An Evaluation of Techniques Based on HMM Speech Synthesis for Using in HTS-ARAB-TALK

M. K. Krichi and A. Cherif
Department of Physics, FST-Faculty of Sciences de Tunis, Campus Universities 2092 - El Manar Tunis, Tunisia
Email: {Krichi_moha, adnen2fr}@yahoo.fr

***Abstract*—This work aims to find the most effective method for natural and good sound quality, after a comparative evaluation, the best method approved by this evaluation is used in our HTS_ARAB_TALK system. HTS is a system speech synthesis based on HMM, which is a new technique relative to other synthesis techniques. Several versions of HMMs are developed, with varying contextual information, algorithms for estimating the parameters of the source-filter synthesis model and extract the coefficients aperiodicity if the STRAIGHT vocoder is used to extract the F0 and obtain the spectrum and autoregressive HMM model. These methods are compared, in a perceptive test, to the naturalness of speech. The evaluation shows that the use of STRAIGHT and MATLAB with HTS significantly improves synthesis naturalness compared to the state of the art.**

***Index Terms*—hidden markov MODEL, autoregressive HMM, speech synthesis, Arabic language, HTS, HTS_ARAB_TALK**

I. Introduction

Since speech is obviously one of the most important ways for human communication, there have been a great number of efforts to integrate speech into human-computer communication environments. Speech synthesis is a technique for generating speech signal from arbitrarily given text (or other) in order to transmit information from a machine to a person by voice. The first speech synthesis systems have a sound quality and naturalness speech problem, but systems are improving nowadays. This fact makes speech synthesis an important field for investigation and improvement for the major languages including Arabic. The progress of speech processing and the development of human-machine interactions are unimaginable: a machine able to analyze, detect and produce. Producing a speech is improved in the last decades but not in all languages. Arabic is the fourth most spoken language in our world with more than 442 million speaker spread in 23 countries as an official language [1]. Furthermore it carries a religious value for more than 1.6 billion Muslim according to [2]. The number of blinds in the Arab World is around 5 million living in a population around 340 million people [3]. Very few studies have been conducted to characterize the voice synthesis of the Arabic language. So, it's an important issue to build Arabic speech synthesis which is reliable, intelligent and user friendly system to give those people a chance to use the technologies like text messages, emails, and web sites using their native language. There have been four major approaches to speech synthesis: articulatory, formant and concatenative and statistic synthesis. Articulatory synthesis tries to model the human articulatory system, i.e. vocal cords, vocal tract, etc. Formant speech synthesizers generate the speech signal entirely from rules on the acoustic parameters, which are derived by human experts from speech data. Since the formants constitute the main frequencies that make sounds distinct, speech is synthesized using these estimated frequencies. Several speech synthesis systems were developed like as vocoder and LPC synthesizers, and PSOLA based systems such as MBROLA synthesizers in [4]. But most of them did not reproduce high quality of synthetic speech when compared with that HMM-Based Speech Synthesis which is the most efficient method able to produce criteria of satisfaction speech and is one of the most popular statistic synthesis techniques nowadays. Given the good performance achieved, in speech, by HMM-based approaches, we decided to explore the potential of HMMs for improving Arabic speech synthesis naturalness.

This paper presents the developed methods and the results of a perceptive evaluation assessing the intelligibility, naturalness, sound quality and pronunciation of the speech synthesized. The paper is organized as follows: different methods based on HMM for speech synthesis presented in Section 2; Section 3 describes the Arabic speech data; Results and evaluations are described in Section 4; finally Section 5 concludes the paper and mentions future works.

II. HMM-Based Speech Synthesis

The major purpose in speech recognition is to find the spoken words in the speech signal. From the feature

Manuscript received July 10, 2014; revised March 30, 2015.

doi: 10.12720/ijsps.4.2.133-138

vectors using the Viterbi algorithm the most probable path through HMMs is finding the spoken words [5]. In speech synthesis, the same procedure name's training part is used with adding the synthesis part. The speech signal can then be synthesized from so generated feature vectors. The basic structure of this system is shown in Fig. 1. Most HMM-based speech synthesizers have similar structures, which are divided into two parts, training and synthesis.

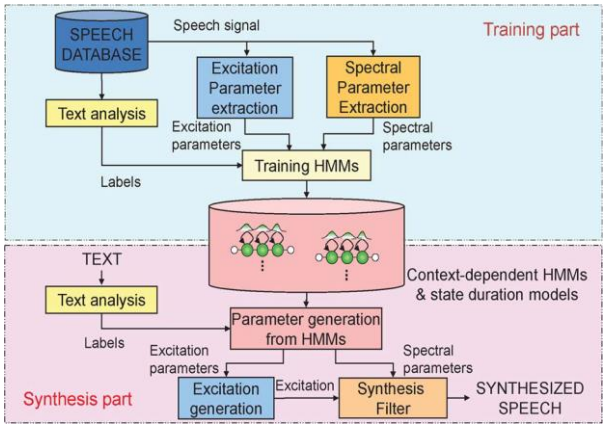

Figure 1. An overview of the basic HMM-based speech synthesis system [6].

Speeches with their description are the inputs to the first part of system: Training part, this description describes duration and symbols for all phonemes with several examples; in addition there exist texts for use in the generation of audio. For each utterance of the speech corpus, excitation and spectral are extracted via HTS-tools, the spectral parameters are often defined by mel-cepstrum coefficients or line spectral frequencies, which are adequate features for statistical modeling. Log F0 is used as an excitation parameter.

A. Basic HMM

The training part

In the training part, context-dependent HMMs are modeled using the phonetic labels and the speech features. To train HMMs, the statistical parameters are calculated using decision trees. A maximum likelihood (ML) criterion is usually used to estimate the model parameters [7] as:

$$\hat{\lambda} = \arg\max_{\lambda} \left\{ p\left(o / \omega, \lambda\right) \right\} \quad (1)$$

The synthesis part

At the synthesis stage, by using input text the context-dependent labels are obtained and they are used by the speech parameter generation algorithm to generate the speech features. The excitation signal is calculated using the excitation features, which then passes through the synthesis filter to have the speech signal. The synthesis filter used in HTS is defined by the spectral features. We then generate speech parameters, o, for a given word sequence to be synthesized, ω, from the set of estimated models, $\hat{\lambda}$ to maximize their output probabilities [7] as:

$$\hat{o} = \arg\max_{o} \left\{ p\left(o / \omega, \hat{\lambda} \right) \right\} \quad (2)$$

The advantage of this approach is in capturing the acoustical features of context-dependent phones using the speech corpora. Synthesized voiced characteristics can also be changed easily by altering the HMM parameters and the system can be easily ported to a new language. In HMM-based speech synthesis, the spectrum, F0 and durations are modeled in a unified framework in [8]. From the HMM model, features are predicted by a maximum-likelihood parameter generation algorithm [6]. Finally, the generated parameters are sent to a parametric synthesizer to generate the waveform. As a first application of this method, we decided to use the canvas provided in the demonstration scripts of HMM-based Speech Synthesis System [6] which is a set of tools used as a patch to HTK1 (HMM Toolkit) and which allows to perform acoustic speech synthesis based on HMMs. The tools used in HTS demonstration scripts is SPTK [9] (Speech Signal Processing Toolkit) for spectrum and Snack for F0. The question number (3) defines a general generative model for sequences of acoustic feature vectors. The question number (3) is a simple form of acoustic model.

$$P(c \mid \theta, \lambda) = \prod_{t} \mathrm{P}(\mathrm{c}_{t|c_t}, \theta_t, \lambda) \quad (3)$$

This question assumes the feature vectors are conditionally independent given the state sequence. Together the Markovian state transition model and this simple acoustic model form a hidden Markov model (HMM) $P(c, \theta \mid \mathrm{l}, \nu, \lambda)$. The Fig. 2 shows the corresponding graphical model.

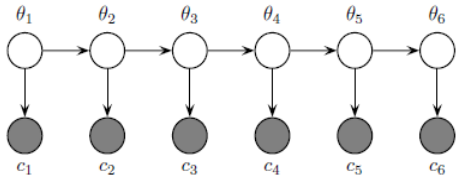

Figure 2. Graphical model for a conventional HMM. Here θ = θ1:6 is the state sequence and c = c1:6 is the feature vector sequence. The dependence on the label sequence l and parameters (ν; λ) is not shown. Note that this is not the model used during training in the standard HMM synthesis framework, which augments the static feature vector sequence with dynamic features.

B. Autoregressive HMM

Autoregressive HMMs [10] have been used before for speech recognition and now for the speech synthesis. The model for parameter estimation and synthesis used in standard HMM is the same model in the autoregressive HMM. The autoregressive HMM extracts the parameter estimation using expectation maximization, in contrast to the standard HMM and also supports a speech parameter generation algorithm not available for the standard HMM [11]. The question number (4) describes a general generative model for sequences of acoustic feature vectors.

$$P(c \mid \theta, \lambda) = \prod_{t} \mathrm{P}(c_{t|c_{t-kt-1}}, \theta_t, \lambda) \quad (4)$$

where K € N is referred to as the order or depth of the model. This acoustic model and the HMM state transition model constitute an autoregressive HMM. Fig. 3 shows a graphical model for the case K = 2.

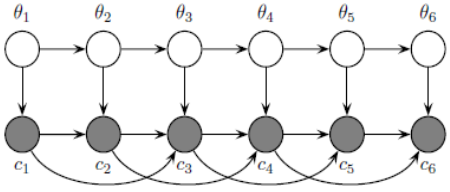

Figure 3. Graphical model for an autoregressive HMM of depth 2. Here θ = θ1:6 is the state sequence and c = c1:6 is the feature vector sequence. The dependence on the label sequence l and parameters (ν; λ) is not shown.

C. STRAIGHT Vocoder

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum) is a high-quality system for speech modification [12]. This system incorporates a mixed excitation model described by [8] which consists on weighting the periodic and noise components using aperiodicity measurements of the speech signal. STRAIGHT vocoder1 extract the spectral envelope and aperiodicity measurements from the speech signal. STRAIGHT represents both the spectrum and aperiodicity of the speech signal by FFT coefficients, which are not suitable for statistical modeling due to their high-dimensionality. Fig. 3 illustrates its overview. It consists on the three main components, i.e., F0 extraction, spectral and aperiodicity measure analysis, and speech synthesis. The STRAIGHT vocoder method is shown in the Fig. 4.

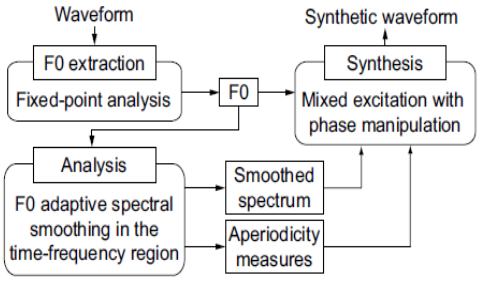

Figure 4. A block diagram of STRAIGHT vocoding method.

III. Arabic Database

As part of our work, we refer to the Arabic language in reference to what is commonly called "Standard Arabic", that is to say, the language of communication in the entire Arab world. It is the language taught in schools, so written, but spoken in the formal framework. Arabic belongs to the Semitic language family. The study of Arabic grammar began early in the 11th century AH and resulted in huge productions, before experiencing a period of stagnation that lasted for several centuries [13]. The phonetic system of Standard Arabic is composed basically by 34 phonemes, which consists of 26 consonants, 3 long vowels, 3 short vowels and 2 semivowels [14].

A. The Diacritics

Short vowels are represented by symbols called diacritics (see Fig. 5). Three in number, these symbols are transcribed as follows:

- The Fetha [a] is symbolized by a small line on the consonant (مَ / ma /)
- Damma the [u] is symbolized by a hook above the consonant (مُ / mu /)
- The kasra [i] is symbolized by a small line below the consonant (مِ / mi /)
- A small round o symbolizing Sukun is displayed on a consonant when it is not linked to any vowel.

B. The Tanwin

The sign of tanwin is added to the end of words undetermined. It is related to exclusion with Article determination placed at the beginning of a word. Symbols tanwin are three in number and are formed by splitting diacritics above, which results in the addition of the phoneme / n / phonetically :

[an]: (عً / AIan /)
[un]: (عٌ / AIun /)
[in]: (عٍ / bin /)

C. The Chadda

The sign of the chadda can be placed over all the consonants non initial position. The consonant which is then analyzed receives a sequence of two consonants identical:

Signe _ / kallama / ("he talked to").

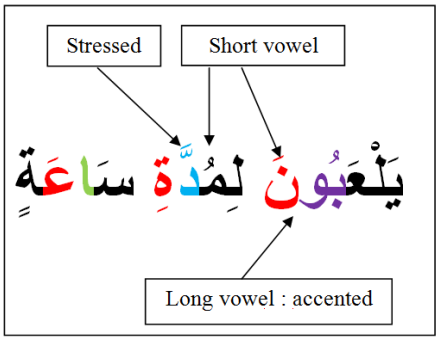

Figure 5. Example of a sentence / jalAIabuuna limuddati saAItin / ("They play for an hour")

The Arabic phonetic system differs from the Latin ones essentially by emphatic and glottal phonemes. The phonetic transcription used for the Arabic consonants and their equivalents are shown in Table I.

The syllabic structures in Arabic are limited in number and easily detectable. Every syllable in Arabic begins with a consonant followed by a vowel which is called the nucleus of the syllable. Short vowels are denoted by (V) and long vowels are denoted by (VV). It is obvious that the vowel is placed in the second place of the syllable. These features make the process of syllabification easier. Arabic syllables can be classified either according to the

length of the syllable or according to the end of the syllable. Short syllable occur only in CV form, because it is ending with a vowel so it is open. Medium syllable can be in the form of open CVV, or closed CVC. Long syllable has two closed forms CVVC, and CVCC. Arabic words are composed at least by one syllable; most contain two or more syllables. The longest word is combined of five syllables. Table II illustrates Arabic syllables. Some of the Arabic words are spelled together forming new long words with 6 syllables like (لُونَهَيَأْكُ), or 7 syllables like (يَسْتَقْبِلُونَهُ). There exist a few Arabic data suitable for HMM-based synthesis, which should ideally include a large number of Arabic databases from a single speaker and corresponding phonetic transcriptions. We used the database [15] in [16], has been phonetically annotated and used in [17]. As HMM-based synthesis requires a lot of training examples.

TABLE I. ARABIC CONSONANT AND VOWELS AND THEIR PHONETIC COMPATIBLE NOTATION OF HTS SYSTEM

Graphemes	symbole	Graphemes	symbole	Graphemes	symbole	Graphemes	symbole
ء	A	ر	r	غ	g	ي	j
ب	b	ز	z	ف	f	◌َ	a
ت	t	س	s	ق	q	◌ً	aa
ث	T	ش	S	ك	k	◌ِ	i
ج	Z	ص	ss	ل	l	◌ٍ	ii
ح	X	ض	dd	م	m	◌ُ	u
خ	x	ط	tt	ن	n	◌ٌ	uu
د	d	ظ	dh	ه	h		
ذ	D	ع	AI	و	w		

TABLE II. ARABIC SYLLABLES TYPES

Syllable	Arabic example		English meaning
cv	لِ	li	to
cvv	في	fii	in
cvc	قُلْ	qul	say
cvcc	بَحْرْ	bahr	sea
cvvc	مَالْ	maAl	money
cvvcc	زَارّ	zaArr	visit

IV. EXPERIMENT

Compared methods

For each Arabic sound, three different methods were compared:

Method 1: In this method, standard HMM framework [18], we use the based HTS, This toolkit is used for implementing HMM-based speech synthesis. HTS-2.1.1 [7] was applied as a patch to HTK-3.4.1. HDecode-3.4.1 for HTK-3.4.1 [19] was also installed. Festival-2.1 [20], speech_tools-2.1, SPTK-3.1 [9], Snack [21], ActiveTcl8.4.19.4 [22], festvox-2.1 [23], and other support software tools were installed in setting up the TTS synthesis system experimentation platform. All the above-mentioned tools are downloadable from their respective websites.

Method 2: In this method, standard HMM framework, we use the same toolkit with STRAIGHT vocoder (version V40 006b) [24] and MATLAB.

Method 3: In this method, autoregressive HMM, we use the same toolkit with STRAIGHT vocoder (version V40 006b) and MATLAB.

A. Evaluation

We used phonetically balanced 200 sentences from Arabic speech database for training. The participants should have the Arabic language as their second language. The group consists of 36 people. The majority of the participants are students at Bourguiba Institute of Languages University Elmanar, Tunisia at the Department of Arabic Linguistics. The level of fluency is varying among the participant, some of them are somehow fluent and the some of them are not very fluent.

- Test ABX

By analyzing the result of ABX listening tests and subjective experiments, it is investigated the characteristics of synthesized speech from HMM set between three system based on HTS and a natural speech.

The evaluation on the similarity is based on to what degree the synthesized emotional speech conveys the identity to the target speaker.

1) Evaluation on similarity

- First test

The method of ABX test, where X is the neutral, A is a synthesized speech by HTS2, and B is a synthesized speech by HTS3. Basically, a listener is asked to decide whether X sounds like the speaker of A or the speaker of B. The participants listened this 12 utterance at random were asked to select either A or B as being the closest the natural speech.

- Second test

The method of ABX test, where X is the neutral, A is a synthesized speech by HTS1, and B is a synthesized speech by HTS2. Basically, a listener is asked to decide whether X sounds like the speaker of A or the speaker of B. The participants listened this 12 utterance at random were asked to select either A or B as being the closest the natural speech. The result represented in the Table III.

TABLE III. ERROR RATE BETWEEN HTS AND HTS1, HTS2, HTS3

HTS1vs. HTS	HTS2vs. HTS	HTS3vs. HTS
78.34%	09.45%	76.34%

- Test One large MOS

We used One large MOS (mean opinion score) listening test was conducted to evaluate the quality of speech obtained.

1) Bad quality
2) Poor quality
3) Fair quality
4) Good quality
5) Excellent quality

The Testing and Evaluation phase of all system is done by the same test group. A questionnaire was designed precisely to assess the intelligibility (clearness), naturalness, sound quality and the pronunciation on the level of phoneme word and sentence.

2) Evaluation on intelligibility

The participant is asked a question "How much you understand the voice?", and is asked to mark how well the voice performs. The results are shown in Fig. 6 below.

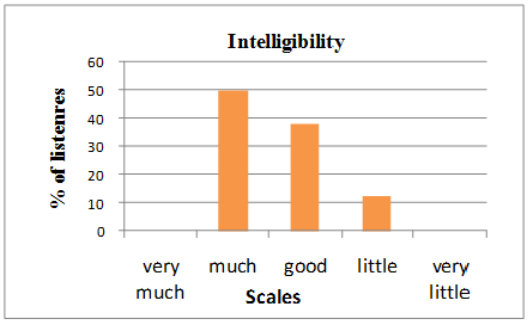

Figure 6. Intelligibility of the voice

3) Evaluation on naturalness

The participant is asked a question "Was the sound natural or not?", and is asked to mark how well the voice performs. The results are shown in Fig. 7 below.

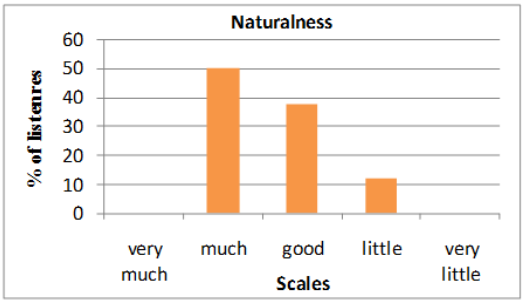

Figure 7. Naturalness of the voice

4) Evaluation on sound quality

The participant is asked a question "What level of quality do you think the synthesizer has?", and is asked to mark how well the voice performs. The results are shown in Fig. 8 below.

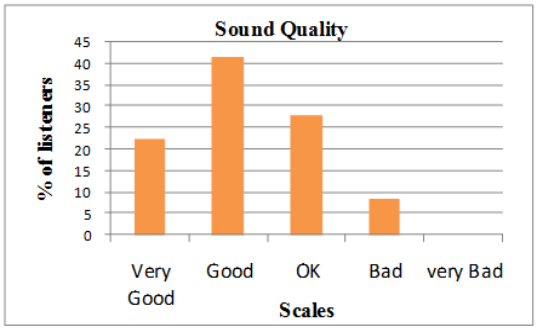

Figure 8. The sound quality of the voice

5) Evaluation on pronunciation

The participant is asked a question "Did you have to concentrate hard to grab the speech?", and is asked to mark how well the voice performs. The results are shown in Fig. 9 below.

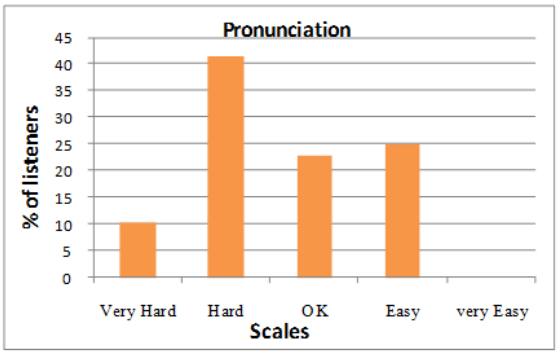

Figure 9. The sound quality of the voice

Method 1 has the minimal score.

In second method, we obtain the best score, the training part and synthesis part with STRAIGHT is the best way to synthesis a good speech. In third method, we use the autoregressive HMM, this method is better than based HTS but not than method 3. Then, Among the 3 synthesis methods, method 2 yields the best results. The obtained best score is clearly better.

6) Result

In the both evaluations, the results show the second method is the best than others. In this work the STRAIGHT version V40 006b was used, because this was the only STRAIGHT version which was publicly accessible (through the following webpage: http://www.wakayama-u.ac.jp/˜kawahara/index-e.html). This version uses a unified approach to estimate the F0, aperiodicity and spectrogram. In third method, we use the autoregressive HMM, this method is better than based HTS but not than method 3. Then, Among the 3 synthesis methods, method 2 yields the best results. The obtained best score is clearly better.

V. Conclusion and Future Works

A HMM-based synthesis Arabic HTS_ARAB_TALK [17] system was developed. Phonemes were the essential elements of the synthesizer, our HTS_ARAB_TALK system is vocabulary independent with intelligible output speech, so it can handle all types of input text. In this paper we have demonstrated and determinate by a comparative study the better method to obtain higher quality and more naturalness in Arabic speech synthesis, adapting methods that have proved efficient in speech synthesis. The proposed method yields significant improvement compared to previous work. The potential of HMM-based speech synthesis has been demonstrated with limited training data. Single speaker, phonetically-annotated speech databases would likely help improving the results. Recording and (automatically) annotating such a database is part of our future work. Several versions of HMM-based synthesis have been implemented and evaluated in other langue but not in

Arabic. The best model obtained includes STRAIGHT predicted by HTS. Comparing with other available Arabic TTS systems, our HTS_ARAB_TALK has small size, high accuracy, and vocabulary independence features which make it in general more reliable than other TTS systems. The system is free for distribution and for development. Deeper statistical analysis of our evaluation data will also be performed, to investigate the influence of age, sex, in speech synthesis. Finally, optimize the HTS_ARAB_TALK to obtain a real time system. We will also improve the prosody modeling by extracting more advanced context features. In conversational speech, naturalness of prosody is still insufficient to properly convey nonverbal information, e.g., emotional expressions and emphasis. To fill the gap between natural and synthesized speech, the statistical approaches are more important in the future.

REFERENCES

[1] M. C. Bateson, *Arabic Lanuage Handbook*, Georgetown University, 2003.
[2] A. Omar, *Dirasat Al-Swat Al-Lugawi*, Cairo: Alam Al-Kutub, 1985.
[3] W. Erwin, *A Short Reference Grammar of Iraqi Arabic*, Washington: Georgetown University Press, 1963.
[4] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vrecken, "The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non commercial purposes," in *Proc. Fourth International Conference on Spoken Language*, 1996, pp. 1393-1396.
[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, 1989.
[6] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis," *Speech Communication*, vol. 11, pp. 1039-1064, 2009.
[7] K. Tokuda, *et al.* The HMM-based speech synthesis system (HTS) version 2.1. [Online]. Available: http://hts.sp.nitech.ac.jp/
[8] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proc. 2nd MAVEBA*, Firenze, Italy, 2001.
[9] Speech signal processing toolkit (SPTK). [Online]. Available: http://sp-tk.sourceforge.net
[10] M. Shannon and B. William, "Autoregressive clustering for HMM speech synthesis," in *Proc. Interspeech*, 2010.
[11] M. Shannon, Z. Heiga, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 587-597, 2013.
[12] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187-207, 1999.
[13] S. Al Ani, *Arabic Phonology: An Acoustical and Physiological Investigation*, The Hague, 1970.
[14] G. Bohas, "Contribution à l'étude de la méthode des grammairiens arabes en morphologie et en phonologie d'après les grammairiens arabes tardifs," thèse de doctorat, Université de Lille 3, 1979.
[15] M. Boudraa, B. Boudraa, and B. Guerin, "Elaboration d'une base de données arabe phonétiquement équilibrée," in *Proc. Actes du Colloque Langue Arabe et Technologies Informatiques Avancées*, Casablanca, Dec. 1993, pp. 171-187.
[16] K. M. Khalil and C. Adnan, "Arabic HMM-based speech synthesis," in *Proc. International Conference on Electrical Engineering and Software Applications ICEESA*, 2013.
[17] K. M. Khalil and C. Adnan, "Optimization of Arabic database and an implementation for Arabic speech synthesis system using HMM: HTS_ARAB_TALK," *International Journal of Computer Applications*, vol. 73, no. 17, Jul. 2013.
[18] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang. (2006). USTC system for blizzard challenge 2006 an improved HMM-based speech synthesis method. *Proc. Blizzard Challenge Workshop* [Online]. Available: http://www.festvox.org/blizzard/blizzard2006.html.
[19] The hidden Markov model toolkit (HTK). [Online]. Available: http://htk.eng.cam.ac.uk/
[20] The festival speech synthesis system. [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival/
[21] The snack sound toolkit (snack). [Online]. Available: http://www.speech.kth.se/snack/
[22] Active Tcl. [Online]. Available: http://wiki.tcl.tk/
[23] The Festvox speech synthesis system. [Online]. Available: http://www. festvox.org/
[24] STRAIGHT, a speech analysis, modification and synthesis system. [Online]. Available: http://www.wakayamau.ac.jp/~kawahara/STRAIGHTadv/index_e.html

Mohamed Khalil Krichi was born in Tunisia in 1984. He received the Master degree from University of Tunis El Manar, FST in Tunisia respectively, all in electrical engineering, specializing in signal processing. He is currently working toward the PhD degree in Arabic speech synthesis with HMM in University of Manar under the supervision of Prof. Adnan. Cherif. His research interests include speech synthesis and analysis. Actually he is an assistant at the Science Faculty of Bizerte

Adnan Cherif: was born in Tunisia, received his engineering diploma from the Engineering Faculty of Tunis and his Ph.D. in electrical engineering and electronics from The National Engineering School of Tunis (ENIT). Actually he is a professor at the Science Faculty of Tunis, Responsible for the Signal Processing Laboratory. He participated in several research and cooperation.