ARABIC TEXT TO SPEECH SYNTHESIS SYSTEM

Aymen EL KADHI[1], Guntram STRECHA[2], Rüdiger HOFFMANN[2] and Hamid AMIRI[1]

1 National School of Engineers of Tunis, University Tunis El Manar
2 Institute of Systems Theory and Speech Technology, TU Dresden, Germany

ay.kadhi@yahoo.fr; guntram.strecha@ias.et.tu-dresden.de; Ruediger.Hoffmann@tu-dresden.de; hamidlamiri@gmail.com

Abstract:

This paper describes in details the construction of an Arabic Text To Speech (TTS) Synthesizer System for the Arabic language. The conversion process from input text into acoustic waveform is performed in tow steps. The first one is text analysis, we developed a module for the conversion of input text into a phonetic representation. The second one is the generation of speech waveforms, where the acoustic output is produced from this phonetic transcription. For this task we used the framework which is named UASR (Unified Approach for Speech Synthesis and Recognition). This system needs training because it is based on HMM speech synthesis.

We trained it with a Standard Arabic Single Speaker Corpus (SASSC), it is composed of more than 7 hours of professional speaker recording.

The results obtained after testing our text-to-speech synthesis system show that the developed system is intelligible and it also obtains an acceptable level of naturalness.

Keywords: Arabic language, speech synthesis, Grapheme to phoneme conversion, HMM

1 Introduction :

Synthesis of speech from text means all treatment allowing a machine to transform a written text into a spoken message. No restriction is made on the nature of words to synthesize (abbreviation, date, number ...), neither on the size of the vocabulary to use.

The contemporary standard Arabic cannot be distinguished from the classical Arabic. It keeps almost in its entirety the same syntax and morphology, it is called literal Arabic. It is the language of the liturgical Islamic religion, press, media, conferences, modern literature, and political speeches. It has a stable and well-regulated writing that is broadcast through a formal education. Standard Arabic retains a monopoly in any official life of the inhabitants of 24 countries whose number exceeds 360 million.

This paper describes the overall architecture of the Arabic TTS, several components of the system, and the linguistic concepts for the Arabic language. A block diagram of a general TTS engine is illustrated in Figure 1.

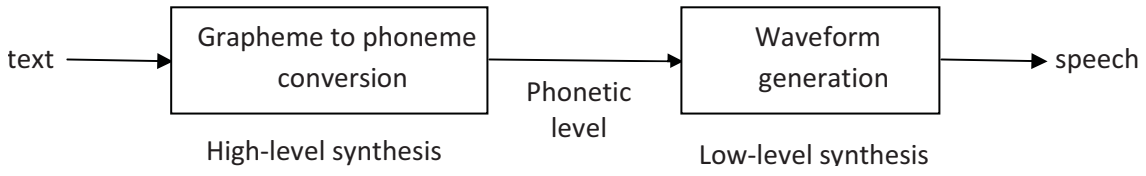

Figure 1- General structure of a TTS synthesizer

At the present time two methods dominate the TTS systems, units selection and parametrics systems. Our work focuses on the parametric speech synthesis by HMM. Hidden Markov Models (HMMs) demonstrated to be an efficient parametric model for speech acoustics in the framework of speech synthesis [1].

This paper is structured as follows. Section 2 describes the Arabic phonology developed to generate phrase targets and phonological features. The description of grapheme-to-phoneme conversion is presented in section 3. The used corpus is presented in section 4. Section 5, describe briefly the TTS model generation and output evaluation procedures. Section 6, summaries our conclusions and an expected future work.

2 ARABIC LANGUAGE

As part of our work we will be dealing with Arabic language in reference to what is commonly known as Modern Standard Arabic language. It is the common official language throughout the Arab world.

The Arabic alphabet is composed of 28 letters. All of them are consonants (Figure 1), and three of them are also used as long vowels (ا و ي). We have 6 vowels which are divided into three short vowels and three long vowels. The duration of a long vowel is about twice the short vowel. These vowels are characterized by the vibration of the vocal cords. The long vowels have similar spectral properties like their short vowel versions. They are represented in the following table:

short vowels	long vowels
/ُ /ِ /َ	ـَا ـِي ـُو

Table 1- Classification of vowels of the Arabic language.

However, the current system has 38 phonetic letters by adding extra phonemes to reflect the effect of the pharyngealized phonemes. Used phonemes are mentioned in the table 2 [2].

An Arabic word is written with consonants and vowels. Vowels are added above or below consonants. The absence of vowels creates some ambiguity at two levels: Meaning of the word and difficulty in identifying its function in the sentence. Arabic sentences are read from right to left. Arabic letters change the presentation form according to their positions in the word.

Syllabification for Arabic language has only six syllable types (CV, CVC, CVV, CVVC, CVCC and CVVCC). Due to their heavy pronunciation the last three types generally appear at the end of phrase. The number of vowels and the number of syllables must be equal in phrase.

Symbol	Phone	Symbol	Phone	Symbol	Phone	Symbol	Phone
a	(فتحة) َ	gh	غ	m	م	T	ط
A	(فتحة مفخمة) ً	h	ه	n	ن	th	ث
aa	ا	i	(كسرة) ِ	pau	pause	TH	ظ
Aa	(مد مفخم) ا	I	ي	q	ق	u	(ضمة) ُ
Ah	ح	j	ج	r	ر	U	و
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

3 Grapheme-to-phoneme conversion :

Grapheme-to-phoneme (G2P) conversion called also letter-to-sound conversion has become an indispensable component in the Text To Speech synthesis system. This conversion is to create the phonetic sequence associated with the sequence of graphemes. A grapheme is the smallest distinctive and meaningful unit of writing, it commonly represents a letter of the alphabetic writing.

There are several techniques of phonetisation: some use explicit knowledge data as a phonetic lexicon or rules of transcription, other must first pass through a training stage from aligned phonetic corpus [3] [4].

The approach we have adopted to implement the grapheme to phoneme conversion

is divided into two phases of language processing. The first phase is the pre-processing module, it organizes the input sentences into convenient lists of words or breathing groups. It also identifies abbreviations, acronyms and numbers in order to transforms them into a full text.

The second phase consists of the phonetic text pretreated using two different methods. The first method is based on the use of a lexicon containing a list of words and abbreviations exceptions, by directly entering the corresponding phonetic words without using the rule-based phonetic transcription, which ensures more speed in processing. The second method is to treat the rest of the text by using a rule-based phonetic transcription.

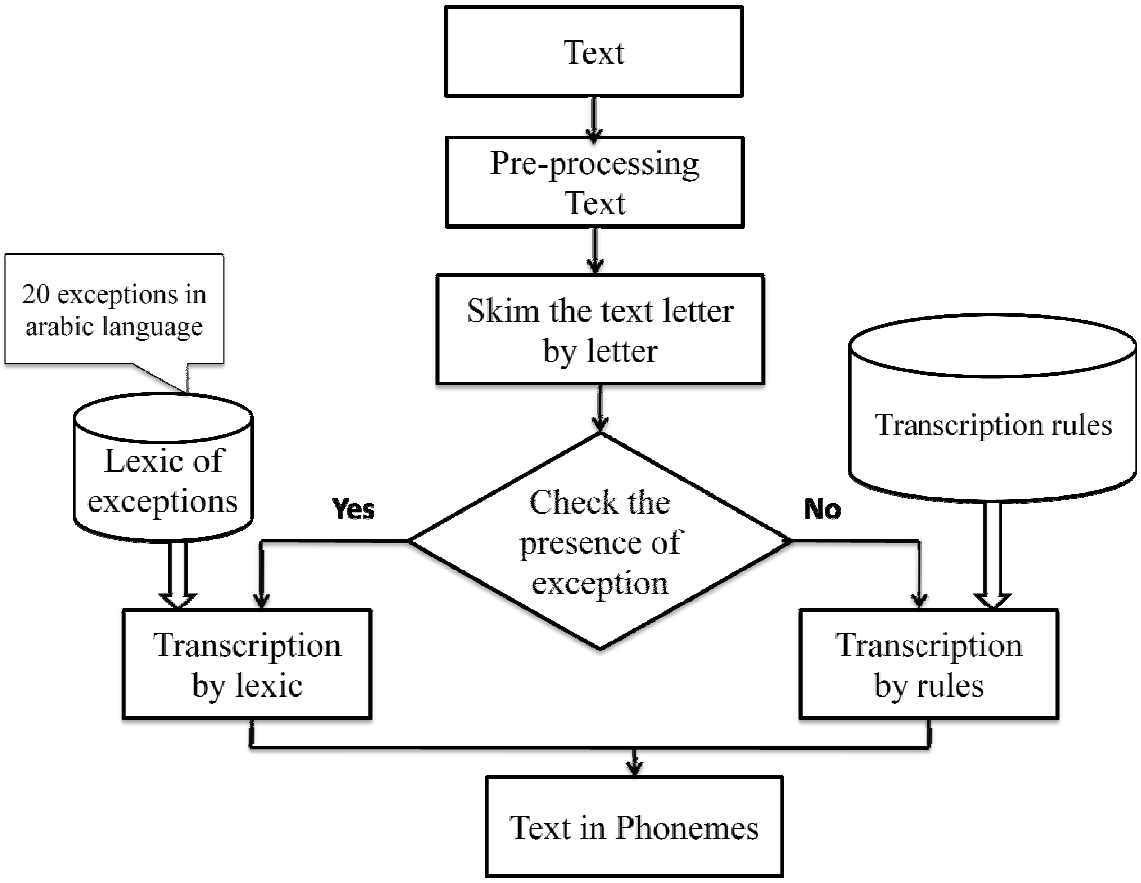

Figure 2 - Architecture of grapheme-to-phoneme conversion system

4 Speech corpus

Us our system is an HMM-based speech synthesis, we need a speech corpus to train the speech model. In state of the art speech synthesis system needs at least a total volume of 5h of speech. Ideally the recordings should cover all phonetic variations as well as all prosodic variations. The convenience of the recorded speech depends on the quality of the speech signal and on the precision with which the transcription was done [5].

In current work the speech corpus called Standard Arabic Single Speaker Corpus (SASSC) is used. It was developed in the computer research institute of King Abdulaziz city for science and technology of Riyadh, Saudi Arabia. It was created to be used as a database for speech recognition and HMM-based speech synthesis, it is freely available for research and educational purposes.

The corpus consists of 51,432 words, it required 7 hours and 20 minutes of audio recording.

The SASSC corpus contains several files that correspond to the different sentences. Each file audio ". wav" (sampling rate of 96 kHz), it matches the base 3 other files that are as follows:

- EGG signal in 96 kHz
- Diacritized text
- Transcription along with the phoneme boundary segmentation.

Owing to the nature and size of the corpus, a professional male news anchor who has worked in several TV programs was selected due to his capacity in preserving his performance for long recording sessions. A linguist was present to correct the speaker in case he skipped or mispronounced. Each recording session was 15 minutes long and no more than four sessions per day were taken [2].

5 Waveform generation and evaluation

This section will describe briefly the TTS model generation and output evaluation procedures.

In current work we used parametric speech synthesis, which is more and more diffused into the field of speech synthesis due to the quality of results insured by this method.

The model is parametric because it describes the speech using parameters, rather than stored exemplars. It is statistical because it characterizes those parameters using statistics (e.g., means and variances of probability density functions) which capture the distribution of parameters found in the training data [6]. The principle is to use the HMM in a generative mode (not as classifier) to find a plausible sequence of parameters for each diphone used in the Arabic language. For this aim we used the Unified Approach to Speech Synthesis and Recognition (UASR) system [7]. First step was the conversion of our speech recordings from 96kHz to 16kHz with a simple bash script. After that we trained the HMM phoneme models.

Therfore we used just diphone carrier words (one word per diphone) to train the models. We used these HMMs to compute the optimal HMM state sequences of all carrier words and chose the diphones whitch have the best recognition log-likelihood for duplicate diphones.

Each selected diphone is represented by a sequence of HMM state indexes. When we synthesis these index sequences will be concatenated according to the wanted phoneme sequence [8].

To rule out the influence of synthetic prosody on the listening test we introduced manually natural prosody in all experiments. And we asked individually 20 Arabic natives speakers evaluators to choice arbitrary 5 audio clips from a directory containing 50 synthesized wave files and rate each audio clip (on a scale from 1 to 5) based on the following qualities :

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

At the end we computed the average: for naturalness score it was 3.1 and for intelligibility it was 3.9.

6 Conclusion

The overall architecture of our Text-To-Speech system for Arabic language has been presented. The exposed results validate the use of HMMs as consistent models of Arabic speech, in the sense that they are capable to product a high-quality synthetic sentences.

In order to increase the naturalness and to have a complete Arabic text to speech synthesis system with the best possible quality of the output speech we need the inclusion of prosodic properties automatically.

References

[1] H. Zen, T. Toda, and K. Tokuda, “The Nitech-NAIST HMM-based speech synthesis system for the BlizzardChallenge 2006,” in Proc. Blizzard Challenge’06, 2006.

[2] Almosallam, A.AlKhalifa, M. Alghamdi, M. Alkanhal, A. Alkhairy SASSC: A Standard Arabic Single Speaker Corpus I. 8th ISCA Speech Synthesis Workshop August 31 – September 2, 2013, Barcelona, Spain.

[3] Paul Taylor. Hidden markov models for grapheme to phoneme conversion. In Proc. of Interspeech, 2005.

[4] Chotimongkol, A., & Black, A. W. (2000, October). Statistically trained orthographic to sound models for Thai. In *INTERSPEECH* (pp. 551-554).

[5] W. Zhu; W. Zhang; S.Qin; F. Chen; H. Li; X. Ma; L. Shen, "Corpus building for data-driven TTS systems," *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop*, pp.199,202, 11-13 Sept. 2002

[6] King, S. : An introduction to statistical parametric speech synthesis, *Sadhana* 36(5) (2011), 837-852.

[7] Hoffmann, R., Eichner, M., Wolff, M.: Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system. In: A. Esposito et al. (Eds.): Verbal and Nonverbal Communication Behaviours. Berlin etc.: Springer 2007 (LNAI vol. 4775), 200 - 218.

[8] Strecha, G., & Wolff, M. : Speech synthesis using hmm based diphone inventory encoding for low-resource devices. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on* (pp. 5380-5383). IEEE.