A State Duration Generation Algorithm Considering Global Variance for HMM-based Speech Synthesis

Shifeng Pan, Jianhua Tao, Yang Wang
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing
E-mail: {sfpan, jhtao, yangwang }@nlpr.ia.ac.cn

***Abstract*— The speech parameter generation algorithm considering global variance (GV) for HMM-based speech synthesis proved to be effective against the over-smoothing problem. In this paper this idea is extended to the generation of state duration. A GV model on syllable duration is proposed and a state duration generation algorithm considering this GV model is presented in details. By improving the GV likelihood on syllable duration, the over-averaging effect on generated state duration is much alleviated. Experimental results are promising which show that the proposed method outperforms the conventional one and the naturalness of synthetic speech is improved.**

I. INTRODUCTION

The Hidden Markov Model (HMM)-based speech synthesis has been widely used in recent years. In this method, pitch, spectrum and duration are modeled simultaneously within a unified framework [1]. By taking account of constraints between the static and dynamic features, smooth speech parameter trajectories can be generated [2]. The synthetic speech is highly intelligible and smooth [3, 4].

Currently, the main drawback of HMM-based speech synthesis is that the synthetic voice does not sound natural enough, including the unsatisfying speech quality and the bland prosody. Besides the influence of vocoder, the over-smoothed spectral parameters generated by HMMs are closely related to the former aspect. As to the second aspect, the over-smoothed pitch and over-averaged state duration generated by HMMs are the main reasons. Many methods have been proposed to improve the naturalness of synthetic voice. Some of them focus on alleviating the over-smoothing effect of generated spectral parameters, such as post-filtering methods [4, 5], incorporating the difference of adjacent LSPs as a stream to HMM feature vector [6]. Some of them focus on integrating multiple-level prosody models to model and hence generate prosody parameters more accurately, such as phone duration model [7], phone and syllable duration model [8], multi-layer F0 model [9], syllable and phrase level F0 model [10]. One of the most successful methods against the over-smoothing problem is the speech parameter generation algorithm considering global variance (GV) [11]. In this method, a GV model is built to model the variation of speech parameter trajectories at utterance level, including trajectories of F0 and spectral parameter. The generated parameter sequence maximizes a likelihood based not only on an HMM likelihood but also on a GV likelihood. The latter likelihood works as a penalty for reduction of the GV of the generated parameter trajectories. This method proved to be effective against the over-smoothing problem and can improve the naturalness of synthetic speech.

In this paper, this method is extended to the generation of state duration. Firstly, a GV model of syllable duration was built. Then, state duration is generated by maximizing a likelihood consisting of both HMM state duration likelihood and GV likelihood. With the penalty of GV likelihood on syllable duration, the over-averaging effect of generated state duration is much alleviated. Experimental results show that the synthetic speech sounds more natural from the view of syllable duration distribution.

The rest of this paper is organized as follows. In section 2, the conventional state duration generation for HMM-based speech synthesis is reviewed. Section 3 describes the proposed GV model on syllable duration and the state duration generation algorithm considering GV in details. In section 4 the evaluation result is presented. The conclusion is given in section 5.

II. CONVENTIONAL STATE DURATION GENERATION

For given HMMs λ, the optimal speech parameter vector sequence O^* is derived as follow based on maximum-likelihood (ML) criterion:

$$O^* = \arg\max_O P(O \mid \lambda), \tag{1}$$

where the state sequence Q (i.e. state duration) is hidden. Though an algorithm based on EM algorithm is proposed in [2] to solve the above problem, it's too complex and time consuming. However, a sub-optimal solution to the problem has been widely used, which is:

$$Q^* = \arg\max_Q P(Q \mid \lambda), \tag{2}$$

$$\hat{O} = \arg\max_O P(O \mid Q^*, \lambda). \tag{3}$$

In this case, the state sequence Q^* is determined independently of speech parameter sequence O, which greatly simplifies the determination of $\hat{O}$. In Eq. (2), the probability of state sequence is:

$$P(Q \mid \lambda) = \prod_{i=1}^{K} p_i(d_i), \quad (4)$$

where d_i is the duration of state i, $p_i(\cdot)$ is the duration probability density function of state i, K is the total state number of the utterance. When single Gaussian distribution is used as duration model, Eq. (4) can be further written as:

$$P(Q \mid \lambda) = \prod_{i=1}^{K} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(d_i - \mu_i)^2}{2\sigma_i^2}\right), \quad (5)$$

where $N(\mu_i, \sigma_i^2)$ is the duration probability distribution of state i. It's easy to see that the above probability is maximized when each state duration d_i is equal to μ_i. Thus the optimal state duration is determined.

III. Proposed State Duration Generation Considering GV

A. GV on Syllable Duration

Considering an utterance with M syllables, the duration vector over M syllables is $\boldsymbol{d} = [d_1, d_2, \cdots, d_m, \cdots, d_M]^{\mathrm{T}}$, where d_m is the duration of syllable m. The GV on syllable duration is defined as follows:

$$v(\boldsymbol{d}) = \frac{1}{M}\sum_{i=1}^{M}(d_i - \overline{d})^2, \quad (6)$$

$$d_i = \sum_{k=1}^{K_i}\sum_{j=1}^{N} d_{ikj}, \quad (7)$$

$$\overline{d} = \frac{1}{M}\sum_{i=1}^{M} d_i, \quad (8)$$

where K_i is the number of HMMs in syllable i, N is the state number of HMM topology, and d_{ijk} is the duration of state j, HMM k, and syllable i.

Fig. 1 shows a sequence of syllable duration extracted from natural Mandarin speech and that generated from HMMs. As can be seen, the variation of syllable duration of natural speech is much larger than that of generated speech. In other words, the syllable duration of synthetic speech is closer to the mean vector of syllable duration. This is partially due to the fact that a state-based HMM is inadequate in modeling a global and hierarchical prosody structure at utterance level, and partially due to the statistical averaging during the estimation of HMM duration model. This over-averaging effect on the sequence of generated syllable duration tends to make the prosody of synthetic speech sound bland.

Since one statistical characteristic of natural speech versus generated speech is that the GV of syllable duration in natural speech is obviously larger than that in synthetic speech, the over-averaging effect on generated syllable duration is expected to be much alleviated by integrating a duration GV model into the conventional generation of state duration based on HMM state duration model.

It should be noted that syllable level is chosen to have duration GV model on in this paper. However, it is quite similar to build duration GV model on other level, e.g. phone level, or multi-level.

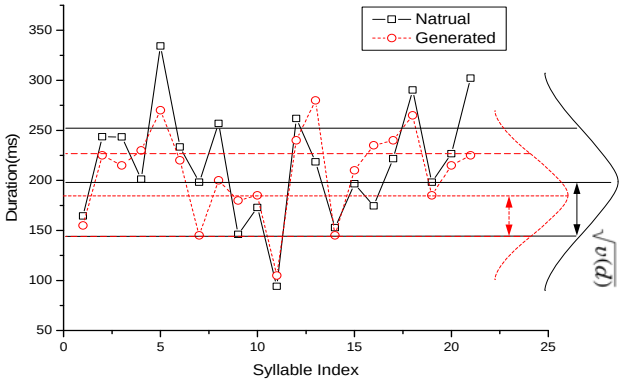

Fig. 1 Syllable duration sequences of natural speech and generated speech. The square root of duration GV is shown.

B. Duration generation considering GV

To integrate duration GV model into the generation of state duration, the proposed likelihood function consists of both conventional HMM state duration likelihood and GV likelihood, which is:

$$L = \log(P(\boldsymbol{d} \mid \lambda_d)P(v(\boldsymbol{d}) \mid \lambda_v)^{\omega}), \quad (9)$$

where λ_d is the conventional HMM state duration model, $\boldsymbol{d} = [\cdots, d_{ikj}, \cdots]^{\mathrm{T}}$ is the vector of state duration over the whole utterance, λ_v is the proposed GV model on syllable duration, $v(\boldsymbol{d})$ is GV calculated on syllable duration over the utterance (see Eq. (6)-(8)), and ω is GV weight. In this paper ω is set to the ratio of the numbers of dimensions between vectors $\boldsymbol{d}$ and $v(\boldsymbol{d})$, i.e., $N\sum_{i=1}^{M} K_i$. A single Gaussian distribution is used here to model the distribution of GV on syllable duration. The above likelihood can be further expanded as

$$L = -\frac{1}{2}\sum_{i=1}^{M}\sum_{k=1}^{K_i}\sum_{j=1}^{N}\sigma_{ikj}^{-2}(d_{ikj} - \mu_{ikj})^2 - \frac{\omega}{2}\sigma_v^{-2}(v(\boldsymbol{d}) - \mu_v)^2, \quad (10)$$

where μ_{ikj} and σ_{ikj}^2 are the mean and variance of state duration Gaussian of state j, HMM k, and syllable i, μ_v and σ_v^2 are mean and variance of duration GV Gaussian. To determine

the optimal state duration vector $\boldsymbol{d}^*$, we can iteratively update $\boldsymbol{d}$ by steepest descent algorithm as follow

$$\boldsymbol{d}^{(i+1)-th} = \boldsymbol{d}^{(i)-th} + \alpha \frac{\partial L}{\partial \boldsymbol{d}}\bigg|_{\boldsymbol{d}=\boldsymbol{d}^{(i)-th}}, \quad (11)$$

where α is the step size. With the likelihood L defined in (10), the gradient in (11) with respect to each d_{ikj} is calculated as

$$\frac{\partial L}{\partial d_{ikj}} = -\sigma_{ikj}^{-2}(d_{ikj} - \mu_{ikj}) - \frac{2\omega\sigma_v^{-2}}{M}(v(\boldsymbol{d}) - \mu_v)(d_i - \bar{d}). \quad (12)$$

As to the initial state duration vector $\boldsymbol{d}^{(0)-th}$ for iteration, there are two kinds of settings. One is the conventional state duration $\hat{\boldsymbol{d}} = [..., \hat{d}_{ikj}, ...]^{\mathrm{T}}$ generated by maximizing HMM state duration likelihood, where $\hat{d}_{ikj}$ is actually the mean of each HMM state duration Gaussian. The other is to use $\boldsymbol{d}' = [..., d'_{ikj}, ...]^{\mathrm{T}}$ which is linearly converted from the conventional one as follows

$$d'_{ikj} = \hat{d}_{ijk} + (d'_i - \hat{d}_i)\frac{\sigma^2_{ikj}}{\sum_{m=1}^{K_i}\sum_{n=1}^{N}\sigma^2_{imn}}, \quad (13)$$

$$d'_i = \sqrt{\frac{\mu_v}{v(\hat{\boldsymbol{d}})}}(\hat{d}_i - \bar{\hat{d}}) + \bar{\hat{d}}, \quad (14)$$

where $\hat{d}_i$ is the syllable duration accumulated from $\hat{\boldsymbol{d}}$ and $\bar{\hat{d}}$ is the mean of $\hat{d}_i$ sequence, which can be calculated according to Eq. (7) and (8). $\hat{\boldsymbol{d}}$ maximizes HMM state duration likelihood, while $\boldsymbol{d}'$ maximizes syllable duration GV likelihood. With setting GV weight as described above, we find that $\boldsymbol{d}'$ usually has a larger value of the proposed likelihood than $\hat{\boldsymbol{d}}$, which usually leads to a better convergency of iteration. Therefore, $\boldsymbol{d}'$ is taken as the initial state duration vector.

IV. EXPERIMENT

A. System Overview

Fig. 2 is the block diagram of HMM-based speech system with the proposed state duration generation. It consists of training part and synthesis part. In the training part context dependent HMMs and GV models are trained separately. In the synthesis part, the input text is firstly analyzed by a text analyzer and context features are extracted. Then the state duration of each context dependent HMM is generated by the proposed state duration generation algorithm considering GV. After that, the speech parameters are generated by the speech parameter generation algorithm with dynamic features. Finally, the synthetic speech is generated by a parametric synthesizer.

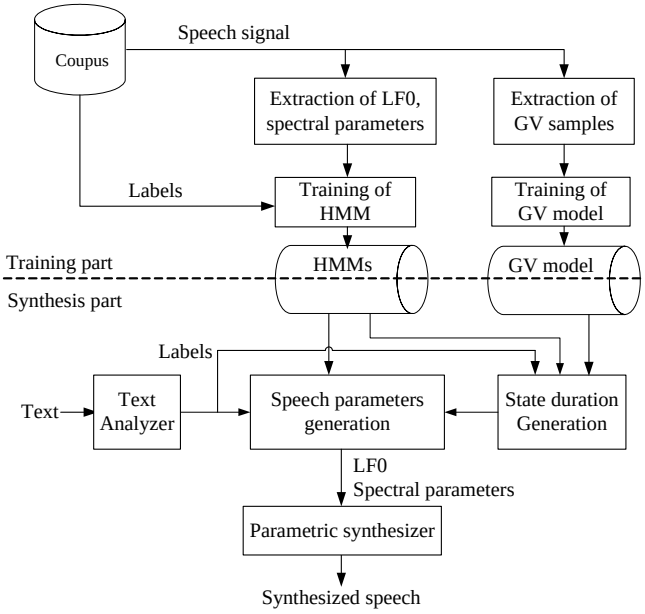

Fig. 2 Block diagram of HMM-based speech system with the proposed state duration generation method.

B. Experimental conditions

We used a 2-hour phonetically balanced Mandarin corpus for training, which consisted of 2000 sentences. Speech signals were sampled at 16kHz/16bit. F0, spectral envelope were extracted by STRAIGHT [12] with a 5ms frame shift. The spectral envelope was then used to extract 24-order LSPs and an extra gain dimension. A 5-state left-to-right with no skip HMM structure was adopted to model each phoneme of Mandarin. The feature vector consisted of log-scaled F0, LSPs, and their velocity and acceleration coefficients. Duration GV model was built on syllable level and a single Gaussian distribution was used to model the distribution of GV. In synthesis part, the generated LSPs were firstly converted to LPCs. Then a LP filter was used to synthesize the speech.

To compare the proposed method with the conventional one, two systems were built in our experiment.

· *Baseline*: An HMM-based speech synthesis system with conventional state duration generation method.

· *Proposed*: An HMM-based speech synthesis system with proposed state duration generation method.

In the iterative updating procedure of state duration generation for *proposed* system, the weight ω was set to the ratio of the numbers of dimensions between vectors $\boldsymbol{d}$ and $v(\boldsymbol{d})$, the step size α was initially set to 0.1, the convergence threshold was set to 0.0001. To increase the convergence

speed, the step size will be increased or reduced according to the polarity of likelihood change after each time of iteration, with a factor 1.2 and 0.5 respectively.

C. *Convergency of Iteration*

Fig. 3 and 4 show two examples of the convergence curves of the likelihood used in the state duration generation of the proposed system, where L denotes the total likelihood and Lgv denotes the likelihood of GV. In Fig. 3, the converted state duration $\boldsymbol{d}'$ described in section 3.2 is used. While in Fig. 4, the unconverted state duration $\hat{\boldsymbol{d}}$ described in section 3.2 is used. As we can see, the initial likelihood L in Fig. 3 is obviously larger than that in Fig. 4. And the convergence of L in Fig. 3 is faster than that in Fig. 4. This indicates that the adopted converted state duration is better for the convergence of iteration, with the GV weight set as described. From Fig. 4, we still can find that the likelihood of GV model increases significantly during the process of iteration. By improving the likelihood of GV model, the over-averaging effect on generated state duration could be alleviated.

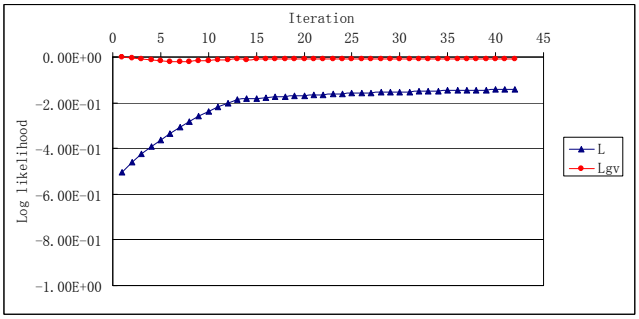

Fig. 3 An example of the convergence curves of the proposed likelihood and likelihood of GV model with the converted initial state duration.

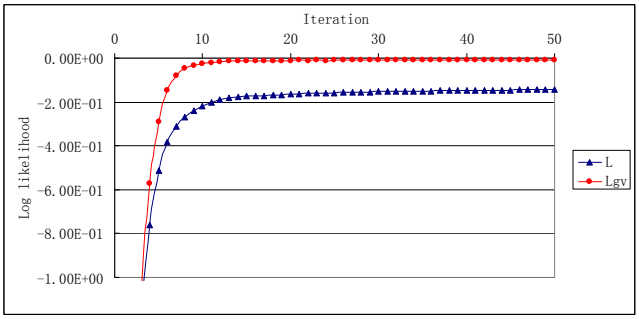

Fig. 4 An example of the convergence curves of the proposed likelihood and likelihood of GV model with the unconverted initial state duration.

D. *Subjective evaluation*

20 sentences out of the training set were synthesized by the two systems respectively. An AB preference test was conducted to evaluate the proposed method. 10 subjects which are all graduated students participated in the test. The subject was required to make a decision for each testing pair if which one sounds more natural or no preference could be made. The results of the test with 95% confidence interval are given in Fig. 5.

As we can see the proposed system outperforms the baseline system perceptually. The subjects indicated that the speech synthesized by proposed method sounds more natural and expressive in many cases from the perception of syllable duration. However, the duration of some syllables in the speech synthesized by proposed method is too long or short in a few cases, which made the synthetic speech sound unnatural.

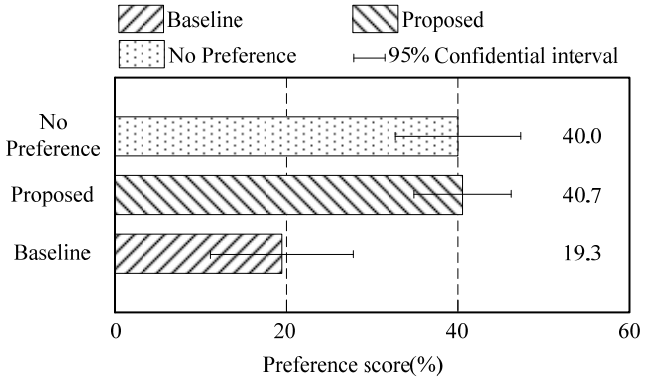

Fig. 5 Preference scores of the two systems.

Still, the syllable duration of speech synthesized by the two systems are quite similar in many cases, which is represented by the “no preference” item shown in Fig. 5.

The occurrence of the above two cases where the baseline voice was preferred and no preference was made is highly related to the accuracy of GV model. In this paper, a single model was trained to describe the distribution of GV on syllable duration. Though it is effective in many cases (about 40.7%), it is still lack of accuracy in other cases. In the case when the mean of GV model distribution is close to the GV of syllable duration generated by maximizing HMM likelihood, the likelihood of GV is close to its maximum. Hence no penalty could be made by the incorporating of GV likelihood into the likelihood function. Still, in the case when the mean of GV model distribution is much greater than the GV of syllable duration generated by maximizing HMM likelihood, the over-adjustment of syllable duration occurs due to the contribution of GV likelihood. Therefore, a more accurate GV model which covers most context environment, e.g., a context dependent GV model, will further improve the naturalness of synthetic speech.

V. CONCLUSIONS

In this paper, a state duration generation algorithm considering global variance (GV) for HMM-based speech synthesis is proposed. A GV model on syllable duration is proposed and a state duration generation algorithm considering this GV model is presented in details. The experimental results are promising which show that the

proposed method outperforms the conventional one by the way of alleviating the over-averaging effect of the generated state duration. We also find the accuracy of GV model adopted in this paper is still not good enough. Therefore, the next step is to build a context dependent GV model on syllable duration and further improvement is expected. Moreover, a multi-level GV model, e.g., GV model on phone duration, syllable duration, or even phrase duration, is worthy of studying.

ACKNOWLEDGMENT

The work was supported by the National Science Foundation of China (No. 60873160, 61011140075 and 90820303) and China-Singapore Institute of Digital Media (CSIDM).

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 *Eurospeech*, pp. 2347-2350, 1999.

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

[3] H. ZEN, T. Toda, and M. Nakamura, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE transactions on information and systems*, vol. E90-D, no. 1, pp. 325-333, 2007.

[4] Z. Ling, Y. Wu, Y. Wang, L. Qin, and R. Wang, "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Eurospeech*, pp. 2263-2266, 2001.

[6] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An HMM-based Mandarin Chinese text-to-speech system," in Proc. ISCSLP, 2006, pp. 223-232.

[7] Y. Wu, and R. Wang, "HMM-based Trainable Speech Synthesis for Chinese," *Journal of Chinese Information Processing*, 75-81, 2006.

[8] B. Gao, Y. Qian, Z. Wu, and F. K. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood," in *Proc. Interspeech*, pp. 2266-2269, 2008.

[9] C. Wang, Z. Ling, B. Zhang, and L. Dai, "Multi-layer F0 modeling for HMM-based speech synthesis," in *Proc. ISCSLP*, pp. 129-132, 2008.

[10] Y. Qian, Z. Wu, and F. K. Soong, "Improved prosody generation by maximizing joint likelihood of state and longer units," in *Proc. ICASSP*, pp. 3781-3784, 2009.

[11] T. Toda, and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inform. Systems*, vol. E90-D, no. 5, pp. 816-824, 2005.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187-208, 1999.