An Automatic Mbrola Tool for High Quality Arabic Speech Synthesis

Abdelkader Chabchoub
Signal Processing Laboratory,
Science Faculty of Tunis,
1060 Tunisia

Adnan Cherif
Signal Processing Laboratory
Science Faculty of Tunis
1060 Tunisia

ABSTRACT
This work describes the Arabic Text-to-speech (TTS) synthesis system. This system uses an automatic tool based on Diphone concatenation with MBROLA 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. 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, Diphone, spectrum analysis, formant, pitch, timbre, MBROLA, Inverse filtering.

1. INTRODUCTION
Speech synthesis is now a technology that enables computers to talk and assist people in learning languages. While existing synthesis techniques produce speech that is intelligible, few people would claim that listening to computer speech is natural or expressive. 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 for MBROLA synthesizer [2].

This paper is organized as follows. In section 2, the morphological model of the Arabic language will be presented with in particular the concepts of word. Section 3, presents a list of phonemes and the corresponding acoustic parameters of each phoneme (duration and F_0). These values are inputted into the Prosodic parameter modification module. This will optimize the parameters. The MBROLA synthesizer signals from the Diphone database generates wave files. Section 4, describes the method of formant extraction [3], the construction of the inverse filter and the re-synthesis voice generation.

2. THE PHONETIC SYSTEM OF ARABIC
Arabic is a Semitic language and it is one of the oldest languages in the world. It is the fifth widely used language nowadays [4]. Although Arabic is currently one of the most widely spoken languages in the world there has been relatively few speech synthesis researches on Arabic compared to other languages.

Standard Arabic has 34 basic phonemes, of which six are vowels, and 28 are consonants [5]. 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 co articulation 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 [6]. 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. In the classical Arabic the following rules determine the case of a general vowel with respect to its predecessor/successor consonants within a syllable such as illustrated next:

- A vowel after a Context Dependent Consonants consonant follows context dependent rules as outlined
- In CV and CW syllables, a vowel after an Always Diluted Consonants consonant should be diluted.
- A vowel after an Always Accentuated Consonants consonant should be accentuated.
- The consonant #sukun# in reading must stand it with Short pause after sukun. The #sukun# is always preceded by vowels CVC.

3. TEXT-TO-SPEECH SYNTHESIZER
General purpose state-of-art diphone TTS systems consist of an NLP (natural language processing) module which converts the input text into a list of phonemes and the corresponding parameters for each phoneme, and of a DSP module which converts the output of the NLP module into a speech signal. The decision was made in this thesis to bypass the implementation of the traditional NLP and DSP module by using a MBROLA PHO (a list of tuples of phoneme label, duration in milliseconds, and an optional series of pairs of pitch position in percent of the segment duration and F0 value in Hz) [7] file as an input and using an external MBROLA binary as the DSP synthesizer.

A conventional Text to Speech (TTS) synthesis architecture has two main components: the Natural Language Processing Component (NLP) and the Digital Signal Processing Component (DSP). In this method of synthesis, information from the annotated speech corpus takes the place of the entire NLP front-end of the TTS system. The main tasks of the NLP front-end are replaced fairly straight forwardly as follows:

1. Phonetisation model: replaced by a phoneme inventory based validation module designed for phonemically annotated corpora, as in the present case; the module presupposes forced alignment pre-processing to provide phoneme-level annotation in the case of orthographic, syllable-sized etc. annotations.

2. Duration model: from the time-stamps of the annotation (details dependent on annotation format).

3. Pitch model: pitch extraction algorithm over the given label time domains.

The modules are cascaded in the order Phonetisation, Duration and Pitch. The input is a pair of a speech signal file and a time-aligned phonemic annotation, followed by phoneme validation, followed by duration extraction, followed by pitch extraction, by integration of the phoneme labels, durations and pitch positions and values into the synthesizer interface format (MBROLA PHO format) and finally generating of re-synthesis model to improve the voice quality modification with inverse filter based formants. The main data flow steps are shown in Figure 1.

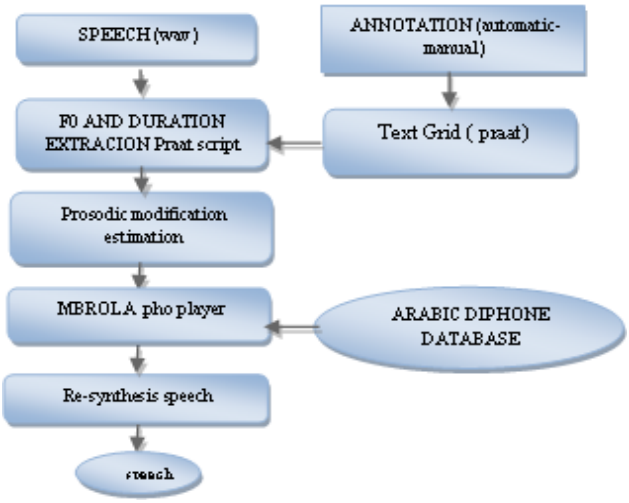

Fig 1: Re-synthesis implementation algorithm.

3.1 Praat pitch extraction

The extraction of pitch is the next step. Copying the phonemes, the durations of the phonemes from the annotation file and measuring the pitch values from the original recording of a human utterance allows best case speech synthesis. To extract pitch from the recordings a Praat script called max_pitch was implemented as in [8] "This script goes through Sound and TextGrid files in a directory, opens each pair of Sound and TextGrid, calculates the pitch maximum of each labeled interval, and saves results to a text file" [9]. The implementation of this script caused another problem and some modifications to the script were made.

The inputs to this script are:

1. WAV files,
2. TextGrid annotation files.

The Praat pitch extraction file produces one TXT file with the pitch values of all the phonemes in the files in the directory. The output "pitchresults.txt" file contains the following information:

1. File names of the files in the directory,
2. Labels,
3. Maximum pitch values of the labeled intervals in Hz.

The pitch results file for one file in a directory is shown in the next example: automatics extract Praat script .pho "أكل الولد ". Equivalent to "Akola el waladou"

_ 78					
a 53	28 167	57 169	85 169		
k 42	36 166	71 160			
k 149	60 152	70 142	81 132	91 130	
a 47	32 128	64 129	96 129		
l 84	18 132	36 138	54 142	71 146	89 149
a 66	23 153	45 154	68 154	91 153	
l 77	19 152	39 150	58 149	78 148	97 148
w 80	19 146	38 145	56 145	75 149	94 153
a 84	18 152	36 150	54 147	71 143	89 141
l 66	23 128	45 121	68 117	91 110	
a 59	25 106	51 105	76 104		
d 96					
u 109	14 111	28 112	41 114	55 116	69 119
_ 6					

3.2 Inclusion of pitch values into MBROLA PHO file

Although at the beginning the extracted pitch values with the use of max_pitch Praat script were put into the MBROLA PHO an automatic inclusion of pitch values into MBROLA PHO files was developed. This procedure takes the pitch results file generated by the modified max_pitch Praat script as an input. As described above the pitch results file contains the names of all the WAV/TextGrid files (the WAV and TextGrid file names are identical) in a directory, labels and the maximum pitch for segments of these files.

The problem was solved by dividing the pitch results file into separate PITCH files in which there are only filenames, labels and pitch values for each file in a directory. Therefore, the PITCH files got the same length as TextGrid files. Similarly, PITCH files and MBROLA PHO files were almost identical. The inclusion script takes:

1. from MRBOLA PHO files with monotone:
   - phonemes,
   - duration of these phonemes,
   - Pitch position equal 50.

2. from PITCH files: pitch values.

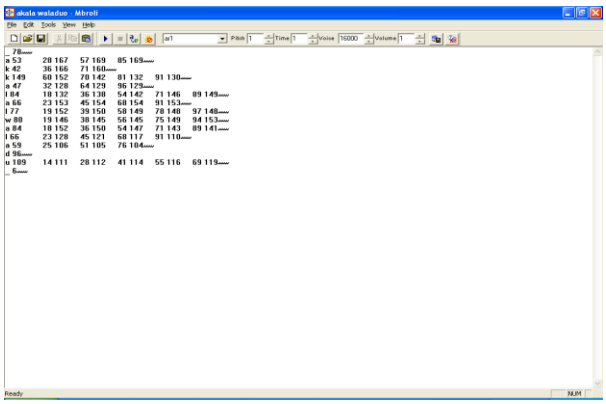

Fig 2: Mbrola tools "Akaka waladu".أكل الولد

Figure 3 shows a waveform and a pitch contour of an original and synthetic speech with our system.

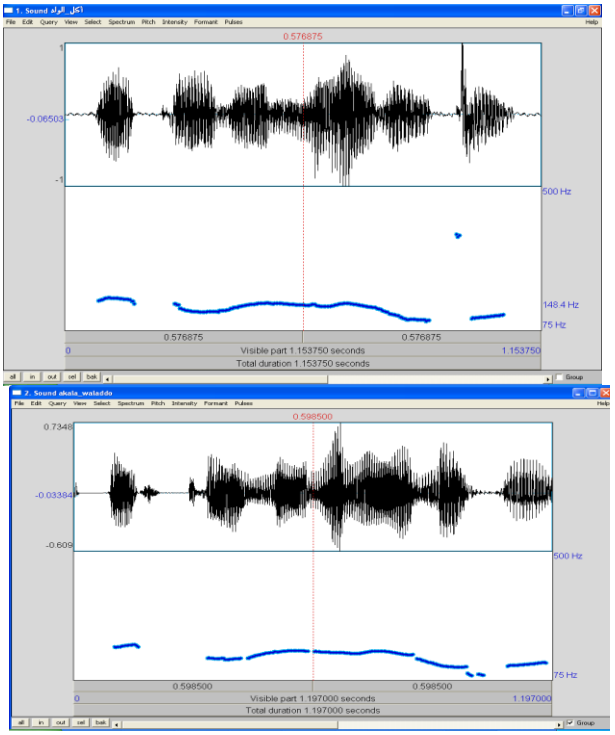

Fig 3: A waveform and a pitch contour of a human utterance and its synthesized equivalent using automatic tools.

4. INVERSE FILTER BASED FORMANTS

4.1 Formant estimation

Formant estimation is the first step in the generating of MS model. As we know, the transfer function of a second-order digital resonator can be written as:

$$H_i(z) = \prod_{i=1}^{k} \frac{1}{\left(1 - z_i / z\right)\left(1 - z_i^* / z\right)} = \frac{1}{1 - \left(z_i + z_i^*\right) z^{-1} + \left(z_i z_i^*\right) z^{-2}} \qquad (1)$$

Thus the transfer function of vocal tract can be written as:

$$H(z) = \prod_{i=1}^{k} H_i(z) \qquad (2)$$

Where z_i is defined by the formant frequency and the bandwidth. For formant estimation, methods based on linear prediction analysis (LPC) and cepstrum analysis have received considerable attention as in Figure 4. In last few years Welling's method [9] and Inverse-Filter Control (IFC) method have been presented and got good result in formant estimation. The following is the method used in this work.

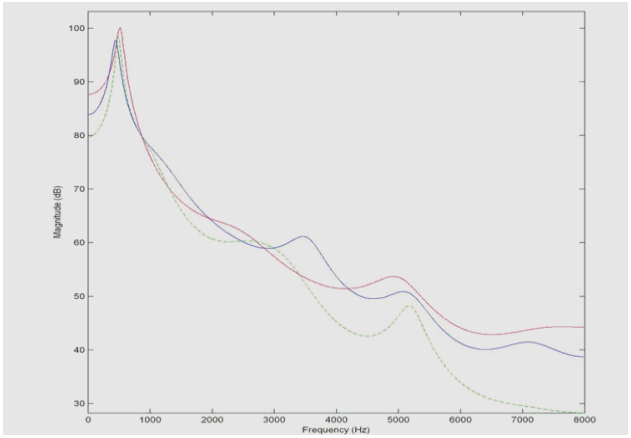

Fig 4: LPC formant for original and synthetic speech

In this method, the short-time power spectrum is decomposed into segments, each of which is modeled by a digital second-order resonator. The segment boundaries are optimized by dynamic programming. An advantage of the method is that an explicit smoothing of the formant frequencies along the time axis does not seem to be necessary. The whole frequency range is divided into K segments with boundaries.

$$\omega_0 = 0, \ldots\ldots, \omega_k, \ldots\ldots, \omega_k = \pi$$

Thus the corresponding second order resonator can be defined as:

$$A_k(e^{j\omega}) = 1 - \alpha_k e^{j\omega} - \beta_k e^{j2\omega} \qquad (3)$$

$$f_k = \arccos\left[-\frac{\alpha_k(1-\beta_k)}{4\beta_k}\right] \qquad (4)$$

$$b_k = \log(-\beta_k) \qquad (5)$$

Where α_k and β_k are the real-valued prediction coefficients. If these prediction coefficients are given, the formant frequency f_k and bandwidth b_k can be obtained according to the formula (4) and (5), and the value of the prediction error is given by.

$$E(\omega_{k-1}, \omega_k / \alpha_k, \beta_k) = \frac{1}{\pi} \int_{k=1}^{k} \left|S_k(e^{jw})\right|^2 \left|A_k(e^{jw})\right|^2 dw \qquad (6)$$

The error of all the segments is

$$E = \sum E_{\min}(\omega_{k-1}, \omega_k) \qquad (7)$$

Formant parameters are then obtained according to formula (4) and (5), if segment boundaries are given.

4.2 Source Estimation

There are lots of methods used for source estimation, such as ARMA analysis method [10], Sum-of-Exponentials method [11], inverse-filtering method, least squares glottal inverse filtering [12] and joint estimation of an AR system with a linear input model.

Voice source can be obtained by matching source model with the inverse-filter result, and we use LF (Liljencrants/Fant) model [13] illustrated in Figure 5. The parameter Top denotes the instant of the maximum glottal flow. T0 is the fundamental period. Tc denotes the ending of the return phase. Ta is the effective duration of the return phase. Ta determines the spectral tilt of the glottal source [14]. The increase of intensity AV will cause the increase of low frequency harmonic components while the decrease of lowest value Ee (shown in Figure 5) will bring the decrease of high frequency harmonic components.

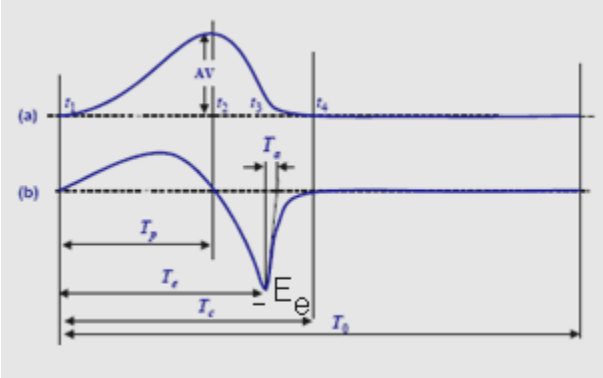

Fig 5: LF model (a) Glottal pulse flow (b) its derivative

The estimation of LF parameter contains two steps: initial parameters estimation and non-linear optimization. A good initial parametric model is critical for nonlinear optimize and finding the global optimum parameters. It's not difficult to find that T_e is one of the easiest acquirable parameters, which can be got from the time when the different voice source signal is minimized. E_e is the value of signal at time Te. T_p can be estimated from the first zero-crossing before T_e, and T_c can be estimated from the first point at which the signal value is below a given threshold after Te. Similarly to T_c, T_0 can be estimated from the first point at which the signal value is below a given threshold before T_p, and the point is limited by open quotient. T_a is the most difficult parameter to get, and it has been discussed in many papers. In our work, we assumes the following relationships between T_a and T_c, T_e,

$$T_a = \frac{2}{3}(T_c - T_e) \qquad (10)$$

In the paper, the non-linear optimized method, which contains dynamic time warping [15], nominalization [16], deciding of dropping points and filter based least squares error, was used to optimize initial parameter estimation. It decreased the error to ±6%, and could be used in LF generation.

5. RESULTS AND EVALUATION

5.1 Objective evaluation

A first look at the results of the system showed that although there were similarities between the natural and synthetic versions, there is a considerable resemblance between the natural and synthetic F_0 contours. Only a few minor differences can be observed, since the F_0 values were extracted only once every 10 ms. Also note the halved F_0 in the creaky parts of the synthetic versions which successfully simulated creak. Similarly for the formant there is small difference with the estimation algorithm. This can be seen in Figure 6.

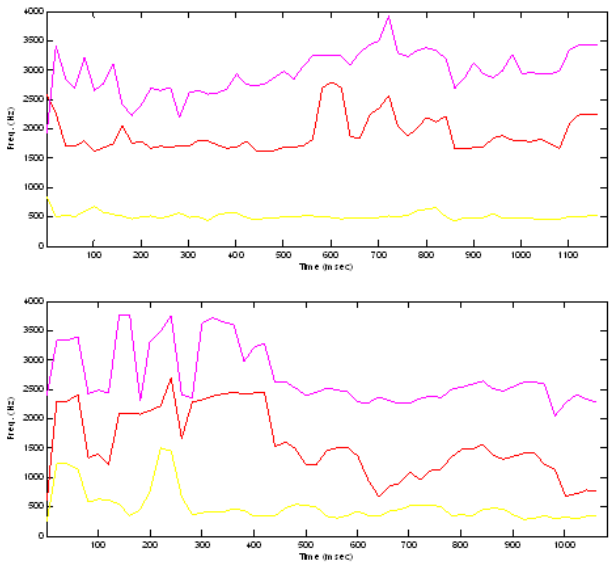

Fig 6: Neutral, and synthesis speech, formants (F1, F2, F3).For Sentence, "أكل الولد"

5.2 Subjective evaluation

Both listening tests were conducted by four Arab adults who are native speakers of the language (4 males). All listeners are 25-35 years old in age, born and raised in the Arab countries. 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 as follows: (5 – Excellent, 4 - Good, 1– Bad). They evaluated the system by considering the naturalness aspect. Each listener did the listening test fifteen times and we took the last ten results considering the first five tests as training. After collecting all listeners' response, we calculated the average values and we found the following results. In the first listening test, the average correct-rate for original and analysis-synthesis sounds were 98% and that of rule-based synthesized sounds was 90%. We found the synthesized words to be very intelligible.

6. CONCLUSION

This paper has introduced a newly high quality Arabic speech synthesis. It is based on estimation of parameters prosodic for MBROLA method and inverse filter based on formants estimation. 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.

7. REFERENCES

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

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

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

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

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

[6] Muhammad, A.1990. "Alaswaat Alaghawaiyah", Daar Alfalah, Jordan, (in Arabic).

[7] Demenko, G., Grocholewski, S., Wagner, A. & Szymański, M. 2006. "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.

[8] Boersma, P. & Weenink, D. 2005. Praat. Doing phonetics by computer. [Computer program]. Version 4.3.04 Retrieved March 31, 2005 from http://www.praat.org/

[9] Bachan, J. & Gibbon, D.2006. "Close Copy Speech Synthesis for Speech Perception Testing"In: Investigationes Linguisticae, vol. 13, pp. 9--24.

[10] L. Welling, L., Ney, H.1998. "Formant Estimation for Speech Recognition", IEEE Trans. On Speech and Audio Processing, Vol.6, No.1.

[11] Fujisaki, H. 1996. "Recent Research towards Advanced Man-Machine Interface through Spoken Language", Elsevier Science.

[12] Krishnamurthy, A.K.1992. "Glottal Source Estimation Using a Sum-of-Exponentials Model", IEEE Trans. On Signal Processing, Vol. 40, No. 3, March 1992.

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

[14] Fant G. 1986. "Glottal flow: models and interaction", Journal of Phonetics, 14, 393-399.

[15] Milenkovic, P. 1986. "Glottal Inverse Filtering by Joint Estimation of an AR System with a Linear Input Model", *IEEE Trans. On Acoustics, Speech, and Signal Processing*, Vol.ASSP-34, No.1.

[16] Jianhua, T., Yongguo, K. 2004 ." multi-source based acoustic model for speech synthesis", 5th ISCA Speech Synthesis Workshop Pittsburgh, PA, USA, 14-16.