High Quality Arabic Concatenative Speech Synthesis

Abdelkader Chabchoub and Adnan Cherif

Signal Processing Laboratory, Science Faculty of Tunis, 1060 Tunisia
achabchoub@yahoo.fr, adnen2fr@yahoo.fr

ABSTRACT

This paper describes the implementation of TD-PSOLA tools to improve the quality of the Arabic Text-to-speech (TTS) system. This system based on Diphone concatenation with TD-PSOLA modifier synthesizer. This paper describes techniques to improve the precision of prosodic modifications in the Arabic speech synthesis using the TD-PSOLA (Time Domain Pitch Synchronous Overlap-Add) method. This approach is based on the decomposition of the signal into overlapping frames synchronized with the pitch period. The main objective is to preserve the consistency and accuracy of the pitch marks after prosodic modifications of the speech signal and diphone with vowel integrated database adjustment and optimisation.

KEYWORDS

Speech processing and synthesis, Arabic speech, prosody, diphone, spectrum analysis, pitch mark, timbre, TD-PSOLA.

1. INTRODUCTION

The synthetic voice that imitates human speech from plain text is not a trivial task, since this generally requires great knowledge about the real world, the language, the context where the text comes from, a deep understanding of the semantics of the text content and the relations that underlie all these information. However, many research and commercial speech synthesis systems developed have contributed to our understanding of all these phenomena, and have been successful in various respective ways for many applications such as in human-machine interaction, hands and eyes free access of information, interactive voice response systems.

There have been three major approaches to speech synthesis: articulatory, formant and concatenative [1] [2] [3] [4]. Articulatory synthesis tries to model the human articulatory system, i.e. vocal cords, vocal tract, etc. Formant synthesis employs some set of rules to synthesize speech using the formants that are the resonance frequencies of the vocal tract[19]. Since the formants constitute the main frequencies that make sounds distinct, speech is synthesized using these estimated frequencies. Several speech synthesis systems were developed such as vocoders and LPC synthesizers [5][6], but most of them did not reproduce high quality of synthetic speech when compared with that of PSOLA based systems [7] such as MBROLA synthesizers[8]. Especially TD-PSOLA method (Time Domain Pitch Synchronous Overlap-Add) is the most efficient method to produce criteria of satisfaction speech [9] and is one of the most popular concatenation synthesis techniques nowadays. LP-PSOLA (Linear Predictive PSOLA) and FD-

DOI : 10.5121/sipij.2011.2403

PSOLA (Frequency Domain PSOLA), though able to produce equivalent result, require much more computational power. The first step of the TD-PSOLA is to perform a pitch detection algorithm and to generate pitch marks through overlapping windowed speech. To synthesize speech, the Short Time signals (ST signals) are simply overlapped and added with desired spacing of the ST-signals.

2. THE ARABIC DATABASE

2.1. 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.

Table 1. Arabic Consonants and Vowels and their phonetic notations SAMPA

Graphemes	Code SAMPA	Graphemes	Code SAMPA	Graphemes	Code SAMPA	Graphemes	Code SAMPA
ء		ر	**r**	غ	**G**	ي	**j**
ب	**b**	ز	**z**	ف	**f**	◌َ	**a**
ت	**t**	س	**s**	ق	**q**	◌ُ	**u**
ث	**T**	ش	**S**	ك	**K**	◌ِ	**i**
ج	**Z**	ص	**s.**	ل	**l**	◌ً	**an**
ح	**X**	ض	**d.**	م	**m**	◌ٌ	**un**
خ	**X**	ط	**t.**	ن	**n**	◌ٍ	**in**
د	**d**	ظ	**z.**	ه	**h**		
ذ	**D**	ع	**H**	و	**w**		

2.2. Database construction

The first step in constructing a diphone database for Arabic is to determine all possible diphone pairs of Arabic. In general, the typical diphone size is the square of the phone number for any language [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 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 diphones database. The proposed sound segments may be considered as "sub-syllabic" units [10]. For good quality, the diphones boundaries are taken from the middle portion of vowels. Because 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 diphones are pronounced with acceptable duration and prosody[20] (i.e. consistent). Ideally, the 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 diphones. 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.

The result of this segmentation provides a significant reduction of unites, all units not exceeding 400 (diphones, phonemes and phones), comparisons with other basis developed by other laboratories; Chanfour CENT laboratory Rabat Faculty of Sciences, a database of diphones S. Baloul thesis LIUM Mans France and a base of diphones by Noufel Tounsi, laboratory TCTS Mons.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 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

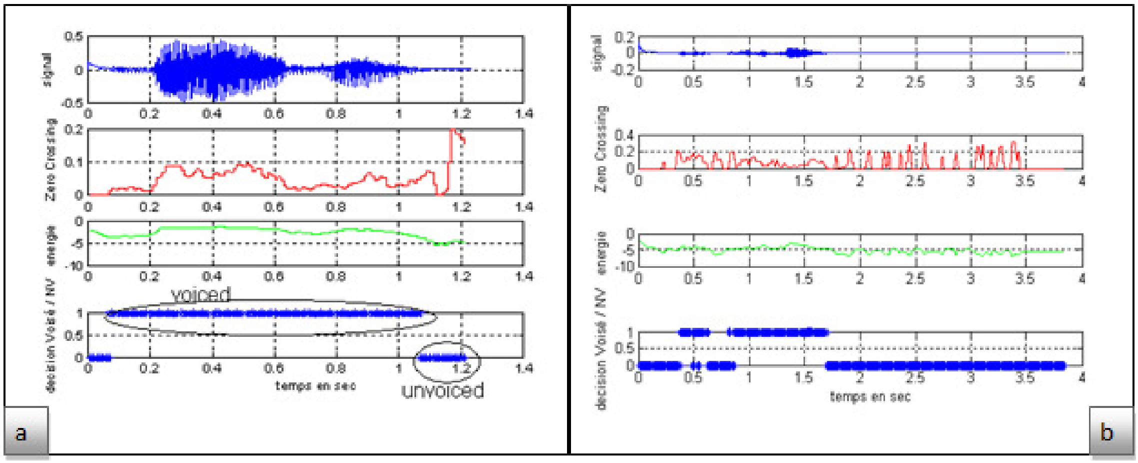


Figure 1. Automatic segmentation of Arabic speech (a) « ;بابbabun» (b) «شمسchamsun» This segmentation is used in order to identify the voiced and unvoiced frames.

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 $t_a(i)$is the sample of the ith 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:

$$J = [j(k)] = j(1)......j(k)......j(K) \qquad (1)$$

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:

$$T_a = [t_a(j(k))] = t_a(j(1))\dots t_a(j(k))\dots t_a(j(K)) \qquad (2)$$

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:

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

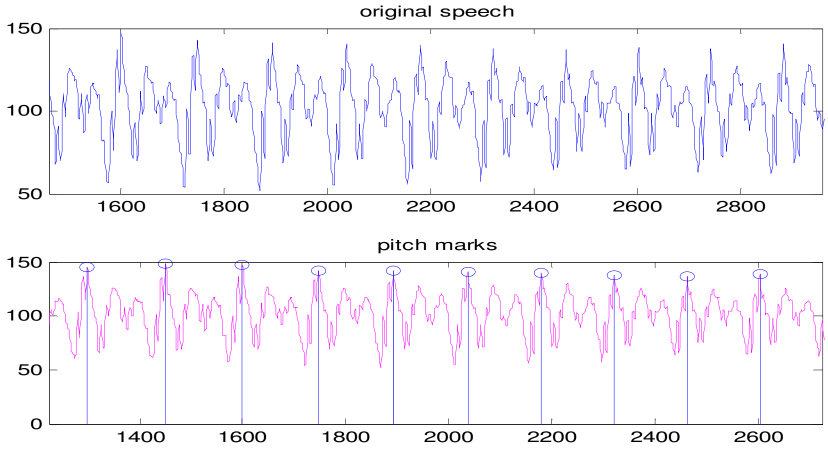
We use the following algorithm for the marking: where l < i. It takes into account the time interval between two marks compared to the pitch period P_a in samples. This criterion returns zero if the two peaks are exactly $P_a(c(l))$ samples away from one another and a positive value if the distance between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

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

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

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

The coefficient δ expresses the compromise between closeness to pitch values and strength of pitch marks. Minimising D is achieved by using dynamic programming. The Pitch marking results is shown in Figure2.



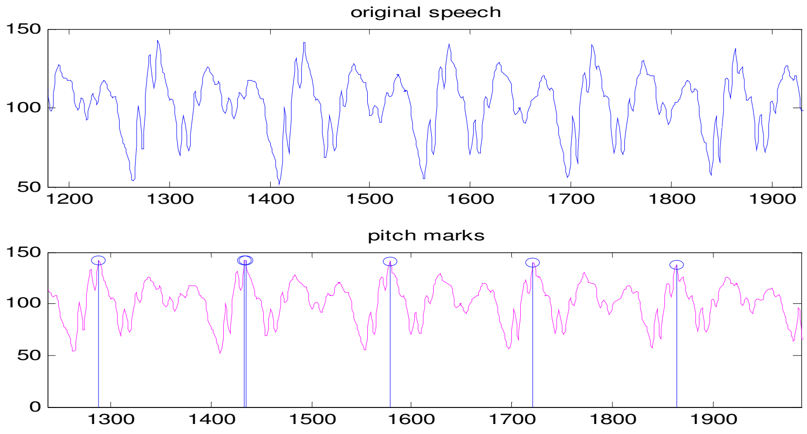


Figure 2. Pitch marks of Arabic speech (a) « ;بابbabun» (b) « akala أكل »

3.1.2.2. Synthesis marks

The OLA synthesis is based on the superposition-Addition of elementary signals $Y_j(n)$, obtained from the $X_i(n)$ 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 $P_a(k)$ are supposed to be known we can deduce $t_s(k)$ as [17];

$$\begin{aligned} t_s(k) &= t_s(k-1) + tscale \cdot P_a(n(k)) \\ n(k+1) &= n_s(k) + tscale \end{aligned} \qquad (6)$$

tscale: coefficient of length modification

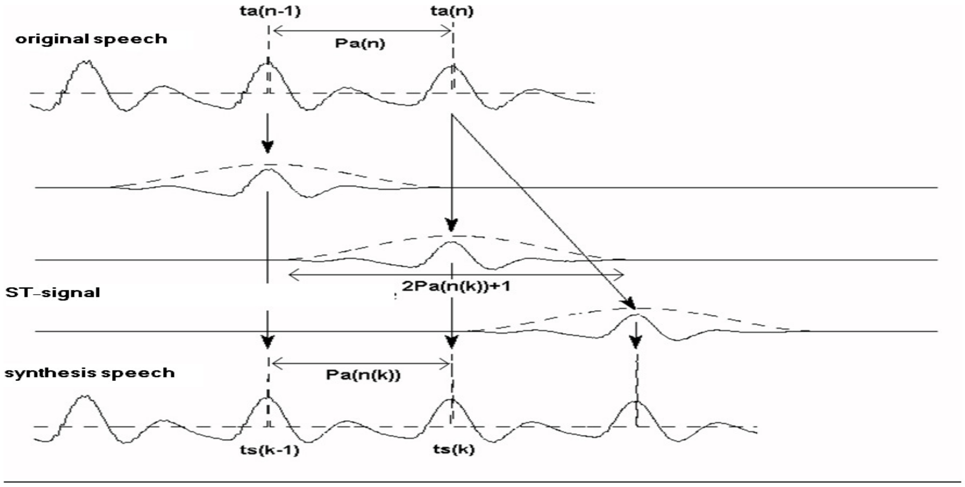


Figure 3. TD-PSOLA for pitch (F_0) modification. In order to increase the pitch, the individual pitch-synchronous frames are extracted, Hanning windowed, moved closer together and then added up. To decrease the pitch, we move the frames further apart. Increasing the pitch will result in a shorter signal, 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]:

$$x(t) = \sum_{n=-\infty}^{\infty} x[n] \sin c\left(\frac{\pi(t - nTs)}{Ts}\right) \qquad (7)$$

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:

$$\begin{aligned} y[m] &= \sum_{n=-\infty}^{\infty} x[n] \sin c\left(\pi fs\left[(mTs - \delta) - nTs\right]\right) \\ &= \sum_{n=-\infty}^{\infty} x[n] \sin c\left(\pi fs\left[(m-n)Ts - \delta\right]\right) \end{aligned} \qquad (8)$$

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

$\sin(\pi fs\left[(m-n)Ts - \delta\right] = \cos(\pi fs\,\delta)\sin(\pi(m-n))$ But $\cos\pi(m-n) = \pm 1$ *and* $\sin\pi(m-n) = 0$ we get

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

As 0 < δ < Ts (resp. -Ts < δ < 0), we define

δ = α Ts, where 0 < α < 1 (resp. -1 < α < 0).

Then the synthesized speech is

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

And it is shown in Figure6

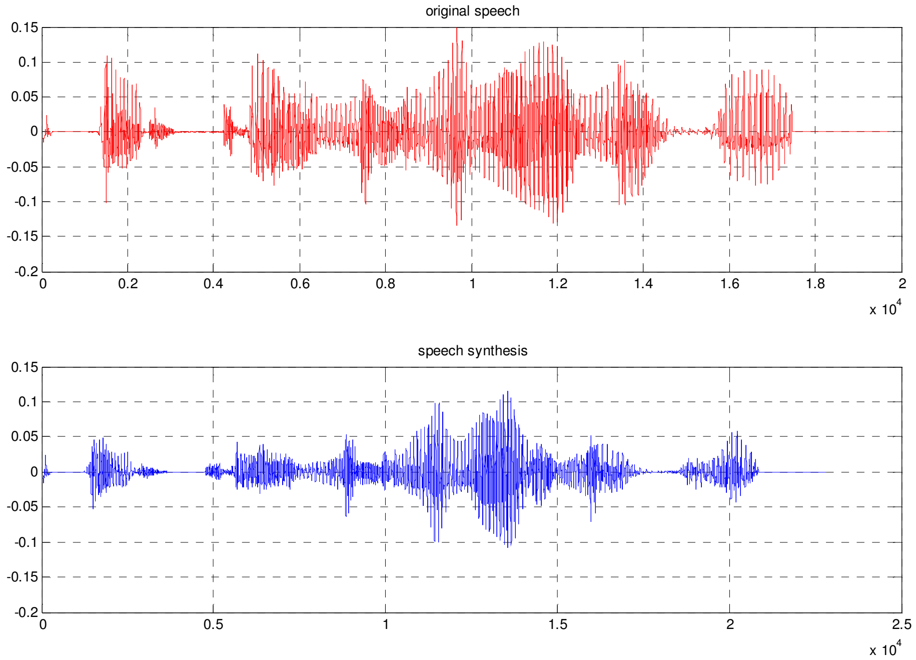


Figure 4. A waveform of a human utterance and its synthesized equivalent using TD-PSOLA tools.« Akala أكل»

4. Results And Evaluation

Two types of tests were applied two evaluate the speech of the developed system regarding the intelligibility and the naturalness aspects. The first test which measures the intelligibility is the Diagnostic Rhyme Test (DRT). In this test, twenty pairs of words that differ only in a single consonant are uttered and the listeners are asked to mark on an answer sheet which word of each pair of the words they think is correct [21]. In the second evaluation test, which is Categorical Estimation (CE), the listeners were asked a few questions about several attributes such as the speed, the pronunciation, the stress, etc.[22] of the speech and they were asked to rank the voice quality using a five level scale. The test group consisted of sixteen persons and the previously mentioned two tests were repeated twice to see whether or not the test results will increase by the learning effect which means that the listeners may become accustomed to the synthesized speech they hear and they understand it better after every listening session . The following tables and charts illustrate the results of these tests.

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 figure5.

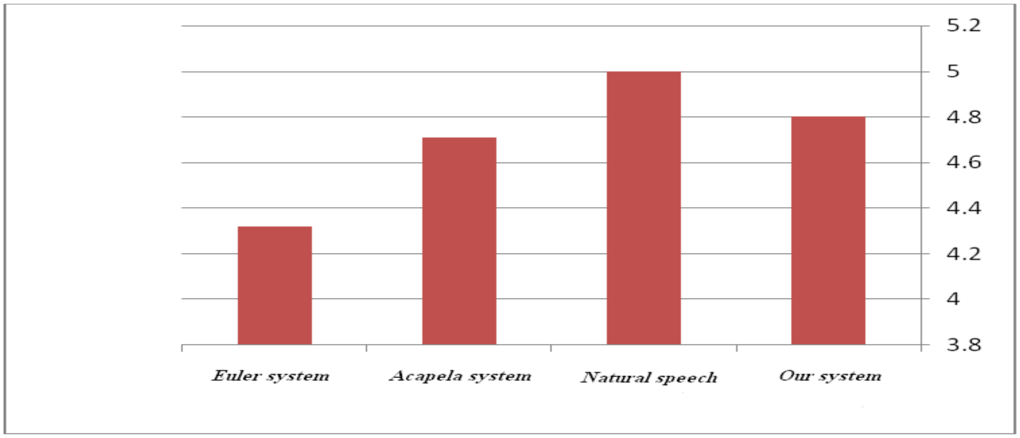


Figure5. Average scores for the first test (system Euler, our system, natural speech and Acapela system. for the intelligibility of speech

5. CONCLUSION

In this work, a voice quality conversion algorithm with TD-PSOLA modifier was implemented and tested under Matlab environment using our 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 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, with a significant reduction in numbers of units of the total base developed. We can see this quality from the listening tests and objective evaluation to compare the original and synthetic speech.

REFERENCES

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

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

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

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

[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. (1996) The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use.

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

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

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

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

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

[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. (1998) "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. (1991) "An algorithm of speech synthesis high-quality" Proceeding of a Seminar SFA/GCP, pp 104-107.

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

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

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

[20] 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.

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

[22] Kraft V., Portele T.,(1995) "Quality Evaluation of Five German Speech Synthesis Systems" Acta Acustica 3, pp. 351-365.