]	792	Tunisian	Digits, words, sentence + spontaneous	Fixed & mobile phones	8 KHz	-
OrienTel Egypt MCA [15]	750	Egyptian	Digits, words, sentence + spontaneous	Fixed & mobile phones	8 KHz	-
OrienTel UAE MCA [16]	880	UAE	Digits, words, sentence + spontaneous	Fixed & mobile phones	8 KHz	-
OrienTel Jordan MCA [17]	757	Jordanian	Digits, words, sentence + spontaneous	Fixed & mobile phones	8 KHz	-
NEMLAR Broadcast News [18]	-	-	News	Radio receiver	16 KHz	

LDC also provides Switchboard 2 Phase I and II databases including NIST evaluation subsets. These databases include large number of speakers recorded in different sessions [21]. The content is spontaneous text material uttered in office and home environments.

A speech database in Castilian Spanish called AHUMADA is developed specifically to consider speaker variability and channel-dependent influences [22]. The database contains the following parameters: microphone and telephone channels; read and spontaneous speech; different speech rates while reading the texts; six different recording sessions; dialectal variations of speakers; fixed utterances and speaker specific utterances; etc. The text materials consist of (a) 24 isolated digits, (b) 10 digit strings consisting of 10 digits each, (c) 10 phonologically and syllabically balanced phrases of 8 to 12 word length, (d) One phonologically and syllabically balanced text of about 180 words, and (e) one minute of spontaneous speech. The speakers were 104 male speakers with age between 28 to 42 years. Both microphones and telephone lines were supplied to a professional DAT device. The sampling rate is 44.1 kHz.

POLYCOST is a telephone speech database consisting of different European languages [23]. There were 134 speakers (74 male and 60 female) from different European countries speaking the following text materials: connected digits uttered in English, sentences uttered in English, and sentences in mother tongue, where one of the prompts was dedicated to free speech. The utterances were recorded in room and office environments. The database contains six sessions per speaker.

Some small databases are developed in Slovene-language at the University of Ljubljana [24]. The databases (K211d, GOPOLIS, VNTV, and VINDAT) contain isolated words, broadcast news, diaphone, etc. K211d is an isolated-word corpus designed for phonetic research studies of the Slovene spoken language. Two hundred and fifty one Slovene words were carefully selected as text prompts. Ten speakers (five female and five male) were selected to participate the recording. The recording was phonetically transcribed and labeled manually. The GOPOLIS corpus is a large speech database containing Slovene dialogues in airline timetable information services. There were 50 speakers (25 male and 25 female) speaking randomly chosen 100 sentences.

There are many other databases dedicated to English, Japanese, Chinese, German, and Spanish. These databases are publicly available either commercially or free. Publicly available databases make research in speech processing and recognition in these languages rich and diverse. Compared to these major languages, Arabic has significantly less number of publicly available speech databases, though Arabic is a major language and an official language in the United Nations.

The most widely recognizable speaker recognition evaluations (SRE) are conducted by the National Institute of Standards and Technology (NIST) [25]. Their projects contribute in finding new directions to the problem of text independent automatic speaker recognition. In the NIST SRE, the speaker recognition performance is measured by means of detection error trade-off (DET) curves and detection cost functions. The NIST releases SRE plans in different years as a part of their ongoing projects. The most recent NIST Year 2010 speaker recognition evaluation plan includes not only conversational telephone speech, but also read and conversational speech recorded in room microphone channel [26].

1.2. Guidelines for developing the database

While developing a speech corpus, the following considerations are usually taken into account by the research team that performs the recording:

Scope of the corpora: The corpora design depends on the application that will use these corpora: phonetic analysis, speech synthesis, speech recognition, or speaker recognition.

Content: In [27], it is observed that text material affects automatic speaker/speech recognition performance to a great extent. The corpus can have a variety of contents, for example, single digit, continuous digits, isolated words, phrases, sentences, paragraphs, etc.

Phonological distribution: The analysis units (words, phrases, sentences, etc.) should be carefully chosen so that the distributions of phones are balanced. Scripts should contain all possible vowels, consonants, co-articulations, etc. [28, 29, 30, 31].

Number of speakers: The total number speakers should be enough to validate the experiment under study. These speakers should speak a sufficient number of utterances. The diversity of speakers (age, education level, etc.) is an important factor to consider [32].

Gender: The corpora may contain almost equal number of male and female speakers [33].

Accent: The speakers can be chosen to cover different types of accents [34].

Speaking style: Based on the target, the corpora may contain read, spontaneous or both types of speech [32, 34].

Environment: The utterances can be recorded in different types of acoustic environments, for example, sound proof room, office room, corridor, restaurant, street, inside vehicle, etc. in order to track the effect of microphone variability on ASR [35, 36, 37, 38, 39].

Recording materials: Data can be collected with different types of microphones and transmission channels, for example, mobile phone, land phone, etc. [35, 36].

Sessions: Data may be collected in different sessions to observe the effect of intersession variability [7, 32].

Partitioning into training and testing data sets: The corpus needs to be large enough to de divided into training and testing sets to account for different types of variability [7, 32]. It is better that the experiments are closed set.

Questions and Answer: A database can contain a question and answer session to get the information of speaker such as his/her name, age, sex, profession or his spontaneous reaction to these questions [7, 27].

It is hard to cover all these points in on one database but we were able to do this. We took into consideration the points mentioned above and designed the database to be rich in many dimensions and beneficial in different applications and studies. The developed database is rich in text, text categories, environment, microphones, channels, nationality, mother language, number of recording microphones, and number of sessions. It can be used in many applications related to speech/speaker recognition and even for Arabic accent classification.

2. Characteristics of the Database

2.1. Richness of the Database

In this section, we describe the different aspects of the richness of our corpus. We also give details of the database and how we designed it.

2.1.1. Richness in Text

The corpus text consist of sentences, words, paragraphs, and answers to questions. In the following subsections, we briefly describe each.

(a) Sentences

Three different types of sentences have been used:

Rich sentences taken from SAAVB: The list given in SAAVB was designed to cover allophones of each phoneme. The list contains 934 sentences. We increased the list to 940 sentences by repeating 6 sentences to get a total number divisible by 20. To divide the 940 sentences into sub lists, each sub list has20 sentences, where each sub list should include all the phonemes with each phoneme repeated as much as possible; we did the following: We divided the 940 list randomly into 47 sub lists (47x20 = 940), each one contains 20 sentences. Each sub list was checked if it contains all the phonemes and the number of occurrence of each phoneme. The randomization was repeated again to find new sub lists, and again we count the occurrence of all phonemes in every sub list. After 20 randomization of the list into sub lists, we selected the randomization that gave optimal sub lists for the recording. Each of these sub lists contains all the phonemes.

Rich sentences from [40]: In this study, 20 lists have been suggested; each list contains 10 phonetically balanced sentences. From the 20 lists we choose 4 lists that are easier to pronounce and do not have something that may be very strange to the speakers or offending or confusing to him. We took the opinion of test speakers in selecting the easy to read lists. We fixed one list for all speakers. A second list was chosen randomly from the remaining three lists.

Accent identifying sentences: we selected two common sentences that are suggested in SAAVB due to their ability to differentiate accents.

(b) Words

Four categories of words have been used:

Rich words: These are rich words suggested in SAAVB. The SAAVB list consisted of 700 words. We divided the list into 35 sub lists randomly. Each sub list was checked for the number of missing phonemes. Randomization was repeated to find new sub lists, and again we checked for the number of missing phonemes in every sub list. The optimal sub lists, obtained after 20 randomizations, were chosen for the recording. The optimal sub list is the one with minimum number of missing phonemes.

Phonetically distinctive words: The words were selected from SAAVB because they contain nasals fricatives, and vowels which are closely related to the speaker characteristics and can help in recognizing the speaker identity.

Common words: This list contains 20 words. We designed it to contain words that are used frequently in the everyday conservation. These words consist of almost all Arabic alphabets except two. Examples of some common Arabic words are the Arabic equivalent of [Hello], [yes], [no], [news], etc.

Numbers: This list contains Arabic digit from zero to nine. These digits contained only 17 Arabic alphabets out of 28 but we included this sub list for its importance in many applications.

(c) Paragraphs

Pronouncing paragraphs are different than pronouncing sentences or words. Therefore, two paragraphs were added in this list. The first paragraph is a verse from Quran (the Holy Book of Muslims). The second paragraph is taken from a book of a famous writer. The paragraph was selected because it included all letters, was easy to read by normal readers, and was appealing to them (it is a feel good paragraph). Each of the verse and the paragraph contains all alphabets.

(d) Question and answers

In this database, it was not easy to record an online conversation. Hence we opted for something similar, which were answers by the volunteers to questions by a team member. This is called spontaneous speech. Samples of these questions are: "What is your name?", "how is the weather today?", "what is your best food?". Of course all the questions are in the Arabic language.

(e) Richness in fixed and variable text

Some of the texts were spoken by all the speakers and some texts were distributed among the speakers.

2.1.2. Richness in Speakers

The speakers were Saudis and non-Saudis. The non-Saudis were Arabs and non-Arabs. The non-Arabs were chosen so that they could read Arabic language at an acceptable level. They were mainly from the fourth level in the Arabic language institute at King Saud University. The non-Saudis represented 28 nationalities. They were chosen to represent clusters of areas or countries.

2.1.3. Richness in Recording Sessions

We achieved three sessions of recording. Every session is verified before recording the next one.

2.1.4. Richness in the Recording Environments

Each speaker is recorded in three different environments: sound proof room (Eckel CL-11), office, and cafeteria. For a reason that will be explained later, the second and third sessions were recorded only in the office and the soundproof room.

2.1.5. Richness in the Recording System

The recording system was similar in the office and the cafeteria, and different in the soundproof room.

(a) Recording systems for office and cafeteria

The office and cafeteria system consisted of the following subsystems:

- Two professional microphones (SHURE Beta 58A) connected to a high quality mixer (Yamaha MW12CX in office, Yamaha MW8CX in cafeteria) to be recorded in stereo.
- Medium quality microphone (Sony Dynamic microphone F-V220) connected to a sound card (Sound Card Creative surrounding 5.1).
- Medium quality microphone connected directly to the computer.
- Mobile (Nokia N97) originating calls to a similar mobile connected to the sound card.

(b) Recording system for soundproof room

The soundproof room system consisted of the following subsystems:

- Professional Side-Address condenser Microphone (SHURE PG42) connected to a high quality mixer (Yamaha MW12CX) to be recorded in stereo.
- Professional quality microphone (SHURE Beta 58A) connected to sound card (Sound Card Creative surrounding 5.1).
- Mobile (Nokia N97) connected to the computer via a sound card.

2.1.6. Summary of Richness of Database

Our database is rich in many aspects, as depicted in Fig.1. This subsection summarizes the richness dimensions or aspects:

(a) Text

- Words, sentences, and paragraphs
- Read text vs. spontaneous text

- Common text vs. uncommon text
- Rich words vs. non-rich words (numbers and common words)
- Fixed text vs. variable text
- Easy to pronounce vs. hard to pronounce
- Quran vs. normal text

(b) Environment

- Very quiet (sound room), quite (office), noisy (cafeteria)

(c) Microphones

- We have many microphones vs. one microphone in some databases
- Medium quality, high quality, and high quality Phantom

(d) Different combinations of microphones and recording equipment

- e.g. low quality recording with Medium quality microphone

(e) Sessions

- 3 sessions

(f) Multi-Nationalities

- 9 Arab nationalities
- 20 non-Arab nationalities

(g) Ethnicity:

- Arabs vs. Non-Arabs
- 5 Different regions of the world.

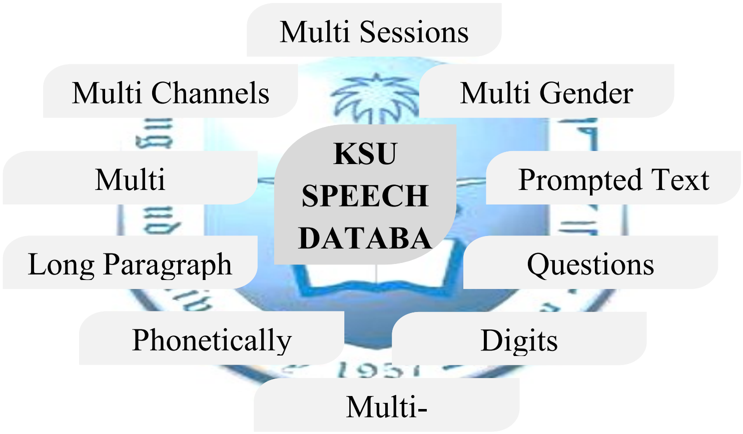


Fig.1: KSU Speech Database Diversity

2.2. Justification of the database specification

Our Corpus vs. SAAVB: SAAVB text was selected by its authors based on a scientific analysis, and since we decided to include sentences and words in our database, it was logical to use the text of SAAVB. But our database is richer because it has more dimensions than SAAVB. For example, SAAVB recorded Saudi speakers from mobile or telephone in one session. Our database recorded many nationalities in three sessions using many microphones within many environments and had more text quantity and variety.

Our Corpus vs. Ref [40]: The authors of [40] designed 20 lists; each list contains 10 sentences so that each list is rich by itself. It had richer text than SAAVB, so we chose it for the same reasons we chose SAAVB. This will make our database richer than both in texts; moreover, our database has more text and has other dimensions as we explained in the case of SAAVB above. Another reason for choosing these lists is that they are not easy to read or pronounce.

Fixed Sentences: These are two common sentences selected from SAAVB. They are designed to differentiate between dialects. We choose these sentences for the same reason.

Common Words: SAAVB sentences and words are not easy; the sentences of [40] are more difficult. Common words were selected by us and were chosen because they will be easier to pronounce, and hence will be more likely to be pronounced correctly.

Numbers: These are important for many applications and are used in daily life.

Rich Words: These words were selected from SAAVB because they have some characteristics that make them useful for speaker recognition as highlighted in section 2.2.1.2.

Paragraphs: Pronouncing paragraphs will have different features than pronouncing sentences and words; hence we included two paragraphs. The first paragraph was selected among many paragraphs because it had the following characteristics: easy to read, it is a feel good paragraph, and it has all the Arabic letters.

The second paragraph is a verse from the Holy Quran. Hence, all speakers were familiar with it. Moreover it has all the Arabic letters. Note that the speakers were asked to read it as normal text and not recite it.

Questions and Answers: The Gulf database was recorded from two speakers speaking to each other. The Babylon Arabic Levantine speech was answers to written questions; so in our database, we included a part similar to Babylon and we consider it as semi questions and answers. It is helpful because it is a different way of speaking than reading from screen or paper, and therefore, may have different characteristics and different effectiveness in recognizing the speaker.

The order of the list: The order of the list was not arbitrary. We made the order of the list such that they will be easy for the speaker.

Silent room: Recording in the silent room produces high quality recording that is very important for language analysis. It can be used to analyze the language, the speaker identity, the native language origin of the speaker.

Office: It is the normal environment in our daily life.

Cafeteria: Recording noisy speech will be helpful in studying the robustness of the speech or speaker recognition methods.

Nationalities: Allow us to investigate effect of the native origin or study characteristic of the speech of a certain nationality.

Ethnicity: Allow us to investigate the effect of ethnicity or to study characteristic of the Arabic speech with respect of ethnic groups.

2.3. The Recording Team

For the recording, and later for verification, the manpower is very critical because understanding the technicality of the computer program and the system equipment is necessary. The recording and verifying were done by a team of six researchers at the college of computer and information sciences, King Saud University. They hold a B.Sc. in computer science\engineering and are native Arabs. They were supervised by two researchers with Master degrees in computer.

2.4. The volunteers

The success of creating the database depends highly on getting volunteers and on the desired number of volunteers. The volunteers had to be Saudis, Arabs, and Non Arabs. The methods used to contact and recruit the volunteers were: Electronic and printed advertisements, presentation in the classes of college of computer, the Arabic language institute, and the personal contacts.

So, all the speakers were literate people, either students at the university, researchers or professors.

2.5. Text Verification

After selecting the text, the next step was to put the text in a displayable form in front of the speaker. The display form was originally a paper form, and then we opted for displaying the text on the screen. Arabic language is unique in some aspects. Indeed, diacritization is sometimes needed to correct pronunciation. Therefore, a great care was taken to make sure that the displayed form was 100% as in the original text.

The written text went into many stages .First of all, it was diacritized, then, rechecked to make sure that it is correct. We asked from volunteers about their opinion for the whole process. The assessment indicates that there is a need for more diacritization of the SAAVB text. Some sentences were confusing and even some words need diacritization to clarify either it is a verb or noun.

Another suggestion was to display the text on the screen. MATLAB did not support reading from word files so we stored the sub lists in RTF format and gave to two professional editors to diacritize it and check it.

2.6. Pilot recording

Before recording the speakers, we tested the system and the whole process with some speakers. Our goals were to test the system (hardware and software), the comfort and technical soundness of the setup, the endurance of the speakers of the recording in one session and in three consecutive sessions, and to measure the time of each step and each session.
From these initial tests we found:

- Some texts needed diacritization.
- Reading from papers was not comfortable; our solution for this was to have the lists displayed on a second screen.
- Speakers were comfortable in all the other points we checked.
- The time range for speaker recording was 5-7 minutes in a location.

3. System Description

3.1. Software description

In order to fulfill the technical specifications, team developed many versions of the program. The flowchart of the main program is illustrated in Fig. 2.The main features of the last version of the program are:

- The generation of the speaker reports, per recording location, containing all the channels, aiming to detect a corrupted wave file, at the end of the recording session of the speaker.
- An automatic visual report of the recorded channels (just the mono version), this method helped the recording team a lot, as they were sure that the channels were recording in an acceptable level.
- A maximum duration of 120 sec has been allocated to each recorded sentence or paragraph, in order to control the length of the speech.

3.2. Hardware description

Table 2 gives the actual hardware configuration in the three locations. Fig. 3 presents the hardware configuration as in the sound proof room and the office. The sampling rate in all of the recordings was 48K sample/sec with 16 bits resolution.

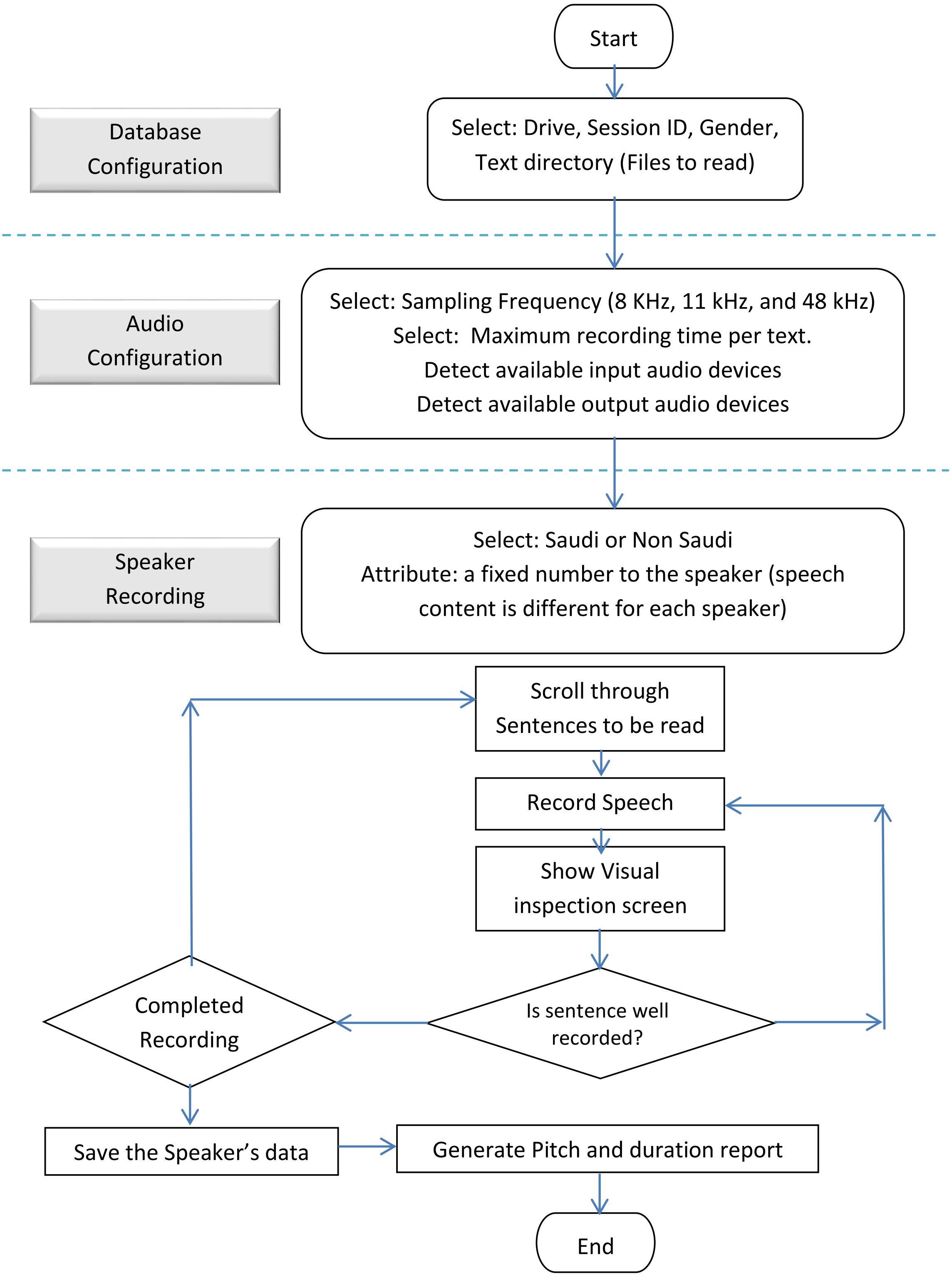
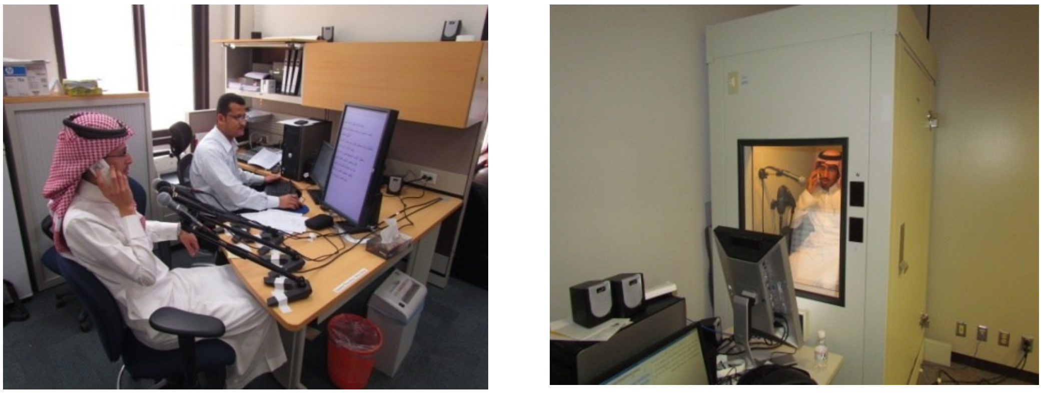

Start
Database Configuration
Select: Drive, Session ID, Gender, Text directory (Files to read)
Audio Configuration
Select: Sampling Frequency (8 KHz, 11 kHz, and 48 kHz)
Select: Maximum recording time per text.
Detect available input audio devices
Detect available output audio devices
Speaker Recording
Select: Saudi or Non Saudi
Attribute: a fixed number to the speaker (speech content is different for each speaker)
Scroll through Sentences to be read
Record Speech
Show Visual inspection screen
Is sentence well recorded?
Completed Recording
Save the Speaker's data
Generate Pitch and duration report
End

Fig. 2. Flowchart of the main program.

Table 2. Hardware configuration

Device	Brand	Office, Café	Soundproof Room
Microphone	SHUR Beta 58A	2	1
Microphone	Sony F-V220	2	---
Mobile	Nokia N97	1	1
Phantom	---		1
Mixer	Yamaha MW-12CX	1	1
Sound Cards	Creative 5.1 Surrounding	2	1



(a) Office Setup

(b) Soundproof Room Setup

Fig.3. Setup for the recording

4. Main Statistics and Results of the Recording

The distribution of the male speakers who recorded in any location and in any session by nationalities is given in Table 3. The female speakers did not include non-Arabs and they were 70 Saudi, 14 Yemenis, 1 Egyptian and 2 Palestinians.

Table 3. List of the recorded male nationalities

Nationality	Number	Nationality	Number	Nationality	Number
Arabs *(Africa & Middle East)*		*Africa Non Arabs*		*Asian Non-Arabs* *Indian Subcontinent*	
Saudi	146	Nigeria	3	India	9
Yemen	15	Uganda	3	Pakistan	8
Egypt	13	Benin	2	Nepal	7
Syria	9	Kenya	2	Afghanistan	4
Tunisia	4	Mali	2	Bangladesh	3
Algeria	4	Central Africa	1	**Total**	**31**
Sudan	4	Guinea Bissau	1	*Asian Non-Arab/East Asia*	
Lebanon	1	Ivory Coast	1	Indonesia	8
Palestine	1	Liberia	1	Philippines	2
Total	**201**	Senegal	1	Thailand	1
East Europe		Togo	1	**Total**	**11**
Serbia	**1**	**Total**	**18**		
Total speakers of all nationalities			**258**		

In verifying the first session, we noticed that the effect of noise was low. So it seems that professional or quality microphones, available nowadays, have a strong noise attenuation capability. Hence the second and third sessions were recorded only in the office and the soundproof room.

The number of volunteers who recorded in the three sessions in the required locations is given in the Table 4. The nationality distribution of male and female volunteers who finished recording in all required locations for the three sessions is provided in Table 5 and Table 6, respectively.

Table 4. Number of speakers in the three sessions

session	No. of male speakers in			No. of female speakers in		
	Office	Cafeteria	Sound proof	Office	Cafeteria	Sound proof
First	253	240	240	87	87	87
Second	206	--	237	77	--	77
Third	136	--	133	64	--	64

Table 5. Male speaker distribution in the three sessions

Session	Saudi	Non-Saudi		Total
		Arab	non-Arab	
First	137	42	61	240
Second	115	41	50	206
Third	55	36	42	133

Table 6. Female speaker distribution in the three sessions

Session	Saudi	Non-Saudi		Total
		Arab	non-Arab	
First	70	17	-	87
Second	61	16	-	77
Third	48	16	-	64

From Tables 4 and 5, we notice that the number of speakers who participated in the third male session is much lower as compared to the first and second session. The reason is that when we started the third male session it was the time of final exams. The majorities of the volunteers were students and were busy in their exams.

Table 7 gives the time duration of the different lists using the mixer data for session 2. Table 7 is divided based on the nationality or race, and it also gives the average duration of the unit of the list.

5. Verification

The recording of the volunteers' speech was followed by verifying the recorded speech. Verification is as vital as the recording itself. For this reason, before starting the recording of any session we verified the previous session. A clear system was designed for the verification and was improved based on our experience. In the following, we briefly discuss the verification stage and shed light on some important findings or ideas.

Table 7. Average Time duration (in Seconds) for the different lists using the mixer data for session 2 (including silence)

Text	Saudi	Arab	Non-Arab	No. of Unit	Avg. /unit
SAAVB sentence_1	32	33	40	Sentence(10)	3.76
SAAVB sentence_2	32	34	41	Sentence(10)	3.80
Fixed sentence	7	7	8	Sentence(2)	4.00
Numbers	10	9	10	Words(10)	1.10
Common words_1	11	11	13	Words(10)	1.26
Common words_2	11	10	12	Words(10)	1.22
Rich words_1	10	10	11	Words(10)	1.13
Rich words_2	10	10	11	Words(10)	1.13
Paragraph_1	29	31	36	Paragraph	34.2
Paragraph_2	48	49	60	Paragraph	56.2
Distinctive words_1	10	9	11	Words(10)	1.11
Distinctive words_2	10	10	11	Words(10)	1.13
Phonetically balanced sent_1	19	19	23	Sentence(10)	2.2
Phonetically balanced sent_2	20	19	24	Sentence(10)	2.29
Q/A _1	20	19	26	Answers(10)	2.34
Q/A_2	17	17	21	Answer(10)	1.98

It is important to mention that the recording system was tested many times in all locations before the actual recording of the volunteers and the team members were selected and trained and supervised. But this cannot substitute the verification of the recording. Moreover the verification stage is more than just verification, it is also commenting on the quality of the recorded speech or documenting the database.

The size of session 1, session 2, and session 3 are 166 GB, 76 GB, and 42.9 GB, respectively. This will give a total database of size 284.9 GB. The number of files for session1, session2 and session 3 are 56217, 22738 and 12924, respectively. Verifying this database is a huge task.

The database is huge, and can actually be looked at as many databases depending on the text, recording system (microphone and digitizing device) the recording environment, and the session number. So human verification of the whole database was a major step by itself and needed a large number of verifiers over a long time. We performed the verification in three stages.

5.1. Stage 1 of verification

This was completed in the first session. The verifiers were given clear instructions in what to do and an excel sheet to fill for every volunteer at each location [41]. The main instructions can be summarized as follows:

- Verify that all channels (or subsystems) of the recording system worked and that there were no missing recording in any channel.

- The mobile channel may have some missing recording due to network quality. This is to be documented, but is not considered error unless it was for a whole sub list or a sizable portion of it (each user read 16 sub lists in each location).
- Not recording any part of any sub list is an error.
- If a letter is missed or substituted then this is to be counted as an error and has to be documented in the sheet. If the replacement or insertion is due to dialect then it is not counted as error but has to be documented.
- Minor stuttering is acceptable but has to be documented
- The verifier also has to comment on the pronunciation correctness

5.2. Stage 2 of verification

After verifying 20% of the first session we were confident that our recording system worked correctly except for rare instances that happened for random sub lists with random users where part of the system, e.g. sound card taking input from medium quality microphone, will not record [may be due to MATLAB having problem with reading from many opened devices]. So the verification was relaxed from verifying all the channels to only verifying the mixer and the mobile channels. This continued for the rest of the first session.

5.3. Stage 3 of verification

This stage was done for the second and third session. From the verification of the first session we were sure of the recording system and that the mobile channel was working except for missing letters or words due to network quality (the sound room is in the basement). So the verification was relaxed to verifying only the mixer channel. Moreover to enhance the verification three improvements were made.

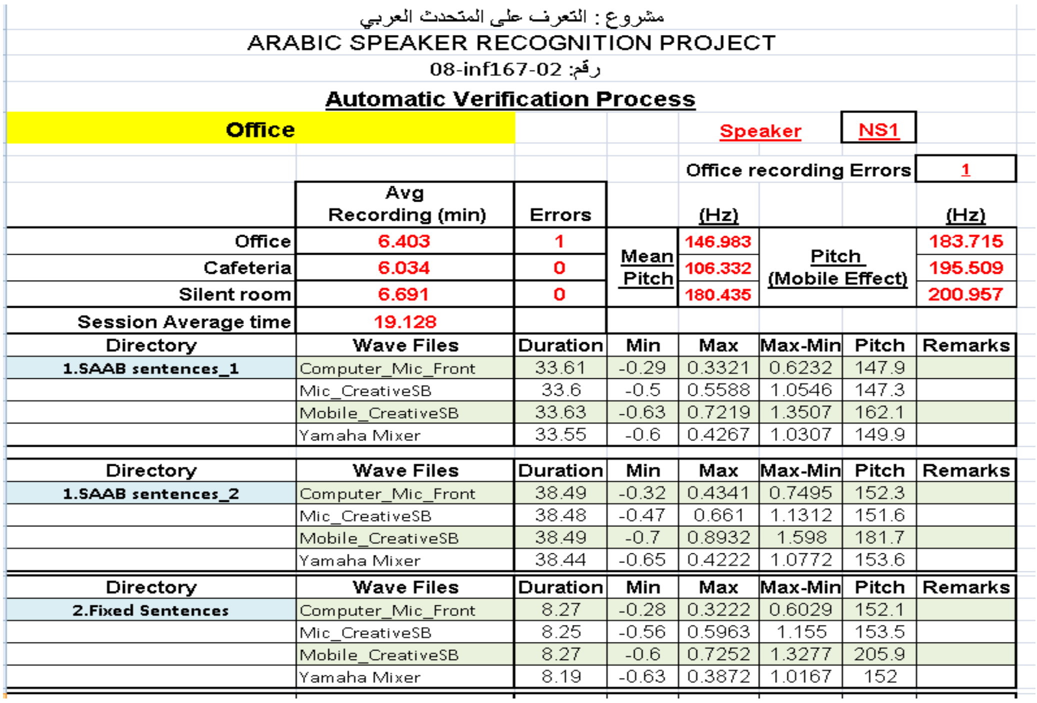
5.3.1. Improvements in stage 3 of verification

These improvements were:

First Improvement: Concatenating the recording of the lists for each volunteer at each location. This simplified the verifier task since he did not have to close and open the recording of 16 sub lists.

Second Improvement: An automatic report was generated through a developed MATLAB program that writes results directly in an excel sheet for each volunteer. The report will flag any lists that may have not been recorded or has a problem, as shown in Fig. 4, and then the verifier has to check the corresponding speech files. The pitch was used to decide if there was a recording or not. The verifier will still do his usual verification but this is an extra help.

Third Improvement: Generation of a graphic display of the recording of all channels while recording and putting the display in a report immediately after finishing the recording of any volunteer, as shown in Fig. 5. This was of great help to catch errors while recording or immediately after recording each volunteer session, so the error can be corrected.



مشروع : التعرف على المتحدث العربي

ARABIC SPEAKER RECOGNITION PROJECT

08-inf167-02 :رقم

Automatic Verification Process

Office			Speaker	NS1
			Office recording Errors	1

	Avg Recording (min)	Errors		(Hz)		(Hz)
Office	6.403	1	Mean Pitch	146.983	Pitch (Mobile Effect)	183.715
Cafeteria	6.034	0		106.332		195.509
Silent room	6.691	0		180.435		200.957
Session Average time	19.128					

Directory	Wave Files	Duration	Min	Max	Max-Min	Pitch	Remarks
1.SAAB sentences_1	Computer_Mic_Front	33.61	-0.29	0.3321	0.6232	147.9	
	Mic_CreativeSB	33.6	-0.5	0.5588	1.0546	147.3	
	Mobile_CreativeSB	33.63	-0.63	0.7219	1.3507	162.1	
	Yamaha Mixer	33.55	-0.6	0.4267	1.0307	149.9	

Directory	Wave Files	Duration	Min	Max	Max-Min	Pitch	Remarks
1.SAAB sentences_2	Computer_Mic_Front	38.49	-0.32	0.4341	0.7495	152.3	
	Mic_CreativeSB	38.48	-0.47	0.661	1.1312	151.6	
	Mobile_CreativeSB	38.49	-0.7	0.8932	1.598	181.7	
	Yamaha Mixer	38.44	-0.65	0.4222	1.0772	153.6	

Directory	Wave Files	Duration	Min	Max	Max-Min	Pitch	Remarks
2.Fixed Sentences	Computer_Mic_Front	8.27	-0.28	0.3222	0.6029	152.1	
	Mic_CreativeSB	8.25	-0.56	0.5963	1.155	153.5	
	Mobile_CreativeSB	8.27	-0.6	0.7252	1.3277	205.9	
	Yamaha Mixer	8.19	-0.63	0.3872	1.0167	152	

Fig. 4. A snapshot of the initial automatic report

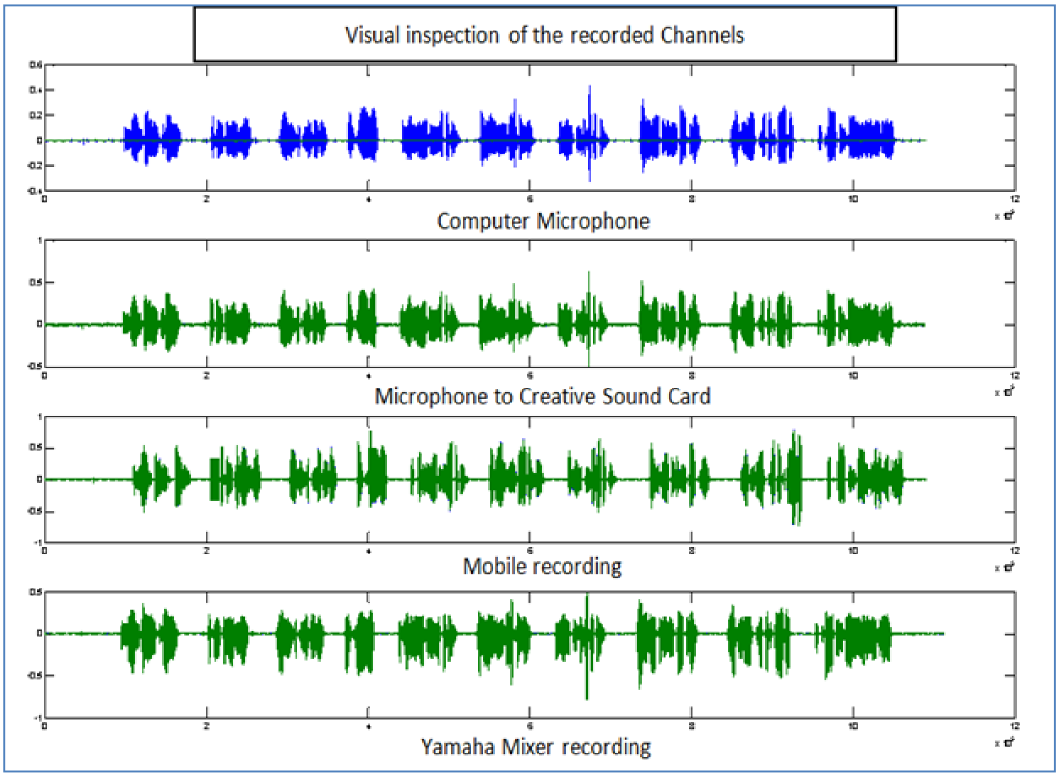


Fig. 5. A snapshot of the visual graphic display checking (office)

5.4. Verification results

Table 8 gives the number of errors detected by verification in all the three environments for the three male sessions. The verification of the female sessions is still going. A recording is considered to have an error if it contains deletion, replacement, or insertion of a character or words. Note that if this insertion or substitution is due to Arabic dialectic or non-Arab pronunciation, then it is documented but not counted as error. An important point to mention is that the severity of the errors is not as presented in the Table 8, because the database is actually 16 sub-lists and error in one of the sub lists will be counted as an error for the speaker for the whole session while in reality it is an error in one of the 16 lists.

Table 8. Verification results for the three sessions

Sessions	Place	Office	Soundproof Room	Cafeteria
First	No. of speakers	253	240	240
	Errors	42	16	24
Second	No. of speakers	206	237	-
	Errors	2	6	-
Third	No. of speakers	136	133	-
	Errors	3	2	-

By comparing the result of verification in the first and second sessions, it became clear that the improvement we made to the recording system greatly reduced the errors.

6. Validation

Validation of the database is a crucial task. Our database is a huge one. We had to select the optimum way to validate it. We are working on that question and conducting some initial experiments. Table 9 gives the results of some of the experiments on subsets of the male database. The attributes of the validation data is as in Table 10.

Table 9. Accuracy of the system with 12 MFCC and 32 Gaussians

Experiment No.	Train	Test	Number of speakers	Accuracy (%)
6 (Saudi)	Paragraph 1	Sentence 2	75	90%
7 (Saudi)	Paragraph 1	Sentence 2	138	83%
8 (Non Saudi)	Paragraph 1	Sentence 2	105	86%
9 (Saudis)	Sentence 1	Sentence 2	140	100%
10 (Non Saudis)	Sentence 1	Sentence 2	105	96%

Table 10. Attributes of the validation data.

Attribute	Value
Training set	Sentence 1
Testing set	Sentence 2
Recording room	Sound proof
Recording channel	Phantom microphone
Recording session	First
Sampling rate	16 kHz (down sampled from the original data)
Window size	20 ms
Frame rate	10 ms
Acoustic features	12 MFCC
Gaussian mixture	32

7. Conclusion

In this paper, we described a very rich and new Arabic speech database dedicated to MSA. We have presented the conditions of making a speech corpus of great quality by researchers in the field. Then we showed that our database satisfied all the conditions. The developed corpus has many dimensions of richness more than any other corpus dedicated to Arabic in the literature. We have also justified in this paper every richness aspect of our database.

Our corpus is huge in size and can be viewed as a collection of different corpora. Nonetheless, we were able to verify its content manually with documented information. We also verified it automatically by tracking the pitch value during recording sessions.

Initial validation of the database was successful and we are working on a more thorough validation.

The goal of our project is to record similar number of female Saudis utterances. We are working on that in the meantime.

The main goal of the database was to be used for speaker recognition. We went many steps ahead and made a rich and versatile database that can be used in many research areas in speech processing. For example, it can be used in the following areas: dialect/accent recognition, speaker nationality recognition, characteristics of the speech of Saudis, Arabs, and non-Arabs, effect of mobile channel in speech and/or speaker recognition, effect of low noise in speech and/or speaker recognition, the use of many channels for speech and/or speaker recognition. The list can go on and these are just examples to appreciate the richness of the database.

8. Acknowledgments

This work is supported by the National Plan for Science and Technology in King Saud University under grant number 08-INF167-02. The authors are grateful for this support. The authors also thank the project consultants Dr. Mansour Alghamdi and Prof. Sid-Ahmad Seloani for their valuable comments.

References

[1] Selouani, S. and J. Caelen, "Arabic phonetic features recognition using modular connectionist architectures. In: Proceedings of the IEEE Interactive Voice Technology for Communication, IVTTA ’98, pp. 155–160, 1998.

[2] Alsulaiman, M. and G. Muhammad, M. Bencherif, A. Mahmood and Z. Ali, “A survey on Arabic speech database”, Archives Des Sciences Journal, 2012. (submitted)

[3] Alghamdi, M., Alhargan F., Alkanhal. M., Alkhairy A., Eldesouki M. and Alenazi A., “Saudi accented Arabic voice bank,” J. King Saud University, Computer and Information Sciences, vol. 20, pp. 45-62, 2007.

[4] Makhoul, J., Zawaydeh B., Choi F., and Stallard D., “2005 BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts,” Linguistic Data Consortium (LDC), Philadelphia, 2005. LDC Catalog Number LDC2005S08.

[5] Harrag, A. and Mohamadi T., " QSDAS: New Quranic Speech Database for Arabic Speaker Recognition", The Arabian Journal for Science and Engineering, vol. 35, no. 2C, pp. 7-13, December 2010.

[6] Abushariah, M. and Ainon R., Zainuddin R., Alqudah A., Ahmed M., and Khalifa O., "Modern standard Arabic speech corpus for implementing and evaluating automatic continuous speech recognition systems", vol. 349, no. 7, Journal of the Franklin Institute, 2011.

[7] Droua-Hamdani, G. and Selouani S. A., and Boudraa M., "Algerian Arabic speech database (ALGASD): corpus design and automatic speech recognition application", The Arabian Journal for Science and Engineering, vol. 35, no. 2C, pp. 157-166, December 2010.

[8] Stephen, Col., A. LaRocca and RajaaChouairi, "West Point Arabic Speech", LDC Catalog LDC2002S02, 2002.

[9] NetDC Arabic BNSC, ELRA Catalog ELRA-S0157.

[10] GlobalPhone Arabic, ELRA Catalog ELRA-S0192.

[11] The Egyptian Arabic Speecon Database, ELRA catalog ELRA-S0308.

[12] A-Speech DB, ELRA catalog ELRA-S0315.

[13] OrienTelMorocco MCA, ELRA catalog ELRA-B0004.

[14] OrienTelTunisia MCA, ELRA catalog ELRA-B0005.

[15] OrienTelEgypt MCA, ELRA catalog ELRA-B0006.

[16] OrienTel United ArabEmirates MCA, ELRA catalog ELRA-B0010.

[17] OrienTel Jordan MCA, ELRA catalog ELRA-B0011.

[18] NEMLAR Broadcast News Speech Corpus, ELRA catalog ELRA-S0219.

[19] Garofolo J. S., Lamel L. F., Fisher W. M., Fiscus J. G., Pallett D. S., and Dahlgren N. L., " DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM", NIST, 1993. Available at http://www.ldc.upenn.edu/Catalog/docs/LDC93S1.

[20] http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html

[21] Graff D., and Walker K., and Canavan A., Switchboard-2 Phase I, II. Linguistic Data Consortium, Philadelphia. Available at www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC98S75; www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S79;

[22] Garcia J. O., Rodriguez J. G., and Aguair V. M., “AHUMADA: A large speech corpus in Spanish for speaker characterization and identification,” Speech Communication, vol. 31, pp. 255-264, 2000.

[23] Petrovska D., Hennebert J., Melin H., and Genoud D., “POLYCOST: a telephone speech database for speaker recognition,” RLA2C, Avignon, France, pp. 211–214, 20–23 April 1998.

[24] Mihelic F., and Gros J., Dobrisek S., Zibert J., and Pavesic N., “Spoken Language Resources at LUKS of the University of Ljubljana,” International Journal of Speech Technology, vol. 6, pp. 221–232, 2003.

[25] The NIST Year 2010 Speaker Recognition Evaluation Plan, available at http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf

[26] http://www.nist.gov/speech/tests/spk/index.htm

[27] Li K., Dammann PJ. E., and Chapman W. D., “Experimental studies in speaker verification, using an adaptive system”, J. Acoust. Soc. Am., vol. 40, no. 5, pp. 966-978. 1966.

[28] Sambur M. R., “Selection of acoustic features for speaker identification,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 23, pp. 176–182, 1975.

[29] Wolf J. J., “Efficient acoustic parameters for speaker recognition,” Journal of the Acoustical Society of America, vol. 51, pp. 2030– 2043, 1972.

[30] Habib S. M. M., Alam F., Rabia Sultana, Chowdhury S. A. and Mumit Khan, “Phonetically Balanced Bangla Speech Corpus”, Conference on Human Language Technology for Development, Egypt, 2011.

[31] Iida A., Campbell N., “A database design for a concatenative speech synthesis system for the disabled”, Fourth ISCA ITRW on Speech Synthesis (SSW-4), Interspeech 2001.

[32] Reynolds D. A., “An overview of automatic speaker recognition Technology,” Proc. IEEE international conference on acoustics, speech and signal processing, ICASSP’02, vol. IV, pp. 4072–4075, 2002.

[33] Doddington G. R., and Przybocki M. A., Martin A. F., and Reynolds D. A., “The NIST speaker recognition evaluation overview: methodology systems, results, perspective,” Speech Communications, 31, pp. 225–254, 2000.

[34] Kersta L. G., “Voiceprint classification for an extended population,” Journal of the Acoustical Society of America (A), vol. 39, pp. 1239, 1966.

[35] Reynolds D. A., “The effects of handset variability on speaker recognition performance: Experiment on the Switchboard corpus,” Proc. IEEE international conference on acoustics, speech, and signal processing, ICASSP’96 pp. 113–116, 1996.

[36] Norton R., “The evolving biometric marketplace to 2006,” Biometric Technology Today, 10(9), pp. 7–8, 2002.

[37] Reynolds D. A., “Gaussian Mixture Models”, Encyclopedia of Biometric Recognition, Springer, Journal Article, February 2008.

[38] Sturim D. E., Campbell W. M., Reynolds D. A., Dunn R. B., Quatieri T. F., “Robust Speaker Recognition with Cross-Channel Data: MIT/LL Results on the 2006 NIST SRE Auxiliary Microphone Task", ICASSP 2007, Apr. 15-20, 2007.

[39] Patil H. A. and Basu T. K., “Development of speech corpora for speaker recognition research and evaluation in Indian languages”, Int. J. of Speech Tech., Springer-Verlag, vol. 11, no.1, pp.17-32, March 2008.

[40] Boudraa M., Boudraa B., and Guerin B., “Twenty Lists of Ten Arabic Sentences for Assessment”, ACUSTICA, ACTA-ACUSTICA, vol. 86, no. 5, 1998.

[41] Alsulaiman, M. and Muhammad G., Bencherif M. and Mahmood A., “Arabic speaker Recognition”, Tech. Report, Research Center, College of Computer and Information Sciences, King Saud University, Sep 2012.

*Corresponding author: Zulfiqar Ali
Speech Processing Lab, College of Computer and Information Sciences,

King Saud University,

Riyadh 11543, Saudi Arabia

Email: zuali @ksu.edu.sa