Standard Arabic Talking Clock Based on Diphone Synthesis

Nadjla BETTAYEB, Mhania GUERTI
Department of Electronics
Signal and Communications Laboratory
Ecole Nationale Polytechnique - Algiers, Algeria
Nedjla.bettayeb@g.enp.edu.dz, mhania.guerti@enp.edu.dz

Abstract— In this paper we present an automatic speech synthesis system by Concatenation in Standard Arabic language. For this purpose, we will describe all the followed steps to design a talking clock application. In this limited vocabulary application, we used the diphone as the basic unit for the synthesis without any modification or processing of the resulting voice, so we can test the quality of such system.

Keywords— Speech Synthesis; Standard Arabic; Concatenative Synthesis; Diphone; Talking Clock

I. INTRODUCTION

Nowadays, technological development and voice interaction with the machine have made the life of humans easier and more comfortable. This interaction requires a speech synthesis system that we can integrate into many commercial applications such as: assistance systems for handicapped persons, languages learning machines, talking devices whether in Domotic or in the industry … etc.

Since the construction of the first synthesizer to the latest research on speech synthesis systems, many methods have been developed and applied to obtain a voice close to the human voice (intelligible and natural synthetic voice)[1]. In recent years, the Arabic language has known a rapid and enormous progress in this area. This can be perceived in the different methods used to improve the quality of synthesis like the concatenative synthesis in [2] and [3], statistical parametric synthesis in [4] and [5].

In this paper we will talk about the main methods of speech synthesis, as well as notions of SA “Standard Arabic”, we will present our system of diphones concatenative speech synthesis in SA and the evaluation that was made to test the system and finally the discussions of the results obtained by this system.

II. SPEECH SYNTHESIS METHODS

There are two main speech synthesis methods:

A. The Parametric Speech Synthesis

This method is based on modeling the speech by contextual rules. These rules and from specific phonetic-prosodic informations (such as: the sequence of phonemes, the duration of each sound to produce, the melodic set point, etc…) allow to determine the path of the parameters of the signal representation model. With the development of research, the current parametric synthesis systems modulate the evolution of the acoustic parameters by statistical models which we called statistical parametric speech synthesis, and the most used statistical model is the HMM (Hidden Markov Model)[6].

B. Concatenative speech synthesis

The second method is based on the juxtaposition of prerecorded units. The choice of the unit depends on the targeted application. In cases of limited vocabulary applications, we can use the word or phrase as a selection unit, but otherwise, it is not easy to record all existing phrases or words in a given language, for that a smaller size unit is necessary (as the phoneme, the diphone or the syllable). Nevertheless, the more we reduce the size of the unit, the more synthesized speech naturalness decreases because of the overlapping areas generated during the concatenation. For this reason, we often integrate to these systems a prosodic processing module (by modeling the pitch and the duration of each unit) to achieve the desired prosody.

This modification step may decrease the quality of the synthetic voice and makes it sound unnatural. Because that problem, a new concatenation by synthesis approach has been applied which is the unit selection synthesis. The latter is also known as corpus based synthesis. It is based on stocking not just units with variable size but also with different linguistic and prosodic contexts for each unit, and then selects the units that best match the desired prosody.

Because the US (Unit Selection) synthesis, which is considered as the best concatenative synthesis method, requires a large memory resource. And the speech quality generated from the first method is not as good as the latter “which uses natural segments to synthesize the speech”[7], the recent research tries to build speech synthesis systems that combine these two methods [8].
III. NOTIONS ON ARABIC LANGUAGE

Arabic is the first language in 23 countries and it is used by 1.62 billion Muslims in the world. Arabic refers to the official Standard Arabic taught in schools. It consists of 40 phonemes: 26 consonants, 3 short vowels, 3 long vowels, 2 semi-vowels and 6 vocalic variants in emphatic context. Orthographically these three short vowels («َ»: [a], «ِ»: [u], «ى»: [i]) are not considered within the word structure, they are added with other diacritical marks like the Sukun “َ” (the silence) and Chada “ِ” (the gemination) to the sentences to distinguish the meaning of each word. In general, we find them only in the language learning books for the beginners and in the Holy Quran. This is the main problem of the Arabic speech synthesis systems.

IV. WORK METHOLOGY

A. The Choice of corpus

The main step in a speech synthesis system by concatenation is the construction of a corpus that contains at least all the possible combinations of the selected units. To choose the diphone «which stretches from the stable part of the first phoneme to the stable part of the second taking into account the transition between them» as a basic unit we need about 1600 pairs for a complete SA speech synthesis system (a system can synthesis any word). However, because our system is a limited vocabulary system our recordings result in 412 pairs of diphones.

B. Registration phase

A female speaker did the recording of our corpus with a sample rating of 44100 Hz, and to reach a maximum number of diphones we recorded 46 phrases and words from an average of 5 to 15 phonemes. We recorded all the possible hours and minutes in addition to the time indicators like [gabaan] "AM", [masaan] "PM"... etc.

C. Segmentation and annotation

Before the sentence’s segmentation, an analysis step was done to determine the boundaries of each unit to segment. We made manual segmentation using the Praat analysis tool [9], based on the audiogram (Fig.1) «by noticing the change in the shape of the speech waveform» and the spectrogram (Fig.1) which is a three-dimensional representation of the three relevant speech parameters (the fundamental frequency $F_0$, the duration and the energy).

After determining the boundaries of each word by listening and spectrographic analysis, a segmentation into phoneme was done firstly with a small adjustment based on listening, and finally into diphones[3] (Fig.2).

The choice of the diphone as the basic unit is made so as not to clutter up our database with the choice of a greater unit, and to avoid the spectral discontinuities generated during the concatenation of smaller units "like phoneme". Because the diphone includes transition from one phoneme to another and the coarticulation (the influence of a phoneme to another) between phonemes is retained.

Then these segments were extracted and saved in the database according to a special code. This label is based on the principle that every Arabic phoneme is represented with a single keyboard character, to facilitate the representation and the call of the target diphone, except the long vowels, which are represented by a double character (TABLE 1).

D. Synthesis

As illustrated in (Fig.3) after specifying the time, the system uses a translation dictionary to convert the hour digits into their written pronunciations depending on the selected code. After this conversion, the system scans each resulting word or phrase, character by character to call in each time the specified diphone until the end of the word.
TABLE I. THE ARABIC CHARACTERS THEIR IAP (INTERNATIONAL PHONETIC ALPHABET) AND CHOSEN CODES

<table>
<thead>
<tr>
<th>Arabic character</th>
<th>API code</th>
<th>Chosen code</th>
<th>Arabic character</th>
<th>API code</th>
<th>Chosen code</th>
</tr>
</thead>
<tbody>
<tr>
<td>أ</td>
<td>[ʔ]</td>
<td>[ʔ]</td>
<td>ع</td>
<td>[ʕ]</td>
<td>[ʕ]</td>
</tr>
<tr>
<td>ب</td>
<td>[b]</td>
<td>[b]</td>
<td>غ</td>
<td>[ɣ]</td>
<td>[ɣ]</td>
</tr>
<tr>
<td>ت</td>
<td>[t]</td>
<td>[t]</td>
<td>ف</td>
<td>[f]</td>
<td>[f]</td>
</tr>
<tr>
<td>ث</td>
<td>[θ]</td>
<td>[θ]</td>
<td>ك</td>
<td>[k]</td>
<td>[k]</td>
</tr>
<tr>
<td>ج</td>
<td>[d]</td>
<td>[d]</td>
<td>ل</td>
<td>[l]</td>
<td>[l]</td>
</tr>
<tr>
<td>ح</td>
<td>[h]</td>
<td>[h]</td>
<td>م</td>
<td>[m]</td>
<td>[m]</td>
</tr>
<tr>
<td>د</td>
<td>[d]</td>
<td>[d]</td>
<td>ن</td>
<td>[n]</td>
<td>[n]</td>
</tr>
<tr>
<td>ذ</td>
<td>[z]</td>
<td>[z]</td>
<td>ه</td>
<td>[h]</td>
<td>[h]</td>
</tr>
<tr>
<td>ر</td>
<td>[r]</td>
<td>[r]</td>
<td>و</td>
<td>[w]</td>
<td>[w]</td>
</tr>
<tr>
<td>ز</td>
<td>[z]</td>
<td>[z]</td>
<td>ي</td>
<td>[i]</td>
<td>[i]</td>
</tr>
<tr>
<td>س</td>
<td>[s]</td>
<td>[s]</td>
<td>ن</td>
<td>[n]</td>
<td>[n]</td>
</tr>
<tr>
<td>ش</td>
<td>[ʃ]</td>
<td>[ʃ]</td>
<td>ج</td>
<td>[g]</td>
<td>[g]</td>
</tr>
<tr>
<td>ص</td>
<td>[s]</td>
<td>[s]</td>
<td>ظ</td>
<td>[θ]</td>
<td>[θ]</td>
</tr>
<tr>
<td>ض</td>
<td>[d]</td>
<td>[d]</td>
<td>ز</td>
<td>[z]</td>
<td>[z]</td>
</tr>
<tr>
<td>ط</td>
<td>[t]</td>
<td>[t]</td>
<td>ق</td>
<td>[q]</td>
<td>[q]</td>
</tr>
<tr>
<td>ظ</td>
<td>[z]</td>
<td>[z]</td>
<td>ك</td>
<td>[k]</td>
<td>[k]</td>
</tr>
</tbody>
</table>

Generally, and in the case of diphones, the search and the call is done by two characters except for some cases like the long vowel, which is itself represented by two characters and the geminated consonant (consonant doubling) that we took the geminate consonants as one phoneme, these exceptions are:

- [CVV]: a consonant followed by a long vowel.
- [VVC]: a long vowel followed by a consonant;
- [VCC]: a vowel followed by a double consonant;
- [CCV]: a double consonant followed by a vowel;
- [CCCV]: a double consonant followed by a long vowel.

E. Evaluation

To evaluate our system, two different tests were applied to the intelligibility and naturalness of synthesized words. This subjective assessment was applied to thirteen students (aged between 20 and 30 years) (eight girls and five boys) who know the SA very well and they are not familiar with the system nor with a synthetic voice [3].

For the first test of measuring the intelligibility, we launch the application’s program to each person who must say what he heard. We note intelligible if the person said correctly what is synthesized and not otherwise. For a better evaluation, the person does not know what this application is, nor the phrase or word to pronounce, and the test is performed at different times in the day.

In the second test for natural, we designed an interface using the MATLAB tool [10] to our application (Fig.4), which allows us to enter any time we want to hear. The mission to give to the persons in this test is to try three different times by entre the hour and minute of the time they want to hear, then press the button listen. After listening to the pronounced sentence they have to evaluate their satisfaction on the application and the voice quality from one to five (5: excellent, 4 very good, 3: good, 2: medium, 1: bad).

F. Discussion

After evaluation of the application, we had a 100% of intelligibility. This percentage was expected because this application is limited in vocabulary (the synthesized words are the same used to build the database), also the synthetic sentences are neither complex nor difficult to be identified by an Arab person.

The natural test (TABLE II) gives a percentage of 73.85% since and according to peoples who have done the test, sometimes they hear a discontinuities in the pronounced sentence (actually they feel the transitions between some diphones) and this may be due to the following causes:

- The recording was not ideal, because we have to ensure that the different words and phrases had been pronounced monotonically and well articulated, a small change in voice prosody can influence in the concatenation phase.
- The difficulty of segmenting some complex sounds: like the liquids that mix with neighboring sounds and the disability to differentiate between them, or the sounds that have adjacent or identical articulation places.
- The coarticulation between sounds, which is a very known problem in the Arabic speech synthesis system, for example we find that the vowels change

Fig.3 The bloc diagram of the talking clock application
slightly their characteristics in the vicinity of certain sounds, as in the case of the emphatics (if an emphatic consonant precedes a vowel that vowel becomes emphatic).

<table>
<thead>
<tr>
<th>Rating</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of votes</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig.4 The main interface of the naturalness test

V. CONCLUSION

In this paper, we have presented all the steps followed to design an application of an Arabic talking clock, which is a speech synthesis system based on diphones concatenation.

After testing this application and analyzing its results, we can judge that the synthesis by diphone concatenation can give an intelligible result. Whereas, for the general public use, or for open applications (not limited vocabulary applications), the result is not entirely satisfactory because the naturalness of the voice of such systems has not the required quality. For these reasons, we suggest the following alternatives:

- Include a prosodic processing module to the speech synthesis system to correct and adjust the synthetic voice;
- Enrich the database with various prosodic examples for each diphone and select the best one in the concatenation phase (concatenation by unit selection);
- Use other concatenation units beside the diphone like the triphone or the syllable especially for the complex sounds.

Moreover, the choice of the speaker and the way of recording the corpus plays very important role in a concatenation by synthesis system.

REFERENCES