Considering Global Variance of the Log Power Spectrum Derived from Mel-Cepstrum in HMM-based Parametric Speech Synthesis

Xiang Yin, Zhen-Hua Ling, Ming Lei, Li-Rong Dai

iFLYTEK Speech Lab, University of Science and Technology of China, P.R.China

byx1030@mail.ustc.edu.cn, zhling@ustc.edu.cn, leiming@mail.ustc.edu.cn, lrdai@ustc.edu.cn

Abstract

This paper utilizes global variance (GV) of the log power spectrum (LPS) derived from mel-cepstrum to improve hidden Markov model (HMM) based parametric speech synthesis. In order to alleviate over-smoothing of the generated spectral structures, an LPS-GV modeling method using line spectral pairs (LSPs) has been proposed in our previous work, where the estimated distribution of LPS-GV was combined with the trained acoustic model to determine the optimal spectral features at synthesis time. In this paper, we extend this method to the condition where mel-cepstral coefficients are used as spectral features. Further, a method of integrating LPS-GV distortions into the criterion of minimum generation error (MGE) model training is proposed in order to avoid high computational complexity of the parameter generation algorithm with GV model. Experimental results show that the parameter generation algorithm using LPS-GV model produces more natural acoustic features than the conventional GV modeling method when mel-cepstrum features are adopted. Besides, integrating LPS-GV distortions into model training criterion achieves similar performance as applying LPS-GV model at synthesis time.

Index Terms: Speech synthesis, hidden Markov model, global variance, log power spectrum

1. Introduction

Hidden Markov model (HMM) based parametric speech synthesis method was proposed in 1990's [1] and has become a mainstream speech synthesis method in recent years. In model training of this method, the spectrum, F0 and duration are modeled simultaneously within a unified framework of HMMs [2]. At synthesis stage, these features are directly predicted from the HMMs through a maximum likelihood parameter generation (MLPG) algorithm under the constraint between static and dynamic features [3]. The speech waveforms are reconstructed by high quality vocoder from the predicted parameter trajectories. This method can synthesize highly intelligible and smooth speech sounds [4].

Although the HMM-based parametric speech synthesis has its advantages, the quality of its synthetic speech degrades due to the over-smoothing issue of the generated acoustic features. Some methods have been proposed to address this problem. A parameter generation algorithm considering global variance (GV) was proposed in [5]. GV represented the temporal variances of spectral parameters in each sentence. In this method, the optimal spectral feature sequences were generated by maximizing the combination of the conventional likelihood of acoustic features and the likelihood of GV vector. Some other methods attempted to improve the model training criterion. A minimum generation error (MGE) training method was proposed in [6], where the acoustic model was estimated by minimizing the distance between predicted and natural acoustic features. Additionally, GV has also been incorporated into the MGE criterion, where an additional component considering global/local (GV/LV) was introduced into the generation error [7].

However, comparing with the spectral features which are used to parameterize the spectral envelope of each frame, the spectral envelope itself is more directly related with the subjective perception in terms of speech quality. Therefore, a GV modeling method for the log power spectrum (LPS) was proposed in our previous work [8] to improve the performance of conventional GV method when line spectral pairs (LSPs) are used as spectral features. Besides LSPs, mel-cepstrum is also widely used in the HMM-based parametric speech synthesis systems. And the conventional GV method only considers the variance of mel-cepstrum dimensions separately, ignoring the inherent correlation among different dimensions. Therefore, we investigate the temporal variances of the LPS derived from mel-cepstrum in this paper. First, we examine the effect of combining the LPS-GV model using mel-cepstrum into the parameter generation criterion, like [5][6]. Then, a method of utilizing LPS-GV at training stage is proposed for the purpose of reducing computation cost at synthesis time. Similar to [7], we redefine the generation error of MGE training by adding a new component, which measures the distortion between the generated and the natural LPS-GV vectors.

This paper is organized as follows. Section 2 introduces the parameter generation algorithm considering the LPS-GV model derived from mel-cepstrum In section 3, the method of incorporating LPS-GV distortion into the generation error of MGE training is described. Section 4 evaluates the performance of our proposed methods by experiments. Section 5 is the conclusion.

2. GV Modeling on Log Power Spectrum Derived from Mel-Cepstrum

2.1. LPS-GV on mel-cepstrum

In order to alleviate the over-smoothing issue of generated spectral envelope, a method to calculate and model the GV on LPS was proposed in [8], where LSPs were adopted as spectral feature and the LPS was derived from LSPs, i.e. LPS-GV on LSPs. In this paper, we study LPS-GV on mel-cepstrum and investigate its applications.

Assume $\boldsymbol{c}=[\boldsymbol{c}_1^{\mathrm{T}},\boldsymbol{c}_2^{\mathrm{T}},...,\boldsymbol{c}_T^{\mathrm{T}}]^{\mathrm{T}}$ denotes the static mel-cepstrum feature sequence for a sentence, where T is the number of frames; $\boldsymbol{c}_t=[c_{t,1},c_{t,2},...,c_{t,D}]^{\mathrm{T}}$ and D is the dimension of $\boldsymbol{c}_t$; LPS sequence calculated from $\boldsymbol{c}$ is $\boldsymbol{s}=[\boldsymbol{s}_1^{\mathrm{T}},\boldsymbol{s}_2^{\mathrm{T}},...,\boldsymbol{s}_T^{\mathrm{T}}]^{\mathrm{T}}$, where $\boldsymbol{s}_t=[s_{t,1},s_{t,2},...,s_{t,K}]^{\mathrm{T}}$ and K is the number of sampling points within frequency range $[0, \pi)$. Based on the definition of mel-cepstrum [9], $s_{t,k}$ at the t-th frame and the k-th frequency point $\omega_k = k\pi/K$ can be calculated as

$$s_{t,k} = \frac{20}{\ln 10}\sum_{d=1}^{D} c_{t,d}\cos(\varpi_k d) \quad , \tag{1}$$

where

$$\varpi_k = \tan^{-1}\frac{(1\text{-}\alpha^2)\cdot\sin\omega_k}{(1+\alpha^2)\cdot\cos\omega_k\text{-}2\alpha} \tag{2}$$

is an approximation to the mel-frequency scale and the frequency warping parameter α was set to 0.42 in our implementation. Then, the GV vector $\boldsymbol{v}(\boldsymbol{s})$ of the LPS $\boldsymbol{s}$ is calculated by

$$\boldsymbol{v}(\boldsymbol{s}) = [v(\boldsymbol{s})_1, v(\boldsymbol{s})_2, ..., v(\boldsymbol{s})_K]^{\mathrm{T}}, \tag{3}$$

$$v(\boldsymbol{s})_k = \frac{\sum_{t=1}^{T}(s_{t,k} - \overline{s}_k)^2}{T}, \tag{4}$$

$$\overline{s}_k = \frac{\sum_{t=1}^{T} s_{t,k}}{T}. \tag{5}$$

Contrast to the conventional GV calculating using mel-cepstrum, the LPS-GV vector $\boldsymbol{v}(\boldsymbol{s})$ considers temporal variance of each frequency point in the spectral envelope.

2.2. Parameter generation with LPS-GV model

In the parameter generation method considering LPS-GV, the optimal features are generated to maximize the combination of the conventional likelihood for acoustic features and the likelihood of the LPS-GV vector.

At training stage, a single Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$ is trained using the LPS-GV vectors of all training sentences to get the LPS-GV model λ_s. This model is incorporated into the MLPG algorithm to predict the mel-cepstrum features at synthesis time. Similar to [8], the criterion of parameter generation can be defined as

$$\begin{aligned} \boldsymbol{c}^* &= \arg\max_{\boldsymbol{c}} L \\ &= \arg\max_{\boldsymbol{c}}\left\{\log\left[p(\boldsymbol{Wc}\,|\,\lambda,\boldsymbol{q})^{w^{(s)}}\cdot p(v(\boldsymbol{s})\,|\,\lambda_s)\right]\right\}, \end{aligned} \tag{6}$$

where $w^{(s)}$ denotes the weight to balance the two likelihood functions; $\boldsymbol{W}$ is a matrix determined by the velocity and acceleration calculation functions [3]; $\boldsymbol{q}$ is the state sequence predicted using state duration probabilities [2].

In order to determine $\boldsymbol{c}^*$, the steepest decent algorithm is used to update $\boldsymbol{c}$ iteratively.

$$\boldsymbol{c}^{(i+1)} = \boldsymbol{c}^{(i)} + \beta\cdot\left.\frac{\partial L}{\partial\boldsymbol{c}}\right|_{\boldsymbol{c}=\boldsymbol{c}^{(i)}}, \tag{7}$$

where i denotes the number of iterations and β is the step size. From the definition of function L in (6), we can calculate the gradient in (7) as

$$\frac{\partial L}{\partial\boldsymbol{c}} = w^{(s)}(-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{Wc} + \boldsymbol{W}^{\mathrm{T}}\boldsymbol{U}^{-1}\boldsymbol{m}) + [\boldsymbol{v}_1{}'^{\mathrm{T}}, \boldsymbol{v}_2{}'^{\mathrm{T}}, ..., \boldsymbol{v}_T{}'^{\mathrm{T}}]^{\mathrm{T}} \quad , \tag{8}$$

where

$$\boldsymbol{v}_t{}' = [v_t{}'(1), v_t{}'(2), ..., v_t{}'(D)]^{\mathrm{T}}, \tag{9}$$

$$v_t{}'(d) = -[\frac{\partial v(\boldsymbol{s})_1}{\partial c_{t,d}}, \frac{\partial v(\boldsymbol{s})_2}{\partial c_{t,d}}, ..., \frac{\partial v(\boldsymbol{s})_K}{\partial c_{t,d}}]\cdot\boldsymbol{\Sigma}_s^{-1}\cdot(\boldsymbol{v}(\boldsymbol{s}) - \boldsymbol{\mu}_s). \tag{10}$$

Based on the definition of LPS-GV in (3)-(5), we have

$$\frac{\partial v(\boldsymbol{s})_k}{\partial c_{t,d}} = \frac{2}{T}\cdot(s_{t,k} - \overline{s}_k)\cdot\frac{\partial s_{t,k}}{\partial c_{t,d}}\cdot(1 - \frac{1}{T}). \tag{11}$$

Further, the $\partial s_{t,k}/\partial c_{t,d}$ can be derived from (1) as

$$\frac{\partial s_{t,k}}{\partial c_{t,d}} = \frac{20}{\ln 10}\cos(\varpi_k d). \tag{12}$$

The generated static feature sequence is updated iteratively until the increasing of criterion function L is smaller than a given threshold ε.

3. MGE Training with LPS-GV

In this section, we investigate the method of utilizing LPS-GV at model training stage. Similar to [8], we adopt MGE training to incorporate LPS-GV measurement. The distortion between the generated and the original LPS-GV vectors is incorporated into the generation error of MGE training for acoustic model estimation.

3.1. Generation error considering LPS-GV

For the n-th sentence in the training set, the distortion between the LPS-GV of the generated mel-cepstrum trajectory $\tilde{\boldsymbol{c}}$ and the natural one $\boldsymbol{c}_n$ is defined as

$$D_v(\sigma(\boldsymbol{s}_n), \sigma(\tilde{\boldsymbol{s}})) = \|\sigma(\boldsymbol{s}_n) - \sigma(\tilde{\boldsymbol{s}})\|^2, \tag{13}$$

where $\boldsymbol{s}_n$ and $\tilde{\boldsymbol{s}}$ are the LPSs corresponding to $\boldsymbol{c}_n$ and $\tilde{\boldsymbol{c}}$ respectively;

$$\boldsymbol{\sigma}(\boldsymbol{s}) = [\sigma(\boldsymbol{s})_1, ..., \sigma(\boldsymbol{s})_K]^{\mathrm{T}}, \tag{14}$$

$$\sigma(\boldsymbol{s})_k = \sqrt{v(\boldsymbol{s})_k}. \tag{15}$$

This LPS-GV distortion is combined with the original distortion of Euclidean distance $D_c(\boldsymbol{c}_n, \tilde{\boldsymbol{c}})$ between the generated spectral features and the natural ones. Thus, the new generation error function is defined as

$$e'(\boldsymbol{c}_n, \tilde{\boldsymbol{c}}, \boldsymbol{\sigma}(\boldsymbol{s}_n), \boldsymbol{\sigma}(\tilde{\boldsymbol{s}})) = D_c(\boldsymbol{c}_n, \tilde{\boldsymbol{c}}) + w^{(t)}D_v(\boldsymbol{\sigma}(\boldsymbol{s}_n), \boldsymbol{\sigma}(\tilde{\boldsymbol{s}})), \tag{16}$$

where $w^{(t)}$ denotes the weight to balance between these two distortions at training stage.

Another option of adopting LPS-GV of mel-cepstrum into model training is to define generation error using only the LPS-GV distortion of (13). However, this kind of generation error may introduce unexpected modification to the model parameters, since (15) only considers LPS-GV of generated acoustic features. Therefore, we combine LPS-GV component with conventional generation error in MGE training, and use a weight parameter to control the contribution of each component.

3.2. Parameter updating

The acoustic model λ is estimated to minimize the generation error in (16) as

$$\hat{\lambda} = \arg\min_{\lambda}\sum_{n=1}^{N} e'(\boldsymbol{c}_n, \tilde{\boldsymbol{c}}, \boldsymbol{\sigma}(\boldsymbol{s}_n), \boldsymbol{\sigma}(\tilde{\boldsymbol{s}})). \tag{17}$$

where N is the number of sentences in the training set. The model parameters contain

$$\boldsymbol{\mu}=[\boldsymbol{\mu}_1^{\mathrm{T}},\boldsymbol{\mu}_2^{\mathrm{T}},...,\boldsymbol{\mu}_M^{\mathrm{T}}]^{\mathrm{T}},\ \boldsymbol{U}=[\Sigma_1^{-1},\Sigma_2^{-1},...,\Sigma_M^{-1}]^{\mathrm{T}}, \tag{18}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{U}$ are the concatenated mean vectors and inverse covariance matrices of all unique Gaussian distributions in the model set λ; $\boldsymbol{\mu}_m$ and $\sum_m$ are the mean vector and covariance matrix of the m-th unique Gaussian distribution; M is the total number of Gaussian distributions.

Here we use probabilistic descent (PD) method [10] to optimize the model parameters. For each training utterance, the model set is updated as

$$\lambda_{n+1}=\lambda_n-\varepsilon_n\boldsymbol{H}_n\left.\frac{\partial e'(\boldsymbol{c}_n,\tilde{\boldsymbol{c}},\boldsymbol{\sigma}(\boldsymbol{s}_n),\boldsymbol{\sigma}(\tilde{\boldsymbol{s}}))}{\partial\lambda}\right|_{\lambda=\lambda_n}, \tag{19}$$

where $\boldsymbol{H}_n$ is a positive definite matrix, and ε_n is the learning rate of the n -th iteration.

Based on $e'(\boldsymbol{c}_n,\tilde{\boldsymbol{c}},\boldsymbol{\sigma}(\boldsymbol{s}_n),\boldsymbol{\sigma}(\tilde{\boldsymbol{s}}))$ in (16), the gradient for the mean parameters can be calculated as

$$\frac{e'(\boldsymbol{c}_n,\tilde{\boldsymbol{c}},\boldsymbol{\sigma}(\boldsymbol{s}_n),\boldsymbol{\sigma}(\tilde{\boldsymbol{s}}))}{\partial\boldsymbol{\mu}}=2\boldsymbol{S}_q^{\mathrm{T}}\Sigma_q^{-1}\boldsymbol{W}\boldsymbol{R}_q^{-1}\boldsymbol{\eta} \tag{20}$$

where $\boldsymbol{S}_q$ is a zero-one matrix [7] used to select the parameters of the sentence HMM from the model set according to the state sequence $\boldsymbol{q}$;

$$\boldsymbol{\Sigma}_q^{-1}=\mathrm{diag}(\boldsymbol{S}_q\boldsymbol{U}), \tag{21}$$

$$\boldsymbol{R}_q=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}_q^{-1}\boldsymbol{W}, \tag{22}$$

$$\boldsymbol{\eta}=(\tilde{\boldsymbol{c}}-\boldsymbol{c}_n)+w^{(t)}\left.\frac{\partial\boldsymbol{s}^{\mathrm{T}}}{\partial\boldsymbol{c}}\right|_{\boldsymbol{c}=\boldsymbol{c}_n}\cdot\boldsymbol{A}(\tilde{\boldsymbol{s}}-\boldsymbol{m}(\tilde{\boldsymbol{s}})), \tag{23}$$

and $\boldsymbol{A}$ is a diagonal matrix, whose diagonal elements are

$$\boldsymbol{A}_{i,i}=1-\frac{\sigma(\boldsymbol{s}_n)_k}{\sigma(\tilde{\boldsymbol{s}})_k},\quad i=(t-1)*K+k. \tag{24}$$

The definitions of function $\mathrm{diag}(\cdot)$ in (21) and function $\boldsymbol{m}(\cdot)$ in (23) can be found in [7]. The elements of $\partial\boldsymbol{s}/\partial\boldsymbol{c}$ in (23) can be obtained according to (12).

The gradient of generation error function for the variance parameters is calculated as

$$\frac{\partial e'(\boldsymbol{c}_n,\tilde{\boldsymbol{c}},\boldsymbol{\sigma}(\boldsymbol{s}_n),\boldsymbol{\sigma}(\tilde{\boldsymbol{s}}))}{\partial\boldsymbol{U}}=2\boldsymbol{S}_q^{\mathrm{T}}\,\mathrm{diag}^{-1}(\boldsymbol{W}\boldsymbol{R}_q^{-1}\boldsymbol{\eta}(\boldsymbol{\mu}_q-\boldsymbol{W}\tilde{\boldsymbol{c}})), \tag{25}$$

where

$$\boldsymbol{\mu}_q=\boldsymbol{S}_q\boldsymbol{\mu}. \tag{26}$$

Applying (20) and (25) to (19), the acoustic model set λ is iteratively updated to minimize the generation error on the training set.

4. Experiments

4.1. Experimental conditions

A Chinese speech database recorded by a female speaker was used in our experiments. It consists of 1,100 sentences together with the segmental and prosodic labels, in which 1,000 sentences were randomly selected for mode training, and the remaining 100 sentences were used for test. Speech waveforms were recorded in 16kHz/16bit format. The acoustic features, including the logarithmized F0, 41-order mel-cepstrum (including 0-th order), were extracted with a 5ms frame shift by STRAIGHT [11] analysis. A 5-state left-to-right with no skip HMM structure was used to train context-dependent phone models. Single-mixture Gaussian distribution was adopted for the modeling of the state phone duration probabilities. Decision-tree-based model clustering [12] was applied in the context-dependent model training to avoid the data-sparsity problem. Besides, the question set for tree splitting was designed considering the characteristics of Chinese. During synthesis, speech waveforms were reconstructed using the spectral envelops derived from the generated mel-cepstrum by STRAIGHT vocoder. The number of frequency points in LPS was set to K =512.

Six systems were constructed in our experiment, including

- *Baseline*: Maximum likelihood based HMM training and conventional MLPG algorithm for synthesis;
- *GV*: Conventional GV method for generation as in [5];
- *LPS-GV*: LPS-GV modeling for generation as introduced in Section 2;
- *MGE*: Conventional MGE training for HMM estimation;
- *MGE_LPS-GV*: MGE training with additional LPS-GV component for generation error as introduced in Section 3;
- *MGE/LPS_GV*: Conventional MGE training plus LPS-GV modeling for generation.

In the *LPS-GV* and *MGE/LPS_GV* systems, the step size β and the convergence threshold ε were set to 0.001 and 0.01; the weight $w^{(s)}$ was set to $K/3DT$. In the *MGE* and *MGE_LPS-GV* systems, 20 iterations were conducted for model parameter updating.

4.2. Experimental results

4.2.1. Objective measurement

Fig. 1 shows the relative reduction of mel-cesptrum feature distortion and the LPS-GV distortion for three dimensions using the *MGE_LPS-GV* system with different value of $w^{(t)}$. When the LPS-GV weight increased, the LPS-GV distortion achieved more decreasing, whereas the feature distortion achieved less decreasing. Furthermore, the effectiveness of $w^{(t)}$ is inconsistent for different dimensions of the mel-cepstral distortion and different frequency points of the LPS-GV distortion, e.g., with the same $w^{(t)}$, the relative reduction of LPS-GV distortion for high frequency point is larger than that for low frequency point.

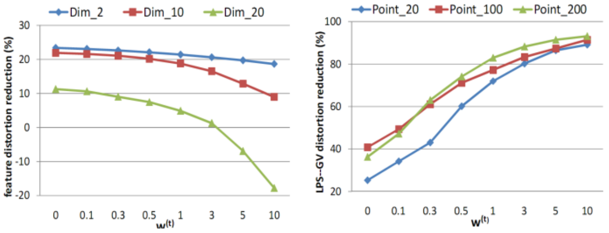

Figure1: Effect of MGE_LPS-GV training with different $w^{(t)}$: relative reduction of generated mel-cepstral distortion (left) and its LPS-GV distortion (right) on test data.

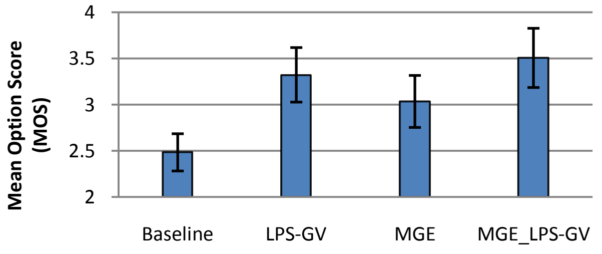

Figure2: Results of subjective evaluation on naturalness.

Table 1: Results of preference tests between the *GV* (a) and *LPS-GV* (b) systems in Test_1, and between the *MGE/LPS-GV* (a) and *MGE_LPS-GV* (b) systems in Test_2.

	a>b	a=b	a<b
Test_1	20.0%	16.4%	63.6%
Test_2	29.3%	24.3%	46.4%

4.2.2. Subjective evaluation

In some informal listening tests, we found that the quality of synthesized speech using the *MGE_LPS-GV* system was affected by the setting of the LPS-GV weight $w^{(t)}$ in (16). Finally, we set $w^{(t)}$ to 3 empirically.

First, 20 sentences were randomly selected from the test set and were synthesized by the *Baseline*, *LPS-GV*, *MGE*, and *MGE_LPS-GV* systems respectively. These synthesized speech were evaluated by 7 Chinese-native listeners, in terms of naturalness. The mean opinion scores (MOS) with 95% confidence interval of the four systems are shown in Fig. 2. From the results, we can conclude that *LPS-GV* system achieved significantly better performance than *Baseline*, which shows the effectiveness of LPS-GV modeling in feature generation algorithm when mel-cesptrum features and maximum likelihood training are adopted. Besides, the quality of synthesized speech after MGE training can also be improved by incorporating LPS-GV distortion into the criterion of MGE training.

Then, we conducted two preference listening tests. The first test was between the *GV* and *LPS-GV* systems, for comparing the performance of LPS-GV and conventional GV modeling when mel-cepstral coefficients are adopted as spectral features. 20 sentences in test set were used and the same 7 Chinese-native listeners participated in the listening test. The results are shown in Test_1 of Tab. 1. We see that LPS-GV using mel-cesptrum can achieve more natural synthetic speech than conventional GV. This is consistent with the experimental results in [8] where LSPs are used as spectral parameters. Another preference test was between the *MGE/LPS-GV* system and the *MGE_LPS-GV* system. The results are shown in Test_2 of Tab. 1. We find that the *MGE_LPS-GV* system performs slightly better than the *MGE/LPS-GV* system. It means integrating LPS-GV into model training criterion can substitute the function of LPS-GV model based parameter generation when MGE training is adopted. A further study on the confidence interval of the preference scores in Test_2 of Tab. 1 shows the difference between these two systems is not significant. The slight superiority of the *MGE_LPS-GV* system in our experiment may be due to the empirically setting of $w^{(t)}$.

5. Conclusions

In this paper, we introduce the LPS-GV of mel-cepstrum into HMM-based parametric speech synthesis by considering the temporal variance of generated LPSs. The LPS-GV of mel-cepstrum is firstly adopted at synthesis stage similar to the conventional GV and the LPS-GV of LSPs. Then in order to reduce the computation cost of parameter generation, LPS-GV of mel-cepstrum is incorporated into MGE criterion at model training stage, where an additional generation error component is introduced by considering LPS-GV distortion. The experimental results show that applying LPS-GV of mel-cepstrum at synthesized stage can improve the naturalness of synthesized speech significantly, compared to the systems without GV and using conventional GV model. Besides, integrating LPS-GV of mel-cepstrum into model training criterion can also improve the naturalness of synthesized speech using conventional MGE training. Finally, the approaches of utilizing LPS-GV at synthesis stage and at training stage are compared. It concludes that adopting LPS-GV in model training is also affective, which avoids extra computational cost at synthesis stage. Comparing LPS-GV on LSPs and mel-cepstrums will be our future work.

6. Acknowledgements

This work was partially supported by“the Fundamental Research Funds for the Central Universities” (Grant No. WK2100060005).

7. References

[1] T. Masuko, K. Tokuda, T. kobayashi, and S. Imai, “Speech synthesis from HMMs using dynamic features,” in *Proc. of ICASSP*,1996,pp.389-392.

[2] 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 ICASSP*, 1999, vol.5, pp.2347-2350.

[3] K. Tokuda, T. kobayashi, and S. Imai, “Speech parameter generation from HMM using dynamic features,” in *Proc.of ICASSP*, 1995, pp. 660-663.

[4] Z.-H Ling, L. Qin, H. Lu, Y. Gao, L.-R Dai, R.-H Wang, Y. Jiang, Z.-W Zhao, J.-H Yang, J. Chen, and G.-P Hu, “The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007,” in *Proc. of Blizzard Challenge workshop*, 2007.

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

[6] Y-J. Wu and R.H Wang, “Minimum generation error training for HMM-based speech synthesis,” in *Proc. of ICASSP*, 2006, vol.1,pp 889-892.

[7] Y.-J. Wu, H. Zen,Y. Nankaku, and K.Tokuda, “Minimum generation error criterion considering global/local variance for HMM-based speech synthesis,” in *ICASSP*, 2008, pp. 4621-4624.

[8] Z.-H. Ling and Y.Hu, and L.–R. Dai, “Global variance modeling on the log power spectrum of LSPs for HMM-based speech synthesis,” in *Interspeech*, 2010, pp. 825-828.

[9] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” in *Proc. of ICASSP*, 1992, pp. 137-140

[10] S. Amari, “A theory of adaptive pattern classifiers,”*IEEE Trans. Electron. Comput*., vol. EC-16, no. 3, pp. 299-307, 1967.

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

[12] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,”*J.Acoust.Soc.Japan (E)*, vol. 21, no. 2, pp. 19-41, 2000