Di-Diphone Arabic Speech Synthesis Concatenation

Abdelkader Chabchoub[1]
Salah Alahmadi[1]
[1]Electronics Department, College of Technology of Medina, KSA

Wahid Barkouti[2]
Adnan Cherif[2]
[2]Signal Processing Laboratory
University of Tunis El Manar Tunisia

ABSTRACT
This work describes the new Arabic Text-to-speech (TTS) synthesis system. This system based on di-Diphone concatenation synthesizer. The quality of a synthesized speech is improved by analyzing the spectrum features of voice source in various F_0 ranges and timbres in detail and new unites concatenation, diphone include the harakāt ⟨حَرَكَات⟩ (vowel marks; singular: harakah ⟨حَرَكَة⟩). It generates speech synthesis based on analysis and estimation of formant by classifying the voice source into different types. The developed model enhances the quality of the naturalness, and the intelligibility of speech synthesis in various speaking environment.

General Terms
Signal processing, analysis and synthesis speech.

Keywords
Arabic speech synthesis, di-diphone, spectrum analysis, formant, pitch, timbre, di-diphone.

1. INTRODUCTION
With the development of the technology in speech processing, the Arabic speech synthesis system (TTS system) has been made rapid progress during last few years, and has been used in various places successfully. But, the results of it is still far away from the high naturalness compared with humans, being lack of the good algorithm of prosodic processing and new unite of concatenation di-diphone (consonance with vowel integrate). In recent years, with the increasing power of modern computers, therefore, in the last few years, research in speech synthesis has focused mostly on producing speech that sounds more natural or human-like in many languages (English and French). That is not the case for the Arabic language. This is due to the difficulty of the Arabic language in terms of structure and co-articulation [1], with traditional methods so processing station based on the estimated formants trained to improve the new Arabic voice by Optimization of the prosodic.

2. THE PHONETIC SYSTEM OF ARABIC

1.1 Introduction for Arabic language

The Arabic language is spoken throughout the Arab world and is the liturgical language of Islam. This means that Arabic is known widely by all Muslims in the world. Arabic either refers to Standard Arabic or to the many dialectal variations of Arabic. Standard Arabic is the language used by media and the language of Qur'an. Modern Standard Arabic is generally adopted as the common medium of communication through the Arab world today. Dialectal Arabic refers to the dialects derived from Classical Arabic [10]. These dialects differ sometimes which means that it is hard and a challenge for a Lebanese to understand an Algerian and it is worth mentioning there is even a difference within the same country.

Standard Arabic has 34 basic phonemes, of which six are vowels, and 28 are consonants [11]. Several factors affect the pronunciation of phonemes. An example is the position of the phoneme in the syllable as initial, closing, intervocalic, or suffix. The pronunciation of consonants may also be influenced by the interaction (co-articulation) with other phonemes in the same syllable. Among these coarticulation effects are the accentuation and the nasalization. Arabic vowels are affected as well by the adjacent phonemes. Accordingly, each Arabic vowel has at least three allophones, the normal, the accentuated, and the nasalized allophone. In classic Arabic, we can divide the Arabic consonants into three categories with respect to dilution and accentuation [12]. Arabic language has five syllable patterns: CV, CW, CVC, CWC and CCV, where C represents a consonant, V represents a vowel and W represents a long vowel.

1.2 Database construction

The first step in constructing a di-diphone database for Arabic is to determine all possible di-diphone pairs of Arabic. [13]. In reality, additional sound segments and various allophonic variations may in some cases be also included. The basic idea is to define classes of di-diphones, for example: vowel-consonant, consonant- vowel, vowel-vowel, and consonant-consonant.

The syllabic structure of Arabic language is exploited here to simplify the required di-diphones database. The proposed sound segments may be considered as "sub-syllabic" units [10]. For good quality, the di-diphones boundaries are taken from the middle portion of vowels. Because di-diphones need to be clearly articulated various techniques have been proposed to extract them from subjects. One technique uses words within carrier sentences to ensure that the di-diphones are pronounced with acceptable duration and prosody [20] (i.e. consistent). Ideally, the di-diphones should come from a middle syllable of nonsense words so it is fully articulated and minimize the articulatory effects at the start and end of the word [14].

The second step is to record the corpus, this recording made by a native speaker of Arabic standard cardioids microphone with a high quality flat frequency response. The signal was sampled at 16 kHz and 16 bit.

Finally Segmentation and annotation, the database registered must be prepared for the selection method has all the information necessary for its operation. The base is first

segmented into phones, in second step to di-diphones Figure1. This was handmade by the studio diphone software developed by the laboratory TCTS of Mons. A correction on the units to ensure quality was made by the software Praat (Boers and my Weening, 2008).Prosodic analysis performed on the corrected signal to determine the pitch and duration of phone.

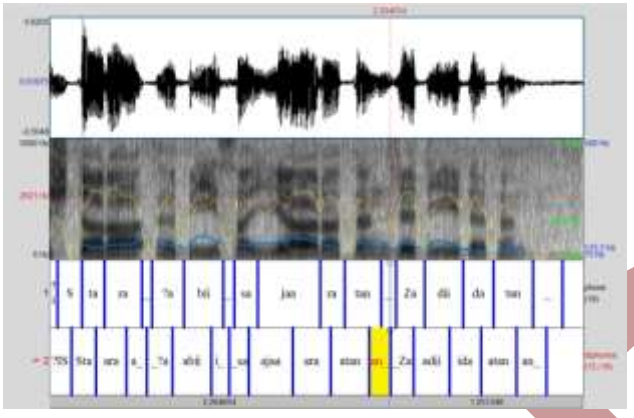

Figure 1. Segmentation of Arabic phone and di-diphone speech this segmentation is used in order to Belding the database

The result of this segmentation provides a new database contain (di-diphones, phonemes and phones).The code SAMPA (Speech Assessment Method Phonetic Alphabet) used for transformation grapheme phoneme.

3. Speech analysis and synthesis

This section will describe the procedures of synchronous analysis and synthesis using TD-PSOLA modifier use for concatenation synthesizer Figure2 presents the block diagram of these two stages.

3.1. Speech analysis

The first step in the speech analysis is to filter the speech signal by a RIF filter (pre-accentuation). The next step is to provide a sequence of pitch-marks and voiced/unvoiced classification for each segment between two consecutive pitch marks. This decision is based on the zero-crossing and the short time energy Figure1. A coefficient of voicement (v/uv) can be computed in order to quantize the periodicity of the signal [15].

3.1.1. Segmentation

The segmentation of a speech signal is used in order to identify the voiced and un-voiced frames. This classification is based on the zero-crossing ratio and the energy value of each signal frame.

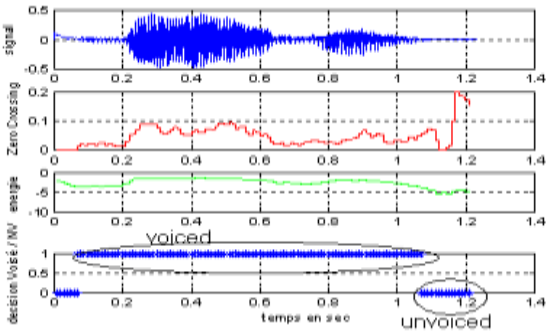

Figure 2. Automatic segmentation of Arabic speech « باب; babun» This segmentation is used in order to identify the voiced and unvoiced frames.

3.1.2. Speech marks

Different procedures of placed $t_a\ i$ are used according to the local features of components of the signal. A previous segmentation of the signal in identical feature zones permits to orient the marking toward the suitable method. Besides results of this segmentation will be necessary for the synthesis stage.

3.1.2.1. Reading marks

The idea of our algorithm is to select pitch marks among local extrema of the speech signal. Given a set of mark candidates which all are negative peaks or all positive peaks:

$$T_a = t_a(i) = t_a(1)...t_a(i).....t_a(N)$$

where t_a is the sample of the peak, and N the number of peaks extracted ([16] explain how these candidates are found).Pitch marks are a subset of points out of T_a, which are spaced by periods of pitch given by the pitch extraction algorithm. The selection can be represented by a sequence of indices:

$$J = j(k) = j(1)..... j(k)..... j(K) \quad (1)$$

With K<N. J has to preserve the chronological order which requires the monotony of j : $j(k) < j(k+1)$.

The sequence of indices along with the corresponding peaks is defined to be the set of pitch marks:

$$T_a = t_a(j(k)) = t_a(j(1))..t_a(j(k))..t_a(j(K)) \quad (2)$$

The determination of j requires a criterion expressing the reliability of two consecutive pitch marks with respect to pitch values previously determined. The local criterion we chose is:

$$d(c(l);c(i)) = |(c(i) - c(l)) - P_a(c(l))| \quad (3)$$

We use the following algorithm for the marking: where l < i. It takes into account the time interval between two marks compared to the pitch period P_a in samples. This criterion returns zero if the two peaks are exactly $P_a(c(l))$ samples away from one another and a positive value if the distance between these peaks is greater or less than the pitch period. The overall criterion is:

$$D = \sum_{K=1}^{K-1} d\left(t_a(j(k)), t_a(j(k+1))\right) - B\left(t_a(j(k+1))\right) \quad (4)$$

Where B is the bonus of selecting an extremum as a pitch mark. In a first time.

$$B(t_a(j(k))) = \delta \left| amplitude(t_a(j(k))) \right| \quad (5)$$

The coefficient δ expresses the compromise between closeness to pitch values and strength of pitch marks. Minimizing D is achieved by using dynamic programming. The Pitch marking results is shown in Figure3.

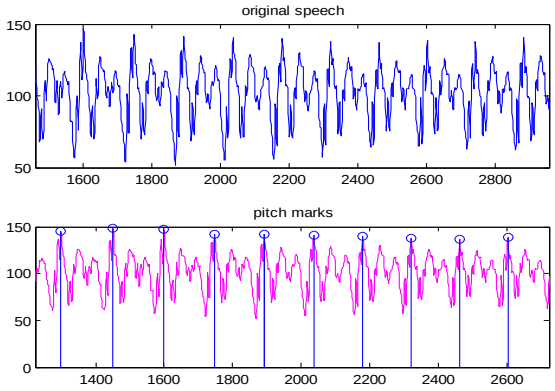

Figure 3. Pitch marks of Arabic speech « akala أكل »

3.1.2.2. Synthesis marks

The OLA synthesis is based on the superposition-Addition of elementary signals $Y_j(k)$, obtained from the $X_i(k)$ placed in the new positions $t_s(j)$. These positions are determined by the height and the length of the synthesis signal. In such synthesis one can modify the temporal scale by a coefficient tscale. The positions $t_s(k-1)$ and the pitch period Pa(k) are supposed to be known we can deduce $t_s(k)$ as [17];

$$t_s(k) = t_s(k-1) + tscale \cdot P_a(n(k))$$
$$n(k+1) = n_s(k) + tscale \quad (6)$$

tscale: coefficient of length modification

In order to increase the pitch, the individual pitch-synchronous frames are extracted, Hanning windowed, moved closer together and then added up. To decrease the pitch, we move the frames further apart. Increasing the pitch will result in a shorter signal, so we also need to duplicate frames if we want to change the pitch while holding the duration constant.

3.2. Synthesis speech

Therefore, given the pitch mark and the synthesis mark of a given frame we use a fast re-sampling method described below to shift the frame precisely where it will appear in the new signal. Let x[n] the original frame, the re-sampled signal is given by A. Oppenheim [18]:

$$x(t) = \sum_{n=-\infty}^{\infty} x[n] \, \mathrm{sinc}\left(\frac{\pi(t - nTs)}{Ts}\right) \quad (7)$$

Where Ts is the sampling period. Calculating the result frame y[m] corresponding to the frame x[n] shifted by a small delay δ amounts to evaluate x (mTs - δ). Therefore, y[m] = x (mTs - δ) i.e:

$$y[m] = \sum_{n=-\infty}^{\infty} x[n] \, \mathrm{sinc}\left(\pi fs\,(mTs - \delta) - nTs\right)$$
$$= \sum_{n=-\infty}^{\infty} x[n] \, \mathrm{sinc}\left(\pi fs\,(m-n)Ts - \delta\right) \quad (8)$$

Where fs is the sampling frequency (1/Ts).Now, by rewriting sinc as $\sin(x)/x$ and by using the following formula:

$$\sin(\pi fs\,(m-n)Ts - \delta) = \cos(\pi fs\delta)\sin(\pi(m-n))$$

But $\cos\pi(m-n) = \pm 1$ *and* $\sin\pi(m-n) = 0$ we get

$$y[n] = \sum_{n=-\infty}^{\infty} x[n] \frac{(-1)^{(m-n+1)} \sin(\pi fs\delta)}{\pi fs\,(m-n)Ts - \delta} \quad (9)$$

As $0 < \delta < Ts$ (resp. $-Ts < \delta < 0$), we define

$\delta = \alpha\, Ts$, where $0 < \alpha < 1$ (resp. $-1 < \alpha < 0$).

Then the synthesized speech is

$$y[n] = \sum_{n=-\infty}^{\infty} (-1)^{(m-n+1)} x[n] \frac{\sin(\alpha\pi)}{\pi} \frac{1}{(m-n)-\alpha} \quad (10)$$

4. Results and Evaluation

4.1. Objective evaluation

To validate our approach we used an objective judgment based on the mean square error and spectral distances. The spectral distances are defined by:

$$d_p = \left[\sum_{k=1}^{K} \left|x_k - y_k\right|^p\right]^{\frac{1}{p}} = \left\|x - y\right\|_p \quad (11)$$

$$d = \frac{1}{16}\left(\sum_{k=1}^{16} \left|x_k - y_k\right|^2\right) \quad (12)$$

These provide to measuring the differences between the prosodic parameters for each natural and synthesized speech and to check the reliability of the system synthesis from a natural speech.

Voix Naturel / Synthétique	a (أكل)	u (أبي)	i (بئر)	a (ذئب)	u (قوتي)	i (بيض)
a (أكل)	0.0366			0.1573		
u (أبي)		0.0897			0.0993	
i (بئر)			0.0130		0.1573	0.1114
a (ذئب)	0.0110			0.0229		
u (قوتي)		0.0725			0.0471	0.0328
i (بيض)			0.0925			0.0913

Table1. The spectral distances of a few Arabic vowels.

This table shows the spectral distances of a few vowels located in the same words spoken by our system and speaker of the natural voice. It can be seen therefore good discrimination between vowels, leading to low squared errors.

4.2. Subjective evaluation

Both listening tests were conducted with 10 adults who have no hearing problem and have a good knowledge of the Arabic language. For both listening tests we prepared listening test programs and a brief introduction was given before the listening test. In the first listening test, each sound was played once in 4 seconds interval and the listeners write the corresponding scripts to the word they heard on the given answer sheet.

In the second listening test, for each listener, we played all 15 sentences together and randomly. Each subject listens to 15 sentences and gives their judgment score using the listening test program by giving a measure of quality. Table2 presents the results of formal listening test medium by MOS (Men opinion score) on a scale of 5.

	Système originale	Notre système	Système Euler	Système Acapela
Le naturel	5	4.43	3.91	4.7
Intelligibilite	5	4.71	4.23	4.8
Qualité globale de la voix	5	4.7	4.5	4.8

Table2. The results of formal listening test averages MOS

From these results, we can conclude that our system is more efficient than the Euler system, but it is not as good as the Acapela system and the original system.

5. Conclusion

In this work, a voice quality conversion algorithm with TD-PSOLA modifier using the new database. The results of perceptual evaluation test indicate that the algorithm can effectively convert modal voice into the desired voice quality. Results of the simulation verify that the quality of the synthesized signal by other Lab. TD-PSOLA with technique depends on the precision of the analysis marking as well as the synthesis marking which must be placed with precision to avoid errors in the phase. Our higher precision algorithm for pitch marking during the synthesis stage increases the signal quality. This gain in accuracy avoids the reduction of deference between original and synthetic signals. We have shown that syllables produce reasonably natural quality speech and durational modeling is crucial for naturalness. We can see this quality from the listening tests and objective evaluation to compare the original and synthetic speech. Perspective, while more research is needed to improve the quality of basic develop dynamic introduce optimizations to choose the best unit in a large database that contains several versions of the units according to the optimization of prosodic parameters.

6. References

[1] Huang, X., A. Acero and H. W. Hon, Spoken Language Processing, Prentice Hall PTR, New Jersey,2001.

[2] Greenwood, A. R "Articulatory Speech Synthesis Using Diphone Units", IEEE international Conference on Acoustics, Speech and Signal Processing, pp. 1635–1638,1997.

[3] Sagisaka, Y., N. Iwahashi and K. Mimura, "ATR v-TALK Speech Synthesis System", Proceedings of the ICSLP, Vol. 1, pp. 483–486,1992.

[4] Black, A. W. and P. Taylor, "CHATR: A Generic Speech Synthesis System", Proceedings of the International Conference on Computational Linguistics, Vol. 2, pp. 983–986,1994.

[5] Childers, D.G. «Glottal source modeling for voice conversion». Speech communication, 16(2): 127-138, 1995.

[6] Childers, D.G., and Lee, C.K. «Vocal quality factors: Analysis, synthesis, and perception». Journal of the Acoustical Society of America, 1991.

[7] Acero A. «Source-filter Models for Time-Scale Pitch-Scale Modification of Speech». IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle, USA, pp.881-884. May, 1998.

[8] Dutoit, T., Pagel, V., Pierret, N., Bataille, and F. & van der Vrecken, O. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use,1996.

[9] Moulines, E., and Charpentier, F. «Pitch-Synchronous Waveform Processing Techniques for TTS Synthesis».Speech communication Vol 9, pp453-467, 1990.

[10] Alghmadi, M., "KACST Arabic Phonetic Database", the Fifteenth International Congress of Phonetics Science, Barcelona, pp 3109-3112. (2003)

[11] Assaf, M., "A Prototype of an Arabic Diphone Speech Synthesizer in Festival," Master Thesis, Department of Linguistics and Philology, Uppsala University, 2005.

[12] Al-Zabibi, M., "An Acoustic–Phonetic Approach in Automatic Arabic Speech Recognition," The British Library in Association with UMI,1990.

[13] Ibraheem A "Al-Aswat Al-Arabia", Arabic title, Anglo-Egyptian Publisher, Egypt,1990.

[14] Maria M., "A Prototype of an Arabic Diphone Speech Synthesizer in Festival", Master Thesis in Computational Linguistics, Uppsala university, 2004.

[15] Laprie, Y. and Colotte, V. "Automatic pitch marking for speech transformations via TD-PSOLA". In IX European Signal Processing Conference, Rhodes, Greece, 1998.

[16] Mower, L., Boeffard, O., Cherbonnel, B. "An algorithm of speech synthesis high-quality" Proceeding of a Seminar SFA/GCP, pp 104-107,1991.

[17] Oppenheim A. V. and Schafer, W. R. Digital Signal Processing. Prentice-Hall, Inc, 1975 .

[18] Oppenheim, A.V. and Schafer R.W. Digital Signal Processing. Prentice-Hall, Inc., New York,1975.

[19] Walker, J., Murphy, P. "A review of glottal waveform analysis. In: Progress in Nonlinear Speech Processing, 2007.

[20] Domino, G., Grocholewski, S., Wagner, A. & Szymański, M. "Prosody Annotation for Corpus Based Speech Synthesis". In: Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology. Auckland, New Zealand, pp. 460-465, 2006.

[21] Maria M., "A Prototype of an Arabic Diphone Speech Synthesizer in Festival",Master Thesis in Computational Linguistics, Uppsala university, 2004.

[22] Kraft V., Portele T. "Quality Evaluation of Five German Speech Synthesis Systems" Acta Acustica 3, pp. 351-365.1995.