GLOBAL VARIANCE MODELING ON FREQUENCY DOMAIN DELTA LSP FOR HMM-BASED SPEECH SYNTHESIS

Shifeng Pan,[1] *Yoshihiko Nankaku,*[2] *Keiichi Tokuda,*[2] *Jianhua Tao*[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, China
[2]Department of Computer Science, Nagoya Institute of Technology, Japan
[1]{sfpan, jhtao}@nlpr.ia.ac.cn, [2]nankaku@sp.nitech.ac.jp, tokuda@.nitech.ac.jp

ABSTRACT

The parameter generation algorithm considering global variance (GV) for HMM-based speech synthesis has proved to be effective against the over-smoothing problem. However, the correlation between dimensions of parameter vector is not sufficiently considered in the current GV model. For some parameters, e.g., Line Spectral Pairs (LSP), the difference of adjacent LSPs has the strong influence on the spectral envelop. Considering this important feature, the paper proposes a GV modeling on the difference of adjacent LSPs, i.e., GV on frequency domain delta LSP. By improving the GV likelihood on frequency domain delta LSP, the over-smoothing effect of generated parameter trajectory is better alleviate than conventional one. The result of a perceptual evaluation shows the proposed method outperforms the conventional one, and the naturalness of synthetic speech is improved.

***Index Terms*—** speech synthesis, hidden Markov model, global variance

1. INTRODUCTION

The Hidden Markov Model (HMM)-based speech synthesis has been widely used in recent years. In this method, the 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].

However, the generated excitation and spectral parameters based on conventional speech parameter generation algorithm [2] are often over-smoothed. The reconstructed speech using over-smoothed speech parameters sounds muffled. Many methods have been proposed to alleviate this muffled effect, such as post-filtering methods [4] [5], incorporating the stream of the difference of adjacent LSPs to HMM feature vector[6] ,the rich context model [7] and the conditional speech parameter generation algorithm [8], etc. A speech parameter generation algorithm considering global variance (GV) was also proposed to solve this problem [9]. In this method, a GV model was trained to model the variation of parameter trajectories at utterance level. The generated parameter sequence maximizes a likelihood based on not only an HMM likelihood but also a GV likelihood. The latter likelihood works as a penalty for reduction of the GV of the generated parameter trajectories. This method is proved to be effective against the over-smoothing problem and can improve the naturalness of synthetic speech [9]. However, the GV likelihood for each dimension of speech parameter is independent in [10]. Though a full covariance matrix of GV model can be trained, it's still too loose to model the correlation between parameters. For spectral parameter like ne Spectral Pair (LSP), there's strong correlation between adjacent LSPs. Better exploiting this property on modeling GV will achieves a better performance. In [10], a GV on power spectrum derived from LSPs is modeled. By establishing the relationship between power spectrum and LSPs, the correlation between LSPs is taken into account to some extent. However, the correlation still has not been sufficiently considered.

In this paper, an improved parameter generation algorithm considering GV using LSPs as spectral parameter is proposed. Considering the property that the difference of adjacent LSPs can greatly affect the shape of spectral envelope, such as formant peak and formant bandwidth, a GV model on the difference of adjacent LSPs is built (for simplification, we call it GV model on frequency domain delta LSP). During the stage of speech parameter generation, this model works as a penalty for the reduction of GV of frequency domain delta LSP. Experimental results show the effectiveness of proposed method.

The rest of this paper is organized as follows. In section 2, the conventional parameter generation algorithm considering GV is reviewed. Section 3 describes the proposed GV model in details. In section 4 the evaluation result is presented. The conclusion is given in section 5.

2. CONVENTIONAL PARAMETER GENERATOIN ALGORITHM CONSIDERING GV

Assume a D-dimensional static feature vector $\boldsymbol{c}_t = [c_t(1), c_t(2), \cdots, c_t(d), \cdots, c_t(D)]^{\mathrm{T}}$ at frame t and a static feature vector sequence $\boldsymbol{c} = [\boldsymbol{c}_1^{\mathrm{T}}, \boldsymbol{c}_2^{\mathrm{T}}, \cdots, \boldsymbol{c}_t^{\mathrm{T}} \cdots, \boldsymbol{c}_T^{\mathrm{T}}]^{\mathrm{T}}$ over T frames, the GV is calculated by

$$\boldsymbol{v}(\boldsymbol{c}) = [v(1), v(2), \cdots, v(d), \cdots, v(D)]^{\mathrm{T}}, \tag{1}$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}(c_t(d) - \bar{c}(d))^2, \tag{2}$$

$$\bar{c}(d) = \frac{1}{T}\sum_{t=1}^{T} c_t(d) \cdot \tag{3}$$

At training stage, the GVs extracted from training sentences are used to train the GV model with single Gaussian distribution. During the stage of parameter generation, the optimum static feature parameter sequence $\boldsymbol{c}^*$ is derived by maximizing the following log-scaled likelihood

$$L = \log\left[P(\boldsymbol{Wc} \mid \hat{\boldsymbol{Q}}, \boldsymbol{\lambda})^{\omega} P(\boldsymbol{v}(\boldsymbol{c}) \mid \boldsymbol{\lambda}_v) \right], \tag{4}$$

where the former probability represents the likelihood of HMM model $\boldsymbol{\lambda}$, the latter one represents the likelihood of GV model $\boldsymbol{\lambda}_v$, ω is the weight to balance these two likelihoods, $\hat{\boldsymbol{Q}}$ is the state sequence determined by maximizing the likelihood of state duration model, and $\boldsymbol{W}$ is a 3DT-by-DT velocity and acceleration matrix.

The incorporation of GV likelihood into the total likelihood penalties the reduction of GV, and therefore can alleviate the over-smoothing effect on speech parameter sequences generated only by maximizing HMM likelihood. However, with the GV modeled on each dimension of static feature vector independently, the contribution of GV likelihood for each dimension is also independent. For spectral parameter like LSP which has strong correlation between adjacent dimensions, the conventional GV model is not idea. Though a full covariance matrix of GV model can be trained, it's still too loose to model the correlation between dimensions.

3. PROPOSED GV MODEL ON FREQUENCY DOMAIN DELTA LSP

3.1. GV model on frequency domain delta LSP

One useful property of LSPs is that the closer two adjacent LSPs are, the more resonant the vocal tract filter is at the corresponding frequency, which means the formant peak is sharper. This property has been exploited to perform the LSP-based formant enhancement [4]. Here, we take the advantage of this property by another way, i.e., building a GV model on frequency domain delta LSP. The generated speech parameters maximize the likelihood including conventional likelihood of HMM and likelihood of GV model on frequency domain delta LSPs. The latter likelihood works as a penalty for reduction of the GV of frequency domain delta LSPs. In this way, the over-smoothing effect on spectral envelope reconstructed by generated LSPs can be better alleviated.

Assume a D-dimensional static feature LSPs vector $\boldsymbol{c}_t = [c_t(1), c_t(2), \cdots, c_t(d), \cdots, c_t(D)]^{\mathrm{T}}$ at frame t, the ($D+1$)-dimensional delta LSPs vector $\delta\boldsymbol{c}_t = [\delta c_t(1), \delta c_t(2), \cdots, \delta c_t(d), \cdots, \delta c_t(D+1)]^{\mathrm{T}}$ is calculated as follows

$$\delta c_t(d) = \begin{cases} c_t(1), & d = 1 \\ c_t(d) - c_t(d-1), & d = 2, 3, \cdots, D \\ \pi - c_t(D), & d = D+1 \end{cases} \cdot \tag{5}$$

Note that the gain of linear predictive analysis is not included in the static feature vector $\boldsymbol{c}_t$ here, and it can be generated by conventional parameter generation method. The GV model on frequency domain delta LSP is defined as follows

$$\boldsymbol{v}(\delta\boldsymbol{c}) = [v(1), v(2), \cdots, v(d), \cdots, v(D+1)]^{\mathrm{T}}, \tag{6}$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}(\delta c_t(d) - \overline{\delta c}(d))^2, \tag{7}$$

$$\overline{\delta c}(d) = \frac{1}{T}\sum_{t=1}^{T} \delta c_t(d) \cdot \tag{8}$$

A single Gaussian distribution $N(\boldsymbol{\mu}_v, \boldsymbol{U}_v)$ is trained to model the distribution of GV on frequency domain delta LSP, using the GV vectors calculated from each training sentence.

3.2. Parameter generation considering GV on frequency domain delta LSP

The speech parameters are generated by maximizing the following likelihood

$$L = \log\left[P(\boldsymbol{Wc} \mid \hat{\boldsymbol{Q}}, \lambda) P(\boldsymbol{v}(\delta\boldsymbol{c}) \mid \boldsymbol{\lambda}_v^{\delta})^{\omega} \right], \tag{9}$$

where $\boldsymbol{\lambda}_v^{\delta}$ is GV model and ω is GV weight. The above likelihood can be further expanded as

$$\begin{aligned} L = &-\frac{1}{2}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\hat{\boldsymbol{U}}^{-1}\boldsymbol{Wc} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\hat{\boldsymbol{U}}^{-1}\hat{\boldsymbol{\mu}} \\ &-\frac{1}{2}\omega\boldsymbol{v}(\delta\boldsymbol{c})^{\mathrm{T}}\boldsymbol{U}_v^{-1}\boldsymbol{v}(\delta\boldsymbol{c}) + \omega\boldsymbol{v}(\delta\boldsymbol{c})^{\mathrm{T}}\boldsymbol{U}_v^{-1}\boldsymbol{\mu}_v + K \end{aligned}, \tag{10}$$

where $\hat{\boldsymbol{\mu}} = [\hat{\boldsymbol{\mu}}_1^{\mathrm{T}}, \cdots, \hat{\boldsymbol{\mu}}_t^{\mathrm{T}}, \cdots, \hat{\boldsymbol{\mu}}_T^{\mathrm{T}}]^{\mathrm{T}}$ and $\hat{\boldsymbol{U}} = diag(\hat{\boldsymbol{U}}_1^{\mathrm{T}}, \cdots, \hat{\boldsymbol{U}}_t^{\mathrm{T}}, \cdots, \hat{\boldsymbol{U}}_T^{\mathrm{T}})$ are the mean vector and covariance matrix of state sequence $\hat{\boldsymbol{Q}}$, and K is independent of $\boldsymbol{c}$.To determine the optimum parameter vector sequence $\boldsymbol{c}^*$, we can iteratively update $\boldsymbol{c}$ by steepest descent algorithm,

$$\boldsymbol{c}^{(i+1)-th} = \boldsymbol{c}^{(i)-th} + \alpha \left. \frac{\partial L}{\partial \boldsymbol{c}} \right|_{\boldsymbol{c}=\boldsymbol{c}^{(i)-th}} , \qquad (11)$$

where α is the step size. With the likelihood L defined in (10), the gradient in (11) can be calculated as

$$\frac{\partial L}{\partial \boldsymbol{c}} = (-\boldsymbol{W}^{\mathrm{T}}\hat{\boldsymbol{U}}^{-1}\boldsymbol{W}\boldsymbol{c} + \boldsymbol{W}^{\mathrm{T}}\hat{\boldsymbol{U}}^{-1}\hat{\boldsymbol{\mu}}) + [\boldsymbol{v}_1^{'\mathrm{T}}, \cdots, \boldsymbol{v}_t^{'\mathrm{T}}, \cdots, \boldsymbol{v}_T^{'\mathrm{T}}]^{\mathrm{T}}, \qquad (12)$$

$$\boldsymbol{v}_t^{'} = [v_t^{'}(1), \cdots, v_t^{'}(d), \cdots, v_t^{'}(D)]^{\mathrm{T}}, \qquad (13)$$

$$v_t^{'}(d) = -\omega \left[\frac{\partial v(1)}{\partial c_t(d)}, \cdots, \frac{\partial v(D+1)}{\partial c_t(d)} \right] \boldsymbol{U}_v^{-1} \left(\boldsymbol{v}(\delta \boldsymbol{c}) - \boldsymbol{\mu}_v \right), \qquad (14)$$

Since that only $v(d)$ and $v(d+1)$ are dependent on $c_t(d)$, (14) can be simplified as

$$v_t^{'}(d) = -\omega \left(\frac{\partial v(d)}{\partial c_t(d)} \boldsymbol{p}_v^{(d)^{\mathrm{T}}} + \frac{\partial v(d+1)}{\partial c_t(d)} \boldsymbol{p}_v^{(d+1)^{\mathrm{T}}} \right) \left(\boldsymbol{v}(\delta \boldsymbol{c}) - \boldsymbol{\mu}_v \right), \qquad (15)$$

where $\boldsymbol{p}_v^{(d)}$ and $\boldsymbol{p}_v^{(d+1)}$ are the *d*-th and (*d*+1)-th column vectors of $\boldsymbol{U}_v^{-1}$ respectively. $\partial v(d) / \partial c_t(d)$ and $\partial v(d+1) / \partial c_t(d)$ are calculated as follows,

$$\frac{\partial v(d)}{\partial c_t(d)} = \frac{\partial v(d)}{\partial \delta c_t(d)} \frac{\partial \delta c_t(d)}{\partial c_t(d)} = \frac{2(T-1)}{T^2} \left(\delta c_t(d) - \overline{\delta c}(d) \right) \qquad (16)$$

$$\frac{\partial v(d+1)}{\partial c_t(d)} = -\frac{2(T-1)}{T^2} \left(\delta c_t(d+1) - \overline{\delta c}(d+1) \right) \qquad (17)$$

4. EXPERIMENT

4.1. Experimental conditions

We used phonetically balanced 450 sentences from ATR Japanese speech database for training. Speech signals were sampled at 16kHz. The F0, spectral envelope and aperiodicity measure [11] were extracted by STRAIGHT [12] with a 5ms frame shift. The spectral envelope was then used to extract 40-order mel-LSPs and an extra gain dimension. A 5-state left-to-right ergodic multi-space probability distribution hidden-semi-Markov model (MSD-HSMM) [13] structure was adopted to model each phoneme of Japanese. The feature vector consisted of log- scaled F0, mel-LSPs, aperiodicity measures, and their velocity and acceleration coefficients. Conventional GV model was trained for log-scaled F0s, gain of linear predictive analysis and aperiodicity measures. The proposed GV model was trained for LSPs. Single Gaussian distribution was used to model the distribution of GV. In the synthesis, firstly, the spectral envelope was reconstructed by generated LSPs. Then the speech was synthesized by STRAIGHT with generated F0, spectral envelope and aperiodicity measures.

Three systems were compared in our experiment.

- *System A*: LSPs were generated by conventional parameter generation method based on HMM likelihood. Then LSP-based formant enhancement was performed on the generated LSPs.
- *System B*: LSPs were generated by parameter generation method with conventional GV model.
- *System C*: LSPs were generated by proposed method.

In system A, the factor of formant enhancement was set to 0.7. In system B, the weight ω was set to 1/3T. In system C, the GV weight was set to 0.3T and iteration step was set to 0.01. To prevent the occurrences of the too close adjacent LSPs in system C, a minimum distance threshold 0.04 was set. The generated LSPs were checked and adjusted to satisfy the threshold.

4.2. Subjective evaluation

53 sentences out of the training set were synthesized by the three systems respectively. 10 out of the 53 test sentences were randomly selected for an AB comparison preference test. 10 Japanese listeners were forced to choose one which sounds more natural from each pair. The results of the preference test with 95% confidence interval are given in Fig. 1.

Though the difference of adjacent LSPs is the target to adjust for both system A and C, the proposed parameter generation method which combines HMM likelihood and GV model likelihood is better than that of LSP-based formant enhancement which is performed only with an empirical formula. Actually, the naturalness of voice C is significantly better than A. For that the training data is very small, which is less than 1hour, sometimes one or several phones are not stable in voice A, which will be also not stable in voice C. However, the contrast between the stable and unstable parts of voice C is more intense than that of voice A. In such cases, many listeners prefer the latter. This is the reason why the preference of voice C to voice A is not as higher as we expected. If the trained HMMs are stable, the result will be further improved.

The result of comparison between C and B indicates that the proposed method outperforms the conventional one. Due to the GV increasing on the difference of adjacent LSPs, the formant of reconstructed spectral envelope by proposed method is more enhanced than that by conventional one, which leads to a more articulate synthetic voice. An example of spectrum sequences generated by the three systems is shown in Fig. 2. As we can see, there’s no clear distinction between spectrums of voice A and B,

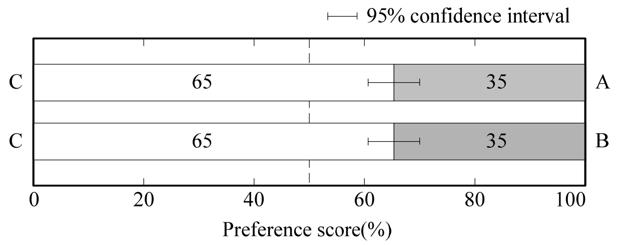

Fig. 1 Preference scores of the three systems.

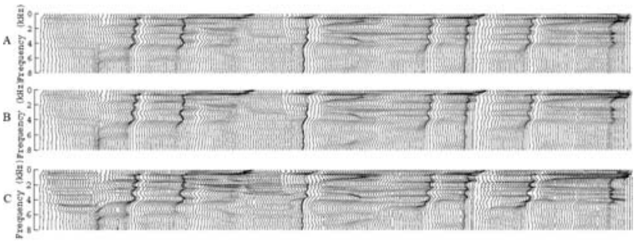

Fig. 2 An example of spectrum sequences generated by the three systems.

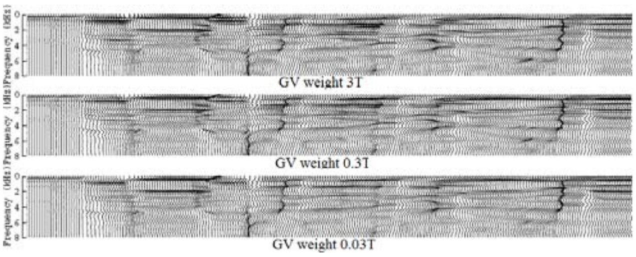

Fig. 3 An example of spectrum sequences generated by the systems C with different GV weight

meanwhile, the formant structure of voice C is more enhanced than A and B.

Another difference we found in the experiment was that the weight of conventional GV likelihood was not sensitive to the shape of generated spectral envelope. However, the weight of proposed GV likelihood was very sensitive. The larger the GV weight is, the more enhanced the formant structure is, as Fig. 3 shows. This also shows the strong effectiveness of proposed GV model on formant structure. However, the synthetic voice also sounds too sharp and unnatural when GV weight is too large. It is necessary to tune GV weight to balance the articulation and unnaturalness of synthetic voice.

5. CONCLUSIONS

In this paper, a GV modeling on frequency domain delta LSP is proposed, and the parameter generation with the new GV model is described in detail. The proposed method can alleviate the over-smoothing effect of generated spectral envelope better than the conventional method. The experimental results are promising and the proposed method outperforms the conventional one. Considering that the improvement is still not significant with the combination of conventional HMM and the proposed GV model, we will further attempt to add the stream of frequency domain delta LSP to HMM feature vector in the future.

6. ACKNOWLEDGEMENTS

The research leading to these results was partly funded by the National Science Foundation of China (No. 60873160), 863 Programs (No. 2009AA01Z320), NSFC-JSPS joint project (No. 61011140075), the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project), and the Strategic Information and Communications R&D Promotion Programme (SCOPE), of the Ministry of Internal Affairs and Communication, Japan.

7. REFERENCES

[1] T. Yoshimura, K. Tokuda, etc, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, ” in Eurospeech, pp. 2347-2350, 1999.

[2] K. Tokuda, T. Yoshimura, etc, “Speech parameter generation algorithms for HMM-based speech synthesis,” in ICASSP, vol. 3, pp. 1315-1318, 2000.

[3] H. Zen, T. Toda, etc, “Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005,” IEICE Trans. Inf. & Syst., vol. E90-D, no. 1, pp. 325－333, 2007.

[4] Z. Ling, Y. Wu, etc, “USTC system for Blizzard Challenge 2006: an improved HMM-based speech synthesis method,” in Blizzard Challenge Workshop, 2006.

[5] T. Yoshimura, K. Tokuda, etc, “Mixed excitation for HMM-based speech synthesis,” in Eurospeech, pp. 2263-2266, 2001.

[6] Y. Qian, F. Soong, etc, “An HMM-Based Mandarin Chinese Text-To-Speech System,” in Proc. of ISCSLP, pp. 223-232, 2006.

[7] Z.-J. Yan, Y. Qian, etc, “Rich context modeling for high quality HMM-based TTS,” in Interspeech, pp. 1755-1758, 2009.

[8] T. Masuko, K. Tokuda, etc, “A study on conditional parameter generation from HMM based on maximum likelihood criterion,” in Autumn Meeting of ASJ, pp. 209-210, 2003.

[9] T. Toda, and K. Tokuda, “Speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” in Proc. of Interspeech, pp. 2801-2804, 2005.

[10] Z.-H. Ling, Y. Hu, etc, “Global variance modeling on the log power spectrum of LSPs for HMM-based speech synthesis,” in Proc. of Interspeech, pp. 825-828, 2010.

[11] H. Kawahara, J. Estill, etc, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight,” in Proc. of MAVEBA, pp. 13-15, 2001.

[12] H. Kawahara, I. Masuda-Katsuse, etc, “Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instantaneous -frequency-based F0 extraction: possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, pp. 187-207, 1999.

[13] H. Zen, K. Tokuda, etc, “A hidden semi-Markov model-based speech synthesis system,” IEICE Trans. Inf. & Syst., vol. E90-D, no. 5, pp. 825-834, May 2007.