A POLYNOMIAL SEGMENT MODEL BASED STATISTICAL PARAMETRIC SPEECH SYNTHESIS SYSTEM

Jingwei Sun[1*,2], *Feng Ding*[1], *Yahui Wu*[1*,3]

[1]Nokia Research, Beijing, P.R. China
[2]Thinkit Speech Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing, P.R. China
[3]School of Information Engineering, University of Posts and Telecommunications, Beijing, P.R. China
snk_thinkit@gmail.cn, feng.f.ding@nokia.com

ABSTRACT

In this paper, we present a statistical parametric speech synthesis system based on the polynomial segment model (PSM). As one of the segmental models for speech signals, PSM explicitly describes the trajectory of the features in a speech segment, and keeps the internal dynamics of the segment. In this work, spectral and excitation parameters are modeled by PSMs simultaneously, while the duration for each segment is modeled by a single Gaussian distribution. A top-down K-means clustering technique is applied for model tying. Mean trajectories acquired from PSMs are used directly to generate speech parameters according to the estimated segment duration. An English speech synthesizer back-end is implemented on CMU Arctic corpus and the performance of the new approach is compared with the classical HMM-based one. Experimental results show that PSM modeling can achieve similar naturalness and intelligence of the synthetic speech as HMM modeling. The system is in the early stage of its development.

Index Terms— Hidden Markov Model, Polynomial Segment Model, statistical parametric speech synthesis, mean trajectory

1. INTRODUCTION

The HMM-based parametric speech synthesis technique (HTS) has been proposed in recent years and it shows to be very effective in generating acceptable speech [1][2]. In the HTS system, spectrum, pitch and duration can be modeled simultaneously in a unified framework of HMM and parameters are generated using dynamic features from HMMs under maximum likelihood criterion [3]. HTS is able to synthesize highly intelligible and smooth speech. Much recent research has been done under the HMM framework to improve the quality of generated speech and many achievements have been made [4].

However, the three limitations of HMM modeling in continuous speech still exist: the weak duration modeling, the conditional independence assumption of observations given the state sequence, and the restrictions on feature extraction imposed by frame-based observations [5]. Many works were carried out to alleviate the above limitations [3][6]. The ideas use features from segments rather than frames, deriving many different models, such as conditional Gaussian HMMs, Stochastic Segment Model (SSM) and Polynomial Segment Model (PSM) [7], etc. In the work of Ostendorf et al [5] a general segment model was defined to cover the different modeling assumptions. Segment models can be thought of as higher dimensional versions of HMM, where Markov states generate random sequences rather than a single random vector observation. These higher order models tend to consume more computational cost than the standard HMM, especially in Large Vocabulary Continuous Speech Recognition (LVCSR). Compared with speech recognition tasks, speech synthesis knows the state sequence. The high-complexity decoding process is not required in speech synthesis tasks. This makes online usage of the segment model for speech synthesis feasible. Dines introduced trended HMM to speech synthesis [8], but it is hard to estimate many polynomial parameters for each HMM state. Furthermore, reliable state-level manual segmentation is not widely available.

In this paper, the Polynomial Segment Model is explored to capture the temporal correlations within a phonetic segment for the parametric speech synthesis. The spectral and excitation parameters for each speech segment are modeled by PSMs simultaneously and the duration of a segment is modeled by a single Gaussian. A top-down K-means clustering technique [9] is used to tie similar models together. During the procedure of parameter generation, the mean trajectories of PSMs are used. A PSM-based speech synthesis back-end is built up to investigate the performance of PSM modeling.

The rest of this paper is organized as follows: Section 2 describes model training and parameter generation algorithms for PSMs. The framework of PSM-based speech synthesis system is described in section 3 and experiments are covered in Section 4. Concluding remarks and some discussion of our future work are presented in the final section.

2. PSM TRAINING AND PARAMETER GENERATION ALGORITHMS

A Polynomial Segment Model, can be defined as,

$$C = Z_N B + E, \tag{1}$$

where C is a $N \times D$ matrix for N frames of D dimensional feature vectors. B is a $(R+1) \times D$ coefficient matrix of a R^{th} order trajectory model and E is the residual error with the same size as matrix C. Z_N is a $N \times (R+1)$ time normalization matrix, which is used to map segments of different durations into a range of between 0 and 1.

2.1. Parameter estimation using existing segmentation

As described in [10], given a set of K segments $S = C_1, ..., C_k$ of model m, the maximum likelihood estimate of the PSM parameter matrix $\hat{B}_m$ and residue covariance $\hat{\Sigma}_m$ are given by

*Work performed as an intern in Nokia Research, Beijing

$$\hat{B}_m = [\sum_{k=1}^{K} Z_{N_k}^T Z_{N_k}]^{-1} [\sum_{k=1}^{K} Z_{N_k}^T Z_{N_k} B_k], \quad (2)$$

and

$$\hat{\Sigma}_m = \frac{\sum_{k=1}^{K} (C_k - Z_{N_k}\hat{B}_m)^T (C_k - Z_{N_k}\hat{B}_m)}{\sum_{k=1}^{K} N_k}. \quad (3)$$

2.2. Log likelihood evaluation

The likelihood of segment C_j against model m can be computed by accumulating the observation likelihoods one at a time, against the corresponding sampling point on the Polynomial Segment Model. The log likelihood, $L(C_j|m)$, with a detailed description given in [7], can be written as

$$\begin{aligned} L(C_j|m) = &-\frac{N_j}{2}[Dlog(2\pi) + log|\hat{\Sigma}_m|] \\ &- \frac{1}{2} tr[(C_j - Z_{N_j}\hat{B}_m)\hat{\Sigma}_m^{-1}(C_j - Z_{N_j}\hat{B}_m)^T]. \end{aligned} \quad (4)$$

2.3. Viterbi-style segmentation and training algorithms

In segment models, segmental boundaries play an important role. Two methods could be used to get the initial segmentation: handcrafted segmentation or HMM-based forced alignment. Usually, handcrafted segmentation is time-consuming and unfeasible for subphoneme level. There are inconsistencies between the segment modeling mechanism and the results from HMM-based forced alignment which is optimized for a standard HMM scheme. In order to get a more accurate segment boundary, the HMM-based aligning method is generalized to form new viterbi-style segmentation and training algorithms for PSM. Boundaries from HMM forced alignment can be used as initialization. To lower the computing complexity, some constraints are applied to reduce the search space. In the following parts, the segmentation and training algorithms are presented respectively.

Define $Y_1^T = \{y_1, y_2, ..., y_T\}$ to be a T-length observation sequence, which is connected to an N-length phone sequence $A_1^N = \{a_1, a_2, ..., a_N\}$. Each phone in the sequence is corresponding to a PSM, and the whole PSM sequence is defined by $M_1^N = \{m_1, m_2, ..., m_N\}$. The end boundary sequence is represented by $B_1^N = \{b_1, b_2, ..., b_N\}$. For each boundary in B_1^N , a candidate boundary set which may contain the underlying accurate boundary is extended from the current boundary by adding a window. The window size d can be adjusted manually. Then the candidate boundary set for each phone in the sequence can be represented by $TB_j : \{b_j - \frac{d}{2}, b_j - \frac{d}{2} + 1, b_j - \frac{d}{2} + 2, ..., b_j + \frac{d}{2}\}, j = 1, .., N$. $L(C_j|m) = \log p(Y_{j_1}^{j_L}|m)$ is used to describe the log likelihood of a segment C_j against the model m, where $C_j = \{y_{j_1}, y_{j_2}, ..., y_{j_L}\}$, represented a L-length observation segment in Y. Finally, define $\delta_t(j)$ to be the log probability of the most likely segmentation sequence ending at frame t for observations $Y_1^t = \{y_1, y_2, ..., y_t\}$ given the label sequence $A_1^j = \{a_1, a_2, ..., a_j\}$. The trace back information is stored in $\Psi_t(j)$. The dynamic programming PSM-based segmenting algorithm can be described as follows:

(a) Initialize:
for $j = 1$,

$$\delta_t(a_1) = \log p(Y_1^t|m_1), \forall t \in TB_1, \quad (5)$$

$$\Psi(a_1) = 0, \quad (6)$$

(b) Iterate:
for $j = 2, ..., N-1$,

$$\begin{aligned} \delta_t(a_j) = &\max_{\forall \tau \in TB_{j-1}} \{\delta_\tau(a_{j-1}) \\ &+ \log[p(y_{\tau+1}^t|m_j)p(a_j|a_{j-1})]\}, \\ &\forall t \in TB_j, \end{aligned} \quad (7)$$

$$\begin{aligned} \Psi(a_j) = &\arg\max_{\forall \tau \in TB_{j-1}} \{\delta_\tau(a_{j-1}) \\ &+ \log[p(y_{\tau+1}^t|m_j)p(a_j|a_{j-1})]\}, \\ &\forall t \in TB_j, \end{aligned} \quad (8)$$

for $j = N$,

$$\begin{aligned} \delta_t(a_N) = &\max_{\forall \tau \in TB_{N-1}} \{\delta_\tau(a_{N-1}) \\ &+ \log[p(y_{\tau+1}^T|m_N)p(a_N|a_{N-1})]\} \end{aligned} \quad (9)$$

$$\begin{aligned} \Psi(a_N) = &\arg\max_{\forall \tau \in TB_{N-1}} \{\delta_\tau(a_{N-1}) \\ &+ \log[p(y_{\tau+1}^t|m_N)p(a_N|a_{N-1})]\}, \end{aligned} \quad (10)$$

(c) Trace back:

$$\begin{aligned} &B_N = T, \\ &B_j = \Psi(a_j), for j = N-1, N-2, ..., 1. \end{aligned} \quad (11)$$

After the boundaries of segments are refined, the PSM of each phoneme in the phone set can be estimated according to equation (2) and (3), as described in 2.1. Define $delta$ to be the allowed minimum difference of two iterations of segmentation process, and I is the maximum iterating times which can be set manually. The viterbi-style PSM training algorithm can be described as follows:

(a) For each phone a in the model list, process all sentences in the corpora with transcription and segmentation, estimate $\hat{B}_a$ and $\hat{\Sigma}_a$ using equation (2) and (3);

(b) For $i = 1, 2, ...I$, refine the segmentation of the speech data using current PSMs, summate $\delta_T(N)$ of all sentences and record the summation in $S(i)$;

(c) If $S(i) - S(i-1) > delta$, go to (b), else terminate the program.

2.4. Parameter generation algorithm

When the label sequence of an utterance to be synthesized is given, a PSM model sequence can be established. The durations are estimated by Gaussian models. For each phone a in the model set, a Polynomial Segment Model m is described by a pair of matrix $\{\hat{B}_m, \hat{\Sigma}_m\}$. represents the polynomial coefficients, and $\hat{\Sigma}_m$ is a global covariance for each frame in the segment. The mean trajectory can be estimated as follows,

$$\hat{C}_m = Z_N \hat{B}_m \quad (12)$$

where Z_N is a design matrix that is known if the segment duration N is given. The mean trajectories of each model are used to generate speech parameters directly.

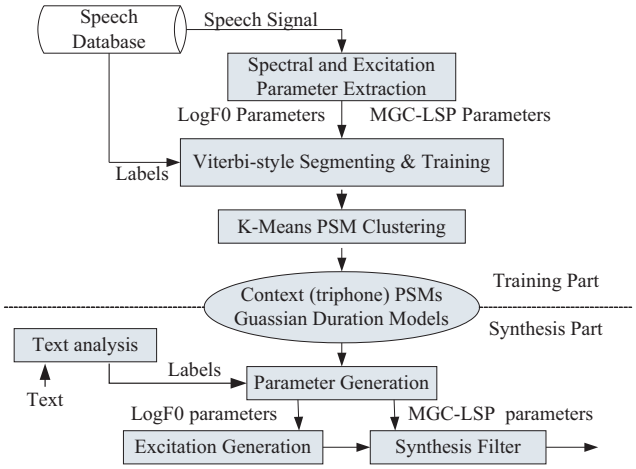

Fig. 1. Overview of a PSM-based speech synthesis system.

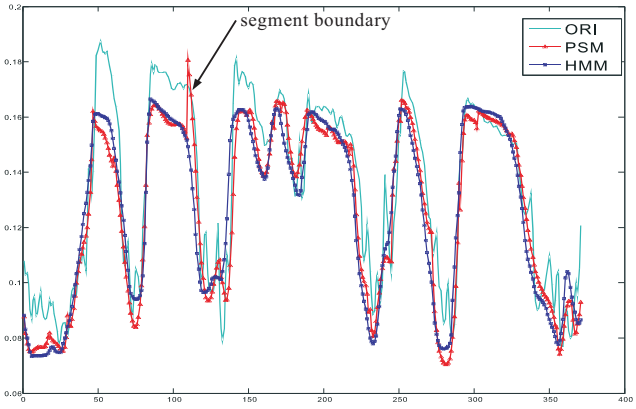

Fig. 2. Original, HMM and PSM based MGC-LSP sequences.

3. FRAMEWORK OF PSM-BASED SPEECH SYNTHESIS SYSTEM

The framework of PSM-based speech synthesis system, including the training part and the synthesis part, is shown in Figure 1.

In the training part, both spectrum (Mel-Generalized Cepstrum-based Line Spectrum Pair [11]) and excitation parameters ($\log F_0$) are extracted from the speech database and modeled by context-dependent PSMs. Each PSM has a single Gaussian duration density to model the temporal structure of speech segment, which is called PSM-based Duration Modeling (PDM) in this work. Then the K-Means based clustering method is used to tie similar models together.

In the synthesis part, a given text is converted to a context-dependent label sequence firstly and then the PSM sequence of the utterance is constructed by concatenating the context-dependent PSMs according to the label sequence. Then, durations of the PSMs are determined using the duration models, and speech parameters are generated directly from the mean trajectories of the PSM sequence. Finally, a speech waveform is reproduced from the generated spectral and excitation parameters using the MLSA filter with binary pulse or noise excitation.

It is worth noting that the segmental boundaries we get from automatic segmentation may be not exactly the $voiced-unvoiced$ boundaries, so it is hard to model $\log F_0$ accurately under PSM framework. The value of $unvoiced$ $\log F_0$ is set as the probable minimum $\log F_0$.

4. EXPERIMENTS AND EVALUATIONS

4.1. Experiments description

CMU ARCTIC corpus, which includes 1132 phonetically balanced sentences, was used to build our systems with 1112 for training and the other 20 for testing. Speech signals were windowed with a 5ms shift, and MGC-LSP coefficients and $\log F_0$ parameters were obtained by speech analysis tools.

HMM and PSM were used respectively for speech parameter modeling. For HMM modeling, speech parameters consisted of 25-order MGC-LSP coefficients, $\log F_0$ parameters, their delta and delta-delta coefficients. A 5-state left-to-right HMM with single diagonal Gaussian output distribution was selected as the model structure. HMM with multi-space distributions was used for F_0 modeling. The decision tree based model clustering technique was applied to the context-dependent phoneme models for state tying. For PSM modeling, speech parameters contained the coefficients of the 25-order MGC-LSP and $\log F_0$ without dynamic coefficients. The polynomial order is set to 4. 9652 triphone PSMs were trained firstly using viterbi-style training algorithm and then clustered into 2672 models using K-means clustering algorithm. HSMM and PDM were used separately for duration modeling. As a result, totally six kinds of models were trained for speech generation: HMM-based MGC-LSP models, HMM-based $\log F_0$models, PSM-based MGC-LSP models, PSM-based $\log F_0$ models, HSMM-based duration models and PDMs.

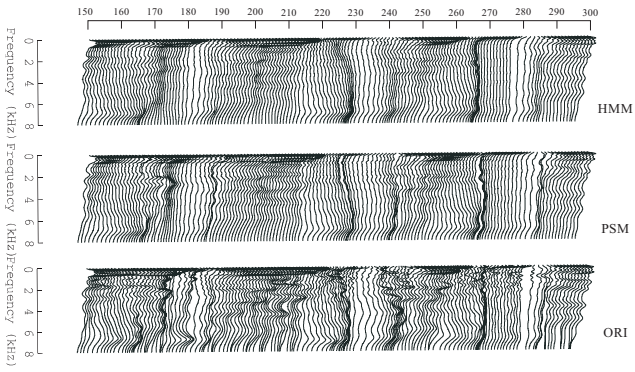

Fig. 3. Original,HMM-based and PSM-based MGC-LSP spectrum.

4.2. Comparison of spectrum

To compare the spectrum simulating performance of these two modeling methods, only MGC-LSP models were used for generation while $\log F_0$ and duration were derived from the original speech parameters. Figure 2 shows a time sequence of the 2nd MGC-LSP parameters of an utterance "Her face was against his breast" extracted from original speech and that generated respectively from HMMs and PSMs. It can be observed that PSM modeling can reconstruct more spectral details within speech segments than HMM modeling. Figure 3 gives an example of spectrum sequences of

generated speech with HMM-based method, generated speech with PSM-based method and original speech. It can be seen that PSM-based method makes generated spectral peaks relatively sharper than HMM-based method. Therefore, by looking at speech on a segmental level rather than on a frame-by-frame basis, we can better capture the temporal structure over the duration of a phone sequence.

4.3. Perceptual evaluation

An ABX opinion test on the naturalness of synthetic speech was conducted to investigate the effects of new modeling for MGC-LSP, $\log F_0$ and duration. Six voices, shown in Table 1, were divided into three pairs and evaluated. "Original" means the related parameters were derived from the original speech. Five trained listeners participated in the test. The sentences were selected from the test set and sent to each listener in a random order.

	MGC modeling	$\log F_0$ modeling	Duration modeling
A	HMM	Original	Original
B	PSM	Original	Original
C	PSM	Original	HSMM-based
D	PSM	Original	PDM
E	PSM	HMM	PDM
F	PSM	PSM	PDM

Table 1. Synthetic voices used for an opinion test.

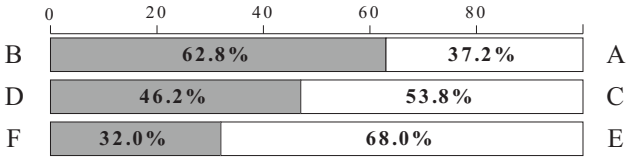

Fig. 4. Preference score of PSM against HMM method.

Figure 4 shows the results of the test. It is observed that PSM modeling can achieve relatively better performance for spectrum than HMM modeling (B vs A). However, PSM modeling for F_0 shows to be bad and deteriorates the overall performance of the system (F vs E). PDM is worse than HSMM-based duration modeling but the gap is not very large (D vs C).

The poor performance of PSM-based F_0 modeling may be due to the awkward processing of unvoiced regions in segments. It is worthy to check the pitch modeling method further under the PSM framework. Another possible reason for the performance deterioration of PSM-based synthesis approach is that no appropriate mechanism is present for PSM to describe the correlation between two neighboring segments. It may lead to discontinuities of synthetic parameters on segment boundaries (as shown in Figure 2). In our experiments, it is found that PSM performed better for long duration segments than short segments. If too many short segments are included in a sentence, they could ruin a listener's flow.

5. CONCLUSION

In this paper, a Polynomial Segment Model based speech synthesis approach was proposed, in which speech spectrum and excitation were modeled simultaneously in a unified PSM framework. New viterbi-style segmentation and training algorithms are developed. Although the system is in its early stage, experimental results show that the PSM appears promising in describing the intra-segmental correlations of time varying speech and deserves further study.

More works are needed to improve the performance of the system. Firstly, an extended forward-backward training algorithm can be implemented to avoid the hard segmentation in viterbi-style training, which is also expected to address the problem of discontinuities at segmental boundaries. Secondly, more detailed contexts (phonetic, linguistic, and prosodic contexts all taken into account) can be used with a relatively larger database, and more advanced clustering techniques such as decision tree based clustering can be implemented. Thirdly, parameter generation algorithms other than the mean trajectory can be investigated to use covariance and dynamic features. Fourthly, voiced-unvoiced boundaries within a segment need further study.

6. 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," in *Proc. of Eurospeech*, pp. 2347-2350, 1999.

[2] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," *in Blizzard Challenge Workshop*,2006.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *Proc. of ICASSP*, pp. 1315-1318, 2000.

[4] A. Black, H. Zen, K. Tokuda, "Statistical Parametric Speech Synthesis, " in *Proc. of ICASSP*, pp. IV-1229, 2007

[5] M. Ostendorf, "From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," IEEE Trans, SAP., vol 4, no.5, pp.360-378, 1996.

[6] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," *in Interspeech, 2004*, pp. 1185C1180.

[7] H. Gish and K. Ng, "A Segmental Speech Model with Application to Word Spotting," Proc. ICASSP 1993.

[8] J. Dines and S. Sridharan, "Trainable speech synthesis with trended hidden Markov models," in ICASSP, 2001, pp. 829-832.

[9] S. Au Yeung, C. Li and M. Siu, "Sub-phonetic Polynomial Segment Model for Large Vocabulary Continuous Speech Recognition," in Proc. of ICASSP, pages 193-196, 2005.

[10] C. Li and M. Siu, "Training for polynomial segment model using the expectation maximization algorithm," Proc. ICASSP, 2004, pp. 841-844.

[11] K. Koishida, T. Kobayashi, T.Masuko, and S. Imai. "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in Proc. of ICASSP, pages 1355-1358, 1997.