3rd International Conference on Automation, Control, Engineering and Computer Science (ACECS'16)

Proceedings of Engineering & Technology (PET)
pp. 141-147

Arabic Speech Synthesis System using HMM: HTS_ARAB_TALK

Krichi Mohamed Khalil[#1], Cherif Adnan [#2]

[#1]*Polytechnic Central Private School of Tunis*

3 street Mohamed V, Elkram Tunis Tunisia

[1] Krichi_moha@yahoo.fr

[#2]*Signal Processing Laboratory, Science Faculty of Tunis*

SFT 1060, Tunisia

[2] adnen2fr@yahoo.fr

Abstract— **this paper describes the Arabic speech synthesis system using HMM (HTS). Our developed synthesis system (HTS_ARAB_TALK) uses phonemes as HMM synthesis unit, which Arabic database was used called the PADAS database. This paper describes a new system of Arabic speech synthesis called HTS_ARAB_TALK which represents a complete architecture system by modifying it the publicly available HTS. The main objective is to maintain the consolidated text coherence which is interpreted by concatenating HMM phoneme. In our experiments, spectral properties were represented by Mel cepstrum coefficients. For the waveform synthesis, a noise or pulse excited corresponding MLSA filter were used. Besides that basic setup, a high-quality analysis/ synthesis system STRAIGHT was employed for more sophisticated speech representation. This method has several advantages. As it is parametric, it is possible to play on the HMM parameters, change the producer voice characteristics. The developed model improves the speech synthesis, naturalness and intelligibility quality in the Arabic language environment.**

***Keywords— HMM, Speech Synthesis, HTS, Arabic Language,* PADAS, *Statistical Parametric Speech Synthesis, Hidden Markov* Model.**

I. INTRODUCTION

Speech is the most important form of communication in everyday life [2]. The goal of synthesis systems is to provide the users with spoken output by generating speech from text. Speech synthesis [1] is used in several applications. Speech synthesis methods can be divided into four categories [23,24, 25, 26, and 27]: Articulatory synthesis, formant synthesis, concatenative synthesis and Statistical Synthesizers. Statistical parametric speech synthesis is a relatively a recent approach summarized by [13,16]. In comparison with formant synthesizers, HMM-based speech synthesizers are also fully parametric and require a small foot print, but they have the advantage that they are fully automatic. These methods consist of mainly two parts:

Procedures for selection and training of basic synthesis units and the synthesis part, where the phonetic and prosodic information are used for speech signal generation. One of the most promising methods is the use of context dependent phone models, modeled with hidden Markov model (HMM) [4,9, 17]. By using our experimental system, two different speech analysis/synthesis methods and speech representations are compared.

- A simple representation by Mel cepstrum coefficients by SPTK toolkit [3].
- More sophisticated speech representation by the high-quality analysis/synthesis method STRAIGHT [15].

Recently trainable synthesis systems have been applied in Japanese [4], English [5, 6, and 7] and in a few other languages [8, 14, and 17]. Arabic HMM-based Speech Synthesis is the state-of-the-art high quality natural TTS systems. *HTS_ARAB_TALK* is one of these systems, which is developed specially for Arabic language. This paper speaks about the overall architecture, several components of the system, and linguistic concepts for Arabic. For that this paper will be devoted to focus on following parts. Section 1, describes the Arabic language form. Section 2 present the HMM system, training part and synthesis part. Section 3 describes the process followed a first realization, Arabic database and file questions. Section 4 presents the results and evaluation of the first realization and section 5 presents the concluding remarks.

1. Introduction to Arabic language

The Arabic language is spoken throughout the Arab world. This means that Arabic is known widely by all Muslims in the world. 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. Standard Arabic has 34 basic phonemes, of which six are vowels, and 28 are consonants [10]. Arabic vowels are affected as well by the adjacent

ISSN: 2356-5608

phonemes. Accordingly, each, of the former 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 [11]. The 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.

The following table represents the Arabic consonants and vowels and their phonetic compatible notion of HTS system.

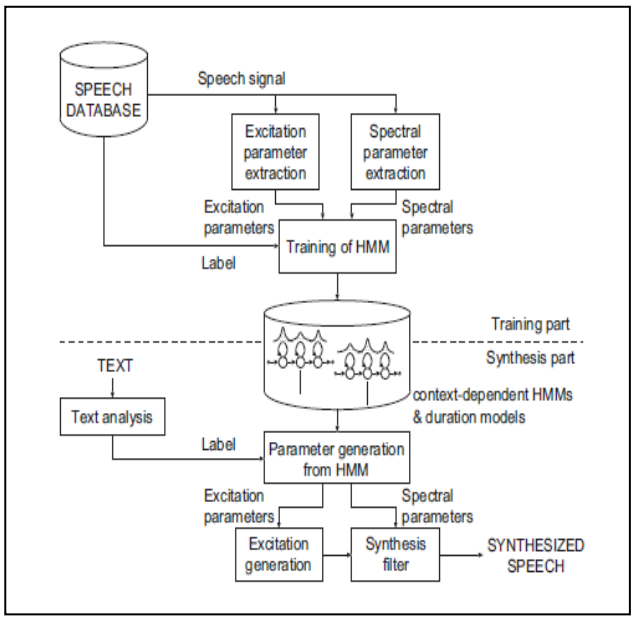
TABLE I. ARABIC CONSONANT AND VOWELS AND THEIR PHONETIC COMPATIBLE NOTATION OF HTS SYSTEM

Graphemes	New code used	Graphemes	New code used	Graphemes	New code used	Graphemes	New code used
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		

II. A SPEECH SYNTHESIS SYSTEM BY HMM

In 2007, an article describing the characteristics of HTS v2.0 is published [21] to provide a set of free tools constituting a speech synthesis system based on HMMs. Since that date, the scientific literature has largely been dominated by the HMMs speech synthesis. This method has several advantages. As it is parametric, it is possible to play on the HMMs parameters to change the generated voice characteristics. If these changes are made wisely, it is possible to synthesize different styles and vocal characteristics from a single natural voice database. Statistical modeling is automatic and therefore, the change in style is even easier. Finally, the real time component can be added as the HMMs are well suited to dynamic changes in style. The overall structure of the speech synthesis system is shown in the following figure.

Fig.1 overview of a typical HMM-based speech synthesis system



It is important to keep in mind certain terms directly related to speech processing field or tools used in this work: definition

- Phoneme

A minimal element, non-segmental, state phonological representation of states, and whose nature is determined by a set of distinctive features [19]. That is to say, it is the smallest constitutive phonation particle. A list of the groups of phonemes groups largest used is given in table I.

- HMMs

This term stands for Hidden Markov Models (hidden Markov models) with or without s at the end as it is put in the singular or plural. It is often used as the acronym in this document representing a modeling theory system under certain conditions.

- HTK

In this paper, HTK [12,20] is used to refer to a set of tools manipulate HMMs. The HTS tools that complement, HTK change so as to make profit for audio synthesis.

- Training

This term refers here to all transactions to form and configure HMMs modeled speech.

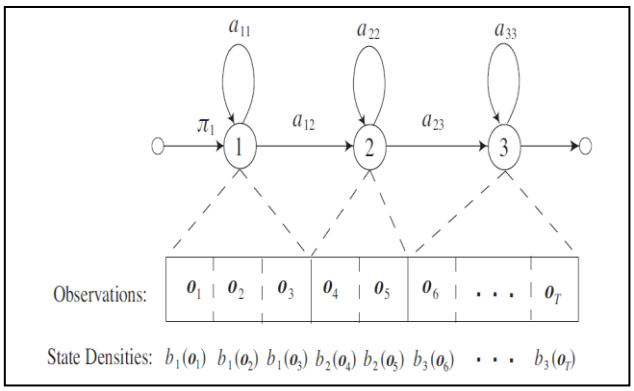
- Synthesis

This term refers to the production parameters or audio signals derived from models HMMs entrained.

A. Notions on hidden Markov models

A hidden Markov model (HMM) is a state machine that changes from state i to state j at each time step. At each time t where the state j is visited, a continuous observation vector ot is generated from the state in question bj (ot) output probability distribution. An HMM is defined by its transition probabilities aij between state i and state j, its observation probability distributions (or emission) bj (o) and the initial states probabilities, that is ie the probability πj that state j is the first state visit. Figure 2 illustrates schematically HMM 3 states manner (from left to right, that is to say we can not return to a state once it has been passed and the state 1 is always first) with transition probabilities between states denoted aij, the probabilities bi (Ot) Ot observed for state i, the observation sequence O and the corresponding state sequence Q.

Fig. 2 A 3-state left-to-right HMM with illustration of an observation sequence and the state output probability distributions associated with each state.



B. synthesis Steps

As shown in Fig.3 synthesis of the two steps can be done by SPTK or STRAIGHT.

A) The extraction and training

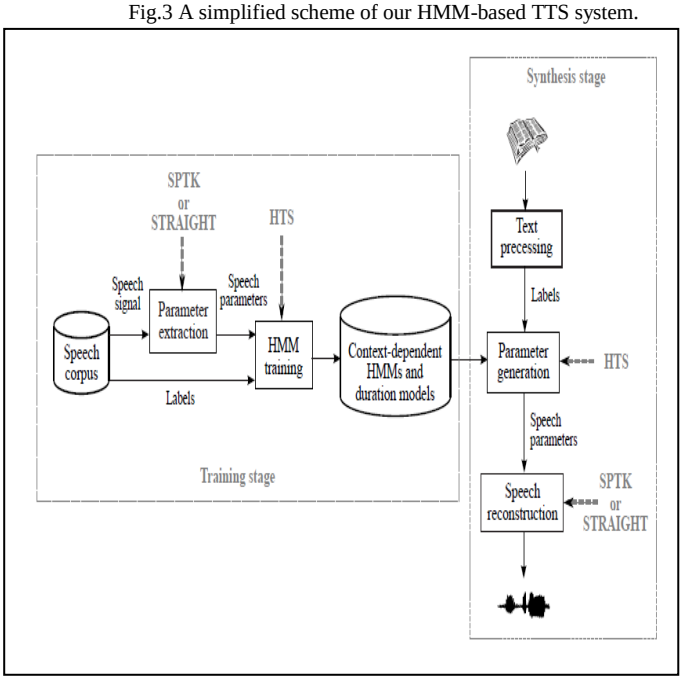
1)As a first step, it is necessary to have a sufficient data base so as to have enough examples of each component to synthesize. In this HTS case, we want synthesize speech whose constituent elements are phonemes.

a) From this database, the extracted characteristics are following: In other words, these characteristics are essentially of two types: the excitation characteristics (related to the speech signal fundamental frequency considered a given moment) and spectral characteristics.

b) It is also necessary that the database is annotated. This is specific operation, for each file in the database, the start and end time of each phoneme present in the file. In the absence of recognition performance system, this must be done manually. For that it'is a time consuming operated. Background information will be added to these notes and durations to provide more information, embodied in transcription files. The labels contain information on the signal phonetic context (duration, current phoneme, phonemes previous / next) and say prosodic information (number of words in the sentence, number of syllables in the word, syllable position vis-à-vis the word ...).

2) These spectral data phonemes, excitation and duration of will serve the HMMs training phase of . After this step, each context dependent,each phoneme HMMs has a model and duration. It is understandable, thatmost important statistically modeled database will be more realistic, provided that each phoneme is represented quite a number of times.

Fig.3 A simplified scheme of our HMM-based TTS system.



1) *In the synthesis phase, it will be necessary to specify the system you hope to synthesize.*

- This text should be analyzed and the phonetic and prosodic contextual information extract to form a transcript file corresponding to the requested text. This step is called generation, so it is an interpretation it is a written text interpretation in a contextual phonetic transcriptions.

2) *Once this file label available, the information contained therein will allow HMMs to generate parameters (counterparts and spectral excitation characteristics mentioned above) according to the modeling done during the training phase.*

3) in this step point, all the synthesis required information for is present. It only remains to generate a waveform (which can be heard) from these parameters. This happens in particular from a model source filter synthesis modeling the human vocal apparatus. In this model, a source is attacking the filter input whose output is the expected waveform.

- The source is either a Dirac pulse train (which representing the vocal cords vibration) is a white noise (that is the turbulence in the vocal tract in the absence of the vocal cords vibration) as the previous steps generated excitation parameters are contained in the information function and that generated.
- The filter, in turn, is also defined in the previous step is generated the spectral parameters. This is the shape of the vocal tract at the time of production.

III. PROCESS FOLLOWED A FIRST REALIZATION

A. Database studying

The ideal is to have a database sufficiently provided with all Arabic phonemes [28,29]. The audio portion of the database is the only one that interested .wav format is PCM coded 16-bits at a sampling frequency of 16 kHz. To adapt the Arabic phonemes with the HTS system, we use a new presentation of phonemes, since for example the HTS system reject "?" so the phonemes will be newly encoded as shown in table I.

1)Generation of transcription files (labels)

The transcription files or labels file are the ones that complement the database [22]. That is to say in these text files, to find phonemes duration's information of, their neighboring phonemes and other information that will allow the machine to know the contents of the audio file.

That is to say that the database provided was accompanied by text files (one for each audio file) in which there were time information to start and end phonemes constituting the audio file. These text files were generated by free software called speech processing Praat. This software allows opening an audio file to listen to pieces and set limits (start time and end time) for phonemes, which corresponds to the annotation phonemes. This operation is time consuming and must be done manually for each phoneme in each audio file. Two types of files .lab needed. One is the so called "mono"and the other is the one called "full".

2)question files

Questions Files are text files that define the questions to the nodes of decision trees for HMM clustering. The questions that are asked for phonemes are directly related to background information provided in the label files (labels).

The following figure 4 is a question file extract.

Fig.4 Extract a file of questions regarding the phonemes is two elements to the left of the current phoneme

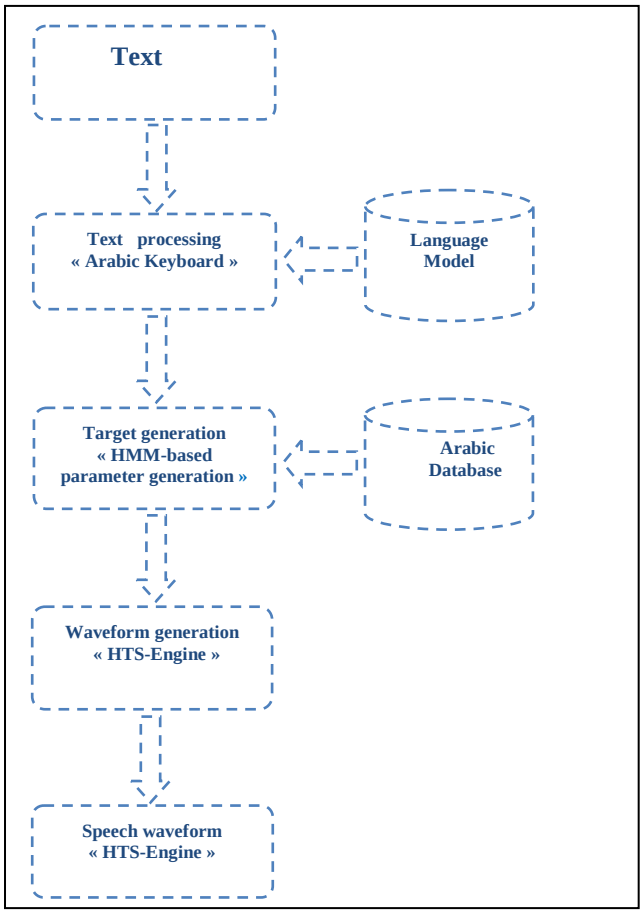
```
QS "LL-Vowel"          {a^*,aa^*,u^*,uu^*,i^*,ii^*}
QS "LL-Consonant"      {b^*,S^*,d^*,D^*,dd^*,f^*,G^*,h^*,


QS "LL-Nasal"          {m^*,n^*}
QS "LL-Fricative"      {z^*,f^*,x^*,D^*,s^*,s'^*,S^*,AI^
QS "LL-Plosive"        {b^*,dd^*,d^*,tt^*,t^*,k^*,q^*,A^
QS "LL-Affrique"           {Z^*}
QS "LL-vibrant"        {r^*}
QS "LL-lateral"        {l^*}
QS "LL-approximant"    {w^*,j^*}
```

IV. DEVELOPMENT OF HTS_ARAB_TALK

Figure 7 shows the architecture of the current system. It is composed of three major components: a HTS-training, a HTS-engine, an Arabic keyboard. In the HTS-training component, we prepare a prosodic Arabic database and construction of the statistical parametric speech. After training part, we send this parameter to HTS-engine. Text is the input of the system.

Fig.5 Block diagram of HTS-ARAB-TALK



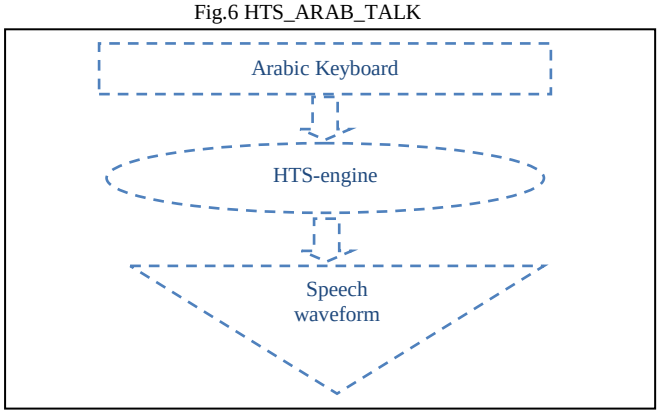
A. *Text segmentation*

Syllable Parser will segment the normalized text to syllable unit according to Arabic rules. The architecture is based on Input, Processing and Output Schematic. This module will convert the symbols input into readable text. Input text may be in the form of paragraphs, sentences, or words. Thus, it is necessary to segment the text in a hierarchal order: higher level structures to paragraphs, paragraphs to sentences, sentences to words and words to syllables and syllable to phonemes. In this research, we limited the input text to paragraph form. A paragraph was segmented into sentences by finding the sentence punctuation marks such as „.", „!" and „?". To segment sentences into words, blank spaces were located in the text that has been classified as a sentence. From the text that has been identified as words, the phonemic representations equivalent to the set of letters of the retrieved word were generated.

B. *Waveform generation*

HTS-engine-API: Since version 1.1, a small stand-alone run-time synthesis engine named HTS-engine has been included in the HTS release. It works without the HTK libraries, and it is released under the new and simplified BSD license; Users can develop their own open or proprietary software based on the run-time synthesis engine and redistribute these source, object, and executable codes without any restriction. In fact, a part of HTS-engine has been integrated into several pieces of software, such as ATR XIMERA [20], Festival [21], and Open MARY [22]. The spectrum and prosody prediction modules of ATR XIMERA are based on HTS-engine. Festival includes HTS-engine [8] as one of its waveform synthesis modules. The upcoming version of Open MARY uses the JAVA version of HTS-engine. The stable version, HTS-engine API version 1.0, was released with HTS version 2.1. It is written in C and provides various functions required to setup and drive the synthesis engine. In this step, we used a HTS-Engine (1.07). The following figure 8 represents the general appearance of the HTS_ARAB_TALK.

Fig.6 HTS_ARAB_TALK



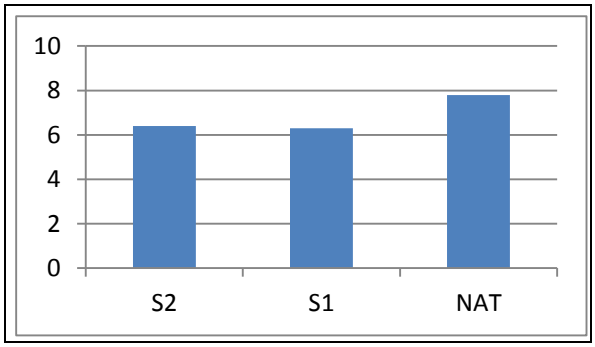
V. EXPERIMENTS AND RESULT

A. *Objective Evaluation*

The goal of the objective evaluation is to assess whether HTS is capable of producing natural hypo and hyperarticulated speech and to which extent. The distance measure considered

here is the mel-cepstral distortion between the both systems, expressed as [29].

Fig. 7 Average scores for the test (HTS-ARAB-TAL (S1), HTS-ARAB-TALK using new database and STRAIGHT vocoder (S2) and natural speech for the intelligibility of speech



B. *Subjective Evaluation*

For our experiments, two different voices (male and female) were used. One large MOS (mean opinion score) listening test was conducted to evaluate the quality of speech obtained. In this test, there are speech representations by using SPTK or STRAIGHT and amount of training data (200 sentences and 398 sentences). In addition, natural source speech is also used in this test. 15 listeners took the test. They listen 16 sentences (in sum) and evaluate them according to the overall quality by using the standard MOS scale.

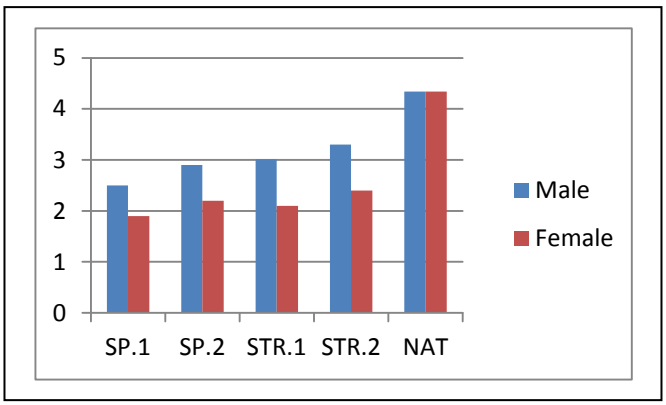
1. Bad quality
2. Poor quality
3. Fair quality
4. Good quality
5. Excellent quality

The test results are presented in table 2 and figure 5 after comparing the two types of speech synthesis, STRAIGHT and SPTK, it is clear that the STRAIGHT use is more efficient compare to SPTK.

TABLE II. MOS TEST RESULT

Gender and number of sentences	**Score for each Speech generation method**		
	SPTK With notation	***STRAIGHT With notation***	***NATURAL With notation***
Male 200 sentences	2,5 SP.M.1	3,01 STR.M.1	4,86 for Male NAT.M
Female 50 sentences	1,9 SP.F.1	2,1 STR.F.1	4,34 for Female NAT.F
Male 398 sentences	2,9 SP.M.2	3,3 STR.M.2	
Female 100 sentences	2,2 SP.F.2	2,4 STR.F.2	

Fig. 8 MOS test results



VI. CONCLUSION

In this paper, first experiments on statistical parametric HMM-based speech synthesis for the Arabic language are presented. A highly intelligible and acceptably natural sounding speech synthesis system in Arabic language has been developed. For speech representation, two different speech analysis/ synthesis are used MLSA filter and STRAIGHT. The listening test subjective evaluation test showed that speech produced by the HMM-based TTS system using STRAIGHT is of comparable quality to speech synthesized by the TTS system using SPTK.

VII. DISCUSSION AND FUTURE WORKS

Text-To-Speech Synthesizer has been developed gradually over the last few decades and it has been integrated into several new applications. For most applications, the intelligibility and comprehensibility of TTS Synthesizer have reached the acceptable level. Nevertheless, in prosodic, text preprocessing, and pronunciation fields there is still much work and improvements to be done to achieve more natural sounding speech. Natural speech has so many dynamic changes that perfect naturalness may be impossible to achieve. However, since the markets of TTS Synthesizer be related applications are increasing gradually, the attention for giving more efforts and funds into this research area is increasing as well. Current TTS Synthesizer Systems are so complicated that one researcher cannot handle the whole system. With good modularity it is likely to divide the system into a number of individual modules whose developing process can be done alone if the communication between the modules is made carefully. Some of the possible improvements that can be made are: Record more sounds in the sound database. More sounds can be recorded to have better performance and more vocabularies. Users can learn more words without much limitation. Build more user friendly interfaces, such as a command to select different voices, for example, voice of a man and voice of a woman. As well as an interface, this will allow users to click on the Arabic words rather than typing them – applicable for users who do not have Arabic keyboard. Adding an animation character (Agent). An agent or mount utterance character can be included to attract user to continue

using this software. Humans are more attracted to animated and attractive interfaces which can create interest and fun in learning. The characters are able to speak the input text, along with the output sound with mouth utterances and gestures. A new high quality Arabic speech synthesis technique has been introduced in this paper. The technique is based on the HMM-based speech synthesis system. This was readily observed during the listening tests based on high quality and objective evaluation when comparing the original with the synthetic speech.

REFERENCES

[1] Cabral, HMM-based Speech Synthesis Using an Acoustic Glottal Source Model, School of Informatics University of Edinburgh, 2010.

[2] R. Boite, H. Bourland, T. Dutoit , J. Hancq and H. Leich "Traitement De La Parole ", vol. 1, 1999, pp.345-435.

[3] Speech Signal Processing Toolkit (SPTK), http://sp-tk.sourceforge.net

[4] K. Tokuda, Speech Parameter Generation Algorithm for HMM-Based Speech Synthesis. HMM, Proc. ICASSP., 2000,Vol. 3, pp. 1314–1318.

[5] K. Tokuda, H. Zen and A. Black, An HMM-Based Speech Synthesis System Applied to English. IEEE TTS Workshop 2002. Santa Monica. California, USA. 2002.

[6] A. Acero, Formant Analysis and Synthesis Using HiddenMarkov Models.EUROSPEECH, Budapest, Hungary, 1999, p.p. 1047-1050.

[7] R. E. Donovan and P. C. Woodland, A hidden Markov-model-based trainable speech synthesizer. Computer Speech and Language, 1999, Vol. 1, pp. 1–19.

[8] M. J. Barros, HMM-based European Portuguese TTS System, INTERSPEECH '05, Lisbon, Portugal, 2005, pp. 2581–2584.

[9] S. Young, The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department, Cambridge, Great Britain, 2002.

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

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

[12] X. Gonzalvo, I. Iriondo, C.J. Socoro, F. Alias, and C. Monzo, "HMM-based Spanish speech synthesis using CBR as F0 estimator", Proc. of NoLISP, 2007.

[13] A. Black, H. Zen, and K. Tokuda, Statistical parametric speech synthesis. In Proc. of ICASSP, Hawaii, USA,2007.

[14] K. Mohamed Khalil, C. Adnan," Improvements of Arabic database and Noise Reduction of Speech Signal using Wavelet for Arabic speech synthesis system using HMM: HTS_ARAB_TALK", Journal of Information Hiding and Multimedia Signal Processing, JIHMSP.

[15] H. Kawahara. "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. Acoustical science and technology" 2006.

[16] X. Gonzalvo, J. Claudi, I. Iriondo, C. Monzo and E. Martinez, "Linguistic and Mixed Excitation Improvements on a HMM-based speech Synthesis for Castilian Spanish", 2007.

[17] H. Zen, and T. Tomoki, "An overview of nitech HMM based speech synthesis system for blizzard challenge 2005", 2005 Proc.of Interspeech, pp.93–96.

[18] S. Krstulovic, A. Hunecke and M. Schroeder, An HMM-Based Speech Synthesis System applied to German and its Adaptation to a Limited Set of Expressive Football Announcements. In: Proc. of Interspeech 2007.

[19] Larousse. http://www:larousse:fr/dictionnaires/francais.

[20] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, XA Liu and G. Moore, J. Odell, D. Ollason, D. Povey, et al. The htk book (for htk version 3.4), 2006.

[21] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda. The HMM-based speech synthesis system (HTS) version 2.0, p. 1, 2006.

[22] K. Oura, An example of context-dependent label format for HMM-based speech synthesis in English, July 2011.

[23] K. Oura, An example of context-dependent label format for HMM-based speech synthesis in English, July 2011.

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

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

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

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

[28] K. Tokuda and H. Zen. Fundamentals and recent advances in hmm-based speech synthesis.Tutorial of INTERSPEECH, 2009.

[29] M. Boudraa, B. Boudraa, B. Guerin, ''Elaboration d'une base de données arabe phonétiquement équilibrée'', Actes du colloque Langue Arabe et Technologies Informatiques Avancées, pp 171-187, Casablanca, Décembre 1993.