FLEXIBLE HARMONIC/STOCHASTIC MODELING FOR HMM-BASED SPEECH SYNTHESIS

Eleftherios Banos , Daniel Erro, Antonio Bonafonte, Asuncion Moreno

TALP Research Center, Universitat Politècnica de Catalunya, Spain
e-mail:{lefteris,derro,antonio,asuncion}@gps.tsc.upc.edu

Abstract

In this paper the preliminary results, of a new approach on speech modeling for statistical parametric HMM-based speech synthesis are presented. The proposed system is based on a flexible pitch-asynchronous harmonic/stochastic model (HSM) [1]. The speech is modeled as the superposition of two components: a harmonic component and a stochastic or aperiodic component. The fact that the synthesis model is pitch-asynchronous allows the direct integration to a HMM-based synthesis system. HTS [2], an open source software toolkit that provides HMM-based speech synthesis was used. The proposed HSM method was compared to the HTS baseline system with the same configurations and database. A number of different experiments were conducted. Results show that high quality of synthesized utterances is reached. A small perceptual test was carried out comparing the two systems on quality of the synthetic voice and similarity to the original voice. HSM outperforms the HTS baseline system in the quality test: HSM $53\,\%$, HTS $35{,}3\,\%$, and undecided $11{,}7\,\%$. Concerning similarity to the original voice, HSM-performed slightly better than HTS: HSM $35{,}3\,\%$, HTS $29{,}4\,\%$, and undecided $35{,}3\,\%$.

1. INTRODUCTION

Unit-selection is the dominant method in speech synthesis [3] due to performance advantages such as high quality, and naturalness of synthetic speech. However unit-selection systems are highly dependent on the database and the quality of the recorded database. Due to this quality dependency, voice modification at the selected units cannot be carried out, and voice conversion/adaptation is a difficult task by the time being, for unit selection systems. Furthermore, databases where perfect recording conditions are not possible to achieve cannot be used. Additionally big storage memory is necessary, which is prohibitory in specific applications. Because of these limitations much research has moved to statistical parametric speech synthesis and mainly to Hidden Markov Models (HMM)-based systems.

Statistical parametric speech synthesis (from now on we refer to HMM method only), cannot offer yet a high speech quality comparing to unit-selection, but definitely overcomes most of the problems listed above offering a very wide area for further research (i.e.polyglot systems). In addition, HMM theory and mathematics are well established in many areas of speech technology. The benefits of applying HMM to speech synthesis are numerous: (i) it is possible to take advantage from techniques tested in different fields and adapt them to a different application (i.e. speech recognition to speech synthesis); (ii) the limitations of HMMs are known; (iii) as the basic concept of HMMs is the same for all applications, high-level implemented systems can be used for different research fields and applications[4].

Moreover, continuous improvement has been observed at HMM-based-text-to-speech systems. To be more specific, according to the Blizzard challenge 2005 [5], 2006 [6], and 2007 [7], HTS system show a significant improvement every year. Although on [8] the organizers of the Blizzard evaluation, provide the results without pointing to each system by name, someone can have information about the evaluation methods. On [9] HTS researchers presented an evaluation of their own system for the three year Blizzard challenge. On 2005, HTS participated with a number of changes on the basic system [10]: a STRAIGHT-based high quality vocoding algorithm used for the F0 extraction, and spectral and aperiodic analysis, resulted to reduce the "buzzy" sound that was produced with the basic vocoding technique. Hidden-semi-Markov models(HSMMs) were used for improvements on duration modeling. Parameter generation from HMMs considering global variance (GV) was applied to reduce the oversmoothing of the generated parameters.

For Blizzard challenge 2006, a semi-tied covariance matrix was used for full-covariance modeling in the HSMMs, and the structure of the covariance matrices for the GV pdfs changed from diagonal to full covariance. The system that was used for the first two Blizzard challenge was a speaker-dependent system.

On 2007, a new speaker-independent system was introduced [7]. The system was guided from speaker adaptation approaches. The general results were satisfactory every year and on some occasions over expectations.

Two main areas of research in HMM-based synthesis are (i) improving the quality of the synthesized speech in terms of naturality and similarity to the original training voice, and (ii) training with a small amount of data. This paper focuses on the first one, which is closely related to the speech parametrization used by the system and its associated reconstruction method. In this paper, a high quality asynchronous (harmonic/stochastic model (HSM) [1]), is applied. The main problem of HSM modeling to be solved can be centralized on the voiced/unvoiced transitions where the separation of the harmonic part and stochastic part is not very precise. Vector generation takes advantage of multi-space distribution HMMs to separate as more precise as possible the Harmonic generation part from the stochastic generation part. Preliminary test show that the synthesized voice has a natural tinge, maintaining the main characteristics of the speaker voice, and outperforms HTS system using the same database and same configurations.

The remaining part of this paper is organized as follows: At Section 2 a technical description of the asynchronous HSM model is given. At Section 3 the integration of the HSM model to HMM system is discussed. Further, at Section 4 a general description of an HMM-based synthesis system is given. At Section 5 the main experiments are presented, and the results of a small perceptual test are discussed. Finally, concluding remarks followed by our future intentions and work plans to improve our model, are presented.

	Harmonic component	Stochastic component
Voiced frames	f0, {Aj}, {φj}	LPC filters
Unvoiced frames	-	

Table 1. *General HSM analysis scheme.*

	Harmonic component	Stochastic component
Voiced frames	f0, LSF+Gain vector	LSF+Gain vector
Unvoiced frames	-	

Table 2. *HMM adopted HSM analysis scheme.*

2. DESCRIPTION OF THE HSM IMPLEMENTATION

The harmonic plus stochastic model (HSM) assumes that the speech signal can be represented as a sum of a number of harmonically related sinusoids with time-varying parameters and a noise-like component. The harmonic component is present only in the voiced speech segments, and it can be characterized at each analysis frame by the fundamental frequency and the amplitudes and phases of the harmonics. The stochastic component models all the non-sinusoidal signal components, caused by the frication, breathing noise, etc. It can be represented at each frame by the coefficients of an all-pole filter. A particular implementation of the HSM was developed at UPC in order to provide a flexible framework for all kind of signal transformations [1], especially speech synthesis and voice conversion. During the next sub-sections, we describe the speech analysis and reconstruction procedures and we discuss some questions related to the integration of the model into a HMM-based system.

2.1. Analysis

The speech signals are analyzed at a constant frame rate of 100 or 125 frames per second. Given a speech frame to be analyzed, frame number k, the fundamental frequency $F0(k)$ is estimated and a binary voicing decision is taken. If the frame is considered to be voiced, the amplitudes $A_j(k)$ and phases $\phi_j(k)$ of all the harmonics below $5KHz$ are calculated by least squares optimization. The cut-off frequency is given a fixed value because spectral envelopes are to be extracted from the harmonic component, as it will be explained later. Once the harmonic component is characterized at every analysis instant, it is interpolated and regenerated from the measured values, using 1st order polynomials for the amplitudes and 3rd order polynomials for the frequencies and phases. Then, the regenerated harmonic component is subtracted from the original signal, and the remaining part of the signal, which is considered to be the stochastic component, is LPC-analyzed at each frame. Table 1 shows the analysis structure of the harmonic plus stochastic model.

2.2. Reconstruction

The signal is reconstructed by overlapping and adding $2N$-length frames, where N is the distance between the analysis frame centres, measured in samples. Each synthetic frame contains a harmonic part, built by summing sinusoids with harmonic frequencies and constant amplitudes and phases, and a stochastic part, generated by filtering white Gaussian noise through the measured LPC-filters. A triangular window is used to overlap-add the frames in order to obtain the time-varying synthetic signal. Being k and l the frame number and the harmonic number, respectively, the following expressions are used to reconstruct the signal $s[n]$:

$$s^{(k)}[n] = \sum_l A_l^2 \cos(2\pi l f_o^{(k)} \frac{n}{f_s} + \phi^{(k)}) + \sigma[n] * h_{LPC}^{(k)}[n] \quad (1)$$

, and

$$s[kN+m] = (\frac{N-m}{N}) s^{(k)}[m] + (\frac{m}{N}) s^{(k+1)}[m-N] \quad (2)$$

where m is in the range $[0, N-1]$. The speech signals reconstructed from the parameters measured during analysis are almost indistinguishable from the original ones.

3. TRAINING HMMS ON THE HSM PARAMETERS

The problem of integrating HSM into a HMM-based speech synthesis system can be faced in two different ways:

1. Training the HMMs directly from the HSM parameters, and generating speech directly from the synthetic parameters. This strategy is problematic for several reasons concerning mainly the harmonic parameters:
   - There is a variable number of harmonics, whereas HMMs require constant length training vectors.
   - The number of harmonics is in general high, which makes the learning process more complicated.
   - The variability of the amplitudes and phases with respect to $F0$ is extremely high.
2. Training the HMMs from spectral envelopes calculated by any method, and using the HSM for reconstructing the speech signals from the synthetic envelopes. The main problem of this approach is the loss of spectral resolution caused by the spectral envelope extraction process. Nevertheless, according to our experience in voice conversion, when both the harmonic component and the stochastic component are represented by all-pole filters, the quality of the resulting synthetic speech is reasonably high.

The strategy followed in the system presented in this paper is the second one. The harmonic all-pole filters are calculated by applying the Levinson-Durbin recursion to the autocorrelation sequence given by

$$R_x[n] = \sum_l A_l^2 \cos(2\pi l f_o \frac{n}{f_s}) \quad (3)$$

Note that in this case the phase information is discarded. Before training the HMMs, the all-pole filters are transformed into their associated line spectral frequencies (LSF), which are reported to have very good properties for this kind of mathematical modeling. Table 2 shows the parameters from which the training vectors of the HMMs are built. During the synthesis process, when new parameter vectors are generated by the system, the LSF vectors are converted back into all-pole filters and multiplied by the predicted gain. The amplitudes to be used in expression (1) are calculated by sampling the harmonic envelope $H(f)$ at multiples of the generated fundamental frequency.

$$A_l^{(k)} = |H^{(k)}(l f_0^{(k)})| \quad (4)$$

The minimum phase response of the harmonic all-pole filter $H(f)$ can be also used for estimating the phases of the harmonics, but a linear phase term α has to be added in order to keep them coherent with those of the previous frame. The recursive expression proposed for the linear phase term α is based on the assumption that the pitch varies linearly from frame $k-1$ to frame k.

$$\phi_l^{(k)} = l\alpha^{(k)} + \arg\{H^{(k)}(l f_0^{(k)})\} \quad (5)$$

,

$$\alpha^{(k)} = \alpha^{(k-1)} + \pi \frac{N}{f_s} (f_0^{(k-1)} + f_0^{(k)}) \quad (6)$$

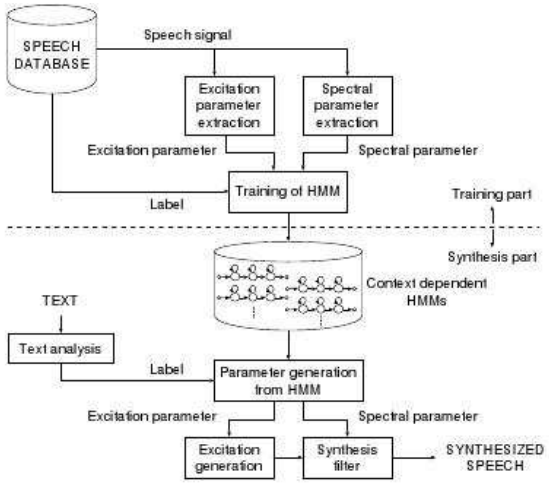

Figure 1. *Overview of a typical HMM-based speech synthesis system.*

4. STATISTICAL PARAMETRIC SYNTHESIS

4.1. Overview of a typical system

Figure 1 illustrates the block diagram of a basic HMM-based TTS system. It is composed of training and synthesis stages. In this system context dependent HMMs (phonetic,linguistic and prosodic context are taken into account) are trained from feature vectors. The feature vectors consists of spectrum (Mel-cepstral) and excitation (F_0) parts, extracted from the speech database. Each HMM has state duration probability density functions(PDFs) to model the temporal structure of speech. Accordingly, TTS models spectrum parameters excitation parameters and durations in a unified framework of HMM [10].

4.2. Training

Context dependent HMMs are trained with feature vectors which consists of spectrum and excitation.The spectrum part includes the spectral parameters and their delta and delta-delta coefficients. Excitation part consists of fundamental frequency ($\log F_0$), its delta and delta-delta coefficients. If spectrum and excitation are trained separately may occur inconsistency problems between them. The $\log F_0$ is composed of one -dimensional continuous (voiced) and zero-dimensional discrete symbol (unvoiced) values. To model such observation sequences Multi-space probability distribution (MSD) HMMS are used. The basic concept of MSD-HMM, is that they can model the sequence of observation vectors with variable dimensionality including zero-dimensional observations [11]. This special kind of HMMs are extremely useful for our work with HSM, because as explained above the harmonic part is composed of continues and discrete values, similar to $\log F_0$ (e.g., multi-dimensional for voiced, and zero-dimensional for unvoiced).

In common with most other continuous density HMM systems, HTS represents output distributions $\{b_{j(o_t)}\}$ by Gaussian Mixture Densities. However, a further generalization is made. Allows each observation vector at time t to be split into a number of S independent data streams O_{st}. The formula for computing $b_{j(o_t)}$ is then

$$b_{j(o_t)} = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} N(O_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s} \quad (7)$$

where M_s is the number of mixture components in stream s, c_{jsm} is the weight of the m'th component and $N(;\mu,\Sigma)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ, that is

$$N(o;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu)} \quad (8)$$

where n is the dimensionality of o. The exponent γ_s is a stream weight. It can be used to give a particular stream more emphasis, however, it can only be set manually.

4.3. Synthesis

In the synthesis part an arbitrarily given text to be synthesized is converted to a context-base label sequence. Then according to the label sequence, a utterance HMM is constructed by concatenating context dependent HMMs. State durations of the sentence HMM are estimated maximizing the likelihood of the state duration densities. According to the duration densities that have been obtained the speech parameter generation algorithm generates the sequence of spectral and excitation parameters (voiced/unvoiced decisions) maximizing the output probabilities [12]. Finally a speech waveform is synthesized using the appropriate speech synthesis filter.

5. EXPERIMENTS AND RESULTS

The main objective of this work is to show the preliminary results of the integration of Harmonic plus Stochastic model in a HMM-based synthesis system.The different experiments depend on the structure of the spectral observation vectors, which can be split into S independent data streams weighted by a stream weight factor (7). Multiple data streams are used to enable separate modeling of multiple information.

The main goal of this specific research is to take advantage of this excellent property in combination with multi-space distribution HMMs to manage to separate as more precise as possible the Harmonic part from the stochastic part. The main problem of HSM modeling can be centralized on the voiced/unvoiced transitions where the separation of the harmonic part and stochastic part is not very precise. Using different streams to model them, in combination with MSD will resolve to a more independent modeling of each one. But while MSD utility of HTS, supports a multi-dimensional to zero dimensional variant vector, does not support multi-dimensional to multi-dimensional vector variability, which is necessary in this case.

To validate the performance of the proposed HSM method, it is compared to the results of the HTS system [6] under the same configurations and database[13]. Mel-cepstral and pitch analysis were substituted by HSM analysis, and MLSA synthesis filter was substituted by the HSM synthesis filter. As described above, for each speech frame k, to be analyzed the fundamental frequency (F_0) was estimated and a voiced/unvoiced frame decision was taken. LSF parameters were extracted for spectral modeling, and $\log F_0$ was used for excitation modeling. The feature vectors were modeled from context dependent HMMs as described for a general HMM system. At most of the experiments 14 LSF parameters were extracted for harmonic or stochastic spectral modeling plus one parameter for the Gain. Some experiments were conducted with a higher number of LSF. Excitation modeling is the same for all experiments and will not be discussed further.

According to the above, the different experiments that were held are:

1. One main vector of 93 features (HSM parameters with their derivatives, $\log F_0$ with its derivatives) was used. When a unvoiced frame is analyzed, a simple mean vector of all the harmonic parts of the voiced frames was used for the Harmonic part. The mean vector showed to perform better than a zero vector. Still some saturation on the synthesized utterances was present mainly at the voiced/unvoiced boundaries.

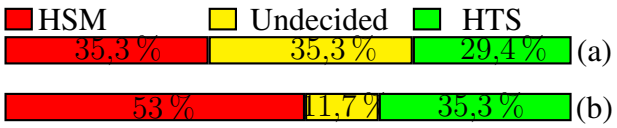

Figure 2. *Detailed results for the two parts of the perceptual test. (a) Similarity and (b) Quality.*

2. One main 30-dimensional feature vector containing static parameters only.The results as expected showed not smooth transitions at the phoneme boundaries, so high lack of natural continuity of the voice observed.
3. Two different vectors for the harmonic and the stochastic parts. The harmonic part was modeled with MSD (15 to 0 dimensions), and the stochastic part was modeled normally. The same experiment was conducted with a different Arctic database (CMU US KSP ARCTIC 0,95). An Indian-English male experienced speaker, and again the naturality of the synthetic voice of our model was significantly good.
4. Few additional experiments have been held using more poles to model spectrum envelope. These experiments kept the same structure as No 1 but 22 features are extracted for spectral modeling. Similar results were taken from this experiment so, the 14-vector size was kept to reduce the computational load and run time.

As expected best results were given by the $3'rd$ method, due to the best modeling. The performance of HSM due to different modeling approaches strengthens our starting point idea: MSD manage to better model Harmonic and stochastic parts and consequently better results are achieved. An perceptual test was given to 17 people where the same utterances were synthesized from HTS and HSM methods. The listeners have a variety of different backgrounds. Four of them are speech synthesis experts, ten listeners have speech processing background, and three listeners don't have experience in speech processing at all. Each listener evaluated 6 sentence pairs, which were presented to them in a random order. The test checks the quality of the synthesized sentences and the voice characteristics similarities to the original training voice. The listeners had to choose between five answers: "A clearly better than B", 'A a bit better than B", "i can't decide", "B clearly better than A","B a bit better than A". In the similarity test, listeners were asked to choose which of the two sentences, A or B, was more similar to the original one. Figure 2 shows the percentage of the number of times each method was preferred. The results show that the proposed method performed slightly better than the baseline HTS. Figure 2 as well shows the percentage of "i can't decide" choices, and actually at the 'similarity' test we can see that although the proposed system performs better, a high rate of the listeners couldn't distinguish the difference between the two systems.

6. CONCLUSIONS

In this work, a preliminary work to integrate an asynchronous Harmonic/Stochastic method for speech modeling, in HTS synthesis system was presented. A perceptual test was performed to compare the proposed system to the HTS system. The results show that the proposed model has good performance for speech synthesis by HMMs. As a future work we will try to use more specific configurations of the HMM-based system according to our model. Furthermore the highest attention will be given to extend the MSD property to manage to model Harmonic and stochastic part more precise. That means to be able to use MSD not only for variable feature vectors of multi-dimensional to zerodimensional but as well to multi-dimensional. We expect that by attempting this approach the performance of our model will improve.

7. ACKNOWLEDGMENT

This work was granted by the Spanish Government ref. AVIVAVOZ TEC2006-13694-C03. As well authors would like to thank the Nagoya Institute of Technology for providing Nitech-HTS synthesis system, giving us the opportunity to conduct an essential research on speech synthesis.

8. REFERENCES

[1] Erro Daniel, Moreno Asuncion, y Bonafonte Antonio, "Flexible harmonic/stochastic speech synthesis," *6th ISCA Workshop on Speech Synthesis (SSW6)*, vol. 6, pp. 194–199, Agosto 2007.

[2] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, y Tadashi Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," *EUROSPEECH'99*, 1999.

[3] A.J. Hunt y A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, pp. 373–376 vol. 1, May 1996.

[4] Olivier Capp, "Ten years of hmms," mar 2001.

[5] Heiga Zen, Tomoki Toda, Masaru Nakamura, y Keiichi Tokuda, "Details of the nitech hmm-based speech synthesis system for the blizzard challenge 2005," *IEICE - Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

[6] Heiga ZEN, Tomoki TODA, y Keiichi TOKUDA, "The Nitech-NAIST HMM-Based Speech Synthesis System for the Blizzard Challenge 2006," *IEICE Trans Inf Syst*, vol. E91-D, no. 6, pp. 1764–1773, 2008.

[7] J. Yamagishi, T.Ñose, H. Zen, T. Toda, y K. Tokuda, "Speaker-independent hmm-based speech synthesis system - hts-2007 system for the blizzard challenge 2007," 2007.

[8] Bennett Christina, L. y Black Alan, W., "The blizzard challenge 2006," *interspeech*, 2006.

[9] J. Yamagishi, T.Ñose, H. Zen, T. Toda, y K. Tokuda, "Performance evaluation of the speaker-independent hmm-based speech synthesis system "hts 2007"for the blizzard challenge 2007," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3957–3960, 31 2008-April 4 2008.

[10] Zen Heiga, Nose Takashi, Yamagishi Junichi, Sako Shinji, Masuko Takashi, Black Alan, W., y Tokuda Keiichi, "The hmm-based speech synthesis system (hts) version 2.0," *6th ISCA Workshop on Speech Synthesis (SSW6)*, vol. 6, pp. 294–299, Agosto 2007.

[11] K. Tokuda, T. Masuko, N. Miyazaki, y T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modeling," *Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on*, vol. 1, pp. 229–232 vol.1, Mar 1999.

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, y T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 3, pp. 1315–1318 vol.3, 2000.

[13] J. Kominek y Black Alan, W., "The cmu arctic speech databases," *SSW5, Pittsburgh, PA*, pp. 223–224, 2004.