Military Technical College
Kobry El-Kobbah,
Cairo, Egypt

6th International Conference on Electrical Engineering

ICEENG 2008

Investigating the effect of speech features and the number of HMM mixtures in the quality HMM-based synthesizers

By

M. S. Barakat* M. E. Gadallah** T. Nazmy*** T. El Arif***

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 the HMM-based speech synthesis system and applies it to Arabic language using small size training speech database as an example, and shows that the resulting model database has the advantage of being small (can be less than 1MB). Experiments show that using Mel-cepstral coefficients as spectral parameters of speech waveforms for training gives better results than using LPC or PARCOR coefficients. Experiments also show that increasing 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.

Keywords:

Hidden Markov Model (HMM), speech synthesis.

* Modern Academy
** Military Technical College
*** Faculty of computer and information sciences, Ain shams University

1. Introduction:

Speech synthesis defined as the process of generating speech signal, this target can be accomplished using many ways. The traditional way is waveform concatenation (ex: PSOLA), this technique has been 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 [18]. But concatenative systems have the disadvantages of limited number of voices and that it uses a lot of memory to use speech waveforms. Another way for speech synthesis is producing speech entirely through software using linguistic rules and models based on analyzing human speech. So this method some times called rule-based synthesis, formant speech synthesis or parametric synthesis since it generates small compact parameters from human speech then it uses these parameters to generate speech signal. DECtalk is still the best commercial formant synthesizer [18]. 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. This paper study the development of HMM-based synthesizer and the effect of changing speech spectral parameters and number of HMM Gaussian mixtures on the synthesized speech and the size of stored HMMs set for Arabic HMM-based synthesis system trained with low size speech training data.

2. HMM-based synthesis system:

Figure 1 shows the training and synthesis parts of the HMM-based TTS system. In the training part, first, 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 sequence. By using parameter generation algorithm, 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 [6] [15] [16]. 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 train 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 on Arabic language in the following sections.

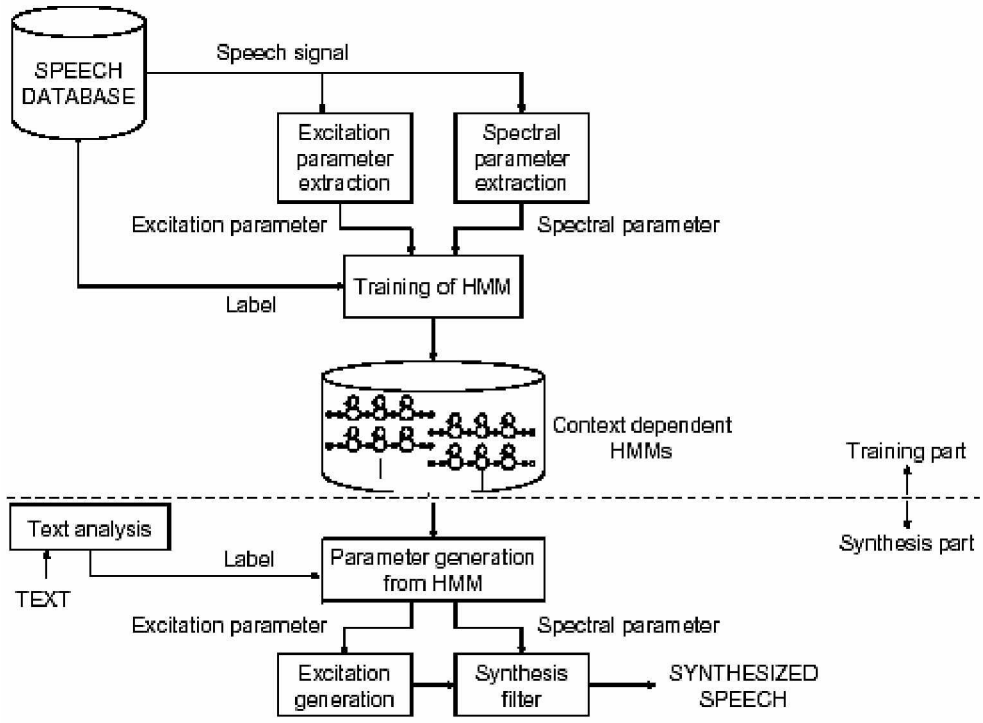


Figure (1): *The scheme of HMM-based synthesis system*

2.1. Training part:

In training part HMMs are built for excitation and spectral parameters for each speech unit. Spectral parameters are modeled using continuous distribution HMMs [3] but excitation parameters modeled using Multi-Space Distribution HMMs (MSD-HMM) to overcome the problem of the voiced and unvoiced regions [4]. 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 [2][5]. HMMs trained using embedded training strategy since embedded training doesn't need phoneme boundaries to be provided in the training data [3], [7] and [12]. State duration densities

are modeled by single Gaussian distributions [10] to model the temporal structure (the speaking rate) of speech data . Now the stored HMMs set are ready to be used for synthesis.

2.2. Synthesis part:

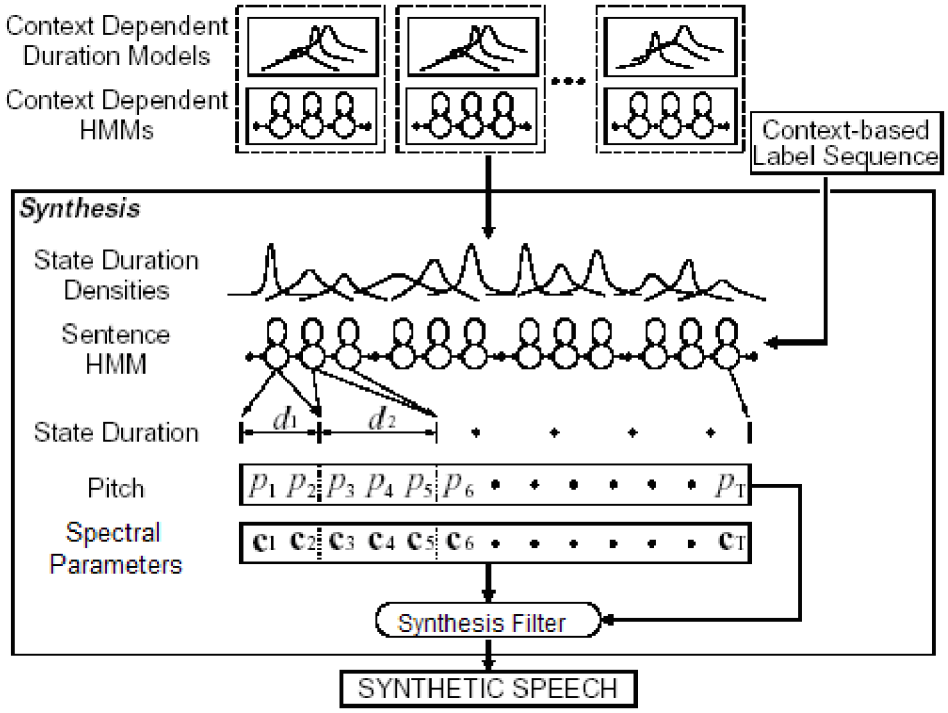


Figure (2): *Synthesis part*

In the synthesis part shown in Figure 2, to synthesize a speech signal a context label sequence that corresponds to the sequence of speech units to be synthesized is constructed first, then, according to the label sequence, a sentence HMM is constructed by concatenating context dependent HMMs. State durations of the sentence HMM are determined so as to maximize the output probability of state durations [10],[3], and then a sequence of spectral parameters and F0 values is determined using the speech parameter generation algorithm. Finally, speech waveform is synthesized directly from the generated mel-cepstral coefficients and F0 values by using the appropriate synthesis filter.

2.2.1 Speech parameter generation algorithm:

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

$$\mathrm{O} = \left[o_1^T, o_2^T, \mathrm{h}, o_{\mathrm{T}}^T\right]^T \tag{1}$$

in such a way that

$$P(O|\lambda) = \sum_{allQ} P(O,Q|\lambda) \tag{2}$$

Is maximized with respect to O, where

$$Q = \{(q_1, i_1), (q_2, i_2), \mathrm{h}, (q_{T,} i_T)\} \tag{3}$$

is the state and mixture sequence, i.e., (q,i) indicates the i-th mixture of state q.
The speech parameter vector $\boldsymbol{O}_t$ consists of static feature vector c_t=[c_t(*1*),c_t(*2*),…,c_t(*M*)]^T (e.g., cepstral coefficients) and dynamic feature vectors Δc_t , $\Delta^2 c_t$ (e.g., delta and delta-delta cepstral coefficients, respectively), that is,

$$o_t = [c_t^T, \Delta c_t^T, \Delta^2 c_t^T]^T, \tag{4}$$

where the dynamic feature vectors are calculated by

$$\Delta c_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau) c_{t+\tau} \tag{5}$$

$$\Delta^2 c_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} w^{(2)}(\tau) c_{t+\tau} \tag{6}$$

First, maximizing P(O,Q| λ) with respect to O for a fixed state and mixture will be calculated for all Qs 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, λ) . Since each state output is modeled by Gaussian distribution , so

$$P(O|Q,\lambda) = N(O|M,U) \tag{7}$$

$$N(O|M,U) = \frac{1}{\sqrt{2\pi^T |U|}} e^{-\frac{1}{2}(O-M)' U^{-1}(O-M)}$$

$$\begin{aligned} \log P(O|Q,\lambda) &= \log N(O|M,U) \\ &= -\frac{1}{2}\left[T \log 2\pi + \log|U| + (O-M)' U^{-1}(O-M)\right] \end{aligned} \tag{8}$$

$$\log P(O|Q,\lambda) = -\frac{1}{2} O^T U^{-1} O + O^T U^{-1} M + K \tag{9}$$

Where

$$U^{-1} = diag[U_{q1,i1}^{-1}, U_{q2,i2}^{-1}, \text{h}, U_{qT,iT}^{-1}] \quad (10)$$

$$M = [\mu_{q1,i1}^{T}, \mu_{q2,i2}^{T}, \text{h}, \mu_{qT,iT}^{T}]^{T} \quad (11)$$

$\boldsymbol{\mu_{qt}}$ and $\boldsymbol{U_{qt}}$ are the 3M × 1 mean vector and the 3M × 3M covariance matrix, respectively, associated with i-th mixture of state qt, and the constant $\boldsymbol{K}$ is independent of O. It is obvious that $P(O|Q, \lambda)$ is maximized when $O=M$ without the conditions (5),(6), that is, the speech parameter vector sequence becomes a sequence of the mean vectors.
Conditions (5), (6) can be arranged in a matrix form:

$$\boldsymbol{O = WC} \quad (12)$$

Where

$$\boldsymbol{C} = [c_1, c_2, \text{h}, c_T]^{\text{T}} \quad (13)$$

And $\boldsymbol{W}$ is matrix that contains the weights used to compute delta and delta-delta in equations (5) and (6). Maximizing $P(\boldsymbol{O}|\boldsymbol{Q},\lambda)$ with respect to $\boldsymbol{O}$ is equivalent to that with respect to $\boldsymbol{C}$. By setting

$$\frac{\partial \log P(WC \mid Q, \lambda)}{\partial C} = 0 \quad (14)$$

The following set of equations will be obtained

$$W^{\text{T}} U^{-1} WC = W^{\text{T}} U^{-1} M^{\text{T}} \quad (15)$$

that can be utilized and solved using Cholesky decomposition and the parameter sequence that maximizes P(O|Q, λ) is obtained [1].

2.2.2 Synthesis filter:

As explained before in case of training HMMs with LPC coefficients or PARCOR coefficients obtained from LPC ones, 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 [9].

3. Experiments:

Using the steps mentioned before, a HMM-based synthesis system for Arabic language developed by building HMMs set for Arabic diaphones [12] using only phonetically balanced 367 sentences from Arabic speech database [11]. Speech signals were sampled at 16 kHz. Feature vectors were constructed by combining two parts, spectral part that consists of 24 static parameter and their delta and delta-delta, excitation part that consist of F0 pattern and its delta and delta-delta. This means that feature vector length is 75. Experiments have been made to see the effect of using LPC coefficients, PARCOR coefficients obtained from LPC ones and mel-cepstral coefficients as spectral parameters during training on the quality of the synthesized speech. Experiments also made to see the effect of increasing the number of Gaussian mixtures in each state of HMMs with those types of spectral parameters. This is done under the condition of using the mentioned low size training speech database and evaluated by objective and subjective experiments.

3.1 Objective experiments:

In objective experiments Entropy and Signal-to-Noise to signal ratio were measured for synthesized speech sentences. Entropy describes information-related properties for an accurate representation of a given signal. In the following expressions, s is the signal and $(s_i)_i$ the sample [13] [14]. The entropy E must be an additive cost function such that E(0)=0 and

$$E(s) = \sum_i E(s_i) \tag{16}$$

Where

$$E(s_i) = s_i^{\,2} \log(s_i^{\,2}) . \tag{17}$$

Noise ratio calculated by removing noise using wavelet technique [13] then dividing the energy of the signal over the energy of the noise. Table 1 show the values of entropy and noise ratio for sentences synthesized from single Gaussian mixture HMMs and 2 Gaussian mixtures HMMs trained by LPC, PARCOR or mel-cepstral coefficients. It is clear that speech generated from HMMs trained with LPC coefficients has higher entropy and Signal-to-Noise ratio than those trained with PARCOR or mel-cepstral coefficients and that increasing the number of Gaussian mixtures generally increases these values in theses types of coefficients. Table 2. shows that increasing the number of Gaussian mixtures increases the size of stored models in all cases and those 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.

Table (1): *entropy and noise ratio between synthesized signal from 1 mix and 2 mix HMMs trained with different features.*

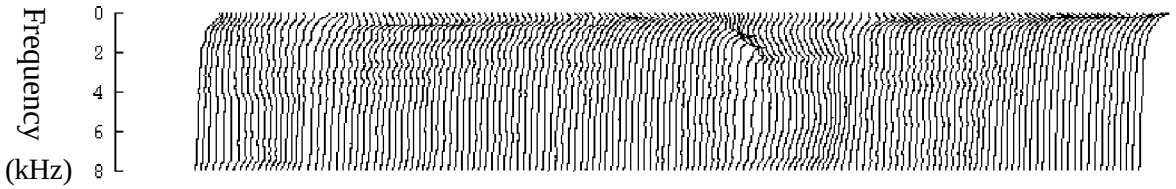
Number of Gaussian Mixtures	Speech generated from HMMs trained by LPC coefficients		Speech generated from HMMs trained by PARCOR coefficients		Speech generated from HMMs trained by mel-cepstral coefficients	
	Entropy	Signal-to-Noise ratio	Entropy	Signal-to-Noise ratio	Entropy	Signal-to-Noise ratio
1	324.79	229.4289	66.99	117.0095	242.99	129.0552
2	332.64	232.5454	70.88	117.2613	245.88	130.3611

Table (2): *Size of trained spectral-excitation and duration for 1 mix and 2 mix models trained with different features.*

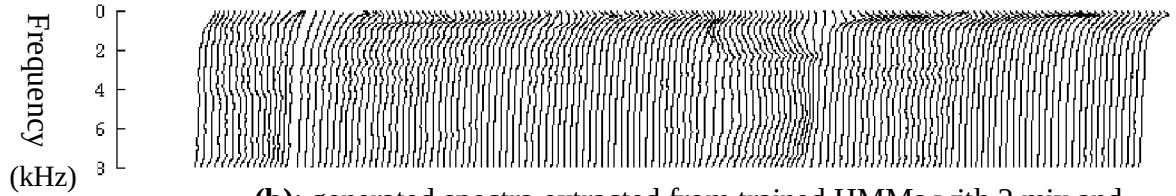
Type of coefficients used for training	1 mixture		2 mixture	
	Spectral and Excitation models	Duration models	Spectral and Excitation models	Duration models
LPC	583.9 KB	19.9 KB	763.8 KB	19.1 KB
PARCORE	576.1 KB	18.7 KB	763.5 KB	19.7 KB
Mel-cepstral	607.6 KB	23.5 KB	912.2 KB	24.4 KB

3.2 Subjective experiments:

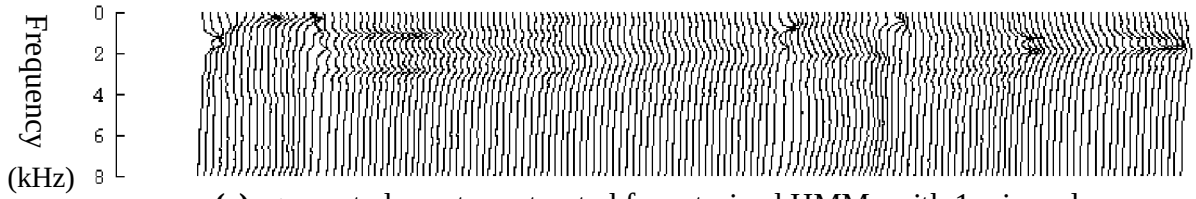
Subjective experiments were done by drawing the spectra and hearing synthesized speech sentences. It is clear from Figure 3 that increasing Gaussian mixtures improves the formant structure of the generated spectra in all cases and that HMMs trained with mel-cepstral coefficients generate spectra better than HMMs trained with LPC coefficients and HMMs trained with PARCOR coefficients obtained from LPC ones. However, by hearing these samples it is observed that the system synthesizes natural sounding speech which resembles the speaker in the training database. But the quality of speech waveforms generated from HMMs trained with mel-cepstral coefficients 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.



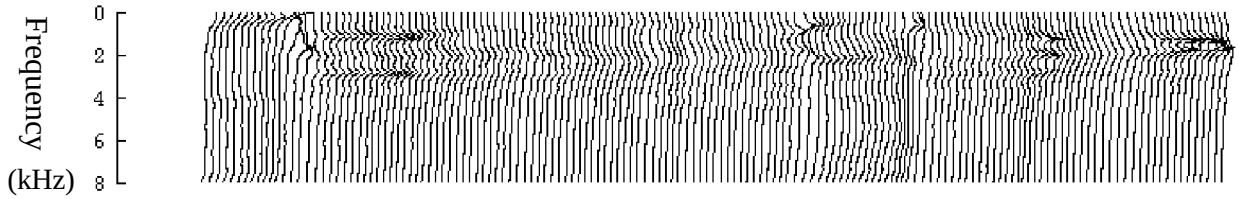
(a): generated spectra extracted from trained HMMs with 1 mix and trained with LPC coefficients



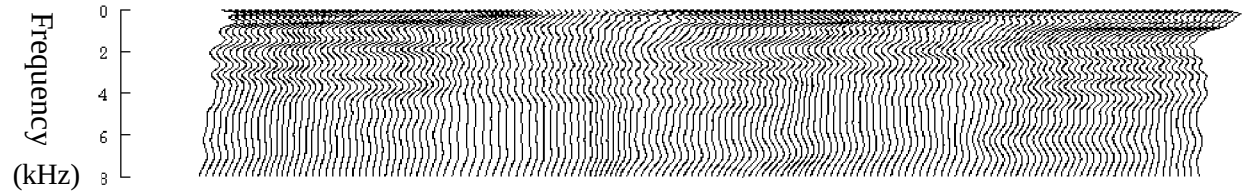
(b): generated spectra extracted from trained HMMs with 2 mix and trained with LPC coefficients



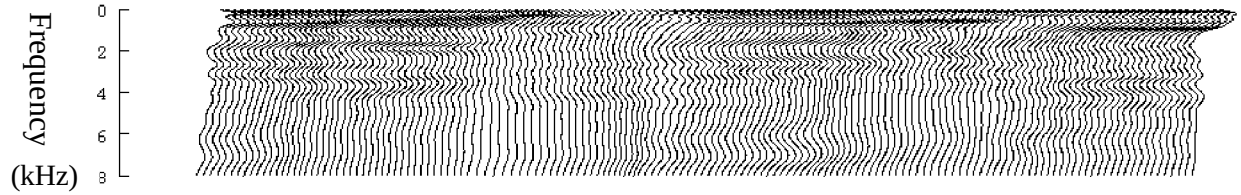
(c): generated spectra extracted from trained HMMs with 1 mix and trained with PARCOR coefficients obtained from LPC ones



(d): generated spectra extracted from trained HMMs with 2 mix and trained with PARCOR coefficients obtained from LPC ones

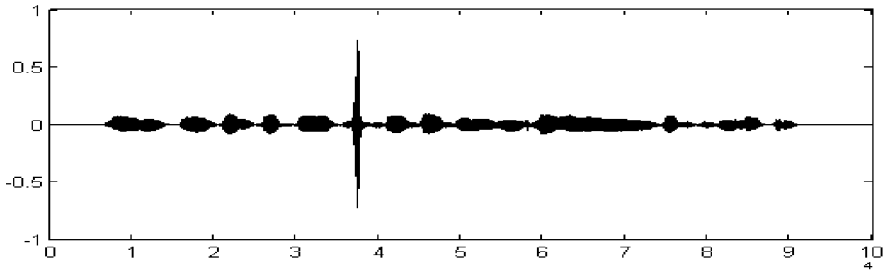


(e): generated spectra extracted from trained HMMs with 1 mix and trained by mel-cepestral coefficients

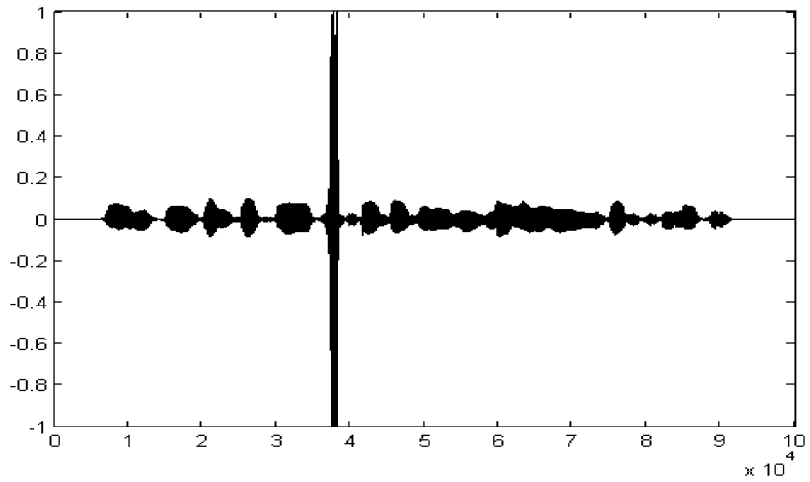


(f): generated spectra extracted from trained HMMs with 2 mix and trained by mel-cepestral coefficients

Figure (3): *Spectra obtained from the same speech sentence generated from 1 mix and 2 mix HMMs trained with different features*



(a): wave signal generated from single Gaussian mixture HMMs



(b): wave signal generated from 2 Gaussian mixture HMMs

Figure(4): *wave signal for a sentence generated from single Gaussian mixture HMMs and 2 Gaussian mixture HMMs.*

This happened because hearing these samples shows that speech generated from HMMs trained by LPC coefficients contains many occurrences of impulsive noise (clicks) in the same utterance. The reason of generating this impulsive noise is the low stability of LPC filter [17] that has the effect of giving false high values of entropy and signal-to-noise ratio because it couldn't be detected by the used noise filter since its duration is very short. Impulsive noise also may occur because of the poor generalization of trained HMMs, this happen if there were not enough repetitions for each model in the training database, so the number of context-dependent models used should be small and their length should be short [12]. Experiments show that increasing Gaussian mixtures has the advantage improving the structure of the synthesized speech units but has the effect of maximizing these clicks since increasing them has the result of more statistical Variances and Means to be updated in each state with this low training data leading to more poorly generalized HMMs as shown in Figure 4.

4. Conclusions:

In this work an HMM-based Arabic speech synthesis system built using low size speech database in which speech is generated from HMMs themselves were introduced. By simultaneous modeling of spectral, excitation parameters and duration densities in unified framework, the speech produced from parameters generated from HMMs by the parameter generation algorithm constraint to be realistic and resemble the speaker in the training database. Objective experiments show that using LPC coefficients as spectral parameters for HMM training produces speech with entropy and Signal-to-Noise ratio higher than using PARCOR or mel-cepstral coefficients, and HMMs trained with PARCOR coefficients has lower size than HMMs trained with LPC coefficients that has lower size than HMMs trained with mel-cepstral coefficients. Objective experiments also show that increasing the number of Gaussian mixtures improves the entropy and the signal-to-noise ratio of the generated speech. However, subjective experiments show that speech generated from mel-cepstral parameters is higher in quality and more natural than the speech generated from LPC and PARCOR coefficients. It shows that speech generated from LPC contains impulsive noise that gives false high values for entropy and signal-to-noise ratio. Listening also show that increasing the number of Gaussian mixtures has the effect of making HMMs poorly generalized resulting in perceivable discontinuities and maximizing clicks in synthesized speech. Surely this will not happen if higher size speech database used for HMMs training.

References:

[1] 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.

[2] 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.

[3] K. Tokuda, H. Zen, and A. Black, “An HMM-based speech synthesis system applied to English,” in IEEE Speech Synthesis Workshop, 2002.

[4] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi,“Multi-space probability distribution HMM,” IEICE VOL.E85-D,NO.3 March 2002.

[5] Takashi Masuko “HMM-Based Speech Synthesis and Its Applications" Doctoral Dissertation, Department of Electrical and Computer Engineering Nagoya Institute of Technology, Nov 2002.

[6] 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.

[7] 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.

[8] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, Prentice-Hall, Englewood Cli.s, N. J., 1993.

[9] 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.

[10] 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.

[11] 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).

[12] M.S.Barakat, M.E.Gadallah, T.Nazmy And T.El Arif "Law Cost Training Of Hmm-Based Speech Recognition Or Synthesis System With Low Size Arabic Labeled Unsegmented Noisy Speech Database" submitted to IJICIS journal.

[13] Donoho, D.L. (1995), "De-noising by soft-thresholding," IEEE Trans. on Inf. Theory, 41, 3, pp. 613-627.

[14] 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.

[15] 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.

[16] Alan W. Black, Heiga Zen, Keiichi Tokuda, "Statistical parametric speech synthesis", Proc. of ICASSP2007, pp.1229-1232, Honolulu, Hawaii, Apr. 2007.

[17] J. G. Proakis and D. G. Manolakis. Digital Signal Processing: Principles, Algorithms, and Applications. Prentice-Hall, Upper Saddle River, NJ, 1996.

[18] Dundee university website www.computing.dundee.ac.uk.