MINIMUM GENERATION ERROR CRITERION CONSIDERING GLOBAL/LOCAL VARIANCE FOR HMM-BASED SPEECH SYNTHESIS

Yi-Jian Wu, Heiga Zen, Yoshihiko Nankaku, Keiichi Tokuda

Nagoya Institute of Technology, Japan

yjwu@sp.nitech.ac.jp, zen@sp.nitech.ac.jp, nankaku@sp.nitech.ac.jp, tokuda@nitech.ac.jp

ABSTRACT

Two techniques, including minimum generation error (MGE) criterion for HMM training, and the parameter generation algorithm considering global variance (GV), had been proposed to improve the quality of HMM-based speech synthesis. In this paper, we incorporate the GV technique into MGE criterion, where an additional generation error component considering global/local variance (GV/LV) is introduced for generation error definition, and the model parameters are optimized to minimize the new generation error function. From the experimental results, the quality of synthesized speech was improved after MGE-GV/LV training, which is similar to the effectiveness of considering GV in parameter generation, however, without introducing any extra computational cost in synthesis process.

Index Terms— Speech synthesis, HMM, minimum generation error, global variance, local variance

1. INTRODUCTION

HMM-based speech synthesis had been proposed for a decade [1]. In this method, the spectrum, pitch and duration are modeled simultaneously in a unified framework of HMMs [2], and the parameter sequence is generated by maximizing the likelihood of the HMMs related to the parameter sequence under the constraint between static and dynamic features [3]. Comparing to other synthesis methods, this method has several advantages as follows: 1) under its statistical training framework, it can learn salient statistical properties of speakers, speaking styles, emotions, and etc., from the speech corpus; 2) voice characteristic of synthesized speech can be easily controlled by modifying acoustic statistics of HMMs [4]; 3) it can generate smooth and stable speech under a small footprint. Due to these, HMM-based speech synthesis gradually became popular both in research and application.

Although current performance of HMM-based speech synthesis is quite good, the quality of synthesized speech still needs to be improved. In recent years, several techniques had been proposed to improve the quality of synthesized speech for HMM-based speech synthesis. In [5], two issues related to maximum likelihood (ML) based HMM training, including the inconsistency between training and application of HMM, and the ignorance of constraint between static and dynamic features, were pointed out, and a minimum generation error (MGE) criterion was proposed to resolve these two issues. Furthermore, it had been applied to the tree-based clustering for context dependent HMMs [6] and the whole HMM training procedure [7]. In [8], a new parameter generation algorithm considering global variance (GV) was proposed to alleviate the over-smoothing problem of generated speech features, where the speech features are generated to maximize not only the conventional likelihood for acoustic feature but also the likelihood for the GV of generated feature trajectory.

The effectiveness of the parameter generation algorithm considering GV indicates that the quality of synthesized speech can be improved when the GVs of generated trajectory become closer to that of natural one. From this point, the HMMs should be trained to reduce the distortion between generated GV and original GV. In this paper, we incorporate this GV concept into MGE criterion, and introduce a more general variance term, called local variance (LV). In MGE-GV/LV training, the generation error is re-defined by introducing an additional generation error component, which measures the distortion between the generated GV/LV and original GV/LV, and the parameters of HMMs are optimized so as to minimize the new generation error function.

The rest of paper is organized as follows. In section 2, we briefly review the minimum generation error (MGE) criterion for HMM training. In section 3, we present the details of incorporating GV/LV component into MGE criterion. In section 4, we describe the experiments to evaluate the effectiveness of MGE-GV/LV training, and present the results. Finally, our conclusion and future work are given in section 5.

2. MINIMUM GENERATION ERROR CRITERION

2.1. Parameter generation algorithm

For a given HMM λ and the state sequence $\boldsymbol{q}$, the parameter generation algorithm is to determine the speech parameter vector sequence $\boldsymbol{o} = [\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \ldots, \boldsymbol{o}_T^\top]^\top$ to maximize $P(\boldsymbol{o}|\boldsymbol{q}, \lambda)$. In order to keep the smooth property of the generated parameter sequence, the dynamic features including delta and delta-delta coefficients $\Delta^{(n)}\boldsymbol{c}_t (n = 1, 2)$ are used, and the parameter vector can be rewritten as

$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta^{(1)}\boldsymbol{c}_t^\top, \Delta^{(2)}\boldsymbol{c}_t^\top\right]^\top. \tag{1}$$

The constraints between static and dynamic feature vector sequence can be formulated as

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}, \tag{2}$$

where $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \ldots, \boldsymbol{c}_T^\top\right]^\top$. Due to limited space, here the details of $\boldsymbol{W}$ is not given, which can be found in [3, 5].

Under this constraint, determining $\boldsymbol{o}$ to maximize $P(\boldsymbol{o}|\lambda, \boldsymbol{q})$ is equivalent to determining $\boldsymbol{c}$ to maximize $P(\boldsymbol{o}|\lambda, \boldsymbol{q})$. By setting $\partial P(\boldsymbol{o}|\lambda, \boldsymbol{q})/\partial \boldsymbol{c} = 0$, we obtain

$$\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{\mu}_{\boldsymbol{q}} = \boldsymbol{R}_{\boldsymbol{q}}^{-1} \boldsymbol{r}_{\boldsymbol{q}}, \tag{3}$$

where

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{W}, \tag{4}$$

$$\boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{\mu}_{\boldsymbol{q}}, \tag{5}$$

and $\boldsymbol{\mu}_{\boldsymbol{q}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{q}}$ are the mean vector and covariance matrix, respectively.

2.2. Measure of generation error

With the generated parameter vector $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$, we need to measure the distortion between the original and generated feature vector, i.e. **feature distortion**. Here, the Euclidean distance was adopted to calculate the distortion

$$D_{\boldsymbol{c}}(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\boldsymbol{q}}) = \| \boldsymbol{c} - \bar{\boldsymbol{c}}_{\boldsymbol{q}} \|^2 . \tag{6}$$

The posterior probability $P(\boldsymbol{q}|\lambda, \boldsymbol{o})$ can be used to weight the distance for all possible state sequence $\boldsymbol{q}$, and the generation error for $\boldsymbol{c}$ is defined as

$$e(\boldsymbol{c}, \lambda) = \sum_{all\ \boldsymbol{q}} P(\boldsymbol{q}|\lambda, \boldsymbol{o}) D_{\boldsymbol{c}}(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\boldsymbol{q}}). \tag{7}$$

As direct calculation of generation error using the above definition is computationally expensive, the representative N-best path can be used to approximate this generation error. In extreme case, we can use the 1-best path, i.e. the optimal state sequence. The generation error is simplified as

$$e(\boldsymbol{c}, \lambda) = D_{\boldsymbol{c}}(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\hat{\boldsymbol{q}}}), \tag{8}$$

where $\hat{\boldsymbol{q}}$ is the optimal state sequence for $\boldsymbol{o}$. In this paper, we use this simplified generation error function in MGE training. In fact, it refers to a Viterbi-type training. In the following part of this paper, we use $\boldsymbol{q}$ to denote $\hat{\boldsymbol{q}}$ by default.

2.3. MGE criterion

With the measure of generation error, we incorporate parameter generation process into HMM training for calculating the total generation errors for all training data, which is

$$E(\lambda) = \sum_{n=1}^{N} e(\boldsymbol{c}_n, \lambda), \tag{9}$$

where N is the total number of training utterances.

Finally, we define the object of MGE criterion, which is to optimize the parameters of HMMs so as to minimize the total generation errors

$$\hat{\lambda} = \arg\min E(\lambda). \tag{10}$$

As direct solution for Eq. (10) is mathematically intractable, probabilistic descent (PD) [9] method was adopted for parameter optimization. The details of updating rules for mean and variance parameters in MGE training can be found in [5].

3. MGE WITH GLOBAL/LOCAL VARIANCE

In this section, we incorporate the GV/LV concept into MGE criterion, where an additional generation error component measuring the distortion between the generated GV/LV and original GV/LV is introduced in generation error definition, and the parameters of HMMs are optimized so as to minimize the new generation error function.

3.1. Global/local variance

In [8], a GV of the static feature trajectory in an utterance is calculated by

$$\boldsymbol{v}_g(\boldsymbol{c}) = [v_g(1), v_g(2), \ldots, v_g(d), \ldots, v_g(D)]^\top, \tag{11}$$

$$v_g(d) = \frac{1}{T} \sum_{t=1}^{T} \left(c_t(d) - m_g(d)\right)^2, \tag{12}$$

$$m_g(d) = \frac{1}{T} \sum_{t=1}^{T} c_t(d), \tag{13}$$

where D is the dimension of static feature vector. The GV is calculated utterance by utterance.

Furthermore, we define a more general variance term, called local variance (LV), which is calculated by

$$\boldsymbol{v}(\boldsymbol{c}) = [v_1(1), \ldots, v_1(D), \ldots, v_T(1), \ldots, v_T(D)]^\top, \tag{14}$$

$$\boldsymbol{m}(\boldsymbol{c}) = [m_1(1), \ldots, m_1(D), \ldots, m_T(1), \ldots, m_T(D)]^\top, \tag{15}$$

$$v_t(d) = \frac{1}{L} \sum_{i=t-L_-}^{t+L_+} \left(c_i(d) - m_t(d)\right)^2, \tag{16}$$

$$m_t(d) = \frac{1}{L} \sum_{i=t-L_-}^{t+L_+} c_i(d), \tag{17}$$

where $L = (L_- + L_+ + 1)$ is the size of the window for variance calculation. It should be noted that the LV is calculated frame by frame. Actually, GV can be regarded as a special case of LV, where the same window covering whole utterance is used for variance calculation for each frame. Therefore, we use LV as example to define the new generation error function, and formulate the updating rules for model parameters.

3.2. Generation error considering GV/LV

In order to normalize the scale of the generation error components for static feature and GV/LV of feature trajectory, we denote

$$\boldsymbol{\sigma}(\boldsymbol{c}) = [\sigma_1(1), \ldots, \sigma_1(D), \ldots, \sigma_T(1), \ldots, \sigma_T(D)], \tag{18}$$

$$\sigma_t(d) = \sqrt{v_t(d)} \qquad (t = 1, ..., T;\ \ d = 1, ..., D), \tag{19}$$

and use $\boldsymbol{\sigma}(\boldsymbol{c})$ instead of $\boldsymbol{v}(\boldsymbol{c})$ to calculate the generation error.

Similar to Eq. (6), the Euclidean distance is adopted to calculate the distortion between the GV/LV of generated trajectory and that of original one, i.e. **GV/LV distortion**

$$D_{\boldsymbol{v}}(\boldsymbol{\sigma}(\boldsymbol{c}), \boldsymbol{\sigma}(\bar{\boldsymbol{c}}_{\boldsymbol{q}})) = \| \boldsymbol{\sigma}(\boldsymbol{c}) - \boldsymbol{\sigma}(\bar{\boldsymbol{c}}_{\boldsymbol{q}}) \|^2, \tag{20}$$

Finally, we combine this GV/LV distortion with the original feature distortion, and define the new generation error for $\boldsymbol{c}$ as

$$e'(\boldsymbol{c}, \sigma(\boldsymbol{c}), \lambda) = D_{\boldsymbol{c}}(\boldsymbol{c}, \bar{\boldsymbol{c}}_q) + w D_{\boldsymbol{v}}(\boldsymbol{\sigma}(\boldsymbol{c}), \boldsymbol{\sigma}(\bar{\boldsymbol{c}}_{\boldsymbol{q}})), \tag{21}$$

where w denotes the GV/LV weight for controlling a balance between these two distortions. As the scales of these two distortions have been normalized, one reasonable value of w could be 1. Further investigation of the effect of w is shown in section 4.

3.3. Parameter updating

With the new generation error function, the MGE training is to minimize the total generation errors

$$\hat{\lambda} = \arg\min E'(\lambda) = \arg\min \sum_{n=1}^{N} e'(\boldsymbol{c}_n, \lambda), \tag{22}$$

with respect to

$$\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top, \ldots, \boldsymbol{\mu}_K^\top\right]^\top, \tag{23}$$

$$\boldsymbol{U} = \left[\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1}, \ldots, \boldsymbol{\Sigma}_K^{-1}\right]^\top, \tag{24}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{U}$ are defined by concatenating the mean vectors and covariance matrices of all unique Gaussian components in the model set λ, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the k-th unique Gaussian component, and K is the total number of Gaussian components in the model set, respectively.

The PD method is adopted here for parameter optimization. For each training utterance $\boldsymbol{c}_\tau$, the parameter set is updated as

$$\lambda_{\tau+1} = \lambda_\tau - \epsilon_\tau \boldsymbol{H}_\tau \frac{\partial e'(\boldsymbol{c}_\tau, \lambda)}{\partial \lambda}\bigg|_{\lambda=\lambda_\tau}, \tag{25}$$

where $\boldsymbol{H}_\tau$ is a positive definite matrix, and ϵ_τ is a learning rate that decrease when utterance index τ increase.

For the mean parameters, the gradient of generation error function is calculated as

$$\frac{\partial e(\boldsymbol{c}_\tau, \lambda)}{\partial \boldsymbol{\mu}} = 2\boldsymbol{S}_{\boldsymbol{q}}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{W} \boldsymbol{R}_{\boldsymbol{q}}^{-1} \boldsymbol{\zeta}, \tag{26}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} = \mathrm{diag}(\boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{U}), \tag{27}$$

$$\boldsymbol{\zeta} = (\bar{\boldsymbol{c}}_{\boldsymbol{q}} - \boldsymbol{c}_\tau) + wA(\bar{\boldsymbol{c}}_{\boldsymbol{q}} - \boldsymbol{m}(\bar{\boldsymbol{c}}_{\boldsymbol{q}})), \tag{28}$$

and A is a diagonal matrix, whose diagonal elements are

$$A_{j,j} = 1 - \frac{1}{L_+} \sum_{i=t-L_-}^{t+L_+} \frac{\sigma_i(d)}{\bar{\sigma}_i(d)}, \quad j = (t-1) * D + d. \tag{29}$$

where $\sigma_i(d)$ and $\bar{\sigma}_i(d)$ are the GV/LV of original and generated feature trajectory, respectively. In the above equations, $\boldsymbol{S}_{\boldsymbol{q}}$ is a $3DT \times 3DK$ matrix whose elements are 0 or 1 determined according to the optimal state sequence $\boldsymbol{q}$ for $\boldsymbol{c}_\tau$. The operation of diag(.) is to convert a $3DT \times 3D$ matrix to a $3DT \times 3DT$ block-diagonal matrix with a block size of $3D$.

For the variance parameters, the gradient of generation error function is calculated as

$$\frac{\partial e(\boldsymbol{c}_\tau, \lambda)}{\partial \boldsymbol{U}} = 2\boldsymbol{S}_{\boldsymbol{q}}^\top \mathrm{diag}^{-1}\big(\boldsymbol{W} \boldsymbol{R}_{\boldsymbol{q}}^{-1} \boldsymbol{\zeta} (\boldsymbol{\mu}_{\boldsymbol{q}} - \boldsymbol{W} \bar{\boldsymbol{c}}_{\boldsymbol{q}})\big), \tag{30}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \boldsymbol{S}_{\boldsymbol{q}} \boldsymbol{m}, \tag{31}$$

and $\mathrm{diag}^{-1}(.)$ is the inverse operation of diag(.).

From Eq. (26) and (30), the computational complexity of MGE-GV/LV training is similar to that of MGE training, since the most computational cost in parameter updating is still related to the calculation of $\boldsymbol{R}_{\boldsymbol{q}}^{-1}$.

4. EXPERIMENTS

4.1. Experimental conditions

We used the phonetically balanced 503 sentences from ATR Japanese speech database (B-set, MHT) in this experiments. The first 450 sentences were used as training data, and the remaining 53 sentences were used for evaluation. Speech signal were sampled at a rate of 16kHz. The acoustic features, including F0 and mel-cepstral coefficients, were extracted with a 5ms shift. Feature vector consists of static features, including 25-th mel-cepstral coefficients and logarithm of F0, and their delta and delta-delta coefficients. A 5-state left-to-right no-skip HMM was used, and MSD-HMM [10] was adopted for F0 modeling. In synthesis, the Mel Log Spectrum Approximation (MLSA) filter [11] was used to synthesize the speech waveform.

The HMM training in this experiment was performed as follows. Firstly, the conventional ML-based HMM training procedure was conducted. Then the optimal state alignment for all training data were obtained using the ML-trained HMMs. With the state alignments, the MGE-GV/LV training was performed to re-estimate the parameters of clustered HMMs, where both spectral and F0 part of model parameters were updated. The window size for LV calculation was set to 50 in MGE-LV training .

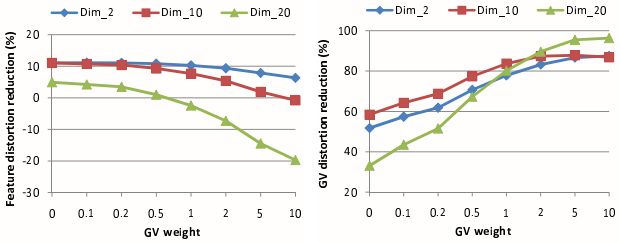

Fig. 1. Effect of MGE-GV training with different GV weights: relative reduction of generated mel-cepstra distortion (left) and its GV distortion (right) on test data

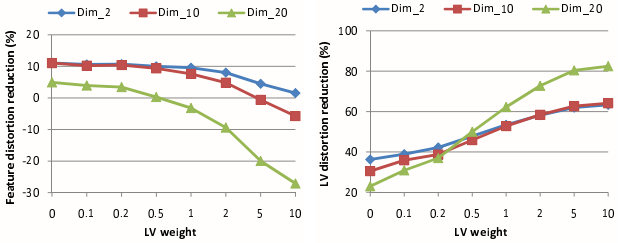

Fig. 2. Effect of MGE-LV training with different LV weights: relative reduction of generated mel-cepstra distortion (left) and its LV distortion (right) on test data

4.2. Experimental results

4.2.1. Objective measure

Fig. 1 shows the relative reduction of generated mel-cepstra distortion and its GV distortion for several typical dimensions after MGE-GV training. It can be seen that when GV weight is set to 0, i.e. MGE training only, both the feature distortion and GV distortion was improved. When the GV weight increased, the GV distortion was improved, whereas the feature distortion increased, which is even worse than the baseline when the GV weight is too large. In addition, the effectiveness of GV weight is different for each dimension of mel-cepstral coefficient. With the same GV weight in MGE-GV training, the relative reduction of GV distortion for high dimension of mel-cepstra coefficients is bigger than that for low dimension of mel-cepstral coefficients, e.g. 96% reduction of GV distortion for 20-th mel-cepstral coefficient when the GV weight set to 10.

Similar effect of MGE-LV training can be found in Fig. 2, where the relative reduction of generated mel-cepstra distortion and its LV distortion after MGE-LV training are shown. Comparing to Fig. 1, the only difference is that the relative reduction of LV distortion is smaller than that of GV distortion, which is reasonable since LV distortion is calculated frame-by-frame, whereas GV distortion is calculated utterance-by-utterance.

Fig. 3 shows the relative reduction of generated F0 distortion and its GV/LV distortion after MGE-GV/LV training. It can be seen that the GV distortion reduced only when the GV weight was smaller than 0.5, and the LV distortion reduced when the LV weight was less than 2. In fact, the GV/LV distortion was always improved on close test (not shown here), which means the model parameters over-fit

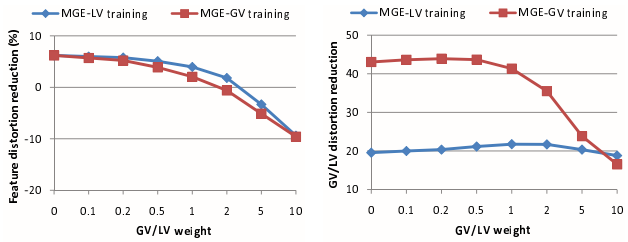

Fig. 3. Effect of MGE-GV/LV training on F0: relative reduction of F0 distortion (left) and its GV/LV distortion (right) on test data

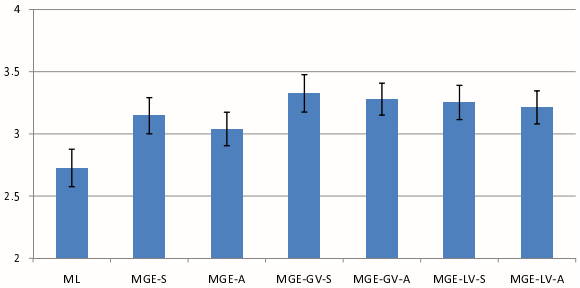

Fig. 4. Results of MOS

to the training data. There are two possible reasons to explain that MGE-GV/LV training for F0 is not as effective as for mel-cepstra. One reason is that the errors in automatic F0 extraction may affect the GV/LVs in training. Another one is that the likelihood of mel-cepstra dominates the Viterbi alignment process, which makes that the state alignments used in MGE-GV/LV training and evaluation are biased to mel-cepstra.

4.2.2. Subjective listening

From the informal listening, we found that the synthesized speech became clearer after applying MGE-GV/LV training. When the GV/LV weight increased, the clearness of synthesized speech was enhanced. However, it simultaneously introduced some artificial effect in speech sound. The GV/LV weight should be properly set for balancing the clearness and naturalness. Here, we set the GV/LV weight to 2.

Finally,we conducted a formal subjective listening test to evaluate the effectiveness of MGE-GV/LV training. Seven kinds of synthesized voice were evaluated, which includes the baseline (ML), MGE training on spectrum (MGE-S), MGE training on spectrum and F0 (MGE-A), MGE-GV/LV training on spectrum (MGE-GV/LV-S), MGE-GV/LV training on spectrum and F0 (MGE-GV/LV-A). Ten Japanese listener participated in the test. Each listener evaluated 15 set of samples consisting of seven synthesized speech samples, and give the MOS on the naturalness. The speech samples were randomly selected for each listener from the 53 test sentences.

Fig. 4 shows the results of listening test. It is obvious that the MGE-GV/LV training worked very well for spectral parameters. The quality of synthesized speech was improved after MGE training for spectrum, and it was improved further after incorporating GV/LV into MGE training. For F0 parameter, the MGE-GV/LV training did not cause significant improvement. Actually, the synthesized quality even had degraded a little after applying MGE training on F0 parameter. By analyzing the synthesized speech, we found that the dynamic range of generated F0 trajectory was enlarged. However, it also introduced some unnatural fluctuation into the F0 contour.

From the results, the difference between MGE-GV and MGE-LV training is insignificant, and MGE-LV training is slightly worse than MGE-GV training, which is not as we expected. Since current experiment only evaluated the effect of MGE-LV training with the window size of 50, we need to conduct more experiments to optimize the window size for MGE-LV training.

Comparing to the results in [8], the effectiveness of MGE-GV/LV training is quite similar to that of considering GV in parameter generation process. Furthermore, one advantage of MGE-GV/LV training is that it does not introduce any extra computational cost in synthesis process.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we incorporate the GV technique into MGE criterion, where an additional generation error component considering global/local variance (GV/LV) is introduced for generation error definition. The experimental results show that MGE-GV/LV training worked well on spectral parameter, but was not effective for F0 parameter. From the subjective listening test, the quality of synthesized speech was improved after MGE-GV/LV training, which is similar to the effectiveness of considering GV in parameter generation process. However, it would not introduce any extra computational cost in synthesis process.

Future work is to investigate the effect of window size for MGE-LV training, and conduct the listening test to compare the MGE-GV/LV training with the GV-based parameter generation technique.

6. ACKNOWLEDGEMENTS

The authors are grateful to Dr. Tomoki Toda of Nara Institute of Technology for helpful discussions.

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] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using mllr," in *Proc. of ICASSP*, 2001, pp. 805–808.

[5] 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.

[6] Y.-J. Wu, W. Guo, and R.H. Wang, "Minimum generation error criterion for tree-based clustering of context dependent HMMs," in *Proc. of Interspeech*, 2006, pp. 2046–2049.

[7] Y.-J. Wu, R.H. Wang, and F. Soong, "Full HMM training for minimizing generation error in synthesis," in *Proc. of ICASSP*, 2007, pp. 517–520.

[8] 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.

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

[10] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. of ICASSP*, 1999, pp. 229–232.

[11] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. of ICASSP*, 1983, pp. 93–96.