

Saudi Accented Arabic Voice Bank

Mansour Alghamdi, Fayeze Alhargan, Mohamed Alkanhal, Ashraf Alkhairy, Munir Eldesouki and Ammar Alenazi¹

¹Computer and Electronic Research Institute, King Abdulaziz City for Science and Technology

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 met these challenges are highlighted. In the project, 1033 speakers speak in Modern Standard Arabic with a Saudi accent. The SAAVB content was analyzed and the results are illustrated. The content was verified internally by the project team and externally by IBM Cairo and can be used to train speech engines such as automatic speech recognition and speaker verification systems.

Key words: Arabic speech database Saudi

Introduction

Speech databases are essential for training automatic speech recognition systems in addition to other applications such as speaker verification, dialect and language identification. Speech databases are also valuable in linguistic studies especially in phonetics, phonology, typology and sociolinguistics. For these reasons, speech databases of many languages have been collected for many years in many countries: English in Australia, Australian National Database of Spoken Language (ANDOSL) Vonwiller, J. P., et. al., (1996); British English speech corpus (WSJCAMO) Robinson et. al. (1995); American English, Texas Instrument and Massachusetts Institute of Technology corpus (TIMIT) TIMIT (1990), Macrophone, Bernstein et. al. (1994); Chinese Spontaneous Telephone Speech Corpus on Flight Enquiry and Reservation (CSTSC-Flight), Zheng et. al. (2002); Mandarin Across Taiwan (MAT), Tseng et. al. (2003); Cantonese, one of the dialects in southern China (Lo et. al., 1998); American Spanish (Voice Across Hispanic America) Muthusamy et. al. (1995); French, French SpeechDat corpus (FRESCO) Langmann et. al. (1996).

Although speech databases have been collected for several languages, Arabic speech databases need more work to cover the dialectal diversity and many Arabic speaking counties remain with almost non-professional speech collections. Saudi Arabia is one of the countries where a speech database

Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics 2008, 28-30 August 2006, Athens, Greece.

that covers its various dialects has not been collected before SAAVB. The area of Saudi Arabia is 1,960,582 sq km. It is located in south west Asia and surrounded from north, east and south by other Arab countries. About 20 million inhabitants live in Saudi Arabia; four fifth of them are native Saudis (Ministry of Economy and Planning, 1999 Census).

A project that aims at collecting 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.

This paper is written to assist the Saudi Accented Arabic Voice Bank (Alghamdi et. al., 2003) users and to document the procedures, specifications and contents of SAAVB for those who will be interested in collecting similar speech data in similar environment.

Database Design and Recording

The procedures to collect the data have four phases: 1) designing the Prompt Sheet, 2) selecting the speakers, 3) recording the speech and 4) transcription.

Each speaker is given a 5 page document. The first page has a code that gives each speaker access to the recording lab to record their speech. The code symbolizes the region, city, gender, age, telephone type and calling environment of the speaker. The code is to be used as the name of all SAAVB files, so, the gender, age and other information related to the speaker can be extracted from the name of the files. The second page has the instructions that help the speaker to log into the recording system and complete the required tasks. The remaining 3 pages are the prompt sheets.

A prompt sheet consists of 59 items: 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, press and official communication in the Arab world. A unique Prompt Sheet for each speaker; no two Prompt Sheets are the same, is prepared. The only two sentences that appear in every Prompt Sheet are Prompt 41 and 42. These two sentences are designed to record dialectic variations among speakers. The total number of prepared prompt sheets is 1059.

Due to the absence of a dialect map for Saudi Arabia, the research team decides to select a sample from every city in the country. One half of the sample is male, and the other half is female. There are 118 cities in Saudi Arabia spread all over the country area, and the target number of speakers have to be divided among them according to the city population.

Results

The total number of the speakers who participated in SAAVB is 1033 distributed among the following categories:

- 523 Male (50.63%), 510 Female (49.37%)
- 725 Cellular (70.18%): 252 Quiet environment (34.76%), 252 Noisy environment (34.76%), 221 Moving vehicle (30.48%).
- 308 Fixed telephones (29.82%): 232 Quiet environment (75.32%), 76 s Noisy environment (24.68%),
- 512, 16-30 years old (50.53%),
- 364, 31-45 years old (35.24%),
- 147, 46-60 years old (14.23%).

The total number of words in the transcription files is 302,107 distributed among 60947 text files with an average of 5 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 the unique words indicates the vocabulary richness of the database where only 8.5% of the transcribed words occur more than 10 times.

The duration of the total recorded speech is 96.37 hours distributed among 60947 audio files (1033 speakers x 59 audio files). This means that the average duration for each speaker is 5.60 minutes and the average duration of each audio file is 5.70 seconds.

Conclusions

This paper presents a description of the Saudi Accented Arabic Voice Bank (SAAVB); how it is collected and its content. SAAVB has been licensed to IBM to be used to train their speech recognition engine. Currently, it is used by KACST speech team to develop an Arabic speech recognition system.

SAAVB includes 1033 directories with a total of 183,518 files. The total size of SAAVB is 2.59 GByte.

SAAVB is now available as a Saudi Arabic voice bank and can be licensed to be used in research or to develop products when a contract with KACST is signed.

Acknowledgements

This paper is supported by KACST through SAAVB project number i-e-6-1. The authors would also like to acknowledge the IBM Egypt team, led by Dr. Ossama Emam, for their cooperation with KACST team and for verifying the database.

References

- Alghamdi, M., F. Alhargan, M. Alkanhal, A. Alkhairi, M. Aldusuqi. Saudi Accented Arabic Voice Bank. Final Report. Computer and Electronic Research Institute, King Abdulaziz City for Science and Technology. 2003.
- Bernstein, J. Taussig, K. Godfrey, J. Macrophone: an American English telephone speech corpus for the Polyphone project. *Acoustics, Speech, and Signal Processing*, 1: I/81-I/84. 1994.
- 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.
- 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.
- Ministry of Economy and Planning: <http://www.planning.gov.sa/docs/045.htm>
- Muthusamy, Y., E. Holliman, B. Wheatley, J. Picone and J. Godfrey. Voice across Hispanic America: A telephone speech corpus of American Spanish. *Acoustics, 1995 International Conference on Speech, and Signal Processing, ICASSP-95*. 1: 85-88. 1995.
- 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.
- TIMIT: Acoustic-Phonetic Continuous Speech Corpus. DMI. 1990.
- Tseng, C., Y. Cheng, W. Lee and F. Huang. *Collecting Mandarin Speech Databases for Prosody Investigations, The Oriental COCOSDA*. Singapore. 2003.
- Vonwiller, J. P., et. al., (Speaker and Material Selection for the Australian National Database of Spoken Language), *Journal of Quantitative Linguistics*, 27, 1996.
- Zheng, T. F., P. Yan1, 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*.