International Journal of Computer and Electrical Engineering, Vol. 2, No.2, April, 2010
1793-8163

The Effect of Speech Features and HMM Parameters on the Quality of HMM Based Arabic Synthesis System

Mohamed Samir Barakat [1] ,Mahmoud El-Said Gadallah [2]

***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. In this paper, the HMM-based speech synthesis system is applied to Arabic language using low size unsegmented speech training database. This technique shows that the resulting HMM set has the advantage of being small (can be less than 1MB) which is very important for communication applications. The basic contribution in this paper is to justify both the HMM parameters and the speech features to be suitable for using small speech database to get the highest quality. The motivation of this work is the starvation of the Arabic speech database. Experiments show that using Mel-cepstral coefficients as spectral parameters of speech waveforms for training gives better results than using LPC or PARCOR coefficients. Also, investigation tests show that increasing the context-dependent models length and the number of Gaussian Mixtures with this relatively small size training data has the disadvantage of poor generalization of HMMs that leads to perceivable discontinuities and clicks in the synthesized speech.**

***Index Terms*—Speech synthesis, Arabic language, Hidden Markov Modeling, and Mel-Cepstral coefficients.**

I. Introduction

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 (ex: PSOLA). 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 [1]. But concatenative systems have the disadvantages of limited number of voices and that they use a large size memory 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 [1]. Formant synthesizers have the advantages of using small memory since the size of the extracted parameters is 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 (less quality than concatenative ones). A new approach that has grown in the last few years is statistical parametric speech synthesis system based on hidden Markov models (HMMs). HMM 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.

The most common problem facing the development of any Arabic speech processing system is the starvation of Arabic labeled and segmented speech training database. So, the purpose of this paper is studying the development of HMM-based synthesizer for Arabic language and to find the length of context-dependent models, the spectral analysis method and the number of Gaussian mixtures that gives best quality and relatively low memory requirements using low size and unsegmented Arabic speech database.

This paper is organized as follows: sections II introduces a brief description for the speech synthesis system based on HMM. In section III, the effects of both HMM parameters and the spectral features of speech signals on the quality of the synthesized speech are investigated and tested. Also, in this section the experimental tests are introduced and the results are analyzed. Finally section IV outlines the important conclusions of this work.

II. HMM-Based Synthesis System

Fig. 1 shows the training and synthesis parts of the HMM-based TTS system [4],[6] and [7]. In the training part, spectral parameters (e.g., LPC, mel-cepstral coefficients, etc…) and excitation parameters (e.g., fundamental frequency) are extracted from speech database. The extracted parameters are modeled by context-dependent HMMs. In the synthesis part, a context-dependent label sequence is obtained and a sentence HMM is constructed by concatenating context dependent HMMs according to the context dependent label

Manuscript received July 19, 2009.

M. S. Barakat, He is now with the Department of Computer Science, Modern Academy, Maadi, Cairo, Egypt (e-mail: msamirhb@hotmail.com).

- M. E. GadAlla, was with Military Technical Collage, Cairo, Egypt. He is now with the Department of Computer Science, Modern Academy, Maadi, Cairo, Egypt (e-mail: mahmoud_mtc@yahoo.com).

sequence. By using parameter generation algorithm [2], spectral and excitation parameters are generated from the sentence HMM. Finally, by using a synthesis filter, speech is synthesized from the generated spectral and excitation parameters [7], [16] and [17]. Spectral and excitation parameters are needed for any synthesis filter to generate speech waveforms so both must be modeled by HMMs. Also HMMs have state duration densities to model the temporal structure of speech. So we need to construct HMMs with spectral and excitation parameters simultaneously to improve the synthesized speech in manner that we can regenerate them in the synthesis phase from the trained HMMs using the parameter generation algorithm. Training and synthesis parts of the system are explained with applying them to Arabic language in the following sections.

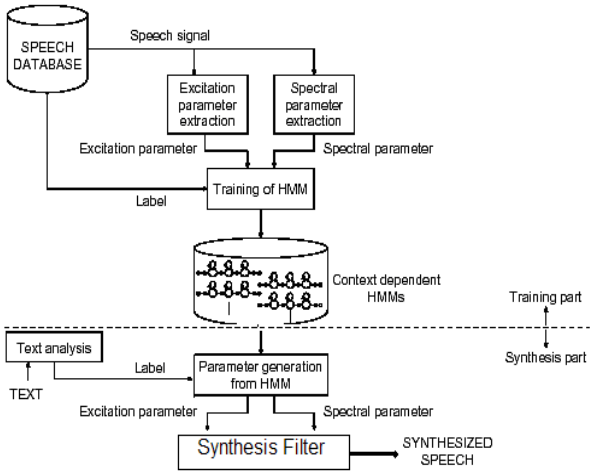


Fig. 1 The scheme of HMM-based synthesis system

A. *Training part*

In the training part, HMMs are built for excitation and spectral parameters for each speech unit. Spectral parameters are modeled using continuous distribution HMMs [4] but excitation parameters modeled using Multi-Space Distribution HMMs (MSD-HMM) to overcome the problem of the voiced and unvoiced regions [5]. If spectrum and pitch models are modeled separately, speech segmentations may be discrepant between them and this may cause discontinuities in the synthesized speech. To avoid this problem, context dependent HMMs are trained with feature vector which consists of spectrum, pitch and their dynamic features (By the inclusion of dynamic coefficients in the feature vector, the dynamic coefficients of the speech parameter sequence generated in synthesis are constrained to be realistic) but in different streams [3][6]. Also, state duration densities are modeled by single Gaussian distributions [11] to model the temporal structure (the speaking rate) of speech data. Now the stored HMMs set are ready to be used for synthesis.

Of course increasing training data will improve the generalization of the trained HMMs but as mentioned before it is very costly specially that there are a starvation in Arabic training data. For the proposed HMM-based Arabic speech synthesis system a database was built for Arabic language. Arabic independent sentences were written using 663 rich and balanced words. The database consisted of only 367 sentences; 2 to 9 words per sentence. The statistical results show that these 367 sentences contain 1835 words and 12940 graphemes [12]. These sentences recorded by a male speaker using personal computer and microphone in normal room not studio. To save effort and time required for segmentation, two issues are considered in the training:

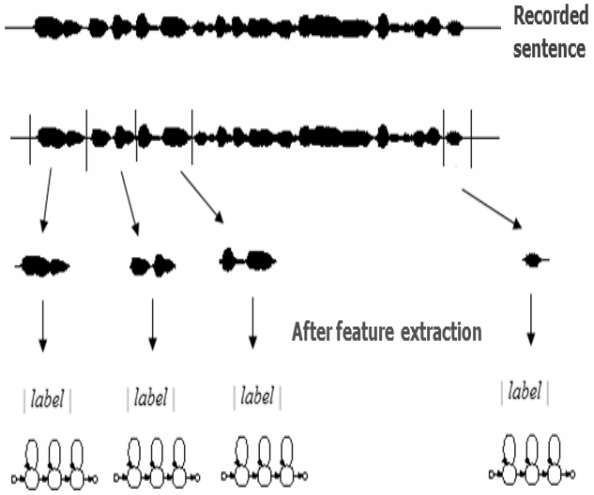
1) Using initial HMM parameters with flat assumptions.
2) Application of *embedded training* (described in the next subsection) instead of isolated unit training [8].

By this way, the transcription file will contain only the label sequence.

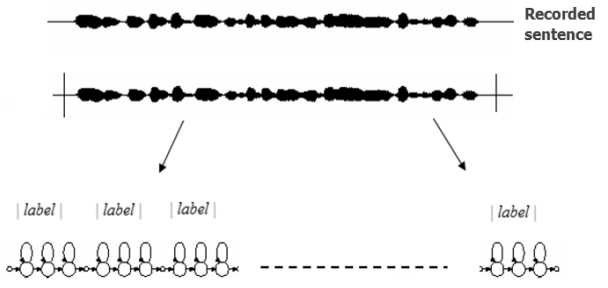
Another important issue in the training part is the length of context-dependent subword units. Which indicates those HMMs will be built for what (i.e., phoneme, diphone ...etc). This issue will be discussed in section 2.

3) Embedded Training

HMMs are built using embedded training strategy (not isolated unit training) since embedded training doesn't need phoneme boundaries to be provided in the training data [4], [8] and [13]. In the isolated unit training HMMs, are built for each speech phoneme and the examples of this phoneme (extracted feature vectors not waveforms) introduced to their corresponding HMMs to be re-estimated by the Forward-Backward algorithm. So the example sentences should be segmented to update each HMM individually. Embedded training does not update the HMM of a particular phoneme using its examples individually, but it simultaneously updates all HMMs in the system using all of the training data. In embedded re-estimation each training file (feature vector) is used with its transcription (label sequence) to construct a composite HMM which spans the whole utterance (sentence). This composite HMM is made by concatenating instances of the phone HMMs corresponding to each label in the label sequence. The Forward-Backward algorithm is then applied and the sums needed to form the weighted averages accumulated in the normal way [8].



(a) Isolated unit HMM training



(b) Embedded HMM training

Fig. 2 Isolated unit and Embedded HMM training

Fig. 2 shows the difference between Isolated unit training and embedded training.

4) Constructing context-dependent models

To take into account the contextual factors, context-dependent subword models should be constructed from the context independent subword models. This is done by cloning each monophone (phoneme) model with its right and left neighbors in the training (in case of triphone construction for example) data and performing embedded training again. This will result in a new set of context-dependent HMMs with new names that comprises the monophones model names. It should be noted that although, as the length of the constructed models increases, the contextual factors are more preserved but the number of models massively increases. For example if the monophone models number is *n* so the numbers of models after converting them to triphone models will be n^3. This increase of the number of new models requires more repetitions of the training data for each model to robustly build it with enough statistical parameters. But in the case of limited speech database (such as the speech database used in this work), the number of repetitions for each model will decrease and for some models, it may reach to only one or two repetitions. So, investigation has been done to get the best context-dependent models length using the available low size speech training database. State tying performed to reduce the total number of parameters without significantly altering the model's ability to represent the different contextual effects [8].

B. Synthesis part

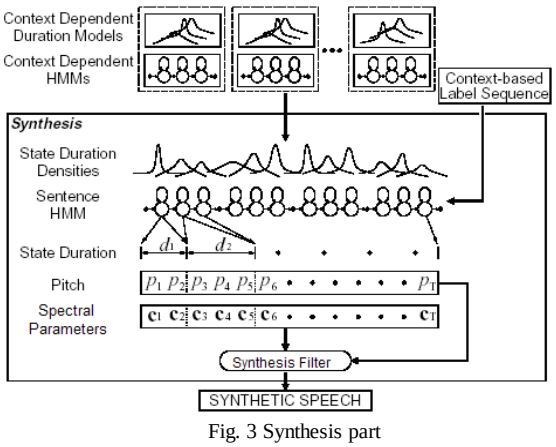


Fig. 3 Synthesis part

In the synthesis part shown in Fig. 3, to synthesize a speech signal a context label sequence that corresponds to the sequence of speech units to be synthesized is constructed first using Grapheme to Phoneme conversion algorithm. According to the determined sequence of labels, the corresponding HMMs are concatenated to form the intended sentence HMM. State durations of the sentence HMM are determined so as to maximize the output probability of state durations [11], [4]. Then a sequence of spectral and excitation parameters is calculated using the speech parameter generation algorithm. Finally, speech waveform is synthesized from these parameters by using the appropriate synthesis filter.

1) Speech parameter generation from HMMs

For a given continuous mixture HMM λ, an algorithm for determining speech parameter vector sequence

$$O = \left[o_1^T, o_2^T, \ldots, o_T^T\right]^T$$

in such a way that

$$P(O \mid \lambda) = \sum_{all\,Q} P(O, Q|\lambda)$$

is maximized with respect to O, where

$$Q = \{(q_1, i_1), (q_2, i_2), \ldots, (q_T, i_T)\}$$

is the state and mixture sequence, i.e., (q,i) indicates the *i*-th mixture of state *q*.

The speech parameter vector $\mathbf{O}_t$ consists of static feature vector $c_t = [c_t(1), c_t(2), \cdots, c_t(M)]^T$

(e.g., cepstral coefficients) and dynamic feature vectors $\Delta c_t, \Delta^2 c_t$ (e.g., delta and delta-delta cepstral coefficients, respectively), that is, $o_t = [c_t^T, \Delta c_t^T, \Delta^2 c_t^T]^T$, where the dynamic feature vectors are calculated by:

$$\Delta \boldsymbol{c}_t = \frac{1}{2}(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1}),$$

$$\Delta^2 \boldsymbol{c}_t = \frac{1}{2}(\Delta \boldsymbol{c}_{t+1} - \Delta \boldsymbol{c}_{t-1}) = \frac{1}{4}(\boldsymbol{c}_{t+2} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t-2}).$$

First, maximizing $P(O,Q|\lambda)$ with respect to $\boldsymbol{O}$ for a fixed state and mixture will be calculated for all *Q*s then, the *Q* that will give maximum probability will be taken, so *Q* is considered to be given and the problem will be maximizing $P(O|Q,\lambda)$. It is obvious that $P(O \mid Q, \lambda)$ is maximized when $\boldsymbol{O}=\boldsymbol{M}$ if Δc_t *and* $\Delta^2 c_t$ were not used that is, the speech parameter vector sequence becomes a sequence of the mean vectors. Computation of observation vector can be arranged in a matrix form:

$$\boldsymbol{O} = \boldsymbol{WC}$$

Where

$$\boldsymbol{C} = [c_1, c_2, \ldots, c_T]^{\mathrm{T}}$$

And $\boldsymbol{W}$ is matrix that contains the constant weights used to compute delta and delta-delta. Maximizing $P(\boldsymbol{O}|\boldsymbol{Q},\lambda)$ with respect to $\boldsymbol{O}$ is equivalent to that with respect to $\boldsymbol{C}$.

By solving this maximization problem mathematically The following set of equations will be obtained

$$W^{\mathrm{T}}U^{-1}WC = W^{\mathrm{T}}U^{-1}M^{\mathrm{T}}$$

Where

$$U^{-1} = diag[U_{q1,i1}^{-1}, U_{q2,i2}^{-1}, \ldots, U_{qT,iT}^{-1}]$$

$$M = [\mu_{q1,i1}^{T}, \mu_{q2,i2}^{T}, \ldots, \mu_{qT,iT}^{T}]^{T}$$

$\boldsymbol{\mu}_{qt}$ and $\boldsymbol{U}_{qt}$ are the $3M \times 1$ mean vector and the $3M \times 3M$ covariance matrix, respectively, associated with i-th mixture of state q_t.

that can be utilized and solved using Cholesky decomposition and the parameter sequence that maximizes $P(\boldsymbol{O}|\boldsymbol{Q}, \lambda)$ is obtained. The detailed algorithm discussed in references [3],[6] and [7].

2) The synthesis filter

The type of the digital filter depends on the type of spectral features used for training the HMMs. In case of training HMMs with LPC coefficients [9] or PARCOR coefficients, LPC filter or lattice filter should be used respectively. And in case of using mel-cepstral coefficients Mel Log Spectrum Approximation filter (MLSA filter) should be used for synthesis. The detailed design of this filter is discussed in reference [10].

III. The Effect Of Hmm Parameters And Speech Features On System Performance

The training part of the HMM-based synthesis system depends on modeling the spectral and excitation parameters using HMMs. This means that the quality of the generated speech and the size of the stored HMMs-set depend on:

1) The HMM parameters such as the length of context-dependent models (diphones, triphones,…etc) and the number of Gaussian mixtures associated with each state.
2) The speech features such as (LPC, PARCOR, MFCC, …) that determines the type of the synthesis filter.

The effects of these parameters on speech quality are investigated for the case of the low size, unsegmented speech database described in section I. In these experiments, the effect of changing HMM parameters is investigated by evaluating the output of the developed Arabic HMM-based synthesizer with different length of context-dependent models then increasing the number of Gaussian mixture used in each state. The effect of speech features is studied by testing the synthesized speech quality for different types of feature extraction techniques.

In these experiments, the length and number of Gaussian mixtures that gave the best results in HMM parameters evaluation are used to examine the effect of the type of speech features parameters. An Arabic HMM-based speech synthesis system has been developed using speech database described in section I where speech signals were sampled at 16 kHz and windowed by a 25ms Blackman window with a 5ms shift. Feature vector contains both spectral and pitch parameters is constructed. Spectral parameters used are 24 LPC coefficients, 24 PARCOR coefficients obtained from LPC ones or 24 mel-cepstral coefficients with their delta and delta-delta. Pitch part consists of log pitch pattern extracted using A Robust Algorithm for Pitch Tracking (RAPT) technique [19] with its delta and delta-delta. A 5 state left to right with no skip continuous density HMMs for spectral parameters modeling and MSD-HMMs for pitch modeling were trained using embedded training. Also state duration models built using continuous density distribution HMMs. Then the tests are done by changing the mentioned parameters and the system is evaluated objectively and subjectively. For objective evaluation of the synthesized speech, the average Entropy has been measured for synthesized speech samples. Entropy describes information-related properties for an accurate representation of a given signal. In the following expressions, s is the signal and (s_i) the ith sample [14]. The entropy E must be an additive cost function such that E(0) = 0 and given by:

$$E(0) = 0 \text{ and } E(s) = \sum_i E(s_i)$$

so

$$E(s_i) = s_i^{\,2} \log(s_i^{\,2})$$

Subjective evaluation of the generated speech quality is done by drawing spectra and the speech signal for original recorded signal and generated signals and also by hearing synthesized speech samples.

C. The effect of HMM parameters

In these experimental tests, the effect of two parameters of HMM has been investigated. Namely, these parameters are the context-dependent models length and the number of Gaussian mixtures of the model states. The effect of the length of the context-dependent models has been investigated by cloning 2 monophones, 3 monophones and 5 monophones models in constructing the models with single Gaussian mixture for each state. The effect of the number of Gaussian mixtures is investigated by testing the synthesized speech quality for different numbers of Gaussian mixtures. To be fair all comparisons in the evaluation done using the same spectral parameters and synthesis filter.

1) Results of the effect of context-dependent models length

a) Objective evaluation

TABLE I: ENTROPY FOR SYNTHESIZED SPEECH FROM CONTEXT-DEPENDENT HMMS OF DIFFERENT LENGTH.

Number of cloned monophones	Entropy
5	278.3

3	259.76
2	242.99

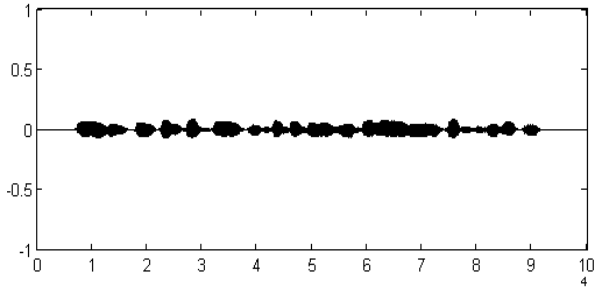
From Table I it is obvious that increasing the length of the context-dependent HMMs the entropy increases. As mentioned in section II, increasing the context-dependent models length preserves more contextual factors leading to more information rate that results in higher entropy in the generated signals. Changing the length of context dependent models also effect on the required memory size as shown in Table II. It is clear increasing the length of context dependent models increases the required memory size significantly since the number of models increases exponentially.

TABLE II: SIZE OF TRAINED SPECTRAL – EXCITATION AND DURATION FOR CONTEXT-DEPENDENT MODELS OF DIFFERENT LENGTH

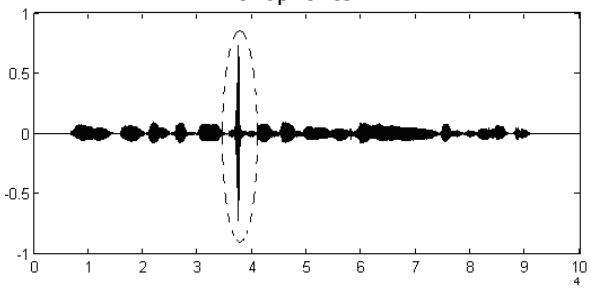
Number of cloned monophones	Spectral and Excitation models size	Duration models size
5	3.5 MB	260 KB
3	1.6 MB	75 KB
2	607.6 KB	23.5 KB

b) Subjective evaluation

Objective experiments show that increasing context-dependent models length leads to improve the entropy of the generated speech. However, listening tests show that increasing context-dependent models length *might* result in perceivable discontinuities and clicks in synthetic speech.



(a) wave signal generated from HMMs of 2 cloned monophones



(b) wave signal generated from HMMs of 5 cloned monophones

Fig. 4 Wave signal for a sentence generated from HMMs of 2 cloned monophones and of 5 cloned monophones

Since increasing the length of context dependent models leads to increasing in the number of models to be re-estimated and the number of repetitions (examples) for each model will be smaller which result in poor generalization of models. This also makes the models sensitive to noise that already may exists in the few repetitions of some models in the training data since it is not recorded in studio and noise removal methods don't remove all the noise and may affect the spectra of recorded speech, all of these reasons help the appearance of these clicks. Fig..4 shows an example of clicks for a speech sentence generated from HMMs trained with 5 cloned monophones and 2 cloned monophones. It is clear that decreasing the context-dependent models length eliminates these clicks.

These clicks contribute in increasing the entropy since its value is very high leading to false entropy values. The advantage of increasing the length of the context-dependent models is preserving contextual factors. In signals that generated from trained context-dependent HMMs of 5 cloned monophones, the same phoneme has been pronounced with many different ways according to its neighbors and how it appeared in the training data. This feature isn't obtained in many cases with HMMs of 2 cloned monophones. Because producing output impulsive noise free is more important than preserving contextual factors the rest of experiments were done using context-dependent models with length of 2 cloned monophones (*diphones*).

2) Results of the effect of the number of Gaussian mixtures

c) Objective evaluation

From Table III it is clear that increasing the number of Gaussian mixtures slightly increases the entropy.

TABLE III: ENTROPY VALUES FOR SYNTHESIZED SPEECH FROM 1 MIX AND 2 MIX HMMS.

Number of Gaussian mixtures	Entropy
1	242.99
2	245.88

TABLE IV : SIZE OF TRAINED SPECTRAL – EXCITATION AND DURATION MODELS FOR 1 MIX AND 2 MIX HMMS

Number of Gaussian mixtures	Spectral and Excitation models size	Duration models size
1	607.6 KB	23.5 KB
2	912.2 KB	24.4 KB

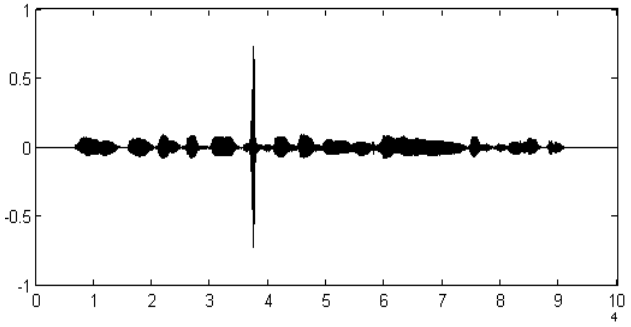
Also from Table IV it is clear that increasing the number of Gaussian mixtures increases the size of the stored HMMs set. This is expected since increasing the number of Gaussian mixtures means more variances and means values to be stored. But the size still small (less than 1MB) since the number of

models didn't increased.

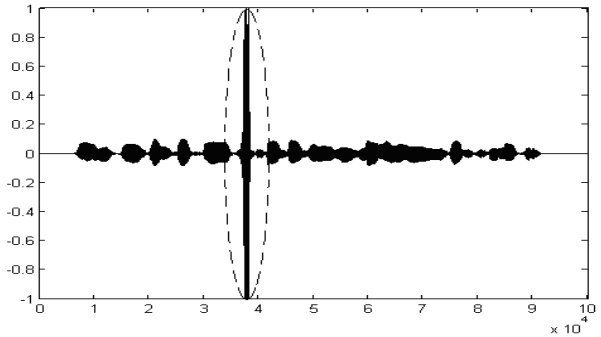
d) Subjective evaluation

Hearing output samples show that increasing the number of Gaussian mixtures results in improving the quality of the synthesized speech by making the pronunciation of each phoneme clearer. On the other hand, it has the disadvantage of maximizing the noise that may exist already in the training data and considering it as a part of the speech signal contents.

Also increasing the number of Gaussian mixtures means that more statistical means and variances in each state to be re-estimated during the training process using the low size training database. This leads to poor generalization of HMMs of specific speech units that had low number of examples in the training data in addition of the noise that may exist in those examples. Fig. 5 shows a waveform generated from single Gaussian mixture HMMs containing a click and the same waveform generated from 2 Gaussian mixtures HMMs. It is clear that increasing the number of Gaussian mixtures maximizes clicks, which implies more noise that will not be detected as happened in the previous case.



(a) wave signal generated from 1 Gaussian mixture HMMs



(b) wave signal generated from 2 Gaussian mixtures

Fig. 5 Wave signal for a sentence generated from single and 2 Gaussian mixture HMMs

D. The effect of spectral parameters Conclusion

The effect of spectral parameters investigated through objective and subjective evaluation of generated speech from single Gaussian mixture HMMs of length 2 cloned monophones trained by LPC coefficients, PARCOR coefficients obtained from LPC ones or mel-cepstral coefficients for the spectral part of feature vectors. Single Gaussian mixture and length 2 cloned monophones were selected based on the evaluation results of HMM parameters discussed in section III.

a) Objective evaluation

TABLE V: ENTROPY FOR SPEECH GENERATED FROM HMMS TRAINED WITH DIFFERENT SPECTRAL PARAMETERS

Evaluation measure	Speech generated from HMMs trained by LPC coefficients	Speech generated from HMMs trained by PARCOR coefficients	Speech generated from HMMs trained by mel-cepstral coefficients
Entropy	324.79	66.99	242.99

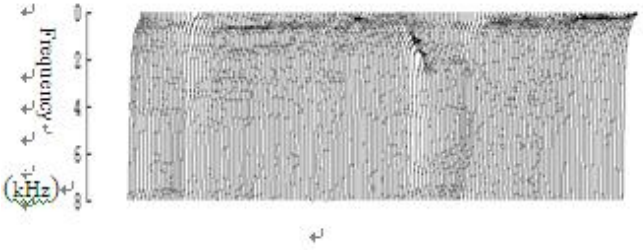
Table V show the values of entropy for synthesized speech samples generated from HMMs trained by LPC, PARCOR or mel-cepstral coefficients. It is clear that speech generated from HMMs trained with LPC coefficients has significantly higher entropy than those trained with PARCOR or mel-cepstral coefficients. The type of spectral parameters effect also in the size of the stored HMMs set. Table VI shows that models trained with PARCOR coefficients has size slightly lower than those models trained with LPC coefficients that has size lower than HMMs trained with mel-cepstral coefficients but the difference is just few kilo bytes.

TABLE VI: SIZE OF HMMS TRAINED WITH DIFFERENT SPECTRAL PARAMETERS

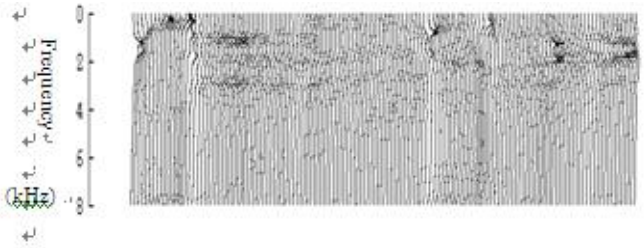
Evaluation measure	Speech generated from HMMs trained by LPC coefficients	Speech generated from HMMs trained by PARCOR coefficients	Speech generated from HMMs trained by mel-cepstral coefficients
Entropy	324.79	66.99	242.99

b) Subjective evaluation

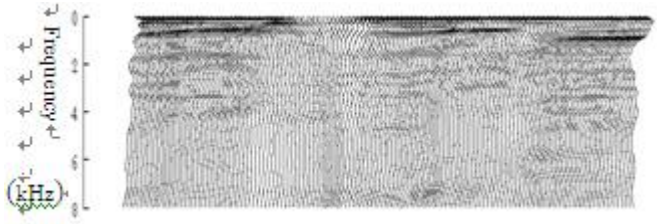
Subjective experiments were done by visual inspection for the spectra of synthesized speech and by hearing synthesized speech samples.



(a) generated spectra extracted from trained HMMs trained with LPC coefficients



(b): generated spectra extracted from trained HMMs trained with PARCOR coefficients obtained from LPC ones



(c): generated spectra extracted from trained HMMs trained by mel-cepstral coefficients

Fig. 6 Spectra obtained from speech generated from HMMs trained with different features for the same sentence

It is clear from Fig.. 6 that HMMs trained with mel-cepstral coefficients generate spectra extremely better than HMMs trained with PARCOR coefficients obtained from LPC ones and HMMs trained with LPC coefficients. In high frequencies the spectra of LPC and PARCOR coefficients is approximately straight line (weak information in high frequencies) but mel-cepestral coefficients produces valuable information in high frequencies. Unlike objective evaluation, subjective evaluation gave other results. Subjective experiments show that the quality of speech waveforms generated from HMMs trained with mel-cepstral coefficients and the synthesized using MLSA filter were significantly better than the quality of those generated from HMMs trained with PARCOR coefficients obtained from LPC ones that gives quality better than HMMs trained with LPC coefficients. The sound was more natural and smoother too and the generated spectra were also better. This happened because the speech generated from HMMs trained by LPC coefficients contains many occurrences of impulsive noise (clicks) in the same utterance (Fig.. 7). The reason of generating this impulsive noise is the low stability of LPC filter [18] that has the effect of giving false high values of entropy.

This explains why LPC coefficients gave high entropy, but the quality of its generated speech resembles the sound of the speaker in the training data like speaking through telephone line (more mechanical) plus the clicks. The quality of speech generated from HMMs trained with PARCOR coefficients like the speech generated from LPC except that it doesn't contain clicks. The speech generated from HMMs trained with mel-cepstral coefficients was more natural and human like.

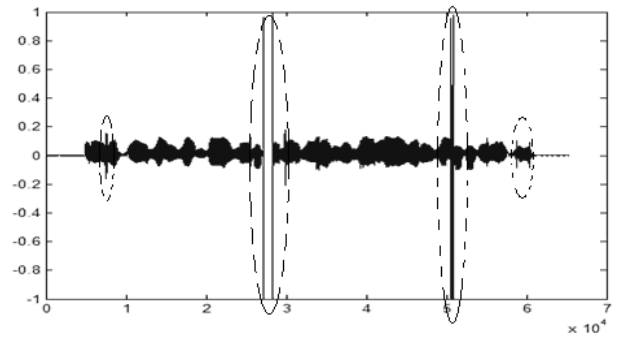


Fig.7 Example of impulsive noise occurred in speech generated from HMMs trained with LPC coefficients

IV. Conclusions

In this work an Arabic speech synthesis system introduced to synthesize speech signals from HMMs. In the training session of the system, simultaneous modeling of spectral, excitation parameters and duration densities in unified framework is performed. HMMs have been trained using *low size unsegmented speech database* by embedded reestimation technique. Investigations for the effect of HMM related parameters (context-dependent models length and the number of Gaussian mixtures) and the type of features have been done via evaluating the quality of the synthesized speech objectively and subjectively. After performing this investigation it is observed that:

1) Implementation of HMM-based synthesizers results in:

Lower effort for preparing speech training DB since no segmentation is needed because of the application of embedded re-estimation.

Very small memory requirements (can be less than 1 MB).

2) In case of using low size training data, increasing the length of context-dependent models and the number of Gaussian mixture may result in impulsive noise due to the poor generalization of HMMs.
3) Mel-cepstral coefficients with MLSA filter give the highest quality.

References

[1] Dundee university website www.computing.dundee.ac.uk.
[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura,"Speech parameter generation algorithms for HMM-based speech synthesis," Proc.ICASSP 2000, pp.1315–1318, June 2000.
[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMMbased speech synthesis," Proc. EUROSPEECH '99, pp.2347–2350, Sep.1999.
[4] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in IEEE Speech Synthesis Workshop, 2002.
[5] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi,"Multi-space probability distribution HMM," IEICE VOL.E85-D,NO.3 March 2002.
[6] Takashi Masuko "HMM-Based Speech Synthesis and Its Applications" Doctoral Dissertation, Department of Electrical and Computer Engineering Nagoya Institute of Technology, Nov 2002.
[7] Takayoshi Yoshimura "Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems" Doctoral Dissertation, Department of Electrical and Computer Engineering Nagoya Institute of Technology, Jan 2002.
[8] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, M. Gales ,V. Valtchev, X. Liu, T. Hain, D. Povey and Phil Woodland, The HTK Book Version 3.4, http://htk.eng.cam.ac.uk/, 2006.
[9] L. Rabiner and B.-H. Juang, Fundamentals of speech recognition, Prentice-Hall, Englewood Cli.s, N. J., 1993.

[10] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. ICASSP-92, pp.137–140, Mar. 1992.
[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura,"Duration Modeling in HMM-based Speech Synthesis System,"Proc. of ICSLP, vol.2, pp.29–32, 1998.
[12] Alghamdi, Mansour, Abdulaziz Alhumayid and Muneer ad-Dusooqee (2003) "Arabic Sound Database: Sentences", Computer and Electronics Research Institute (HK-28), King Abdulaziz City for Science and Technology, Riyadh (in Arabic).
[13] M.S.Barakat, M.E.Gadallah, T.Nazmy And T.El Arif "Training Hidden Markov Models With Low Cost Speech Database For Hmm-Based Speech Systems" submitted to IJICIS journal.
[14] Donoho, D.L. (1995), "De-noising by soft-thresholding," IEEE Trans. on Inf. Theory, 41, 3, pp. 613-627.
[15] Coifman, R.R.; M.V. Wickerhauser (1992), "Entropy-based Algorithms for best basis selection," IEEE Trans. on Inf. Theory, vol. 38, 2, pp. 713-718.
[16] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black, Keiichi Tokuda, "The HMM-based speech synthesis system version 2.0", Proc. of ISCA SSW6, pp.294-299, Bonn, Germany, Aug. 2007.
[17] Alan W. Black, Heiga Zen, Keiichi Tokuda, "Statistical parametric speech synthesis", Proc. of ICASSP2007, pp.1229-1232, Honolulu, Hawaii, Apr. 2007.
[18] J. G. Proakis and D. G. Manolakis, " Digital Signal Processing: Principles, Algorithms, and Applications " Prentice-Hall, Upper Saddle River, NJ, 1996.
[19] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," Chapter 14 in Kleijn and Paliwal, "Speech Coding and Synthesis".