International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 12, December - 2013

HTS_ARAB_TALK: A new implementation of an Arabic Speech Synthesis System

Mohamed Khalil Krichi, Adnan Cherif
Science Faculty of Tunis SFT 1060, Tunisia

Abstract

A statistical parametric speech synthesis system based on hidden Markov models (HMMs) has grown in popularity over the last few years. In this approach the system simultaneously models spectrum, excitation, and duration of speech using context-dependent HMMs and generates speech waveforms from the HMMs themselves. This paper describes a new system of Arabic speech synthesis called HTS_ARAB_TALK which represents a complete architecture system by modifying the publicly available HTS. This new implementation is based on statistical parametric speech synthesis which is a relatively new approach to speech synthesis. A brief description of statistical parametric speech synthesis system based on hidden Markov models (HMMs) is presented with some emphasis on the feature that is relevant to the Arabic language. Finally, a mean opinion score for the synthesized speech is presented. These results were supported by subjective evaluation.

1. Introduction

HMMs have been used for ASR since several decades ago. Meanwhile, HMM-based speech synthesis is a more recent application of the HMM in speech technology. Although many speech synthesis systems can synthesize high quality speech, they still cannot synthesize speech with various voice characteristics such as speaker individualities, speaking styles, emotions, etc. To obtain various voice characteristics in speech synthesis systems based on the selection and concatenation of acoustical units, a large amount of speech data is necessary. However, it is difficult to collect store such speech data. In order to construct speech synthesis systems which can generate various voice characteristics, the HMM-based speech synthesis system (HTS) [1] was proposed. Speech synthesis is defined as the process of generating speech signal by machine. This target can be accomplished using many ways. The traditional way is waveform concatenation such us PSOLA [1]. This technique has shown to synthesize high quality, typically more natural sounding speech, now the RealSpeak from Nuance and AT&T Labs Text-to-Speech (TTS) are famous concatenative commercial speech technology systems for TTS [2]. But concatenative systems have the drawbacks of limited number of voices and large size memory used to store speech waveforms. Another way for speech synthesis is through software using linguistic rules and features based on analyzing human speech. So this method sometimes called rule-based synthesis, formant speech synthesis or parametric synthesis since it generates small compact parameters from human speech and uses them to synthesize speech signal. DECtalk is still the best commercial formant synthesizer [2]. Formant synthesizers have the advantages of using small memory since the size of the extracted parameters are less than the size of the speech signal in waveform and the easy customization of synthesized voices. But they have the disadvantage in the generated sound that it is more mechanical sounding so it results less quality than concatenative ones. Statistical parametric speech synthesis system based on hidden Markov models (HMMs). HMM is a new approach that has grown in the last few years witch has been proved as a powerful tool in speech recognition since the models produced from the training process contain statistical data that models; the input speech signal and these models have small size. 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 describes the overall architecture, several components of the system, and linguistic concepts for Arabic. This paper is structured as follows. Section 2 describes the HMM-based Speech Synthesis. Section 3, describes the development of HTS_ARAB_TALK. Section 4, describes the evaluation result. Finally, section 5 summaries our conclusions and an expected future work.

2. HMM-based Speech Synthesis

This section describes an example of the HMM-based speech synthesis system based on the algorithm for speech parameter generation from HMM with dynamic features described in previous sections, and several results of subjective experiments on quality of synthesized speech.

2.1 System Overview

Figure 1 shows a block diagram of the HMM-based speech synthesis system. The system consists of two stages; the training stage and the synthesis stage. First, in the training stage, mel-cepstral coefficients are obtained from speech database by mel-cepstral analysis [27]. Dynamic features, i.e., delta and delta-delta mel-cepstral coefficients, are calculated from mel-cepstral coefficients. Then phoneme HMMs are trained using mel-cepstral coefficients and their deltas and delta-deltas. In the synthesis stage, an arbitrarily given text to be synthesized is transformed into a phoneme sequence. According to this phoneme sequence, a sentence HMM, which represents the whole text to be synthesized, is constructed by concatenating phoneme HMMs. From the sentence HMM, a speech parameter sequence is generated using the algorithm for speech parameter generation from HMM. By using the MLSA (Mel Log Spectral Approximation) filter [28], [29], speech is synthesized from the generated mel-cepstral coefficients.

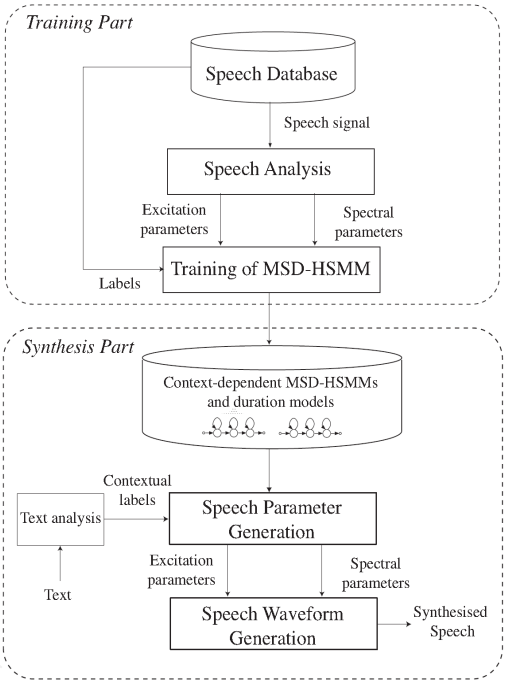

Fig.1 Synthesis stage of HMM speech synthesis.[5]

2.2 Speech Database

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 [17]. Arabic vowels are affected as well by the adjacent phonemes. The ideal is to have a database sufficiently provided with several examples phonemes [15]. The database used in [16] is without prosodic information. Our target is to improve the database in [16].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 "?". In this work, we added in each labels file a lot of prosody information.

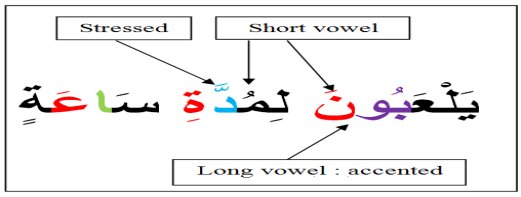

Fig.2 Example of a sentence / jalAIabuuna limuddati saAItin / ("They play for an hour")

The Arabic letters are written from right to left and most of them are attached to one another. Most Arabic words can be reduced to a root which often consists of three letters. Modifying this root by adding prefixes and/or suffixes and changing the vowels results in many word patterns. The target is to obtain a label file for each sentences, in each file, there are once sentences. In the following next figure, we explain the information involves in each utterance.

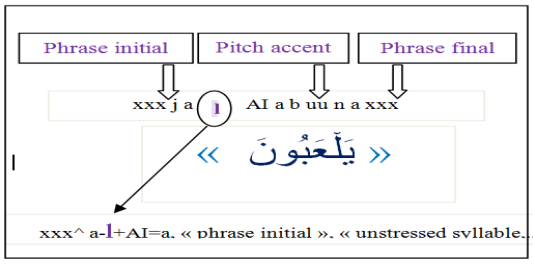

Fig.3 sentences Arabic with prosodic information

In Arabic, word–stress and its placement are predictable because if we take the structural patterns of the word, then rules can be formulated so as to pinpoint the syllable on which stress falls. Word–stress, therefore, is non–phonemic in Arabic [23] Arabic stress does not produce a distinction in meaning. Most linguists and orientalists, nevertheless, have distinguished three degrees of non–phonemic stress: primary, secondary and weak. [24] For instance, stating a general rule of word–stress placement in Arabic, maintains that: stress falls on the long syllable nearest to the end of the word. In the absence of a long syllable, the stress falls on the first syllable and on the third syllable from the end in words of three or more syllables [25] extensively discusses accentuation and other phonological phenomena related to syllable structure in classical Arabic.

2.3 HMM-Training

All HMMs used in the system were left-to-right models with no skip. Each state had a single Gaussian distribution with the diagonal covariance. Initially, a set of monophone models was trained. These models were cloned to produce a triphone models for all distinct triphones in the training data. The triphone models were then reestimated with the embedded version of the Baum-Welch algorithm. All states at the same position of triphone HMMs derived from the same monophone HMM were clustered using the furthest neighbor hierarchical clustering algorithm [30]. Then output distributions in the same cluster were tied to reduce the number of parameters and to balance model complexity against the amount of available data. Tied triphone models were reestimated with the embedded training again. Finally, the training data was aligned to the models via the Viterbi algorithm to obtain the state duration densities. Each state duration density was modeled by a single Gaussian distribution. In this work, we use a first step in HTS that a training part. All of parameters are obtained by a HTS system. After this training, we send all of parameters in HTS-Engine but this step need other file.

2.4 Question file

Questions file 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 6 is a question file extract.

```
QS "R-Word_GPOS==0"          {*/F:0_*}
QS "R-Word_GPOS==aux"        {*/F:aux_*}
QS "R-Word_GPOS==cc"         {*/F:cc_*}
QS "R-Word_GPOS==content"    {*/F:content_
QS "R-Word_GPOS==det"        {*/F:det_*}
```

Fig.4 Extract a file questions

2.5 Speech Synthesis

In the synthesis part, an arbitrarily given text to be synthesized is converted to a phoneme sequence. Then triphone HMMs corresponding to the obtained phoneme sequence are concatenated to construct a sentence HMM which represents the whole text to be synthesized. Instead of the triphones which did not exist in the training data, monophone models were used. From the sentence HMM, a speech parameter sequence is generated using the MLSA (Mel Log Spectral Approximation) filter [23], [24], speech is synthesized from the generated mel-cepstral coefficients directly.

3. Development of HTS_ARAB_TALK

Figure 5 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 [31].

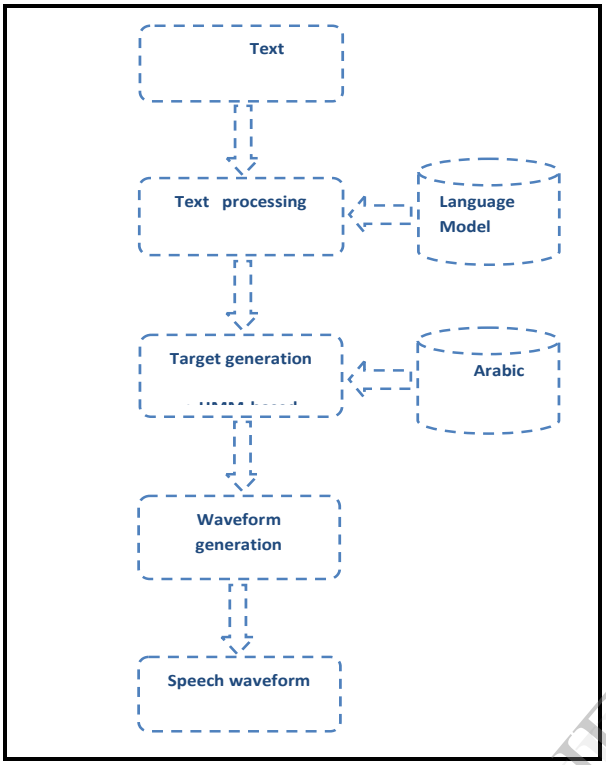

Fig.5 Block diagram of HTS-ARAB-TALK

3.1 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 text in hierarchal order: higher level structures to paragraphs, paragraphs to sentences, sentences to words and words to syllable 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.

3.2 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 releases. 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 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.

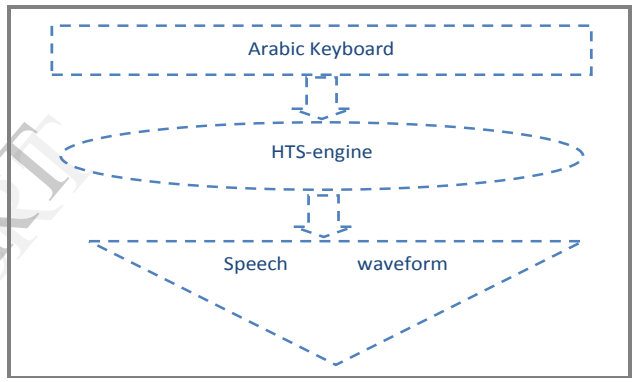

Fig.6 HTS_ARAB_TALK

4. Subjective evaluation

The only concern when choosing the test group is that they should be non-speaking of the Arabic language. In order to decide what a good command is, it was decided that the participants should have the Arabic language as their second language. The group consists of 12 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. The main goal of this evaluation test is to determine how much of the spoken output one can understand is. Evaluation consists of a subjective comparison between the 4 models. A comparison category rating (CCR) test was used to compare the quality of the synthetic speech generated by our system, Euler system, Acapela system and natural speech models. The participants were asked to attribute a preference score according to the quality

of each of the sample pairs on the comparison mean opinion score (CMOS) scale [33]. Listening test was performed with headphones. After collecting all listeners' response, we calculated the average values and we found the following results. These results are shown in fig.7

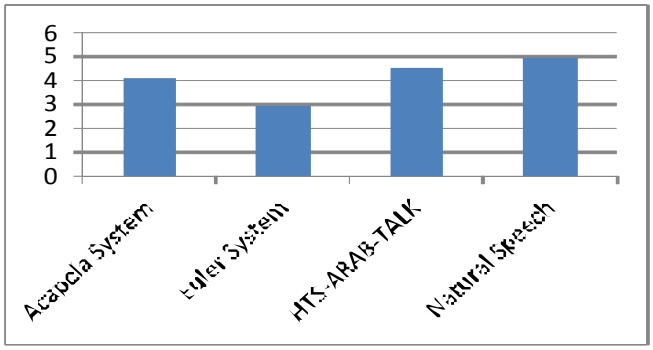

Fig.7 Average scores for the first test (system Euler, our system, natural speech and Acapela system. for the intelligibility of speech.

5. 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 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.

6. References

[1] Cheng-Yuan, L. and Jang, J. "A two-phase pitch marking method for TD-PSOLA synthesis" ICSLP, 2004.
[2] Yorozu, Y. Hirano, M. Oka, K. Tagawa, and Y. "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[3] Fukada, T. Tokuda, K. Kobayashi T. and Imai, S. "An adaptive algorithm for mel-cepstral analysis of speech," Proc. of ICASSP'92, vol.1, pp.137–140, 1992.
[4] http://kt-lab.ics.nitech.ac.jp/˜tokuda/SPTK/.
[5] Tokuda, K. Masuko, T. Miyazaki N. and Kobayashi, T. "Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling," Proc. of ICASSP, 1999.
[6] Toth, B. and Nemeth G. "Optimizing HMM Speech Synthesis for Low-Resource Devices", November 15, 2011.
[7] Tokuda, K. Zen, H. and Black, A.W. "An HMM-based speech synthesis system applied to English", in IEEE Speech Synthesis Workshop, 2002.
[8] Assaf, M. "A Prototype of an Arabic Diphone Speech Synthesizer in Festival," Master Thesis,Department of Linguistics and Philology, Uppsala University, 2005.
[9] Zen, H. Tokuda, K. and Kitamura, T. "Decision tree based simultaneous clustering of phonetic contexts, dimensions, and state positions for acoustic modeling", in Proc. Eurospeech, 2003b, pp. 3189-3192.
[10] Tokuda, K. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, T. "Speech parameter generation algorithms for HMM-based speech synthesis", in Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2000, Vol. 3, pp. 1315-1318.
[11] Fukada, T., Tokuda, K., Kobayashi, T., Imai, S., "An adaptive algorithm for mel-cepstral analysis of speech", in Proc. Of ICASSP92, 1992, vol.1, pp.137-140.

[12] Shichiri, K., Sawabe, A., Tokuda, K., Masuko, T.,Kobayashi, T., Kitamura, T., "Eigenvoices for HMM-based speech synthesis", in Proceedings of International Conference on Spoken Language Processing,2002, pp. 1269–1272.
[13] Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T., "Adaptation of pitch and spectrum for HMM-based speech synthesis using mllr", in Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2001, Vol. 2, pp. 805808.
[14] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., "Speaker interpolation in HMM-based speech synthesis system", in Proceedings of European Conference on Speech Communication and Technology97, 1997, Vol. 5, pp. 2523-2526.
[15] 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.
[16] K. Mohamed Khalil, C. Adnan," Arabic HMM-based Speech Synthesis ", in International Conference on Electrical Engineering and Software Applications ICEESA 2013.
[17] M. Assaf, "A Prototype of an Arabic Diphone Speech Synthesizer in Festival," Master Thesis,Department of Linguistics and Philology, Uppsala University, 2005.
[18] Eriwn, W. M. (1963) A Short Reference Grammar of Iraqi Arabic.Washington: Georgetown University Press.
[19] Mitchell, T F (1975) Principles of Firthian Linguistics. London: Longman.
[20] Kawai, H. Toda, T. Yamagishi, J. Hirai, T. J. Ni, Nishizawa, T., Tsuzaki, M. and Tokuda. K. XIMERA: A concatenative speech synthesis system with large scale corpora. IEICE Trans. Inf. Syst. (Japanese Edition), J89-D(12):2688–2698, Dec. 2006.
[21] Black, A.W. Taylor, P. and Caley, R. The festival speech synthesis system .http://www.festvox.org/festival/. Young M., The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
[22] Schroder M. and Trouvain J. The German text-to-speech synthesis ¨ system MARY: A tool for research, development and teaching. International Journal of Speech Technology, 6:365–377, 2003.
[23] Omar, A. (1985) Dirasat Al–Swat Al–Lugawi.Cairo: Alam Al– Kutub.
[24] Eriwn, W. M. (1963) A Short Reference Grammar of Iraqi Arabic.Washington: Georgetown University Press.
[25] Mitchell, T F (1975) Principles of Firthian Linguistics. London: Longman.
[26] Tokuda, K. Zen, H. and Black A., An HMM-Based Speech Synthesis System Applied to English. IEEE TTS Workshop 2002. Santa Monica. California, USA. 2002.
[27] T. Kato, T. Masuko, T. Kobayashi, and K. Tokuda, "Pitch pattern generation using paralell HMMs with multi-space probability distribution," ASJ Spring meeting, 1-7-19, pp.217–218, Mar. 1998 (in Japanese).
[28] K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mel-cepstral analysis using frequency transformation based on second-order all-pass function," ASJ Spring meeting, 3-7-13, pp.279–280, Mar. 1998 (in Japanese).
[29] T. Wakako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Evaluation of mel-cepstral analysis using frequency transformation based on second-order all-pass function in speech analysis-synthesis," ASJ Spring meeting, 3-7-14, pp.281–282, Mar. 1998 (in Japanese).
[30] F. Takahashi, T. Masuko, K. Tokuda, and T. Kobayashi, "A study on performance of a very low bit rate speaker independent HMM vocoder," ASJ Spring meeting, 2-P-23, pp.313–314, Mar. 1999 (in Japanese).
[31] K. Mohamed Khalil, C. Adnan," Optimization of Arabic database and an implementation for Arabic speech synthesis system using HMM: HTS_ARAB_TALK". International Journal of Computer Applications (0975 – 8887) Volume 73– No.17, July 2013.
[32] K. Tokuda, H. Zen, A.W. Black, An HMM-based speech synthesis system applied to English, Proc. of 2002 IEEE SSW, Sept. 2002.
[33] K. S. Rao and B. Yegnanarayana, (4-8 October 2004) "Intonation modeling for Indian languages", in Proccedings of Interspeech'04,Jeju Island, K0rea, pp733-736.