An-Najah Univ. J. Res. (N. Sc.), Vol. 19, 2005

Text-to-Speech Synthesis by Diphones for Modern Standard Arabic

تركيب الكلام من النص بواسطة ثنائيات الأصوات للغة العربية المعاصرة

Nader Abu Ghattas, & Hanna Abdel Nour

Department of Electronic Engineering. Faculty of Engineering. Al-Quds University. Jerusalem, Palestine

E-mail: qudsengi@planet.edu

Received: (29/5/2001), Accepted: (22/11/2005)

Abstract:

An unlimited vocabulary text-to-speech synthesis by diphones system is used to generate Modern Standard Arabic speech: the system is the PSOLA algorithm; the diphones are obtained from the permutation of 44 phones (phonemes and allophones). The diphonic combinations were introduced in carrier words and recorded by a selected speaker. A dictionary of diphones was established by means of a process of segmentation that abided by certain rules. Evaluation of the system was undertaken to assess the accuracy on word and sentence levels. The results showed high perception levels.

Index Terms: Modern Standard Arabic, Speech Synthesis by Diphones, Segmentation, Evaluation.

ملخص

استُخدم نظام "تركيب الكلام من النص بواسطة ثنائيات الأصوات" لإنتاج كلام باللغة العربية الحديثة دون حدود للمفردات: النظام هو برمجيات PSOLA وحصلنا على ثنائيات الأصوات من تباديل ٤٤ صوتا مختلفا. أدخلت هذه الأزواج على كلمات "ناقلة" وسجلت بصوت قارئ مختار، ومن ثم تم تجزئة هذه التسجيلات حسب قواعد محددة للحصول على "قاموس" من ثنائيات الأصوات. قُيِّم هذا النظام على مستويين: دقة الكلمة ودقة الجملة. دلت النتائج على درجة عالية من الوضوح.

Introduction

Modern Standard Arabic (MSA) is the language commonly understood in all Arab countries. Speech Synthesis of MSA has already been realized using Klatt synthesis [1]. This method uses the formants as its basic building blocks to generate sounds. But this synthesizer rigidly classifies the excitation models as an entirely voiced or unvoiced signal and hence smooth transitions between phonemic boundaries are difficult to achieve [2]. An alternative method is synthesis by concatenating units of prerecorded speech. When the unit of sound is a phone, the concatenation process proves to be unsuccessful because of the lack of knowledge of the transients that link up two successive phones. Larger speech units, such as diphones and demisyllables, which preserve the transients in their middle give smooth speech as concatenation is done in the stable zone [3]. This method presents real practical advantages and it already gave high quality synthesized speech. In the case of Arabic speech, a first method was utilized by El-Imam [4]; it employed a limited number of diphones. The reported understandability tests gave 86% word accuracy. Another approach was that of Ghazali et al. [5]. The system took into account a complete inventory of diphones that were gathered into six categories [6]. Those same categories were used in this research.

The chosen synthesizer was the PSOLA[1] algorithm developed by the CNET/France [8]. It is operational for several European languages as the algorithm is language independent [8]. We used this algorithm for the unlimited vocabulary Text-to-Speech (TtS) synthesis of MSA text that is written with diacritics.

TtS synthesis by diphones consists of a chain of processes, which are: text transformation, grapheme-to-phoneme conversion, prosody generation, and finally segmental concatenation. This paper is concerned with the last stage, which is the production of MSA speech from a diphones inventory. The starting point is to identify the phones of MSA, and then chose an appropriate speaker whose utterance of MSA sound-units, here the diphones, will be used by the synthesizer.

[1] PSOLA is a CNET/FRANCE TELECOM registered trademark.

The number of MSA phones is 44 [6] (see annex 1). All possible diphonic combinations were introduced in nonsense words. An appropriate speaker, whose voice was suitable for synthesis as well as for analysis procedures, was chosen. All the nonsense words were recorded in a professional studio[2]. Sampled at 16kHz, the corpus of words was processed by PSOLA speech engine. Next, the segmentation of the diphones was manually done according to certain rules that were defined by the researchers. These rules took into account the particularities of the MSA [9] and the restrictions imposed by the used synthesizer. With this speaker's voice, a dictionary of diphones was thus produced. The resultant diphonic units of speech signals were integrated into the PSOLA synthesizer; synthesizing speech of unlimited vocabulary in MSA could then be carried out.

In order to evaluate the work done a corpus of sentences was produced. The evaluation of TtS by diphones was done using 238 sentences, which covered 100% of the total number of phones and 53% of the total number of diphones of the MSA. The test consisted of a dictation of this corpus by a personal computer to a group of six listeners.

Building-up the Dictionary of Diphones

A- Selection of the Speaker

As a first step an appropriate speaker is carefully selected since the quality of the used voice is important for the analysis and synthesis processes. This speaker was chosen out of three persons that had different Arabic dialectical origins: Palestinian, Syrian, and Moroccan. The best test that could be done was to synthesize sentences by the three voices and then evaluate the quality of the synthesized speech. A group of seven sentences was built for this purpose (see annex 2). The constituent diphones were then identified and introduced into nonsense

[2] Recordings had been performed in the Journalism Studio of the IUT 'B' / Bordeaux 3 university, Bordeaux, France.

words. Three limited dictionaries of diphones were then developed. The seven sentences were synthesized by each one of the three dictionaries.

The whole process had been repeated for the complete set of the 1665 diphones of the MSA with the voice of the selected speaker [6]. They were introduced into nonsense words of the structures /TA++ASA/ and /TA++SA/, the plus signs indicate the location of the diphone. These two structures were chosen to cover all the syllabic components of MSA [9]. The recorded nonsense words were then sampled at 16kHz, coded into 16 bits each, and the obtained digital speech was processed by PSOLA speech engine. In each of the nonsense words, the diphone was then manually delimited.

B- Segmentation of Diphones

The marking of the diphones present in the carrier nonsense words was performed manually and according to certain rules. The rules took into consideration the relevant aspects of the MSA as well as the characteristics of the synthesizer. The segmentation was done by specifying the instants of the beginning and of the end of a diphone, in addition to an instant in the transition region. The instants of beginning and end were of course chosen in the middle of the stable regions of the constituent phones of the diphones.

With regard to the PSOLA synthesizer there were only two constraints to consider: there had to be at least three pitch-marks between any two manually imposed segmentation marks, and a duration of 200ms had to be kept inside the limits of any diphone that enclosed a silence.

The segmentation was performed in the time domain. The instants to mark were identified by sight on the signal and its spectrogram. The validity of the segmentation was examined by listening to the resultant diphone. This verification was done to ensure that the two phones of a diphone were present, and that there was no interference from adjacent phones.

Considering the characteristics of the MSA the following guidelines were used:

- The transition mark is put at the beginning (end) of a voiced phone when the later is on the right (left) side.
- When a vowel is short (long) 45 msec (75 msec) of it are included in the diphone.
- In the case of CC diphones, where one of the consonants is voiced and the other unvoiced, the transition mark is put slightly at the beginning of the voiced consonant.
- From a plosive found to the right (left) of a diphone, silence (burst) is taken.
- A small part the phoneme /ε/ is integrated into the diphone regardless of its order. But when followed by a silence all of it has to be incorporated.
- The location of a glide is identified from the knowledge of the durations of the adjacent phones. The segmentation mark is put in the middle of its time interval.
- The full duration of the consonant /ħ/ is considered to be at most 100 msec. Otherwise when synthesized it will be heard as if geminated. The same is to be respected for the consonants /x/, /s/, /ʃ/, /f/ and /h/.
- The segmentation of the diphones for which /b/ is its second constituent is done at the middle of the periodic interval of the voiced labial.
- In the case of /d/ the burst is clear after the periodic part of the signal. Integrate the entire periodic interval with the previous phoneme of the diphone when this voiced alveolar is to the right.
- /l/ is similar to /d/.
- /r/ appears to be divided into two parts separated by a small amplitude signal. In the frequency domain the small amplitude signal has very low energy (a valley of energy between it and the next vowel). Segmentation is done in the middle of this low energy, small amplitude part, and if possible at a passage by zero.
- The segmentation of the *hamza* /?/ is relatively difficult. In fact it is extremely influenced by its neighborhood. Its segmentation requires that each realization should be considered as a special case. The *hamza* is an occlusive consonant; therefore it must begin with a silence and end with a burst. In reality it is not as simple as that.

When /?/ is between two vowels, V/?/V, the observed phenomenon is a transition period between two vowels similar but longer in time compared to the case VV. The Short-Time FFT spectrum shows that the interval of occlusion has the same spectrum as the following vowel but with energies that are much smaller. In fact the vowels modulate the consonant /?/ because the occlusion is not perfect, that is the closure of the glottis is not complete.

Evaluation of the TtS Synthesis by Diphones

This was a preliminary evaluation done to ensure that the produced speech was audible and clear. Therefore a corpus was produced to assess the intelligibility of the synthesized speech at the sentence and word levels. The sentences were semantically correct and meaningful but in several cases contained uncommon words. Naturalness was by no means a goal because nothing was yet done regarding intonation. The corpus was made up of 238 sentences that covered 100% of the total number of phones and 53% of the total number of diphones of MSA [8].

The researchers selected a group of six naive listeners who were not at all familiar with the system. The listeners were not formerly trained on machine voice. Previously synthesized, the sentences were dictated to the audience by a personal computer. Each sentence was uttered only once, and listeners had to write down what they heard. The test was done in an ordinary, not isolated acoustically room.

The objective results that were obtained from this test were as follows: the average percentage of entirely correct sentences (sentence accuracy) was 67%, the average percentage of correct words (word accuracy) was 87%. The average number of words per sentence was 3.

Conclusion

The main part of TtS synthesis, which is building a dictionary of diphones for MSA using PSOLA algorithm was accomplished. Two factors were of importance for this work: the choice of the speaker and the definition of segmentation rules. As a result the system generated

clear utterances. The next stages will deal with a thorough evaluation, automatic text-to-phone conversion and naturalness.

Aknowledgement
This project was done in cooperation with Elan Informatique, Toulouse, France. We would like to thank Professor M. Najim who gave us the opportunity and the means to work on this subject. Special credit must be given to Ms. Amneh Oreiqat who made an excellent work in building up the corpus of test sentences and in performing the test. We are thankful to DRIC / French Ministry of Higher Education for partially supporting this work.

References

[1] Rajouani, A. et al., "An Arabic text-to-speech system based on rules", Proc. ICEMCO'96, Cambridge, UK (1996), pp. 65-69.
[2] Lam, K-f et al., "Interpolating V/UV mixture functions of a harmonic model for concatenative speech synthesis", Proc. IEEE ICASSP-96 (1996), pp. 393-396.
[3] O'Shaughnessy, D., "Speech Communications, human and machine", 2nd edition, IEEE press, USA (2000).
[4] El-Imam, Y. A., "An unrestricted vocabulary Arabic speech synthesis system", *IEEE Transactions on Acoustics*, (1989), pp. 1829-1844.
[5] Ghazali, S. et al., "Synthèse de l'arabe standard à partir du texte par TD-PSOLA : le traitement des processus phonologiques", Proc. 19th JEP Conference, Brussels, (1992), pp. 89-93.
[6] Abdel Nour, H. et al, "Vowel analysis of spoken Palestinian Arabic," *Bethlehem University Journal*, **vol. 18** (1999), pp. 71-81.
[7] Moulines, E. et al, "A real-time French text-to-speech system generating high quality synthetic speech," Proc. IEEE ICASSP-90 (1990), pp. 309-312.
[8] Bigorgne, D. et al, "Multilingual PSOLA text-to-speech system," Proc. IEEE ICASSP-93, **vol. II** (1993), pp. 187-190.
[9] Rajouani, A. et al., "Consonant duration from an Arabic speech database", Proc. IEEE ICASSP-96 (1996), pp. 1104-1107.

Annex 1

Phones of the MSA

A- Consonants:

Arabic Symbol	ء	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص
Phonetic Symbol	/ʔ/	/b/	/t/	/θ/	/ʒ/	/ħ/	/x/	/d/	/ð/	/r/	/z/	/s/	/ʃ/	/s̲/
Arabic Symbol	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	هـ	و	ي
Phonetic Symbol	/d̲/	/t̲/	/ð̲/	/ɛ/	/ɣ/	/f/	/q/	/k/	/l/	/m/	/n/	/h/	/w/	/j/

B- Vowels:

Arabic Symbol	ـَ	ـُ	ـِ	ـا	ـو	ـيـ	ـَ	ـُ	ـِ	ـا	ـو	ـيـ
Phonetic Symbol	/a/	/u/	/i/	/aa/	/uu/	/ii/	/â/	/û/	/î/	/ââ/	/ûû/	/îî/
Arabic Symbol	ـُ	ـِ	ـو				Pharyngealized مفخمة					
Phonetic Symbol	/o/	/e/	/oo/									

See Abdel Nour et al., 1999.

Annex 2 Test sentences for the selection of a speaker

1) /#ɛallamakafannalmuusiiqââ#/ عَلَّمَكَ فَنَّ الموسيقى
2) /#jarwijabittamaami#/ يَرْوِيَ بِالتَمامِ
3) /#s̲ââd̲âfâxâruufajni#/ صادَفَ خَروفَيْنِ
4) /#juzaɣridumunħanijan#/ يُزَغْرِدُ مُنْحَنِياً
5) /#xutimatfað̲âhârât#/ خُتِمَتْ فَظَهَرَتْ
6) /#tât̲îîrû s̲oħbâtâhâânaħwaʃʃamsi#/ تَطيرُ صُحْبَتَها نَحْوَ الشَمْسِ
7) /#uktobɛanħaadiθaten#/ أُكْتُبْ عَنْ حادِثَةٍ