# Saudi Accented Arabic Voice Bank

**Mansour Alghamdi, Fayez Alhargan, Mohammed Alkanhal,**
**Ashraf Alkhairy, Munir Eldesouki and Ammar Alenazi**
*Computer and Electronic Research Institute*
*King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia*

**Abstract.** The aim of this paper is to present an Arabic speech database that represents Arabic native speakers from all the cities of Saudi Arabia. The database is called the Saudi Accented Arabic Voice Bank (SAAVB). Preparing the prompt sheets, selecting the right speakers and transcribing their speech are some of the challenges that faced the project team. The procedures that meet these challenges are highlighted. SAAVB consists of 1033 speakers speak in Modern Standard Arabic with a Saudi accent. The SAAVB content is analyzed and the results are illustrated. The content was verified internally and externally by IBM Cairo and can be used to train speech engines such as automatic speech recognition and speaker verification systems.

*Keywords:* Arabic; speech recognition; database; speech sounds; Saudi dialects

## 1. Introduction

Speech databases are essential for training automatic speech recognition systems [1] in addition to other applications such as speaker verification [2, 3, 4], accent [5] and language identification [6]. Speech databases are also valuable in linguistic studies especially in the areas of phonetics, phonology, typology and sociolinguistics [7]. For these reasons, speech databases of several languages have been collected during the last few decades in different countries: English in Australia, Australian National Database of Spoken Language (ANDOSL) [8]; British English, Spoken English Corpus (SEC) [7], British English speech corpus (WSJCAMO) [1]; American English, Texas Instrument and Massachusetts Institute of Technology corpus (TIMIT) [9], Macrophone [10]; languages and their dialects in North and Latin America, SpeechDat across Latin America (SALA-I and SALA-II) [11]; Multilingual database, GlobalPhone [12]; Dutch, The Spoken Dutch Corpus [13]; Chinese Spontaneous Telephone Speech Corpus on Flight Enquiry and Reservation (CSTSC-Flight) [14]; Mandarin Across Taiwan (MAT) [15, 16]; Cantonese, one of the dialects in southern China [17]; American Spanish,

Voice Across Hispanic America (VAHA) [18]; French, French SpeechDat corpus (FRESCO) [19]; Indian languages: Tamil, Telugu and Marathi [20]; SpeechDat Car includes nine EU languages: Danish, English, Finnish, Flemish/Dutch, French, German, Greek, Italian and Spanish [21].

One well known and widely used speech database is TIMIT [9]. It consists of 630 native speakers of American English, of whom 70% are male and 30% female. Each speaker reads 10 sentences in one session that takes approximately 30 seconds. The sentences are 2 dialect sentences, 450 phonetically compact sentences and 1890 phonetically diverse sentences. The utterances are orthographically transcribed, for example: '0 58164 These exclusive documents must be locked up at all times'. The two numbers at the beginning are the start and end sample numbers. The utterances are also phonetically transcribed using the English alphabet. For example, the transcription of the first word of the above sentence is: '0 2190 h#2190 2450 dh2450 4610 iy4610 5880 z' where 'h#' represents the silence before the utterance. The two numbers before the phones, dh iy z, are the start and end sample numbers. The utterances are recorded in a soundproof chamber

at 16-kHz sample rate. TIMIT has been used in different experiments including speech [22] and speaker [2, 3, 4] recognition systems.

For the Arabic language and its dialects, spontaneous telephone dialogues have been collected from Arabic speakers from Egypt, Lebanon, Syria, Palestine and Jordan by Linguistic Data Consortium [23]. European Language Resources Association (ELRA) has read and spontaneous Modern Standard Arabic speech databases spoken by Moroccans (530 speakers), Tunisians (598 speakers) and Egyptians (750 speakers) [24]. The speakers utter around 49 read and spontaneous items recorded over fixed and mobile telephone networks. The uttered items include: isolated and sequence digits, currency money amounts, dates, time phrases, spelled words, yes/no questions, phonetically rich words and sentences and spontaneous items. Appen Gulf Arabic database has 150 speakers from Saudi Arabia and the United Arab Emirates [25]. The 75 speakers from Saudi Arabia are randomly selected, and the speech recording is desktop based. In the GlobalPhone speech database, the linguistic materials are selected from political and economic topics in newspapers [12]. The speakers are from Tunis and Sfax, Tunisia. They speak Modern Standard Arabic. One hundred and seventy speakers (equal number of both sexes) read for approximately 20 minutes each. Each speaker speaks in one session. The recording is carried out in quiet rooms and public places. The data are recorded using a digital tape recorder at 48-kHz sampling rate. The database is validated by a native speaker who listens to the speech and follows the text that is supposed to be read. The NEMLAR Arabic Broadcast News Speech Corpus consists of 40 hours of transcribed Standard Arabic speech from 259 speakers [26]. The data are recorded from four different radio stations. Each recording contains 25–30 minutes of news and interviews.

Speech corpus is usually collected spontaneously [14, 27], canonically [1, 7, 9, 12, 25, 28], or both [8, 11, 15, 18, 19, 23, 24]. A spontaneous speech corpus is collected from real world human–human/human–machine communication. In this case, the speaker does not read prompts, rather he or she speaks naturally to convey a message and/or get information. The speech is expected to be natural and not affected by the reading habits. A canonical speech corpus, on the other hand, is designed for speakers to follow

certain procedures, including the reading of prompts, for the purpose of collecting specific speech sounds. The read speech content can be words [28], sentences such as application phrases and phonetically rich sentences [9, 27] or text such as news, lectures, articles and stories [1, 7, 12]. Although the former is suitable for language understanding and dialogue design, it tends to include unneeded frequently repeated words and utterances but does not necessarily include all the sounds of the language under investigation. However, a canonical speech corpus tends to be phonetically rich, i.e. all the sounds of the language are presented in various phonotactic positions. Most of the recently developed speech databases, SAAVB included, belong to the third type, which contains spontaneous and read speech.

Although speech databases have been collected for several languages, Arabic speech databases need more work to cover the dialectal diversity. Many of the 22 Arabic-speaking countries still remain with almost non-professional speech collections. Saudi Arabia is one country where a speech database that covers its various dialects has not been collected before this project except for the one mentioned above which consists of a limited number of speakers [25].

The area covered by Saudi Arabia is 1,960,582 sq km. It is located in south-west Asia and is surrounded in the north, east and south by other Arab countries. Saudi Arabia has about 20 million inhabitants and four-fifths of them are native Saudis [29]. Its location means that some parts of its population share linguistic features with the neighboring Arab countries. For example, Saudis who live in the Eastern Region speak to the inhabitants of the other Gulf countries in the Gulf dialect. People who live in the north of the country share some linguistic characteristics with the people of Jordan and Iraq. Similar cases exist between the Saudis who live in the south and west of Saudi Arabia and the Yemenis and Egyptians, respectively.

A project that aims to collect speech database faces several obstacles. Finding the right speakers that represent the population is one example. Another example is choosing the linguistic materials that are suitable for both the speaker's culture and useful for training speech recognition systems. Speech

transcription is yet another challenge.

This paper has been written to assist SAAVB users and to document its procedures, specifications and contents for those who are interested in collecting similar speech data in a similar environment [30, 31].

## 2. Methods and Procedures

The procedures that are applied to collect SAAVB consist of four phases: 1) designing the prompt sheet, 2) selecting the speakers, 3) recording the speech and 4) transcription (Figure 1). Each phase needs careful planning and execution.
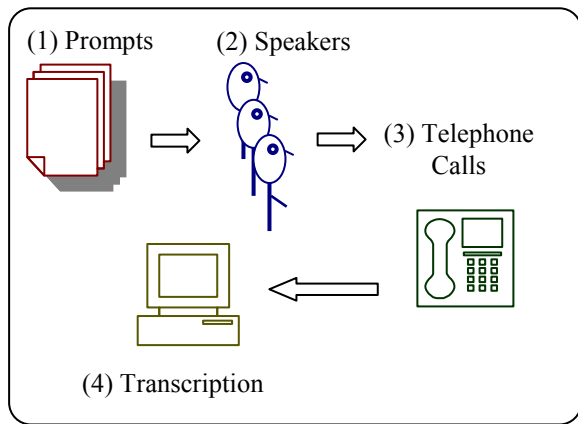


**Figure 1. SAAVB Procedures.**

### 2. 1. Designing the Prompt Sheet

Each speaker is given a five-page document. The first page has a code that gives each speaker access to the recording laboratory to record his/her speech. The code symbolizes the region, city, gender, age, telephone type and calling environment of the speaker. The code consists of six digits (the six digits from the left in Figure 2) and is to be used as the name of the directory of the speaker, so that the gender, age and other information related to the speaker can be extracted from the name of the directory of each speaker. The files in each directory have the code as part of their names with two more digits to represent the item number (Figure 2). The second page contains the instructions that help the speaker to log into the recording system and complete the required tasks. The remaining three pages are the prompt sheets.
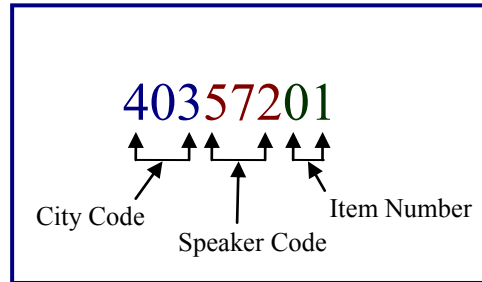


**Figure 2. An example of a file name: the first three digits represent the city; the second three digits represent the speaker (the first digit shows how many speakers are identical in terms of specifications (5 means one and 7 means two); the second digit represents the age and gender (0=16–30 year male, 1=31–45 year male, 3=46–60 year male, 5=16–30 year female, 7=31–45 year female, 9=46–60 year female); the third digit represents the environment and telephone type (0=quiet fixed, 2=noisy fixed, 4=quiet cellular, 6=noisy cellular, 8=vehicle cellular); the last two digits represent the prompt item number.**

Speech databases differ in the number of utterances for each speaker:

| Number of Utterances | Speech Database | Reference |
|---|---|---|
| 10 | TIMIT | [9] |
| 37 | FRESCO | [19] |
| 44 | SALA | [11] |
| 45 | Macrophone | [10] |
| 66 | MAT | [16] |
| 110 | WSJCAMO | [1] |

The number of utterances for each speaker in SAAVB is 59 (Appendix I). They consist of 49 read items (83%) and 10 elicited spontaneous responses (17%). All the read items are written in Modern Standard Arabic which is widely used in the media and press, and for official communication in the Arab world. However, the prompts are written without diacritics to allow for accent variation, at least in the vocalic pronunciation. The overall content of the prompt sheet is shown in Table 1.

The result is a unique prompt sheet for each speaker; no two prompt sheets are the same. The only two utterances that appear in every prompt sheet are Prompts 41 and 42. These two sentences are designed to record accent variations among speakers. The total number of prepared prompt sheets is 1059.

Statistical analysis of the content of all prompt sheets shows that the total number of words is 270,828. This means that the average word number per prompt sheet is 263 words. The prompt dictionary consists of 12,024 words: 4,610 of them occur only once (38%), 1,022 words occur twice (8.5%), 305 words occur three times (2.5%) and the remaining 6,087 words occur more than three times (51%). The average reappearance of each word in the dictionary is 22.5 times. The most repeated words are the digits (0, 1, 2, 3, 4, 5, 6, 7, 8 and 9) which occupy 39% of the total word number. The second most frequent words are the prepositions (في (in), و (and), إلى (to), من (from)) which occupy 4% of the words. Therefore, a high proportion of the words in the dictionary occur only once while 24% of the words occur more than 10 times (see Figure 3).

In addition to the Arabic alphabet, the prompts include email addresses written in the English alphabet. The total number of character clusters (a character cluster is a group of letters and numbers that are separated by @ or a dot in the email) is 8118. The English alphabet dictionary consists of 1898 character clusters. Segments such as 'com' and '@' occur 1771 and 1904 times, respectively.

**Table 1: The frequency and description of the prompt sheet content.**

| Frequency | Description |
|---|---|
| | **Digits** |
| 1 | One digit (read) |
| 1 | Four digits (read) |
| 1 | Five digits (read) |
| 1 | Six digits (read) |
| 1 | Two to seven digits (read) |
| 1 | Seven digits (read) |
| 1 | Ten digits (read) |
| 1 | Mobile telephone number (read) |
| 1 | Credit card number (read) |
| | **Money amounts** |
| 1 | Saudi Riyals (read) |
| 1 | US Dollars (read) |
| 1 | Euros (read) |
| | **Yes/no questions** |
| 1 | Expected 'yes' answer (spontaneous) |
| 1 | Expected 'no' answer (spontaneous) |
| | **Dates** |
| 1 | Hijrah (spontaneous) |
| 1 | Hijrah (read) |
| 1 | Gorgonian (read) |
| | **Reference point in time** |
| 1 | Read |
| | **Times** |
| 1 | (Spontaneous) |
| 1 | (Read) |
| | **Cities** |
| 1 | National city (spontaneous) |
| 1 | National city (read) |
| 1 | International city (read) |
| | **Names** |
| 1 | National company (read) |
| 1 | Government institution (read) |
| 1 | Personal name (read) |
| 1 | Personal name (spontaneous) |
| | **Spelling** |
| 1 | Phonetically rich word (read) |
| 1 | Personal name (read) |
| 1 | City name (read) |
| | **Phonetically rich words and sentences** |
| 4 | Sequence of four words (read) |
| 9 | Sentences (read) |
| | **Emails** |
| 1 | (Spontaneous) |
| 1 | (Read) |
| | **Location** |
| 1 | (Spontaneous) |
| | **Telephone types** |
| 1 | (Spontaneous) |
| | **Accent variation sentence** |
| 1 | Question (read) |
| 1 | Statement (read) |
| | **Application words/phrase** |
| 4 | Commands (read) |
| 4 | Questions (read) |
| | **Spelled out number** |
| 1 | One or two words (read) |
| | **Speaker's statement about his/her city** |
| 1 | Sentence (spontaneous) |
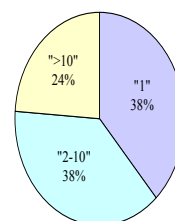
**Figure 3. The word-frequency percentage in the prompt sheet: 1=percentage of words that occur only once, 2–10=percentage of words that occur from two to ten times, >10=percentage of words that occur more than 10 times.**

## 2. 2. Selecting the Speakers

One of the obstacles that prevents identification of

the right subjects in Saudi Arabia and similar countries is the lack of a dialect map. A dialect map is essential in order to get a sample that represents at least the major dialects in a country. To overcome this obstacle, the research team decided to select a sample from every city in the country. There are 118 cities in Saudi Arabia spread all over the country area (Figure 4). The target number of speakers is divided among them according to the city population.
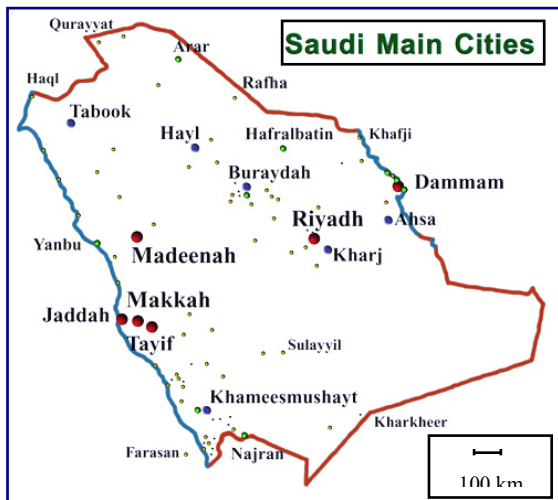


**Figure 4. Map of Saudi Arabia showing the locations of its main cities.**

Our objective was to determine a model for the number of speakers per city as a function of the population of the city. In order to do this, we needed to develop a parameterized structure for the relationship, and then compute the parameter values based on constraints, such as the maximum number of speakers that could be part of the experiment.

First, we investigated the statistics concerning the population of the cities and the factors involved in the experiment design. The first two columns of Table 2 list the population sizes of the cities of Saudi Arabia and the number of cities, as determined from the Saudi census data of 1999 [29]. The remaining two columns will be explained later in this section.

The factors in the experiment design, along with the target representation in the experiments, are as follows: the target representations are indicative of the breakdown of applications built using SAAVB. The percentages below are based on rough estimation

of the telephone users within the Kingdom.

Gender:
- 50% male
- 50% female

Age:
- 50%  16–30 years
- 35%  31–45 years
- 15%  46–60 years

Telephone type
- 70% cellular telephones
  - 35% quiet environment
  - 35% noisy environment
  - 30% vehicle
- 30% fixed telephones
  - 75% quiet environment
  - 25% noisy environment

**Table 2: The number of speakers with reference to city populations. Cities are clustered to the nearest population figure.**

| Population | Number of Cities | Log10 (Population) | Num Cities * Log(Pop) |
|---|---|---|---|
| 3,000,000 | 1 | 6.48 | 6.48 |
| 2,000,000 | 1 | 6.30 | 6.30 |
| 1,000,000 | 1 | 6.00 | 6.00 |
| 500,000 | 3 | 5.70 | 17.10 |
| 250,000 | 3 | 5.40 | 16.19 |
| 220,000 | 3 | 5.34 | 16.03 |
| 140,000 | 4 | 5.15 | 20.59 |
| 110,000 | 3 | 5.04 | 15.12 |
| 90,000 | 2 | 4.95 | 9.91 |
| 70,000 | 4 | 4.85 | 19.38 |
| 50,000 | 48 | 4.70 | 225.55 |
| 5,000 | 45 | | |
| **Excl 5,000** | **Σ = 73** | | **Σ = 358.64** |

The number of factors is three: gender, age and telephone type. If each factor is only allowed to have two values (binary), then the least number of experiments needed to investigate the factors is four [32]. Hence, the minimum number of speakers in a city is four. This applies to cities with a  population of 5,000. Of course, the number of values for each factor is not necessarily two.

To study the minimum number of experiments needed to investigate the values of the factors, we reasoned as follows. The gender factor has two values

(male and female), implying a single binary factor (a factor with two possible values). The age factor, which has three values, can be considered as a composite of two factors, with each factor having two values. This results in four values, which can be collapsed into three by considering two of the values to be a single value – in accordance with common practice in experimental design. Hence, Age is investigated using two binary factors. Telephone type is nested, with the top decomposition being binary (cellular and fixed). Since the fixed telephone has two values (quiet and noisy), it can be handled with a single binary factor. The number of values for the cellular telephone is three (quiet, noisy and vehicle), implying two binary factors. Thus, in total, the number of binary factors for the Telephone variable is four. In summary, the total number of binary factors needed to incorporate age, telephone type and gender fully is seven (one for gender, two for age, four for telephone type). This implies that the minimum number of experiments needed to investigate all aspects is eight. Hence, cities with a population of 50,000 need eight experiments (speakers).

We can now model the functional relationship between the number of speakers per city, s, and the city population, CP, as follows:

$$s = \alpha + \beta \log(CP) \qquad CP \geq 50{,}000$$

$$s = 8 \qquad\qquad CP = 50{,}000$$

$$s = 4 \qquad\qquad CP = 5{,}000$$

A log relationship was utilized to allow the number of speakers to increase at a decreasing rate as the city population increased. Such a mapping allows a more weighted representation of the variety of dialects that are contributed by the smaller cities, thereby providing a richer voice bank.

The parameters $\alpha$ and $\beta$ are determined based on two constraints. First, the number of speakers equals eight for a single city with a population of 50,000. Second, the total number of speakers is 1056. Hence, the total number of speakers, excluding cities of population size 5,000 is $1056 - 45 \times 4$ speakers. We should mention here that we chose the upper limit to be 1,000 in order to complete the recordings within the schedule. We are then able to increase the total

number of speakers, TT, to 1,056.

The above two constraints result in the following two equations, based on the data giving the number of cities of different population sizes:

$$876 = 73\,\alpha + 358.6442769\,\beta$$

$$8 = \alpha + (\log 50000)\,\beta$$

The solution to the above simultaneous equations is $\alpha = -79.8455$, $\beta = 18.6946$. The predicted numbers of speakers, based on the resulting model, are shown in Table 3.

**Table 3: Model and actual number of speakers per city.**

| Population | Number of Cities | Model Num of Speakers per city | Actual Num of Speakers per city |
|---|---|---|---|
| 3,000,000 | 1 | 41.242 | 38 |
| 2,000,000 | 1 | 37.95 | 36 |
| 1,000,000 | 1 | 32.322 | 32 |
| 500,000 | 3 | 26.695 | 30 |
| 250,000 | 3 | 21.067 | 22 |
| 220,000 | 3 | 20.029 | 20 |
| 140,000 | 4 | 16.359 | 16 |
| 110,000 | 3 | 14.401 | 14 |
| 90,000 | 2 | 12.772 | 12 |
| 70,000 | 4 | 10.732 | 10 |
| 50,000 | 48 | 8 | 8 |
| 5,000 | 45 | 4 | 4 |

The actual number of speakers planned for the cities is shown in the rightmost column of Table 3. The deviation from the model is a result of rounding to integers and making adjustments needed to meet constraints related to the practicalities of acquiring the data.

The total number of speakers (1,056) is distributed among the categories mentioned above, keeping the same proportion for each category. The results are shown in Figure 5. SAAVB target number of speakers is within the average of other similar reported speech databases:

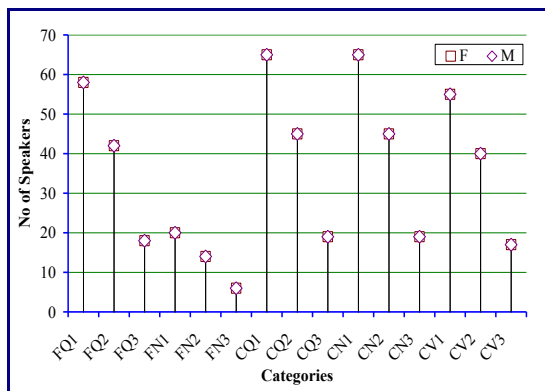| Number of Utterances | Speech Database | Reference |
|---|---|---|
| 90 | Speech Database for VOYAGER | [27] |
| 100 | GlobalPhone, Cantonese | [12, 17] |
| 140 | WSJCAMO | [1] |
| 500 | VAHA | [18] |
| 630 | TIMIT | [9] |
| 1253 | FRESCO | [19] |
| 1300 | PhoneBook | [28] |
| 2000 | SALA | [11] |
|  | Macrophone | [10] |



**Figure 5. Number of male (M) and female (F) speakers in each category. The symbols on the x-axis represent: F=fixed telephone, Q=quiet environment, N=noisy environment, C=cellular telephone, V= vehicle, 1=aged 16–30, 2=aged 31–45 and 3=aged 46–60.**

There are several options available to approach the target speakers and record their voices. One of them is through advertisements which can be on television, the Internet or in newspapers or other available media. Another option is to search for speakers with the target dialects in large cities such as Riyadh where thousands of Saudis have moved from different cities. The option the SAAVB team chose is to record the speakers in their natural dialect environment, i.e. the speakers should be living in the cities they linguistically represent. Moreover, all the speakers in SAAVB spent at least their first five years

in the same cities from which they have been selected and recorded.

Coordinators were hired to travel to the selected cities in the Kingdom. They carried with them the number of speakers they are supposed to find in each city, their categorical specifications and their prompt sheets. The coordinators organized meetings with local people and gave presentations about the project and how speakers' voices are recorded and used in research. With a lot of hard work and determination, the coordinators received all the support they needed and were able to locate almost all the targeted speakers. Each speaker who agreed to participate in the project signed an agreement that his/her recorded speech would be recorded and used to develop speech-related systems. Compensation was paid to the speaker when the speech was recorded and approved by the transcribers.

**2. 3. Speech Recording**
Some speech databases are recorded directly from the speakers or through a broadcasting medium such as radio or television [7, 8, 12, 13, 15, 17, 27, 26]. The other speech databases, SAAVB included, are recorded through a telephone system [10, 11, 13, 14, 16, 19, 24, 28, 33]. Each of the above categories has its own objectives and applications where the speech content and the recording procedures are different. The first is usually recorded at a high sampling rate: 16 kHz [12], 22 kHz [26], 48 kHz [17]. The latter has a lower sampling rate, 8 kHz, due to the telephone system bandwidth [14, 16, 19, 24, 33].

A data acquisition laboratory (DAL) is set up to receive and record telephone calls and other information related to the speakers. It consists of four PCs with a Dialogic card (Dialogic D/41ESC) installed in each one. The Dialogic cards are connected to 11 extension telephone lines and one direct telephone line. The direct telephone line is used to acquire the speaker's telephone number. The system uses the number to call the speaker back. This prevents the speakers from bearing the calling cost since a toll-free service telephone number was not available for cellular phones in the Kingdom at the time. This option allows the speakers to speak normally without being worried about call costs. SAAVB recording took place during the period from August 2002 to September 2003.

The objective of the DAL is to automate acquisition of the data by receiving the telephone calls, interacting with the speakers, recording their speech and managing the database automatically. The recording process is as follows:

1. A speaker is selected by a coordinator according to specific criteria.
2. The speaker calls the DAL by dialing the number of the direct telephone line.
3. The speaker disconnects the call after hearing the telephone second ring.
4. The speaker waits for the system to call back within seconds.
5. DAL calls the speaker and plays the instructions until it says, 'Now read the first item after hearing the tone'.
6. The speaker reads the first item and presses the button '1' once completed.
7. Steps 5 and 6 are repeated for all remaining items.
8. When the speaker completes all the items, DAL plays a message of appreciation and disconnects the call.
9. DAL calls the coordinator and plays a recorded message that says the speaker has completed the recording.
10. The coordinator calls the system and listens to all the recorded items and approves them or deletes any inappropriate ones.
11. The coordinator calls the speaker and asks him/her to call the system again and record any inappropriate items. When the coordinator is satisfied, he asks the system to save all the recorded items of the speaker.
12. The system saves the speaker's information (the telephone number(s) he used, time(s) and date(s) of the call(s)) in a text file and every linguistic prompt (Appendix I) in a separate file encoded using a μ-law 8 bit pulse-code modulation (mono PCM format). Each speaker's files carry the extension of the speaker's code followed by a sequential number. All files belonging to the same speaker are saved in an independent directory that carries the speaker's code (Figure 2).
13. The speaker's files are saved on one of the PCs and a backup copy is made on every other PC (three backup copies).
14. Once a speaker has completed all the prompts, the system does not allow the same speaker to login again unless the coordinator deletes the files needed to be re-recorded.

DAL performs other functions such as detecting power failure, disconnection of a telephone line and the volume of the speaker's voice, in addition to monitoring the security of the laboratory.

## 2. 4. Transcription

Speech transcription can be done in IPA, SAMPA, Worldbet [34] or the language orthography: English [1, 9], Chinese [14], Mandarin [15], Arabic [23]. Orthographic transcription is widely used for the reason that it is more convenient for the transcribers.

Arabic orthography is used to transcribe the audio files of SAAVB. The Arabic alphabet consists of the letters in Table 4 in addition to those in Table 5. The letters in Table 5 are used in writing but they are either other symbols of the consonants in Table 4, symbols of a vowel or represent a combination of a consonant and a vowel.

**Table 4. Arabic orthography (AO) and their representations in International Phonetic Alphabet (IPA) as the consonant inventory of Modern Standard Arabic.**

| AO | IPA | AO | IPA | AO | IPA | AO | IPA |
|----|-----|----|-----|----|-----|----|-----|
| ب | " | ذ | ? | ط | #$^{\surd}$ | ل | W |
| ت | # | ر | 6 | ظ | ?$^{\surd}$ | م | . |
| ث | > | ز | A | ع | M | ن | 0 |
| ج | C | س | @ | غ | K | ه | N |
| ح | L | ش | B | ف | < | و | f |
| خ | J | ص | @$^{\surd}$ | ق | + | ي | U |
| د | $ | ض | $$^{\surd}$ | ك | ) | ء | - |

In addition to the orthographic symbols for consonants, there are symbols for vowels (Table 6). Arabic vowels, although rarely used in writing, are written as diacritics above or below the letters. Arabic diacritics include symbols other than the vowels (Table 7). When an Arabic text is fully diacritized, every letter must be followed by a diacritic except 'ى, ا, آ, assimilated ل (*Allam Ashamsiyah*)' and the long vowels 'و and ي'.

**Table 5. Additional Arabic orthographic symbols (AO) and their IPA representations.**

| AO | IPA | AO | IPA |
|----|-----|----|-----|
| ى | • | ئ | - |

| AO | IPA | AO | IPA |
|---|---|---|---|
| أ | - | ؤ | - |
| إ | - |  |  |
| ا | /-•/ "utterance initial as in "العلم"" | | |
| ا | /•/ "preceded by /a/ within a word as in "عالم"" | | |
| ا | ∅ "word initial but not utterance initial as in "في العلم"" | | |
| ة | /N / "utterance final as in "سماء صافية"" | | |
| ة | /#/ "else as in "معرفة الإنسان"" | | |

| Diacritic | Definition |
|---|---|
| ِ | Kasrah: represents the high front vowel /i/. |
| ّ | Shaddah: the preceding consonant is geminate. |
| ْ | Sukoon: the preceding consonant is neither followed by a vowel nor geminate. |
| ٌ | Tanween Dham: /-un/ comes as word final. |
| ً | Tanween Fateh: /-an/ comes as word final. |
| ٍ | Tanween Kasr:; /-in/ comes as word final. |

The symbols in Tables 4–6 are used in SAAVB transcription with some modifications. For example, 'ة' is used when the speaker pronounces it as /t/, but 'ه' is used when it is pronounced as /h/. Each consonant is followed by a diacritic except for ' ا و ي'when they represent long vowels. When a letter immediately follows another letter without a diacritic between them, the two letters represent one phoneme that does not have a symbol in Arabic orthography (Table 8). This procedure solves the occurrence of a consonant that is not represented in Arabic orthography.

Vowel production variations are clustered around the Arabic three main vowels (/a u i/). For example, '◌ُ' is used to represent /u/ and /o/ and '◌ِ' is used to represent /i/ and /E/.

**Table 6. Arabic short (left) and long (right) vowels.**

| AS | IPA | AS | IPA |
|---|---|---|---|
| ◌َ | • | ا | •á |
| ◌ُ | TM | و | TMá |
| ◌ِ | ' | ي | 'á |

**Table 7. Arabic diacritics. The horizontal line represents an Arabic letter.**

| Diacritic | Definition |
|---|---|
| ◌َ | Fathah: represents the low vowel /a/. |
| ◌ُ | Dhammah: represents the high back vowel /u/. |

**Table 8. More symbols for other Saudi dialect consonants. /v/ and /p/ are the sounds that occur in foreign words especially in email addresses. The other sounds are variations of some modern standard Arabic.**

| MSA | AO | IPA | MSA | AO | IPA |
|---|---|---|---|---|---|
| -- | فف | = | k | تش | #B |
| -- | بب | ! | k | تس | #@ |
| C | جج | $C | q | جق | * |
| ?√ | زظ | A√ | q | دز | $A |

The transcribers were carefully selected and trained how to transcribe SAAVB files by applying transcription rules. To make their work convenient, a code was written to create a transcription interface (Figure 6). The interface is meant to minimize the time and effort taken by the transcribers. The transcription procedure is as follows:

1. A transcriber logs into the system using his user name and password.
2. Once in the transcription window, he selects a speaker and one of the speaker's files.
3. The content of the file appears as a text in the upper window (Figure 6). This text is the one that the speaker had to read on the prompt sheet.
4. The transcriber listens to the whole audio file or part of it as many times as needed and whenever needed. He can listen to the utterance and display its waveform and spectrogram to have a closer look at the signal or to play a certain portion of it.
5. The transcriber then writes the transcription of the audio file in the lower window below the original

text. The transcription in this window must by fully diacritized.

6. When the transcription is completed and reviewed by the same transcriber, the transcription is saved by clicking the 'save' button.

7. A reviewer would login to review the transcription of the previous transcriber. He has the option to alter the previous transcription or approve it. Once he has gone over all the details of the transcription, he saves his work.

8. The system archives all files including different revisions for reference and evaluation.

The transcription is unpunctuated [7] and includes marks for non-speech sounds such as background and telephone line noise. It has two layers: the orthographic and the phonetic [25, 21, 26, 33]. The orthographic represents the writing system that is used in reading. The phonetic transcription represents the actual sounds of the speech.
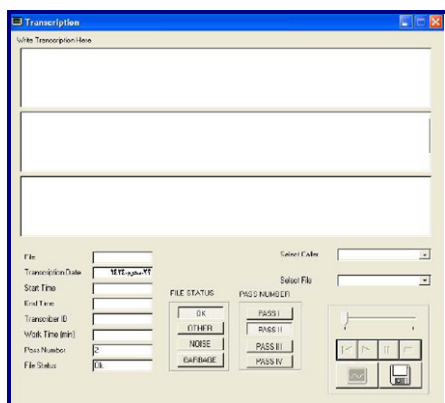


**Figure 6. The interface that the transcribers use for transcription: the upper window shows the text that the speakers read, the middle window contains the transcription of the first transcriber, the lower window contains the transcription of the second transcriber (reviewer).**

### 3. Results

During a period of 10 months, the SAAVB was completed, including collecting the speech and carrying out the transcription (Figure 7). The details of its contents are illustrated below.
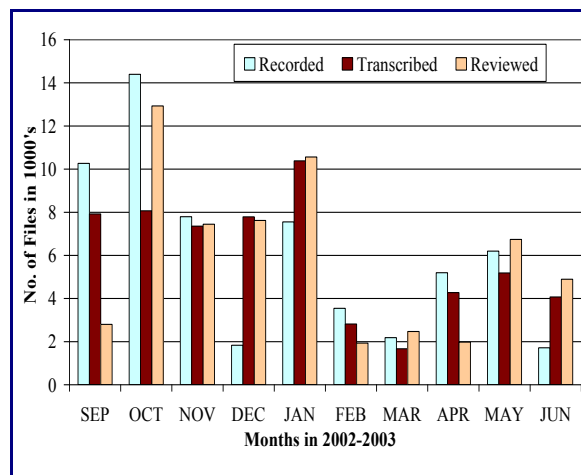


**Figure 7. The number of files stored as speech recorded files, transcribed files and reviewed files during the period of data collection and transcription.**

### 3. 1. Speakers

The number of speakers recorded in each category mentioned in Section 2.2. above is as follows:

- Total number of recorded speakers:
  o 1033 speakers,
- Gender:
  o Male: 523 speakers (50.63%),
  o Female 510 speakers (49.37%),
- Telephone type:
  o Cellular telephones: 725 speakers (70.18%),
    - Quiet environment: 252 speakers (34.76%),
    - Noisy environment: 252 speakers (34.76%),
    - Moving vehicle: 221 speakers (30.48%).
  o Fixed telephones: 308 speakers (29.82%),
    - Quiet environment: 232 speakers (75.32%),
    - Noisy environment: 76 speakers (24.68%).
- Age:
  o 16–30 years 512 speakers (50.53%),
  o 31–45 years: 364 speakers (35.24%),
  o 46–60 years: 147 speakers (14.23%).

As seen from the results above and the target figures in Section 2. 2., the deviation percentage is less than 1%. This is due to the tremendous effort and

determination made by the coordinators to search for the target speakers. Table 9 gives more details on the number of speakers at each end category. The speakers are distributed between categories according to the percentages above.

Figure 8 shows the number of missing speakers in each category. The maximum number of missing speakers in a category is two, which is very low when compared with the numbers in Figure 5.

**Table 9. The number of speakers (SS) and the percentage (%) of each end category.**

| Gender | Telephone Type | Acoustic Environ. | Age | SS | % |
|---|---|---|---|---|---|
| Male | Fixed Telephone | Quiet | 16–30 | 58 | 5.61 |
| | | | 31–45 | 41 | 3.97 |
| | | | 46-60 | 17 | 1.96 |
| | | Noisy | 16–30 | 20 | 1.94 |
| | | | 31–45 | 14 | 1.36 |
| | | | 46–60 | 5 | 0.48 |
| | Cellular Phone | Quiet | 16–30 | 65 | 6.29 |
| | | | 31–45 | 45 | 4.36 |
| | | | 46–60 | 19 | 1.84 |
| | | Noisy | 16–30 | 65 | 6.29 |
| | | | 31–45 | 45 | 4.36 |
| | | | 46–60 | 18 | 1.74 |
| | | Moving Vehicle | 16–30 | 55 | 5.32 |
| | | | 31–45 | 40 | 3.87 |
| | | | 46–60 | 16 | 1.55 |
| Female | Fixed Telephone | Quiet | 16–30 | 58 | 5.61 |
| | | | 31–45 | 41 | 3.97 |
| | | | 46–60 | 17 | 1.65 |
| | | Noisy | 16–30 | 20 | 1.94 |
| | | | 31–45 | 12 | 1.16 |
| | | | 46–60 | 5 | 0.48 |
| | Cellular Phone | Quiet | 16–30 | 63 | 6.10 |
| | | | 31–45 | 43 | 4.16 |
| | | | 46–60 | 17 | 1.65 |
| | | Noisy | 16–30 | 64 | 6.20 |
| | | | 31–45 | 43 | 4.16 |
| | | | 46–60 | 17 | 1.65 |
| | | Moving Vehicle | 16–30 | 54 | 5.23 |
| | | | 31–45 | 40 | 3.87 |
| | | | 46–60 | 16 | 1.55 |
| Total | | | | 1033 | 100 |
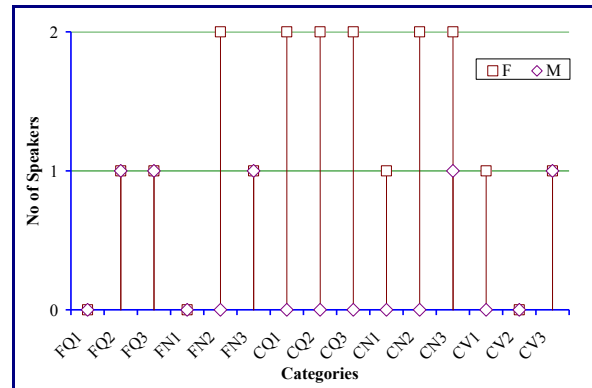


**Figure 8. Number of male (M) and female (F) speakers in each category who did not participate in SAAVB. The symbols on the x-axis represent: F=fixed telephone, Q=quiet environment, N=noisy environment, C=cellular telephone, V=vehicle, 1=age 16–30, 2=age 31–45 and 3=age 46–60 years.**

### 3. 2. Transcription

The total number of words in the transcription files is 302,107 distributed among 60,947 text files with an average of five words in a file. The dictionary includes 34,961 words. 19,941 words are unique, i.e. they occur only once in the database. The high percentage of unique words indicates the vocabulary richness of the database where only 8.5% of the transcribed words occur more than 10 times. Figure 9 shows the vocabulary frequency in the transcription files compared to that in the prompt sheet. The reason for the words in the prompt sheets being more frequently repeated is due to the fact that they are mostly undiacritized while all the words in the transcription files are fully diacritized. An undiacritized word could have different diacritization. For example, the undiacritized word 'جمالهم' *their camels* can be transcribed as:

جِمَالَهُمْ  /Cima:lahum /

جقْمَالْهِمْ  /*.a:lhim /

ججْمَالَهُمْ  /dC.a:lhum/

Undiacritized words, such as the above one, would allow for such variation in accents to appear in the speakers' speech. The accent would include the vowels and the consonants, in addition to the phonotactic rules such as the appearance of consonant clusters at word initial position which MSA does not allow.

The total number of the phonemes is 1,843,282. This means that the average number of phonemes per word is 6.1. The most frequent phoneme is the lower vowel /a/, which often follows many Arabic consonants. Arabic possesses single and geminate sounds. The single sounds in the SAAVB are more frequent (Table 10) while the geminates are less frequent and most of the non-MSA consonants do not appear in the database as geminates (Table 11). The geminates are written in Arabic orthography using the diacritic '◌ّ' which follows a letter.

As illustrated in Table 8, the Saudi consonants are variants of the MSA consonants. The Saudi consonant /*/ is a variant of the MSA consonant /q/, /\$ C/ is a variant of /C/, /ts/ and /dz/ are variants of /k/, /A $^{\sqrt{}}$/ is a variant of /? $^{\sqrt{}}$/. The /v/ and /p/ phonemes occur only in the foreign words such as those in email addresses.
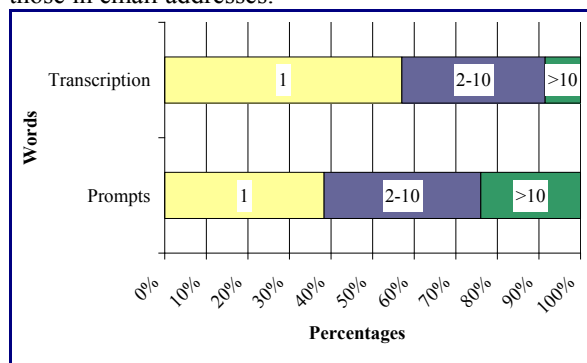


**Figure 9. A comparison of word-frequency percentage in the prompt sheet and in the transcription files: 1=percentage of words that occur only once, 2–10=percentage of words that occur from two to ten times, >10=percentage of words that occur more than 10 times.**

**Table 10. The frequency (F) of SAAVB single phonemes written here in International Phonetic Alphabet (IPA).**

| IPA | F | IPA | F | IPA | F | IPA | F |
|---|---|---|---|---|---|---|---|
| V o w e l s | | M | 62,442 | C | 15,329 | Saudi Consonants | |
| • | 405,773 | @ | 57,654 | + | 13,691 | * | 2,222 |
| ʾ | 125,221 | # | 48,678 | ) | 13,508 | ϖ | 2 1 2 |
| ™ | 40,737 | " | 47,681 | @ √ | 13,186 | \$C | 1 4 1 |
| MSA Consonants | | f | 45,401 | B | 12,775 | π | 1 0 7 |
| W | 106,263 | > | 39,138 | # √ | 6,935 | #@ | 4 |
| . | 79,191 | \$ | 35,184 | A | 6,078 | \$ ζ | 1 |
| 0 | 72,958 | < | 35,095 | ? √ | 5,848 | A √ | 1 |
| N | 72,842 | U | 24,134 | ? | 5,101 | | |
| - | 72,159 | L | 22,185 | K | 4,635 | | |

| IPA | F | IPA | F | IPA | F | IPA | F |
|---|---|---|---|---|---|---|---|
| 6 | 62,890 | J | 19,132 | \$ √ | 1,104 | | |

**Table 11. The frequency (F) of SAAVB geminate phonemes written in International Phonetic Alphabet (IPA).**

| IPA | F | IPA | F | IPA | F | IPA | F |
|---|---|---|---|---|---|---|---|
| V o w e l s | | f | 1,707 | \$ √ | 4 0 6 | Saudi Consonants | |
| • | 135,088 | 0 | 1,652 | " | 3 9 1 | * | 3 1 |
| ʾ | 75,048 | . | 1,619 | + | 3 8 2 | #@ | 0 |
| ™ | 19,008 | B | 1,500 | ) | 3 4 5 | \$C | 0 |
| MSA Consonants | | ? | 1,288 | C | 2 9 2 | ϖ | 0 |
| # | 10,316 | W | 1,181 | < | 1 8 7 | \$ ζ | 0 |
| @ | 3,634 | @ √ | 1,177 | J | 1 1 1 | A √ | 0 |
| 6 | 2,427 | U | 1,163 | K | 9 9 | π | 0 |
| ? √ | 2,277 | # √ | 1,151 | N | 7 9 | | |
| \$ | 2,238 | L | 4 9 9 | M | 4 5 | | |
| > | 1,822 | A | 4 4 7 | - | 3 6 | | |

### 3. 3. Speech

The duration of the total recorded speech is 96.37 hours distributed among 60,947 audio files (1033 speakers × 59 audio files). The average duration for each speaker is 5.60 minutes and the average duration of each audio file is 5.70 seconds. The total duration for males and females is very similar, 48.81 and 47.56 hours, respectively.

**Table 12. Three files extracted files from SAAVB: The first one is the wave file of the word /sittah/ "six", the second is the written prompt "6", and the third is the transcription.**

| File Name | Content |
|---|---|
| 20250801.wav |  |
| P20250801.txt | ٦ |
| R20250801.txt | سِتَّة |

### 4. SAAVB Contents

The total size of the SAAVB is 2.59 GByte. It consists of 1033 directories with 183,518 files. Every directory represents a speaker and contains 178 files distributed as follows:

1 text document that has all the information needed about the speaker (gender, telephone type, age and acoustic environment).
59 text files of the prompts. The names of these files start with 'p' plus the code mentioned in Figure 2 with the .txt extension, for example, P40357201.txt. These files are the orthographic transcription.
59 text files of the speech transcription. The names of these files start with 'r' plus the code mentioned in Figure 2 with the .txt extension, for example, R40357201.txt. These files are the phonetic transcription.
59 audio files (8-bit µ-law modulation in PCM format). The names of these files have the code mentioned in Figure 2 with the .pcm extension, for example, 40357201.pcm.

Since SAAVB contains several variables including speech samples from different acoustic environments, it is left open for researchers and application developers to select the training and testing sets according to their needs [27, 40]. Other available divided speech databases, for example TIMIT, indicate that its training and testing sets are only 'suggested', which means that the users may select their own training and testing sets [9].

SAAVB can be used in a speech recognition system which has already been used by IBM. It has also been used for speaker verification where Gaussian Mixture Model (GMM) is employed [37].

### 5. Validation

A speech database can be validated internally, externally or both [10, 12, 17, 19, 35]. Validation involves documentation, file format, signal quality, transcription quality, lexicon and speaker and environment distribution [36]. SAAVB is internally and externally validated. Its content is internally checked through the following steps: 1) the recording system does not record an item if its amplitude is very low, 2) it does not add the recorded files of a speaker to the database until they are approved by the coordinator, 3) the transcribers eliminate all the speaker's files if any of them do not follow the specifications including speech quality and prompt contents and 4) transcription reviewers validate the transcription. SAAVB was externally validated by IBM Egypt Branch on 11 December 2003. It has also been licensed to IBM to be used to train their speech

recognition engine. IBM has added the Saudi acoustic model developed from the SAAVB to its WebSphere. Currently, WebSphere is used in an automatic telephone directory at KACST.

### 6. Conclusions

The description of the Saudi Accented Arabic Voice Bank and how it is was amassed provides important information for researchers who intend to use it and for those who would like to collect similar speech databases. The paper demonstrated a new approach to selecting a sample from a population to compile a speech database when there is no dialect map available for the population. It also gives a guideline to a method for transcribing speech sounds that are not represented in Arabic orthography.

The SAAVB is rich in terms of its speech sound content and speaker diversity within Saudi Arabia. It has been used to train an automatic speech recognition engine and can be used to train other systems for speaker, gender, accent and language identification.

### 7. Acknowledgements

**Appendix I**

Linguistic Material Form ([R] = Read; [S] = Spontaneous):

1. One-digit number (10 randomized items), example, 'ᷲ' (7). [R]

2. Ten-digit number (1033 different items), example, '٥۰۹۸۳۲٦٥۷۳' (5098326573). [R]

3. Five-digit number (1033 different items), example, '۳۷۰۹۱' (37091). [R]

4. Seven-digit number (1033 different items), example, '٤۸٦۲۰۱٤' (4862014). [R]

5. Mobile telephone number (1033 different items), example, '۰٥٥٤٦۰۸۸٥' (055460885). [R]

6. Credit card number (1033 different items), example, '۳٤۰٦ ٥٥۰۰ ۱۱۷۰ ٥۲۳۸' (3406 5500 1170 5238). [R]

7. Four-digit number (1033 different items), example, '۳۳۷۷' (3377). [R]

8. Two- to seven-digit number (1033 different items), example, '٥٦۳۲۹' (56329). [R]

9. An amount in Saudi riyals (1033 different items), example, ' ٤۷۹٥٤ ريالا ' (47954 SR). [R]

10. An amount in US dollars (1033 different items), example, ' ۳۷٥۹ دولار ' (3759 dollar). [R]

11. An amount in euros (1033 different items), example, ' ۱۰۳٦۹۰ يورو ' (103690 euros).[R]

12. A question where a 'yes' answer is expected (the same item for all), 'هل أنت سعودي؟' (Are you Saudi?) Expected answer, 'نعم' (Yes). [S]

13. A question where a 'no' answer is expected (the same item for all), 'هل أنت برازيلي؟' (Are you Brazilian?) Expected answer, 'لا' (No).[S]

14. Speaker's birth date, example, '۱۳۹۲هـ' (1392 H). [S]

15. A date in Hijrah (1033 different items), example, '۲۱ رجب ۱۳۲٥هـ' (21 Rajab 1325 H).[R]

16. A reference point in time (369 items), example, 'البارحة' (last night).[R]

17. Time of calling the system by the speaker, example, '۹:۳۰ مساء' (9:30 pm).[S]

18. Time (1033 different items), example, الساعة ۱۱ و ۲۰ دقيقة صباحا (The time is 11:20 am).   R]

19. Speaker's birthplace, example, الرياض (Riyadh).      [S]

20. National city (118 randomized and repeated items), example, الدمام (Dammam).    [R]

21. International city (182 randomized and repeated items), example, لندن (London).   [R]

22. National company name (1033 different items), example, سابك (Sabic).     [R]

23. Government institution (1033 different items), example, وزارة الصحة (Ministry of Health).       [R]

24. Arabic personal name (1033 different items), ناصر حمود المعلم (Nasir Humoud Almualim).   [R]

25. Speaker's name, example, خلود علي الباني (Khuloud Ali Albani).      [S]

26. Spelling of a phonetically rich word (671 randomized and partially repeated items) [38], example, ضوء /dᵛaw-/.      [R]

27. Spelling of a city name (Prompt 20 and 21), example, مكة /mi:m ka:f ta:- marbutᵛah/ (Makkah).      [R]

28. Spelling of a personal name (260 randomized and repeated items), example, ناصر /nu:n -alif sᵛa:d ra:-/ (Nasir).      [R]

29-32 Four phonetically rich words (from the same list as that of Prompt 26), example, قَضْم ثَغْر ظِلْف خَتْم /qadᵛm, >aKr, ?ᵛilf, Jatm/ (bite mouth hoof stamp) [38].     [R]

33-36, 51-55. A sentence (including 366 phonetically rich sentences [39]), example, خَرَجْتُ فَإِذَا وَحْشٌ عِنْدَ البَابِ /JaraCtu fa-i?a waLBun Minda –albabi/ (I went out and there was a monster at the door); and 558 normal sentences, example, تحمي الأظافر الأصابع /taLmi: ala?ᵛa:fir alasᵛa:bi-/ (Nails protect fingers).   [R]

37. Speaker's email address, example, mnal@hotmail.com      [S]

38. Email address, example, twali@yahoo.com

39. Place of calling, example, المنزل (at home).

40. Used telephone type, example, هاتف ثابت (fixed telephone).                              [S]
41. Accent variation sentence (the same sentence for all). لماذا سافرت إلى الخارج في العيد؟ /lima:?a: sa:fart -ila: alJa:riC fi: alMi:d/ (Why did you travel abroad during the festival?)            [R]
42. Accent variation sentence (the same sentence for all). جاء الضيوف الثلاثة بالذهب قبل الظهر . /Ca:- ald√yu:f al>ala:>ah bial?ahab qabl al?√uhr/ (The three guests came with the gold before noon.)                                       [R]
43-46. Command (109 randomized and repeated items), example, أعد التعليمات /aMid altaMlima:t/ (Repeat the instruction).           [R]
47-50. Question (185 randomized and repeated items), example, أين أقرب مستشفى؟ /-ayna -aqrab mustaBfa:/ (Where is the nearest hospital?)                                                   [R]
56. Number written in Modern Standard Arabic (234 randomized and repeated items), example, سِتُّون /sittu:n/ (sixty)         [R]
57. Speaker's sentence about his/her city, example, مدينتي عاصمة المملكة العربية السعودية /madi:nati: Ma:simatu almamlakati alMarabi:ti assaMu:diyyah/ (My city is the capital of the Kingdom of Saudi Arabia).                  [S]
58. Gregorian date (1033 different items), example, ١٩ سبتمبر ١٩٨٣م (19 September 1983).                                          [R]
59. Six-digit number (1033 different items), example, ٥٩٧٦٤٨ (597648).                        [R]

# References

[1] Robinson, T., J. Fransen, D. Pye, J. Foote and S. Renals. WSJCAMO: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition. 1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95. 1: 81–84. 1995.

[2] Hong, Q.Y. and S. Kwong. A Discriminative Training Approach for Text-independent Speaker Recognition. Signal Processing. 85: 1449–1463. 2005.

[3] Wildermoth, B. and K.K. Paliwal. GMM Based Speaker Recognition on Readily Available Databases. Proceedings of the Microelectronic Engineering Research Conference, Brisbane, Australia. 2003.

[4] Sanderson, C. and K.K. Paliwal. Joint Cohort Normalization in a Multi-feature Speaker Verification System. Proceedings of the 10th IEEE International Conference on Fuzzy Systems, Melbourne, Australia. 232–235. 2001.

[5] Kat, L.W. and P. Fung. Fast Accent Identification and Accented Speech Recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing. 1: 221–224. 1999.

[6] Niesler, T. and D. Willett. Language Identification and Multilingual Speech Recognition Using Discriminatively Trained Acoustic Models. ISCA Workshop on Multilingual Speech and Language Processing. Stellenbosch, South Africa. 2006.

[7] Taylor, L.J. and G. Knowles. Manual of Information to Accompany the SEC Corpus: The Machine Readable Corpus of Spoken English. Available from: http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM. Consultation date: 28 November 2007.

[8] Vonwiller, J.,, I. Rogers, C. Cleirigh and W. Lewis. Speaker and Material Selection for the Australian National Database of Spoken Language. Journal of Quantitative Linguistics, 2: 177–211. 1995.

[9] TIMIT: Acoustic-Phonetic Continuous Speech Corpus. DMI. 1990.

[10] Bernstein, J., K. Taussig and J. Godfrey. Macrophone: an American English telephone speech corpus for the Polyphone project. Acoustics, Speech, and Signal Processing, 1: I/81–I/84. 1994.

[11] Moreno, P., O. Gedge, H. Heuvel, H. Höge, S. Horbach, P. Martin, E. Pinto, A. Rincón, F. Senia and R. Sukkar. SpeechDat Across all America: SALA II. Project website: http://www.sala2.org. Consultation date: 28 November 2007.

[12] Schultz, T. Globalphone: A Multilingual Speech and Text Database. Proceedings of the International Conference of Spoken Language Processing. Denver, USA. 2002.

[13] The Spoken Dutch Corpus: http://www.elis.rug.ac.be/cgn/. Consultation date: 28 November 2007.

[14] Zheng, T.F., P. Yan, H. Sun, M. Xu and W. Wu. Collection of a Chinese Spontaneous Telephone Speech Corpus and Proposal of Robust Rules for Robust Natural Language Parsing. Joint International Conference of SNLP-O-COCOSDA, Hua Hin, Thailand: 60–67, 2002.

[15] Tseng, C., Y. Cheng, W. Lee and F. Huang. Collecting Mandarin Speech Databases for Prosody Investigations. The Oriental COCOSDA. Singapore. 2003.

[16] Wang, H., F. Seide, C. Tseng and L. Lee. MAT-2000-Design, Collection and Validation of a Mandarin 2000-Speaker Telephone Speech Database. ICSLP 2000, Beijing. 4: 460–463. 2000.

[17] Lo, W.K., T. Lee and P.C. Ching. Development of Cantonese Spoken Language Corpora for Speech Applications. Proceedings of the First International Symposium on Chinese Spoken Language Processing. 102–107. Singapore. 1998.

[18] Muthusamy, Y., E. Holliman, B. Wheatley, J. Picone and J. Godfrey. Voice Across Hispanic America: A Telephone Speech Corpus of American Spanish. Acoustics. International Conference on Speech and Signal Processing. 1: 85–88. 1995.

[19] Langmann, D., R. Haeb-Umbach, L. Boves and E. den Os. FRESCO: The French Telephone Speech Data Collection - Part of the European SpeechDat(M) Project. FRESCO. The Fourth International Conference on Spoken Language Processing. Philadelphia. 1: 1918–1921. 1996.

[20] Gopalakrishna, A., R. Chitturi, S. Joshi, R. Kumar, S. Singh, R.N.V Sitaram and S.P. Kishore. Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems. Proceedings of International Conference on Speech and Computer (SPECOM), Patras, Greece, October 2005.

[21] Heuvel, H. and L. Boves. Annotation in the SpeechDat Projects. International Journal of Speech Technology. 4: 127–143. 2001.

[22] Glass, J., J. Chang, and M. McCandless. A Probabilistic Framework for Feature-based Speech Recognition. Fourth International Conference on Spoken Languages. Philadelphia, USA. 4: 2277–2280. 1996.

[23] Linguistic Data Consortium: http://www.ldc.upenn.edu. Consultation date: 28 November 2007.

[24] European Language Resources Association: http://www.elra.info. Consultation date: 28 November 2007.

[25] Appen: http://www.appen.com.au. Consultation date: 28 November 2007.

[26] Khalid Choukri, Khalid, Salah Hamid and N. Paulsson. Specifications of the Arabic Broadcast News Speech Corpus. NEMLAR 2005: http://www.nemlar.org. Consultation date: 28 November 2007.

[27] Zue, V., N. Daly, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, S. Seneff and M. Soclof. The Collection and Preliminary Analysis of a Spontaneous Speech Database. Proc. DARPA Speech and Natural Language Workshop: 126–134. 1989.

[28] Pitrelli, J.F., C. Fong, S.H. Wong, J.R. Spitz and H.C. Leung. PhoneBook: A Phonetically-rich Isolated-word Telephone-speech Database. 1995 International Conference on Speech, and Signal Processing, ICASSP-95. 1: 101–104. 1995.

[29] Saudi Ministry of Economy and Planning: http://www.planning.gov.sa/docs/045.htm. Consultation date: 28 November 2007.

[30] Saudi Accented Arabic Voice Bank. King Abdulaziz City for Science and Technology. 2003.

[31] Alghamdi, M., F. Alhargan, M. Alkanhal, A. Alkhairi and M. Ad-Dusooqee. Saudi Accented Arabic Voice Bank. Final Report. Computer and Electronic Research Institute, King Abdulaziz City for Science and Technology. 2003.

[32] Montgomery, D.C. Design and Analysis of Experiments, 6th Edition. John Wiley and Sons, Inc. 2004.

[33] Maamouri, M., T. Buckwalter and C. Cieri. Dialectal Arabic Telephone Speech Corpus:

Principles, Tool Design, and Transcription Conventions. Paper presented at the NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, September 22–23, 2004.

[34]   Altosaar, T.,   M. Karjalainen  and M. Vainio.   A Multilingual Phonetic Representation and Analysis System for Different Speech Databases. 3: 1914–1917. 1996.

[35] Van den Heuvel, H., L. Boves, K. Choukri, S. Goddijn and E. Sanders. SLR Validation: Present State Of Affairs And Prospects. Proc LREC'00. Athens. 2000.

[36] Iskra1, C., B. Grosskopf, K. Marasek, H. van den Heuvel and F. Diehl. SPEECON – Speech Databases for Consumer Devices: Database Specification and Validation. In Proceedings LREC. 2002.

[37] Alkanhal, M., M. Alghamdi and Z. Muzaffar. Speaker Verification Based on Saudi Accented Arabic Database. International Symposium on Signal Processing and its Applications in conjunction with the International Conference on Information Sciences, Signal Processing and its Applications. Sharjah, United Arab Emirates. 12–15 February 2007.

[38]   Alghamdi, M., M. Basalamah, M. Seeni and A. Husain. Database of Arabic Sounds: Words, 15th National Computer Conference, Dammam. 2: 797–815. 1997.

[39]   Alghamdi, M., A. Alhumayid and M. Ad-Dusooqee. Arabic Sound Database: Sentences, Computer and Electronics Research Institute (HK-28), King Abdulaziz City for Science and Technology, Riyadh. 2003.

[40] Pearce, D. and H.-G. Hirsch. The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions. 6th International Conference on Spoken Language Processing. Beijing, China. 16–20 October 2000.

# قاعدة بيانات الأصوات العربية الهاتفية للمتحدثين السعودية

**منصور الغامدي، وفايز الحرقان، ومحمد الكنهل، وأشرف الخيري، ومنير الدسوقي، وعمار العنزي**

*معهد بحوث الحاسب والإلكترونيات، مدينة الملك عبدالعزيز للعلوم و التقنية*

*ص. ب. ٦٠٨٦، الرياض ١١٤٤٢، المملكة العربية السعودية*

**ملخص البحث.** تقدم هذه الورقة قاعدة بيانات للكلام العربي الهاتفي المنطوق من قبل متحدثين سعوديين من جميع مدن المملكة العربية السعودية. وتعرض الورقة أبرز التحديات التي واجهت فريق العمل التي منها إعداد عبارات الكلام، واختيار المتحدثين المناسبين وتسجيلهم، والكتابة الصوتية للكلام العربي. وتذكر الورقة الحلول التي سلكها فريق العمل لمواجهة هذه التحديات. وتضم قاعدة البيانات ١٠٣٣ متحدثا باللكنة السعودية للغة العربية المعاصرة. وتتولى الورقة عرض وتحليل محتويات قاعدة البيانات التي أجيزت من قبل شركة آي بي إم واستخدمتها في بناء محرك للتعرف الآلي على الكلام العربي. ويمكن استخدام هذه القاعدة في تدريب واختبار نظم حاسوبية مختلفة منها: التعرف على الكلام، واللهجات، واللغة العربية، والجنس، إضافة إلى التحقق من المتحدث.

**الكلمات المفتاحية:** *اللغة العربية؛ قاعدة بيانات؛ التعرف على الكلام؛ الأصوات اللغوية؛ اللهجات السعودية*