SASSC: A Standard Arabic Single Speaker Corpus

Ibrahim Almosallam, Atheer AlKhalifa, Mansour Alghamdi,
Mohamed Alkanhal, Ashraf Alkhairy

The Computer Research Institute
King Abdulaziz City for Science and Technology
Riyadh, Saudi Arabia
{ialmosallam,aalkhalifa,mghamdi,alkanhal,alkhairy}@kacst.edu.sa

Abstract

This paper describes the process of collecting and recording a large scale Arabic single speaker speech corpus. The collection and recording of the corpus was supervised by professional linguists and was recorded by a professional speaker in a soundproof studio using specialized equipments and stored in high quality formats. The pitch of the speaker (EGG) was also recorded and synchronized with the speech signal. Careful attempts were taken to insure the quality and diversity of the read text to insure maximum presence and combinations of words and phonemes. The corpus consists of 51 thousand words that required 7 hours of recording, and it is freely available for academic and research purposes.

Index Terms: Text-to-Speech, Arabic Speech Corpus

1. Introduction

In the last few years, an increasing number of research projects in Natural Language Processing (NLP) have developed an interest in the Arabic language. Arabic is the official language in more than 21 countries and has two major forms: Standard Arabic and Dialectical Arabic. While dialectical Arabic is the native language of many Arabic speakers nowadays, Standard Arabic is the formal language used in education, culture and media. Standard Arabic consists of: Modern Standard Arabic (MSA) and Classical Arabic. Though MSA is derived from Classical Arabic , the language of Qura'n (Islam's Holy Book), it is more simplified [1].

This paper describes the process of building an Arabic speech corpus, with the aim of collecting large amounts of Arabic speech data for research purposes. The availability of such corpora without copyrights restrictions is important for Arab and non-Arab researchers who are looking for speech resources of contemporary Arabic. MSA was the language of choice to represent the literary standard across the Arab world, while the existing of an Arabic corpora that are specified for speech synthesis and recognition research purposes is still insufficient. Despite the existence of several Arabic speech corpora developed in the last few years, a diverse fully transcribed and segmented MSA speech corpus is a necessity for Text-To-Speech (TTS) and Automatic Speech Recognition (ASR) research and development.

The presented corpus was designed to contain a large selection of text collected from different modern text resources. These resources were chosen to capture the phonetic distribution of the Arabic language. Although the overwhelming majority of the corpus is in MSA form, a designated subcategory was recorded in traditional or classical Arabic. The data was then revised and manually diacritized (Tashkeel) by linguists. Afterward, the voice and pitch of the speaker were captured using specialized equipments and recorded in high quality formats by a professional speaker. The transcriptions were revised again to match the pronunciation of the spoken records. The recordings were then divided into phrases by each sentence's pause and stored in separate files. Finally, the data was phonetically segmented and labeled in preparation for TTS and ASR usages.

The paper is organized as follows: In section 2, we present and discuss some related work. Section 3 outlines the constitution of the corpus and section 4 will describe the recording procedure. Then, the corpus evaluation procedure will be described in section 5 . Finally, in section 6 we conclude and discuss the availability of the dataset.

2. Related Work

The quality of speech synthesis systems highly depends on the corpus's phoneme set distribution. According to [1], an overlook on the existing Arabic speech datasets shows the lack of available public domain, raw-data and single speaker MSA corpus. Similar corpora are either copyrighted, accented [2] [3] and or of a multi-speaker [4]. Most available datasets are designed for speech recognition proposes, whereas TTS systems mostly rely on commercial corpora or record the designated speech database according to the system's needs [5].

The Linguistics Data Consortium (LCD) has a collection of MSA and dialectical Arabic speech and text datasets for research and development purposes. The content varies between annotated news, conversational telephone speech [6] [7] and others; scripted and unscripted while recorded in different conditions. On the other hand, Speech Ocean offers an Arabic speech synthesis database (King-TTS-004) recorded in a studio [8]; It contains a single speaker recordings by a native professional broadcaster with sampling rate of 16kHz and two channels for speech and Electro Glotto Graph (EGG) signals. It includes 12 hours of pure recording time of 3930 sentences with sub database for currency, time, numbers, English alphabets and so on. However, these databases are licensed and can only be used for a fee.

Access to fully available, transcribed and segmented speech corpora is essential in speech and language research. As an example, [9] built 7 speech databases of 7 Indian languages and was released without restriction for commercial and non-commercial usage. The text was selected from Wikipedia articles in Indian languages due to the lack of public domain text corpora. Afterwards, a set of 1000 phonetically balanced sentences were nominated using Festvox script that applies certain criteria to achieve the optimal selection. The data was recorded by native speakers of these languages in a professional recording studio, and the transcription was altered to accommodate any mistakes were done while recording. The databases were between one to two hours long, and the recordings were automatically segmented afterward using Zero Frequency Filtering (ZFF) technique.

Another important source of data in speech and language research is the Electro Glotto Graph signal or (EGG). EGG , also known as Laryngograph signal, is the measurement of vocal folds vibrations during speech. These measurements are sensed by electrodes attached to the speaker's throat . It is important in aiding medical research such as voice disorders or laryngeal mechanism. Moreover, in NLP it plays an important role in speech recognition and synthesis. For example, it can act as an additional source of information in word recognition [10] or as a mean of detecting glottal pulses (pitch-marks) to assure that the synthesized speech is in a consistent matter.

Likewise, [11] introduced a high-quality, Romanian speech corpus called RSS. It is also available for academic uses to help promote Romanian speech technology research. The data was recorded at 96 kHz sampling frequency then down-sampled to 48 kHz based on an over-sampling method with the intention of noise reduction. However, the total length of the recorded data is only about 3.5 hours based on 3500 sentences. Additionally, the article discussed the effect of sampling frequency on the speaker similarity of synthesized voices. It found that the use of lower sampling frequencies such as 16kHz lowers the similarity of the synthesized voice to the original speaker.

3. Constitution of the Corpus

3.1. Text Selection

The text corpus included a variety of genres and writing styles from multiple sources to insure a representative sample. Furthermore, diversity of pronunciation was also taken into consideration during the text selection process to capture the entire phoneme spectrum that exist in the Arabic language. The corpus is divided into sub-categories to capture a wide variety of vocabulary and to enable users of the data set to perhaps extract ”exact” relevant phrases if necessary. For example, the use of numbers, date and time are very common in most applications. The dataset is divided into nine such categories:

1. Dates and Time: Phrases about time such as dates, time of the day, days of the week and months of the year, In both Hijri and Gregorian calendars .
2. Numbers: The speaker uttered several numbers in isolations as well as in combined forms. For example the number 123 was pronounced as (one two three) and as (one hundred twenty three).
3. Financial: Phrases about money and currency.
4. Customer Service: Phrases common in IVR systems (question and answer format).
5. Names: A list of common Arabic names (first, middle and last).
6. Story: Excerpts from novels and stories,
7. Traditional: Text selected from classic sources such as old books.
8. Miscellaneous: A collection of phrases from various domains and writing styles
9. News: Excerpts from local and foreign news

	Total	Unique	Median
Syllables	333,981	627	2
Tri-phones	324,225	8,689	7
Bi-phones	335,769	1,076	38
Mono-phones	347,313	38	5,651.5

Table 3: Syllables, Tri-phones, Bi-phones and Mono-phones frequencies

The text corpus consists of 51,432 words, amongst which 21,556 were unique and required 7 hours and 20 minutes of audio recording. It contained 627 unique syllables out of a total of 333,981 syllables. The break down by category is provided in table 1. All of the above mentioned categories were recorded in a normal tone of voice. However, a part of the miscellaneous text was recorded using different expressions such as sadness, joy, surprised and questioned for 15 minute each.

3.2. Transcription

A professional linguist was closely involved during both corpus selection and recording to verify the quality of the text and pronunciation. Furthermore, unlike most other languages, Arabic can only be unambiguously pronounced when diacritized. Due to this problem, most applications rely on automatic Arabic diacritization as a pre-processing step to transcription [12]. Therefore, the text was manually diacritized and revised to insure unambiguity. An automatic text-to-transcription algorithm was used to transcribe the text [13]. However, the entire corpus was revised manually. The total number of phonemes used to transcribe the corpus is 38, shown in table 2. The Statistics about the phonemes' and syllables' distribution in the corpus is provided in table 3. Furthermore, the frequency distribution for the phonemes are provided in figure 1.

4. Recording Procedure

4.1. Speaker selection

Due to the nature and size of the corpus, a professional speaker who can maintain his performance for long recording sessions was hired to record the corpus. Although diacritization reduces ambiguity, it makes the reading experience more challenging. Diacritized text forces the reader to pay closer attention to the symbols at the character level which slows down the reading process and makes it more difficult. In normal situations, readers would predict the words they are reading based on the content and experience. Ten candidates were screened and tested to insure their proficiency in the Arabic language, clarity of pronunciation and voice. The panel of linguists voted on the first two criteria while the pleasantness of the voice was aided by online voting. The speaker, selected based on the previous criteria, is a professional male news anchor who has worked in several TV programs and has experience in recording poetry books.

4.2. Recording environment

The recording sessions took place in a sound proof studio to minimize undesired noise. Both the speech and the EGG signals were recorded and stored synchronized in a stereo wave file (left=speech, right=EGG) using a 96 kHz sampling rate and a 16-bit resolution. The intensity of the signal was carefully monitored to remain equal across different sessions by request-

	Number of Words	Unique Words	Recording Time	Number of Utterances
Dates & Time	900	227	00:09:06	226
Numbers	967	133	00:13:17	224
Financial	2080	272	00:21:22	451
Customer Service	2643	928	00:22:58	400
Names	5503	1451	00:43:05	887
Story	4085	2490	00:44:42	151
Traditional	7797	4027	01:10:06	485
Miscellaneous	10927	7025	01:45:48	1011
News	16530	8208	02:44:16	537
Full Corpus	51432	21556	07:20:28	4372

Table 1: General statistics about the text and audio corpora in each sub-category

Symbol	Phone	Symbol	Phone	Symbol	Phone	Symbol	Phone
a	(فتحة) ـَ	gh	غ	m	م	T	ط
A	(فتحة مفخمة) ـَ	h	هـ	n	ن	th	ث
aa	ا	i	(كسرة) ـِ	pau	pause	TH	ظ
Aa	(مد مفخم) ا	I	ي	q	ق	u	(ضمة) ـُ
Ah	ح	j	ج	r	ـِر	U	ـو
Az	ذ	Jn	ء	R	ر	w	و
b	ب	JU	ع	s	س	y	ي
d	د	k	ك	S	ص	z	ز
D	ض	kx	خ	sh	ش		
f	ف	l	ل	t	ت		

Table 2: Phonemes and their corresponding labels in the transcription files

ing the speaker to repeat a phrase and compare it with a reference phrase recorded at the first session. The positions of the microphone, the speaker and the EGG electrodes were adjusted in every session to match the reference phrase. A linguist was present to follow the speaker and correct him in case he skipped, mispronounced or could not recognize a word. Each session was 15 minutes long and no more than four sessions per day were taken.

4.3. Editing and preparation

The recorded data was further edited to remove unwanted segments such as errors in pronunciation, long pauses, etc. The sessions were then split based on pauses and the EGG signals were exported into separate files. The corpus is divided into utterances by pauses and an HMM aligner was built to label and segment each utterance on the phoneme level. The Cambridge University HMM toolkit was used to build the speaker-specific aligner [14]. The phoneme symbols, the beginning and the ending timestamps are provided in the transcription files. Although the corpus was segmented and aligned automatically by an HMM system, over half an hour of speech was manually aligned by computational linguists. The final corpus is divided into four sets of files:

1. Wave: The speech waveform for each utterance in 96 kHz and 16-bit resolution.
2. EGG: The corresponding EGG signal in 96 kHz and 16-bit resolution.
3. Text: The corresponding diacritized text.
4. Label: The corresponding transcription along with the phoneme boundary segmentation.

The above set of files are provided for each expressive version and each utterance was labeled by their corresponding category discussed in section 3. The labels are provided in two formats (mono and full) in accordance with the HTS standard format [1]. The mono labels provide only the mono-phoneme sequence along with the beginning and ending timestamps. The full labels however provide the penta-phone sequence, each phone with its two proceeding and two preceding phonemes. Moreover, they are more detailed and provide more information such as the number of phonemes, syllables and words in each utterance. The position and number of phonemes, syllables and words are also indicated in the full label file. In the case of syllables, the files also indicate whether a syllable is stressed or accented. It is worth mentioning here that there are two types of syllables in the Arabic language: open syllables CV and CV:, and closed syllables CVC, CV:C and CVCC. Where C stands for consonant, V for vowel and V: for long vowel.

5. Corpus Evaluation

In order to evaluate the quality of the final dataset, the corpus was used to build a text-to-speech model. This model will put to test various quality aspects of the data, such as: size, structure, format and consistency. This section will describe briefly the TTS model generation and output evaluation procedures. It is important to emphasize that the purpose of the test and evaluation is not to compete with other Arabic TTS systems but to evaluate the dataset and to illustrate its usage.

The HMM-based Speech Synthesis System (HTS) was used to build the model. HTS is a free tool to build HMM-based TTS, which is basically a patch code for the HTK. It provides scripts and code that uses HTK as well as other free speech tools to train TTS models. The version used to evaluate the dataset in this paper was HTS 2.2. The format of the dataset was de-

[1] HMM-based Speech Synthesis System (HTS): http://hts.sp.nitech.ac.jp/

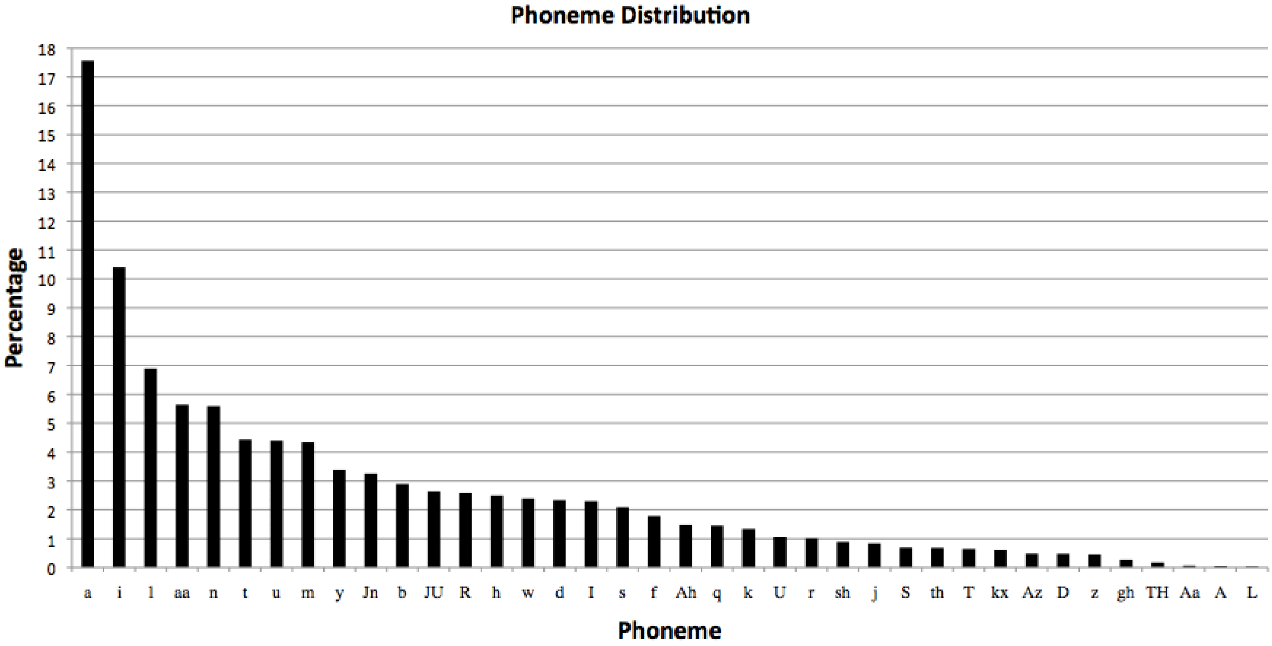

Figure 1: Distribution of the Arabic phonemes in the corpus

signed to be compatible with HTS and the default parameters were tuned for the data. The training procedure completed successfully and five sentences were synthesised to be evaluated. Two well known metrics were measured to evaluate the synthesized speech, namely naturalness and intelligibility. Some of the group members as well as other researchers (total of 10) were asked to listen to the five generated samples and provide their evaluation on each one. Evaluators were asked individually to write down what they have heard, and to rate each audio clip (on a scale from 1 to 5) based on the following qualities:

1. Naturalness: How close was the synthesised speech to being natural? (1: Very robotic sound - 5: Close to natural)
2. Intelligibility: How much effort was taken to understand the content of the sample? (1: Had to focus - 5: Easy to understand)

The average naturalness score was 3.58 and the average intelligibility score was 3.9. The original sentences were compared to what the evaluators have written and the word-error rate was 2.13%. The numbers suggest that the performance of the TTS system on the provided corpus was good on naturalness and very good on intelligibility. The synthesized samples used in the evaluation process can be found at http://cri.kacst.edu.sa/SASSC/samples.zip.

6. Conclusion and Future Work

In this paper, the processes for collecting and recording the SASSC dataset was described. The text was designed to capture a wide variety of usages and as much of the Arabic phonetic spectrum as possible. The text is fully diacratized, transcribed and was recorded in a high quality format in a soundproof studio by a professional speaker along with the pitch signal. Moreover, a TTS system based on the corpus has been evaluated which achieved promising results.

SASSC as large repository for Arabic speech, provides several avenues for future work. These may include a study on the minimum amount of speech required to produce an acceptable voice. Invest more in recording different speech expressions and evaluate their results. Moreover, linguists can find the corpus as a resource for studying different Arabic structures and pronunciations of MSA.

In order to promote and facilitate research and development for Arabic language in the NLP field, the corpus will be freely available for research and educational purposes only, a permission has to be obtained from KACST for any commercial or non-academic use of the corpus. The dataset is available at http://cri.kacst.edu.sa/SASSC/index.html

7. Acknowledgment

The authors would like to acknowledge and thank Dr. Mohsen Rashwan, Dr. Sherif Abdou, Dr. Ahmad Ragheb and Mostafa Shahin from Research and Development International (RDI) at Cairo, Egypt for providing their help and expertise in various phases during the corpus collection and recording procedures.

8. References

[1] N. Habash, *Introduction to Arabic Natural Language Processing*. Morgan and Claypool Publishers, 2010.

[2] M. Alghamdi, F. Alhargan, M. Alkanhal, A. Alkhairy, M. Eldesouki, and A. Alenazi, "Saudi accented arabic voice bank," *Experimental Linguistics ExLing*, p. 9, 2008.

[3] G. Droua-Hamdani, S. A. Selouani, and M. Boudraa, "Algerian arabic speech database (ALGASD): Corpus design and automatic speech recognition application," *Arabian Journal for Science and Engineering*, vol. 35, no. 2C, p. 158, 2010.

[4] M. Algamdi, "KACST Arabic phonetics database," in *The 15th International Congress of Phonetics Science*, 2003, pp. 3109–3112.

[5] F. Chouireb and M. Guerti, "Towards a high quality arabic speech synthesis system based on neural networks and residual excited vocal tract model," *Signal, Image and Video Processing*, vol. 2, no. 1, pp. 73–87, 2008.

[6] M. Maamouri, D. Graff, and C. Cieri, "Arabic broadcast news transcripts," LDC - Linguistic Data Consortium, December 2006. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T20

[7] Appen-Ltd, "Levantine Arabic conversational telephone speech, transcripts," LDC - Linguistic Data Consortium., Jan 2007. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T01

[8] SpeechOcean, "Arabic speech synthesis database I (Male)," April 2013. [Online]. Available: http://www.speechocean.com/en-TTS-Corpora/371.html

[9] K. Prahallad, N. Kumar, V. Keri, R. S, and A. W. Black, "The IIIT-H indic speech databases," in *INTERSPEECH*, 2012.

[10] P. S. Dikshit and R. W. Schubert, "Electroglottograph as an additional source of information in isolated word recognition," in *Proceedings of the Fourteenth Southern Biomedical Engineering Conference*, 1995, pp. 1–4.

[11] A. Stan, J. Yamagishi, S. King, and M. Aylett, "The Romanian speech synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate," *Speech Communication*, vol. 53, no. 3, pp. 442–450, 2011.

[12] M. Rashwan, M. Al-Badrashiny, M. Attia, S. Abdou, and A. Rafea, "A stochastic Arabic diacritizer based on a hybrid of factorized and unfactorized textual features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 166–175, 2011.

[13] M. Attia, "Theory and implementation of a large-scale arabic phonetic transcriptor, and applications," Ph.D. dissertation, Dept. of Electronics and Electrical Communications, Cairo University, Cairo, Egypt, 2005.

[14] S. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The htk book version 3.4," *Cambridge University Engineering Department*, 2006.