Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **1**, January **2015**

Improvements of Arabic database and Noise Reduction of Speech Signal using Wavelet for Arabic speech synthesis system using HMM: HTS-ARAB-TALK

Mohamed Khalil Krichi

Department of Physics
FST-Faculty of Sciences de Tunis
Campus Universities 2092 - El Manar Tunis, Tunisia

Adnan Cherif

Department of Physics, FST-Faculty of Sciences de Tunis
Campus Universities 2092 - El Manar Tunis, Tunisia

Received May, 2014; revised August, 2014

Abstract. *In this paper, we present an optimization of the database for Arabic speech synthesis system using HMM: $HTS - ARAB - TALK$. Our developed synthesis system uses phonemes as HMM synthesis unit. The database used in speech synthesis based on Hidden Markov model consists of two types of data, audio and descriptive file (labels). Our objective is to improve database used in denoising wav files and add prosody information in labels. Statistical parametric speech synthesis is a relatively new approach to speech synthesis. It generates speech synthesis based on Hidden Markov model , the techniques in this approach, has been demonstrated to be very efficient in synthesizing high quality, natural and expressive speech. The developed model improves the quality of the naturalness, and the intelligibility of speech synthesis in various speaking environment.*

Keywords: HMM, Speech Synthesis, Text to Speech, Arabic Language, Statistical Parametric Speech Synthesis, Hidden Markov Model, Wavelet.

1. **Introduction.** HMM based speech synthesis [1] is a new technique relative to other synthesis techniques, and it seems promising. In this approach, context-dependent HMMs are estimated from databases of natural speech or database made in a more comfortable environment but denoising step is extremely necessary, and speech waveforms are generated from the HMMs themselves, but the addition of a filter stage is a help to get a more natural signal. Then to synthesize speech, parameters are generated from these HMM models according to the input text, then speech is synthesized from these parameters. A text to speech (TTS) synthesizer is a computer based system that should be able to read any text aloud. Typical TTS systems have two main components, text analysis and speech waveform generation, which are sometimes called frontend and backend, respectively. In the text analysis component, given input text is converted into a linguistic specification consisting of elements such as phonemes. In the speech waveform generation component, speech waveforms are made from the produced linguistic specification. These systems are only applicable when a limited vocabulary is required, and when sentences

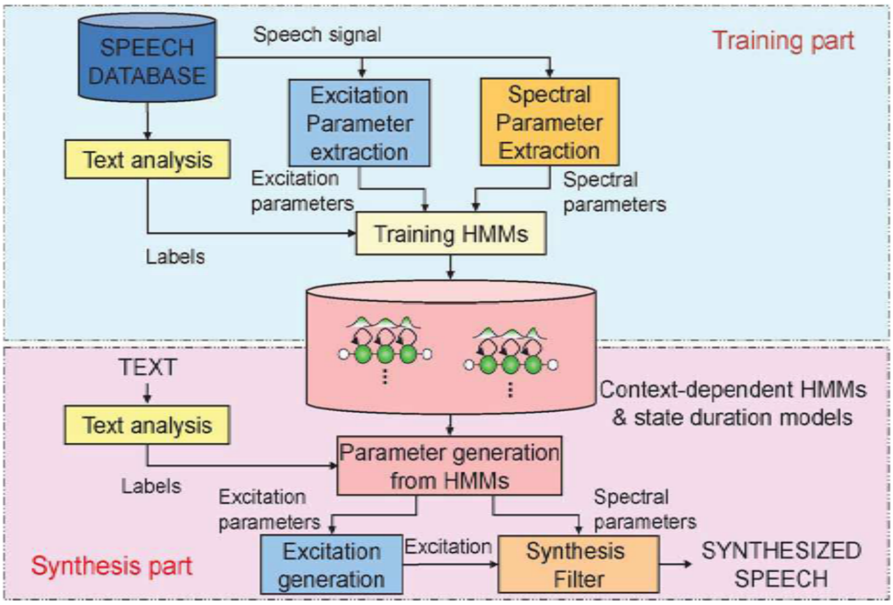


Figure 1. An overview of the basic HMM-based speech synthesis system [5]

to be pronounced have a very restricted structure, as in the case for the announcement of arrivals of train on a railway station for instance. In the context of TTS synthesis, it is impossible to record and store all the words of the language. It is thus more suitable to define TTS as the automatic production of speech [2]. Database preparation is one of the most necessary parts in TTS systems. The ideal is to have a database sufficiently provided with several examples phonemes [3]. The database used in [4] is without prosodic information. Our goal is to improve the database in [4].The audio portion of the database is the only one that interested as in this study wav format is PCM coded 16-bits at a sampling frequency of 16 kHz. But the database used in our system is done in a noise-free environment, but the existence of audio noise in the files is on. Therefore the addition of a filtering stage is necessary. The use of wavelet-based filtering is the most widely used thanks to their simplicity and efficiency.

The rest of this paper is organized as follows. Section 2 summarizes the previously proposed HTS (HMM-based speech system). Sections 3 describes the environment preparation and optimization, noise reduction for database and improve the descriptive file for database (labels) and explain another method of extraction of features. Results from objective and subjective evaluations are presented in Section 4, and concluding remarks and our plans for future work are presented in the final section.

2. **HTS System.** The HTS system is a popular HMM-based speech synthesizer, which is available online [5]. The basic structure of this system is shown in Figure 1. Most HMM-based speech synthesizers have a similar structure, which can be divided into the analysis, training and synthesis parts. During analysis, excitation and spectral parameters are extracted for each utterance of the speech corpus. For example, logF0 is generally used as an excitation parameter. The spectral parameters are often defined by mel-cepstral coefficients or line spectral frequencies, which are adequate features for statistical modeling. The phonetic labels can be obtained from the text, e.g. by using a text analyzer. Typically, they also have context information, such as phone identity, phone boundaries, syllable, etc. The time label boundaries do not need to be estimated if the speech database is phonetically labeled or if a flat-start training of the HMMs is to be used. Otherwise,

they can be calculated from the recorded utterances and their text transcriptions using a time alignment technique, such as the Viterbi algorithm [4]. In the training part, the phonetic labels and the speech features are used to model context-dependent HMMs. In this process the statistical parameters of the HMMs are calculated. Then, decision trees which describe all the contextual factors are used to cluster the trained HMMs. In a typical statistical parametric speech synthesis system, we first extract parametric representations of speech including spectral and excitation parameters from a speech database and then model them by using a set of generative models (e.g., HMMs). A maximum likelihood (ML) criterion is usually used to estimate the model parameters as [1]:

$$\hat{\lambda} = \arg\max_{\lambda} \left\{ p(\frac{o}{\omega}, \lambda \right\} \tag{1}$$

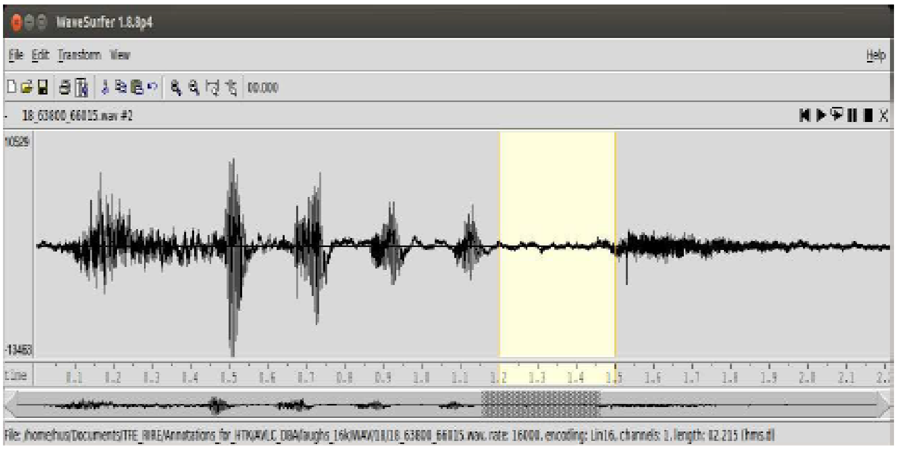
where λ is a set of model parameters, O is a set of training data, and ω a set of word sequences corresponding to O. The synthesis part of the system is shown in the lower part of Figure 1. It first converts a given text to be synthesized into a sequence of context-dependent labels. According to the label sequence, a sentence-level HMM is constructed by concatenating context-dependent HMMs. The duration of each state is determined to maximize its probability based on its state duration probability distribution [7]. Then a sequence of speech parameters including spectral and excitation parameters is determined so as to maximize its output probability in (2) using the speech parameter generation algorithm [8]. 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 as [1]:

$$O = \arg\max_{O} \left\{ p(\frac{o}{\omega}, \lambda \right\} \tag{2}$$

Finally, a speech waveform is reconstructed from the parametric representations of speech. Although any generative model can be used, HMMs have been widely used. Statistical parametric speech synthesis with HMMs is commonly known as HMM-based speech synthesis [9].

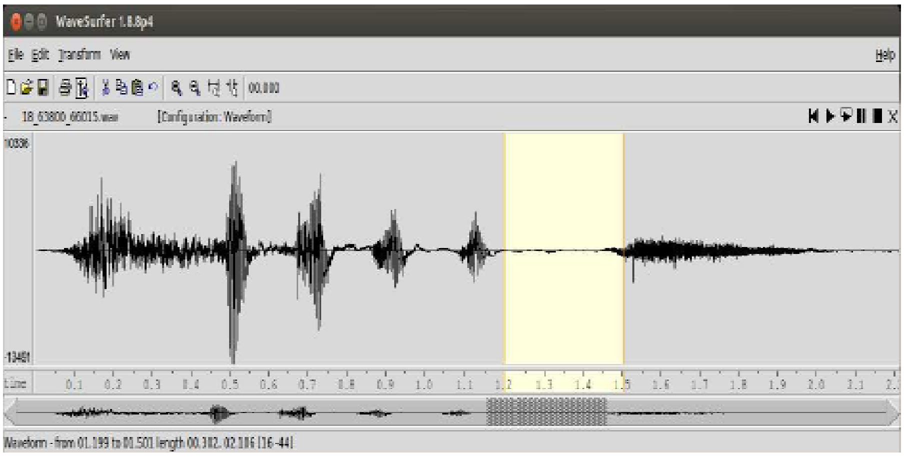
3. **Environment preparation and optimization.** Standard Arabic is the language used by media and the language of Quran. Modern Standard Arabic is generally adopted as the common medium of communication through the Arab world today. Standard Arabic has 34 basic phonemes, of which six are vowels, and 28 are consonants [10]. The databases compose two parts, one is all of sentences in wave form and the seconds is a description phoneme by phoneme name's labels. 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. The first step is to use a stage for reduce noising.

3.1. **Noise Reduction.** Degradation of signals by noise is an omnipresent problem [11]. In almost all fields of signal processing the removal of noise is a key problem. The wavelet transform is striking for its great variety of different types and modifications. A whole host of different scaling and wavelet functions (or scaling and wavelet coefficients) provide plenty of possible adjustments and regulating variables [12]. The audio recordings were noisy with a continuous background noise. Our goal is to reduce this undesirable component. Figures 2 and 3 shows the time signal before and after filtering for a particular audio file. We note in particular that the zone of silence highlighted is closer to zero in the filtered signal in the original signal release.



File: /home/hus/Documents/TFE_RIRE/Annotations_for_HTK/AVLC_DB/laughs_16k/WAV/18/18_63800_66015.wav, rate: 16000, encoding: Lin16, channels: 1, length: 02.215 (hms.d)

Figure 2. Example for original speech of database file

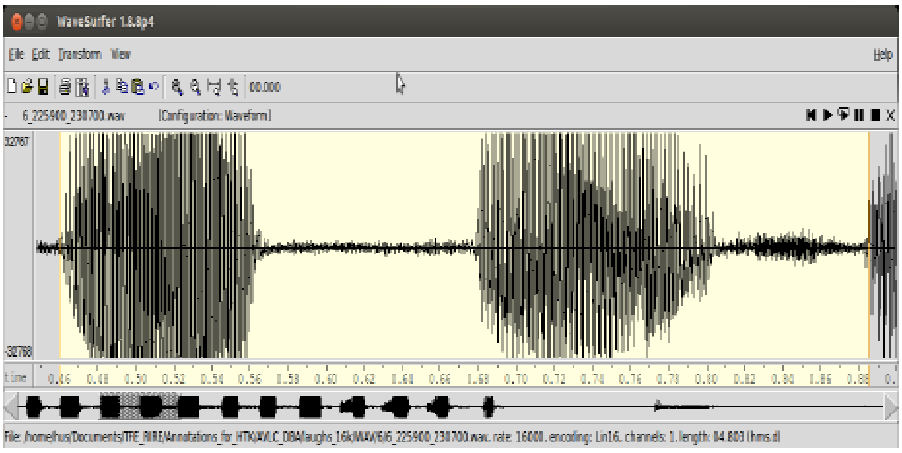


Waveform - from 01.199 to 01.501 length 00.302, 02.186 [16-44]

Figure 3. Example for the same speech denoising

3.2. **Reconstitution saturated Files.** In addition, some files, including a relatively large amount for one of the subjects we studied, showed unwanted saturation due to the very strong expressiveness of sound and tenderness high e microphone when recording. This problem has also been treated to replenish saturated signals [13]. Figures 4 and 5 shows the time signal before and after reconstruction for audio file. Represents the signal also suffered filtering noise preceding paragraph. We can notice that the quiet zones are mitigated and that the two areas High energy is more saturated

3.3. **Prosody information.** In Arabic, wordstress 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. Wordstress, therefore, is nonphonemic in Arabic [14] Arabic stress does not produce a distinction in meaning. Most linguists and orientalists, nevertheless, have distinguished three degrees of nonphonemic stress: primary, secondary and weak. [15] For instance, stating a general rule of wordstress 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 [16] extensively discusses accentuation and other phonological phenomena related to syllable structure in classical Arabic. His approach is prosodic. To him, a final syllable of the word is stressed if it is



File: /home/hus/Documents/TFE_BIRE/Annotations_for_HTK/AVLC_DBA/laughs_16k/WAV/6/6_225900_230700.wav. rate: 16000. encoding: Lin16. channels: 1. length: 04.803 (hms.d)

Figure 4. Example for original speech of database file

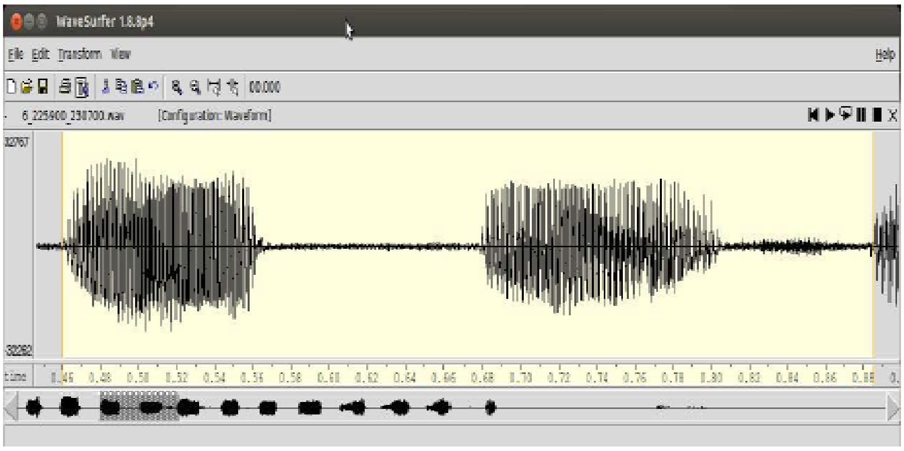
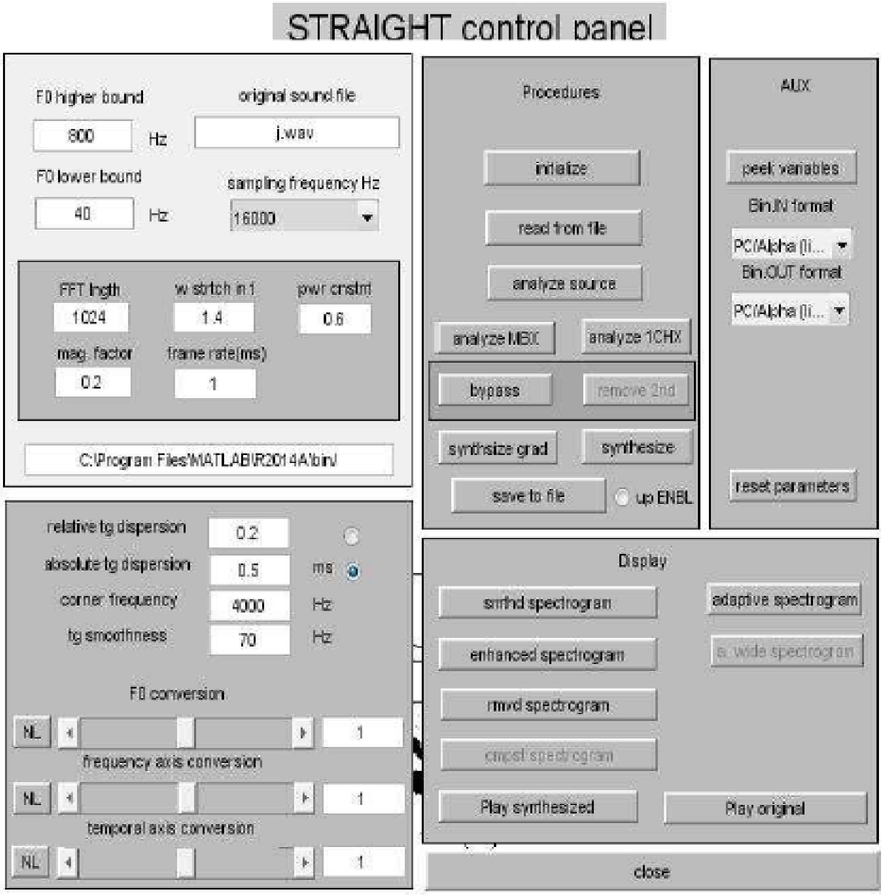


Figure 5. Example of the same sound without saturation after reconstitution

long, i.e. VVC(C) or VCC. He does not consider VV# as a long vowel, e.g.
/Aakaltuu/ I eat
/Aamshi/ I go
If the pre-final syllable is closed, that is of CVC, or CVVC it will be stressed, e.g.
/kataba/ he wrote
/AiTnaani/ these two
But if the prefinal syllable is open, i.e. of the form CV, then either that syllable or the syllable preceding it is stressed, e.g. / katabta / You wrote
/ kaabaltu / he corresponded with
/ qattalat / she murdered
/ Aakalataa / they (feminine) eat
Another method of extraction of features
In the foregoing, the methods used in the demo HTS were taken, possibly with adaptations of parameters to perform feature extraction. As a reminder, the extraction of MFCCs is using SPTK - 3.4.1[17] software while extracting the fundamental frequency is using the Snack 2.2.10 software [18]. Other software can perform such operations. One of these, which is recognized in the field of speech processing, is STRAIGHT [19], developed by Professor Hideki Kawahara (Auditory Media Laboratory - Department of Design Information Sciences Faculty of Systems Engineering, Wakayama University). So



STRAIGHT control panel

F0 higher bound 800 Hz

original sound file j.wav

F0 lower bound 40 Hz

sampling frequency Hz 16000

FFT lngth 1024

w strtch in1 1.4

pwr cnstrnt 0.6

mag. factor 0.2

frame rate(ms) 1

C:\Program Files\MATLAB\R2014A\bin\

relative tg dispersion 0.2 ○

absolute tg dispersion 0.5 ms ●

corner frequency 4000 Hz

tg smoothness 70 Hz

F0 conversion 1

frequency axis conversion 1

temporal axis conversion 1

Bin.IN format PC/Alpha (li...

Bin.OUT format PC/Alpha (li...

Figure 6. STRAIGHT-Tools for extraction parameter

we could make first use of the tools provided by STRAIGHT using extraction of MFCCs and F0, the rest of the scripts being unchanged for the execution of training and synthesis. STRAIGHT by extraction provides better results than previous methods. We could expect this conclusion because in the case of speech, the tools provided by STRAIGHT significantly improve the quality of the synthesis. For speech analysis, it extracts the spectral envelope, F0 and aperiodic parameters of the speech signal with STRAIGHT (V40) [20]. Figure 6 represents a toolkit, which uses STRAIGHT to extract several data, all necessary information to get a good sound quality

4. **Experiments.** For our tests we used the HTS-ARAB-talk that is provided in [21]. Speech was sampled at 16 kHz and we used 3-state left-to-right HMMs. At run time, we use full label files, where each line is an alpha-numerical character sequence encoding all the information listed above for one phoneme, which is then input into the system. We synthesized speech produced by both system HTS-ARAB-TALK optimized and old system. We evaluated the quality of speech synthesized by using both objective and subjective measurements. Cong Thuc

Where mcd are mel-cepstral coefficients generated by the two different systems, and D is the mel-cepstral, S2 is HTS-ARAB-TALK with optimization and S1 is a simple HTS-ARAB-TALK coefficient order. We applied this metric to the mel-cepstral coefficients generated by two test system and the results are presented in Table.

4.1. **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

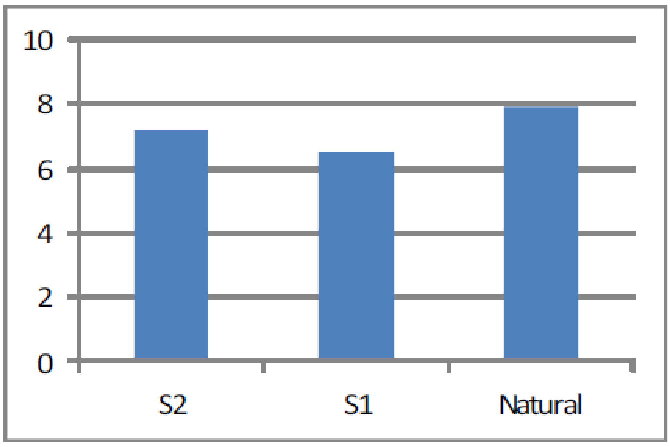


FIGURE 7. Average scores for the test (HTS-ARAB-TAL (S1), HTS-ARAB-TALK using new database and STRAIGHT vocoder (S2) and natural speech for the intelligibility of speech

extent. The distance measure considered here is the mel-cepstral distortion between the both systems, expressed as [22]:

4.2. **Subjective Evaluation.** In order to confirm the objective evaluation conclusion, we performed a 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 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. For this evaluation, the listener was asked to compare three sentences: A, the original (natural); B, the sentence synthesized by HTS-ARAB-TALK (S1); C, the sentence synthesized by HTS-ARAB-TALK using new database and STRAIGHT vocoder (S2). He was asked to score, on a 9-point scale, the overall speech quality of C in comparison with A and B. C was allowed to vary from 0 (= same quality as A) to 9 (= same quality as B). Therefore this score should be interpreted in terms of a ’distance’ between B and A and C: the lower the score, the more C ’sounds like’ A and thus the better the quality, and conversely. The test consists of 15 sentences. Before starting the test, the listener was provided with some reference sentences covering most of the variations to help him familiarize with the scale. During the test, he was allowed to listen all of sentences as many times as he wanted, in the order he preferred. Howmore B ’sounds like’ A and thus the better the quality, and conversely.

5. **Conclusion and future works.** In this work, we proposed a HMM-based speech synthesis system for Arabic language using a change HTS-ARAB-TALK. HMM-based speech synthesis has started to be used in daily life, e.g., cell-phones, Smartphone, in-car navigation systems, and call centers. With this modification the naturalness obtained by HMM-based synthesis using a phase model like STRAIGHT can be approached with a better quality with few parameters. It also allows an intuitive control of the voice quality which is of great interest for expressive speech synthesis or to quickly synthesize different speaker personalities with various voice qualities from the same voice. The main objective

is to use the HTS to establish a new system TTS but in Arabic language. The objective and subjective evaluation showed that speech produced by the HMM-based TTS system, while more research is needed to enhance the quality of basic (databases so the development of new optimizations techniques are necessary) to choose the best unit in a large database that contains several versions of the units according to the optimization of prosodic parameters. Future research includes the improvement of the analysis stability and robustness (e.g. against high frequency artefact). 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] H. Zen, K. Tokuda, Black, A., Statistical parametric speech synthesis, speech communication, vol.5111, pp. 1039-1064, 2009.

[2] K. Tokuda, H. Zen and A.W. Black, An HMM-based speech synthesis system applied to English, *IEEE Speech Synthesis Workshop*, 2002.

[3] K. Mohamed Khalil, C. Adnan, Arabic HMM-based speech synthesis, *International Conference on Electrical Engineering and Software Applications*, ICEESA 2013.

[4] A. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Information Theory*, vol. IT-13, pp. 260-269.

[5] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, Speech parameter generation algorithms for HMM-based speech synthesis, *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1315-1318, 2000.

[6] H. Kawahara, STRAIGHT, a speech analysis, modification and synthesis system, http: //www.wakayama-u.ac.jp/$\tilde{k}$awahara/STRAIGHTadv/in dex_e.html

[7] M. Assaf, A Prototype of an Arabic diphone speech synthesizer in festival, *Master Thesis*, Department of Linguistics and Philology, Uppsala University, 2005.

[8] A. Omar, Dirasat AlSwat AlLugawi.Cairo: *Alam AlKutub*

[9] L. Hadjileontiadis et S. Panas. Separation of discontinuous adventitious sounds from vesicular sounds using a wavelet- based filter, *IEEE Trans. Biomed. Eng.*, vol. 44, no. 7, pp. 876-886, 1997.

[10] S. Mallat. A wavelet tour of signal processing, *Academic Press*, 1999.

[11] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method, *Proc. Blizzard Challenge Workshop*, 2006.

[12] W. Eriwn, A Short Reference Grammar of Iraqi Arabic, *Washington: Georgetown University Press.*

[13] T F. Mitchell, Principles of Firthian Linguistics. *London*: Longman.

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

[15] The Snack Sound Toolkit (Snack), http://www.speech.kth.se/snack/

[16] H. Kawahara, STRAIGHT - TEMPO: A universal tool to manipulate linguistic and para-linguistic speech information. *In Proc, IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Florida, USA.

[17] 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*, vol. 73, no. 17, pp. 0975-0987, July 2013.

[18] H. Kawahara, STRAIGHT, a speech analysis, modification and synthesis system, http://www.wakayama-u.ac.jp/ kawahara/STRAIGHTadv/in dex_e.html

[19] B. Picart, T. Drugman, T. Dutoit, Analysis and Synthesis of Hypo and Hyperarticulated Speech, *Proceedings of the Speech Synthesis Workshop 7* (SSW7), NICT/ATR, Kyoto, Japan, pp. 270-275, 2010.

[20] M. Boudraa, B. Boudraa, B. Guerin, Elaboration dune base de donnes arabe phontiquement e'quilibree, *Actes du colloque Langue Arabe et Technologies Informatiques Avancees*, pp. 171-187, Casablanca, Decembre 1993.