# Improving Arabic HMM Based Speech Synthesis Quality

*Ossama Abdel-Hamid*[1],[3], *Sherif Abdou*[1],[3], *Mohsen Rashwan*[2],[3]

[1] Department of Information Technology, Cairo University, Giza, Egypt
[2] Department of Electronics and Communications, Cairo University, Giza, Egypt
[3] Research & Development International (RDI), Giza, Egypt
{ossama_a, sabdou, mohsen_rashwan}@rdi-eg.com

## Abstract

HMM based speech synthesis, where speech parameters are generated directly from HMM models, is a new technique relative to other speech synthesis techniques. In this paper, we propose some modifications to the basic system to improve its quality. We apply a multi-band excitation model. And we use samples extracted from the spectral envelop as spectral parameters. In the synthesis, the voiced and unvoiced speech parts are mixed according to bands voicing parameters. The voiced part is generated based on a harmonic sinusoidal model. Experimental tests performed on Arabic dataset show that the applied modifications improved the quality.

**Index Terms**: speech synthesis, HMM, MBE.

## 1. Introduction

HMM based speech synthesis [1] is a new technique relative to other synthesis techniques, and it seems promising. In this technique, HMM models are used to simultaneously model different speech parameters. Then to synthesize speech, parameters are generated from these HMM models according to the input text, then speech is synthesized from these parameters.

The basic system is similar to the system described in [2]. It was found that it had some problems in the synthesized speech quality. This degraded quality comes from the usage of features similar to the ones used in speech recognition. In speech recognition it's desired to get rid of the small details that differentiate a user from another, and only keep as little information as possible that discriminate between different phonemes. The case is different in speech synthesis, where it's desired to generate synthesized speech with the full details of the original voice.

Also using a hard decision for the frame on being either voiced or not is a limiting factor in the system, as some phonemes are of mixed excitation type, and on most phonemes there is some noise on the speech signal. So considering the excitation as either pulse train with no noise or white noise with no periodicity is not suitable. So it's more suitable to represent excitation not only by one voicing parameter, but using a mixed-excitation technique.

In our proposed approach, we use HMM to model more detailed speech parameters set, in order to increase the output speech quality. The used speech parameters include voicing of each band to apply an MBE (Multi-Band Excitation) technique [3], where the frame bandwidth is divided into a number of sub-bands, and each band is marked as either voiced or unvoiced. Also instead of using mel-cepstral coefficients to represent spectral envelop, we use a fixed number of spectral envelop samples which are modeled directly in the HMM models. In synthesis, the voiced and unvoiced speech parts are mixed according to bands voicing parameters. The voiced part is generated based on a harmonic sinusoidal model [4][5][6]. While the unvoiced part is generated as a filtered white noise.

We applied the modified HMM based speech synthesis system to Arabic language. Arabic has the problem of diacretization, where input text should be diacretized to be able to convert the text into phonemes sequence. The Arabic language analysis and diacretization is based on RDI® language analysis tools for Arabic, and is out of scope of this paper.

In the following sections, the basic HMM based speech synthesis system is presented in section 2, then the modified system is described in section 3. Results are presented in section 4, and conclusion is presented in section 5.

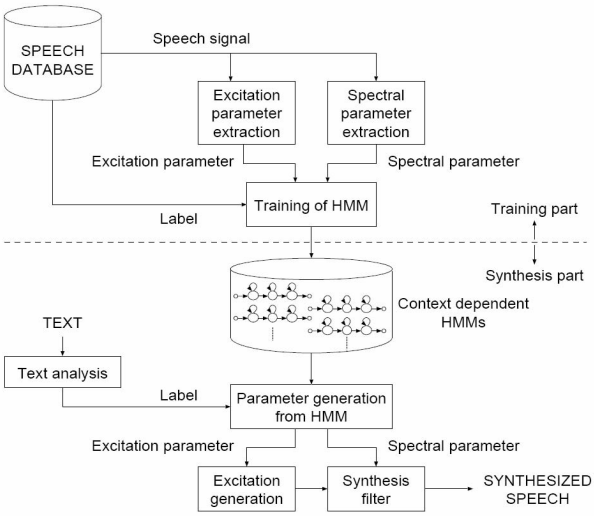


Figure 1. *An overview of the basic HMM-based speech synthesis system.*

## 2. Basic system

Figure 1 illustrates an overview of the basic HMM-based speech synthesis system [1]. In this system, the feature vector consists of spectrum and *F*0 parts. The spectrum part consists of mel-cepstral coefficients, their delta and delta-delta and the *F*0 part consists of log *F*0, its delta and delta-delta.

In the training phase, feature vector sequences are modeled by context-dependent HMMs. The training procedure of the

context-dependent HMMs is almost the same as that in speech recognition systems. The main differences are that not only phonetic contexts but also linguistic and prosodic ones are taken into account and state duration probabilities are explicitly modeled by single Gaussian distributions [7].

In the synthesis phase, first a given text to be synthesized is converted to a context-dependent label sequence and a sentence HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Secondly, state durations maximizing their probabilities are determined. Thirdly, mel-cepstral coefficients and log $F0$ sequences maximizing their output probabilities for a given state sequence are generated by the speech parameter generation algorithm (case 1 in [8]). Finally, speech waveform is synthesized from the generated mel-cepstral coefficients and log $F0$ sequences using a Mel Log Spectrum Approximation (MLSA) filter.

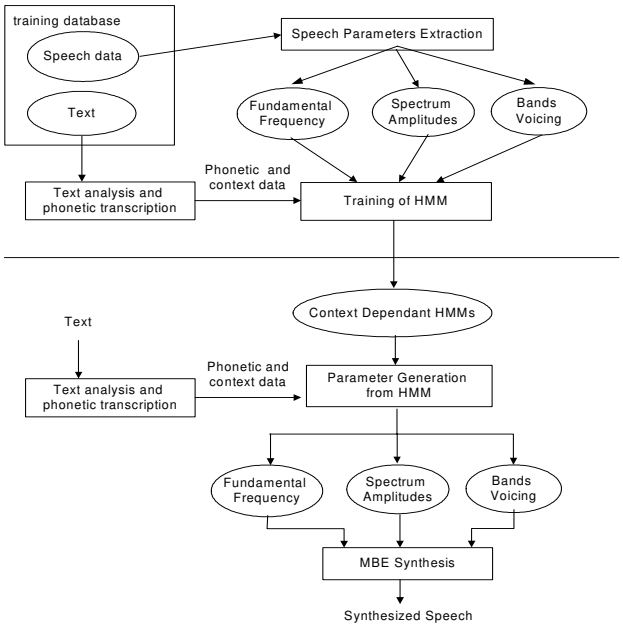


Figure 2 *Proposed HMM based synthesis system.*

# 3. Modified system

We proposed several modifications to the basic HMM-based speech synthesis system [1] to improve its quality. First to remove the buzziness in the voice we used MBE technique [3], where the speech frequency band is divided into a number of sub-bands. A voicing ratio is estimated in each band to indicate how much the band is voiced. And in synthesis, voiced and unvoiced sources are mixed according to the voicing ratio.

The usage of a limited number of mel-cepstral features as a representation of the spectral envelop may be a source of quality degradation degradation, and when we increase the number of mel-cepstral parameters, they tend to represent speech formants. So in the modified system these features were replaced by a larger number of spectral envelop samples. In synthesis phase, these spectral envelop samples are used to compute sinusoids amplitudes of the voiced part, or to filter noise excitation for unvoiced part.

So the set of parameters used in our proposed model are: spectral envelop samples, bands voicing, and fundamental frequency as shown in figure 2. These parameters augmented with the delta and delta-delta parameters are used to train context-dependant HMM models.

In the synthesis phase, speech parameters are generated from HMM models and these parameters are used to generate the final speech output. Speech is re-synthesized from these parameters as described in section 3.4, where the voiced bands are synthesized based on the sinusoidal harmonic model, by obtaining the harmonics amplitudes from the spectral envelop, while unvoiced bands are generated as white noise filtered by the spectral envelop. The details of parameters extraction and speech synthesis are described in the next subsections.

## 3.1. Speech parameters estimation

To model speech parameters in HMM models well, it is required to make the three parameters types independent of each other. So in our model we used a fixed number of spectrum envelop samples computed to represent either the speech harmonics amplitudes for voiced bands, or the noise energy for unvoiced bands, in a form independent from both the fundamental frequency and the voicing parameters. To compute the parameters, first the fundamental frequency (pitch) is estimated, then other parameters are computed based on the estimated pitch. The details of computations are described in the following subsections.

## 3.2. Spectral envelop estimation

To estimate the spectrum envelop, the amplitudes of speech harmonics are estimated using peak picking. Some speech frequency bands may be unvoiced and there's no harmonic peak in that band. So to make the analysis unified for both voiced and unvoiced bands, amplitudes or spectral envelop value for that frequency is estimated by computing the energy of the frequency band around each harmonic. To find speech harmonics, first fundamental frequency should be estimated accurately for voiced frames, and assumed to be some arbitrary value for unvoiced frames. The energy in each band is estimated, and its square root is computed from STFT as follows:

$$a_i = \sqrt{\sum_{k=\lfloor (i-0.5)*h+0.5 \rfloor}^{\lfloor (i+0.5)*h+0.5 \rfloor} s_k^{\,2}} \qquad (1)$$

where $i$ is the index of the harmonic, $S_{k:}$ is the $k$th STFT sample. $h$ is the number of STFT samples per harmonic, and:

$$h = 2 f_0 N / s_r$$

where $f_0$ is the fundamental frequency, $N$ is the window size, and $s_r$ is the sample rate.

Then because the number of bands depends on the fundamental frequency, these energy values are interpolated to get a fixed number of spectral envelop samples.

So the method depends on finding the total energy in each harmonic band. This method doesn't depend on the voicing of each band, it just sums the total energy in each band, which is centered around a multiple of the fundamental frequency. Then

a fixed number of samples are computed using piecewise linear interpolation of the computed values.

### 3.3. Voicing estimation

Voicing parameters specifies whether the band is voiced or not. The speech frame is divided into a number of frequency bands, and voicing is estimated for each band. Voicing of each band is used in synthesis to mix between harmonic and noise like speech according to the voicing degree.

Voicing is estimated by computing the ratio between the energy around the hypothesized harmonics peaks, and the energy in the valleys around them, as shown in Figure 3.

To compute voicing for a certain band, the total energy in the band is summed as *b*, and the energy for the frequencies around the harmonics in the band with a distance less than or equal to $f_0/4$ (the gray area) is summed and named *v*. The voicing value is the ratio *v* / *b*. The ratio *v/b* is expected to be around 0.5 for unvoiced bands, and near one for voiced bands as the energy will be concentrated around the harmonics.

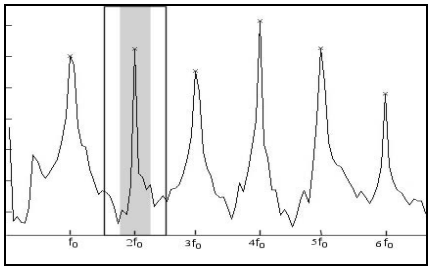


Figure 3. *Voicing computation*.

### 3.4. Generating output speech from speech parameters sequence

The synthesized speech consists of voiced and unvoiced parts mixed according the voicing parameters. The voiced part is generated based on the sinusoidal harmonic model [5], and the unvoiced part is generated as a filtered white noise as shown in figure 4.

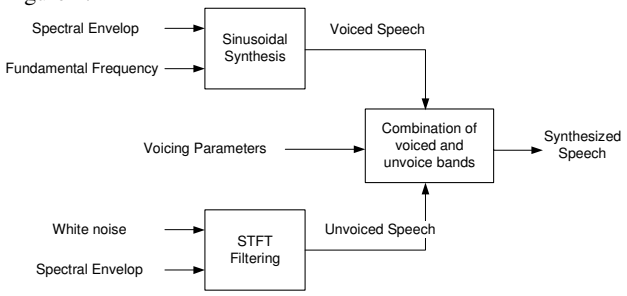


Figure 4 *Speech synthesis from parameters*

The voiced speech is generated for each frame, and can be described as a summation of a set of sinusoids as:

$$s_i(t) = \sum_{h=1}^{N_i} a_{i,h}(t)\sin(\theta_{i,h}(t)) \tag{2}$$

where *i* is the frame index, *t* is the time from the beginning of the frame, $a_{i,h}(t)$ is the amplitude track of the *h*th harmonic, $\theta_{i,h}(t)$ is the phase track of the *h*th harmonic.

So to generate voiced speech the amplitudes and phases tracks are needed. Amplitude track is generated as a linear interpolation of the amplitude at the beginning and end of the frame as:

$$a_{i,h}(t) = a_{i,h} + \frac{t(a_{i+1,h} - a_{i,h})}{T}, \tag{3}$$

Where $a_{i,h}$ is the amplitude of the *h*th harmonic at the beginning of frame *i*, *T* is the total frame length.

The amplitudes can be computed from the generated spectral envelop samples by linear interpolation. The spectral envelop samples are computed at fixed frequencies, while harmonics frequencies change according to the fundamental frequency. So at the beginning of the frame, the harmonics frequencies are multiples of the fundamental frequency at the beginning of that frame. So the amplitudes can be computed at these frequencies.

Phase tracks can be computed based on the frequency track and the phase at the beginning of the frame, frequency track of each sinusoid is a multiple of the fundament frequency track, so phase track can be represented as:

$$\theta_{i,h}(t) = \theta_{i,h}(0) + \int_0^t h\delta_i(\tau)d\tau \tag{4}$$

Where $a_{i,h}$ is the amplitude of the *h*th harmonic at the beginning of frame *i*, $\delta_i(\tau)$ is the fundamental frequency track.

Fundamental frequency track is estimated as a linear interpolation between the fundamental frequencies at the beginning and end of the frame, and represented as:

$$\delta_i(\tau) = f_{0,i} + \frac{t(f_{0,i+1} - f_{0,i})}{T} \tag{5}$$

Where $f_{0,i}$: is the fundamental frequency at the beginning of frame *i*.

While there are no information about phase is modeled in HMM models, the phase at the beginning of the frame is considered as the phase at the previous frame end, and beginning with a zero phase at the first frame. So:

$$\theta_{i,h}(0) = \theta_{i-1,h}(T) \tag{6}$$

$$\theta_{0,h}(0) = 0 \tag{7}$$

Substituting from (5) (6) and (7) in (4):

$$\theta_{i,h}(t) = h\theta_{i-1,0}(T) + 2\pi h\left(f_{0,i}t + \frac{t^2(f_{0,i+1} - f_{0,i})}{2T}\right) \tag{8}$$

The generation of unvoiced part is done through generating a random noise signal filtered using a STFT filter. The filter is constructed by interpolating the spectral envelop samples generated from the HMM.

The mixing of the voiced and unvoiced sources is done through STFT filtering according to the voicing parameters generated from the HMM, then adding the filtered voiced and

unvoiced parts. The voicing parameters can be used in two ways, either by setting a threshold around .75 for example, and consider the bands above the threshold to be voiced, and unvoiced otherwise. That's because unvoiced values are around 0.5 and voiced bands values will be near to 1 as described in section 3.3. The other way is to mix the voiced and unvoiced sources for each band in ratios proportional to the voicing value, which was applied in the experimental system.

## 4. Experimental results

We ran an experiment to compare three systems. The first was the basic system as described in section 2. 24 cepstral coefficients were used plus the zeroth coefficient. The second system was the same as the first except that MBE was applied by computing the voicing parameters as described in section 3, and in the synthesis phase voicing parameters were used to mix bands from voiced and unvoiced excitations. 17 voicing bands were used. The third system was the full modified system as described in section 3. It is the same as the second system except that spectrum is represented as spectral envelop samples, and synthesis is done as described in section 3. 80 samples are taken to represent the spectral envelop.

The three systems were trained on the same dataset. Training data consisted of one hour of Arabic speech of female voice. Context features are nearly the same as described in [2] but applied to Arabic.

To evaluate the three systems a MOS (Mean Opinion Score) test was conducted. 28 same sentences were synthesized from each system, totaling in 84 speech utterances. The sentences were not present in the training data. Seven listeners evaluated the sentences.

The MOS test result is shown in figure 5. It is clear that the modified system is much better than the basic system. The addition of MBE alone improved the quality a little, but the biggest improvement is due to the usage of the new features set with the described analysis/synthesis techniques. Some samples that show the different versions performance are available on our website at: “http://www.rdi-eg.com/tts/samples.htm”

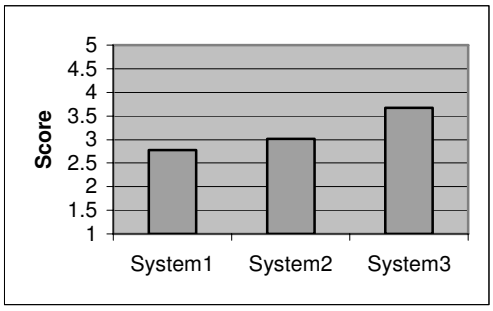


Figure 5. *MOS test result*

## 5. Conclusion

We proposed an HMM-based speech synthesis system that uses different parameters set instead of the cepstral coefficients. The parameters used in the system are based on the sinusoidal representation of speech, where the voiced speech harmonics are represented as harmonically related sinusoids. Sinusoidal based analysis synthesis methods on general, generates higher quality re-synthesized speech than cepstral based methods.

The proposed new features set for an HMM based speech synthesis system improved the quality a lot, although the models have grown in size. This is due to that speech synthesis needs storing of more details of speech to produce higher speech quality.

Also it is shown that the usage of multi-band excitation model improves the quality and allows for a better modeling of mixed excitation phonemes, hence it solves the buzzy voice problem.

In the current system the spectrum phase is completely omitted, although it's reported that this decreases the quality, so future research may find suitable modeling of phase. Also the number of spectral envelop samples may be reduced by using an auditory scale such as Bark scale instead of the used linear scale.

## 6. Acknowledgements

Special thanks are posed to Research & Development International (RDI) for its support of this research, and we gratefully acknowledge the speech and language teams for their help in building the system.

## 7. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM Based Speech Synthesis," Proc. of EUROSPEECH, vol.5, pp.2347–2350, 1999.

[2] Keiichi Tokuda, Heiga Zen, Alan W. Black, "An HMM-based speech synthesis system applied to English," 2002 IEEE Speech Synthesis Workshop, Santa Monica, California, Sep. 11-13, 2002.

[3] Daniel W. Griffin and Jae S. Lim, "Multiband Excitation Vocoder", IEEE Transactions on acoustics, speech, and signal processing, vol. 36, no. 8, August 1988.

[4] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 34, no. 4, pp. 744-754, 1986.

[5] D. O’Brien and A. Monaghan, "Concatenative Synthesis Based on a Harmonic Model," IEEE Transactions on Speech and Audio Processing, vol. 9, no. 1, pp. 11-20, 2001.

[6] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," IEEE Trans. Speech and Audio Processing, vol. 9, no. 1, pp. 21--29, Jan. 2001.

[7] 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.

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. of ICASSP 2000, vol.3, pp.1315–1318, June 2000.

[9] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling," Proc. of ICASSP, 1999.