Analysis of Spectral Enhancement Using Global Variance in HMM-Based Speech Synthesis

Takashi Nose and Akinori Ito
Graduate School of Engineering, Tohoku University
tnose@m.tohoku.ac.jp, aito@spcom.ecei.tohoku.ac.jp

Abstract
This paper analyzes the problem of the spectral enhancement technique using global variance (GV) in HMM-based speech synthesis. In the conventional GV-based parameter generation, spectral enhancement with variance compensation is achieved by considering a GV pdf with fixed parameters for every output utterances through the generation process. Although the spectral peaks of the generated trajectory are clearly emphasized and subjective clarity is improved, the use of the fixed GV parameters results in a much smaller variation of GVs among the synthesized utterances than that of the natural speech, which sometimes causes undesirable effect. In this paper, we examine the above problem in terms of multiple objective measures such as variance characteristics, spectral and GV distortions, and GV correlations and discuss the result. We propose a simple alternative technique based on an affine transformation that provides a closer GV distribution to the original speech and improves the correlation of GVs of generated parameter sequences. The experimental results show that the proposed spectral enhancement outperforms the conventional GV-based parameter generation in the objective measures.

Index Terms: HMM-based speech synthesis, global variance, parameter generation, over-smoothing, variance compensation

1. Introduction
To realize a flexible text-to-speech (TTS) system that can express and control various speaker characteristics, emotions, and speaking styles, statistical parametric speech synthesis based on hidden Markov models (HMMs) has become one of the promising approaches [1, 2]. In a general HMM-based speech synthesis system, speech parameter sequences, i.e., spectral and excitation parameters such as mel-cepstrum and F0, are modeled by phonetically and prosodically context-dependent HMMs, and model parameters for an input text are predicted by decision trees constructed in the model training. Although this approach can generate smooth and stable speech parameters, the generated trajectories are generally over-smoothed due to the context clustering, and the quality and intelligibility of the synthetic speech degrade.

One effective approach to alleviate the over-smoothing problem is spectral enhancement based on variance compensation of the generated parameter sequence such as postfiltering [3], global variance (GV) or norm constraint in parameter generation [4, 5], and histogram equalization [6]. In these techniques, the GV-based parameter generation has been widely studied, and there are several variations of the implementation manner of the GV-based variance compensation [7–11]. In [7] and [8], the GV constraint was introduced in the model training as the frameworks of minimum generation error (MGE) training [12] and trajectory model [13], respectively. In [9] and [10], focused on the GV compensation for LSP parameters because the conventional GV-based parameter generation did not work well for LSPs. As a more complicated approach, Zen et al. modeled the GV as a framework of product of experts [11]. There is also a different approach based on local variance compensation to capture the time-variant variance characteristics [7, 14]. One of the limitations of these techniques is increase of the computational cost in the training or the synthesis process compared to the original GV-based parameter generation, which would be an obstacle in some applications.

Although the conventional GV-based parameter generation [4] has an advantage in terms of the total computational cost compared to the more complicated ones described above, there is a problem that the variance of GVs of generated parameter sequences becomes much smaller than that of the original speech when the GVs are modeled by a single Gaussian pdf. This is because the target GV value given by the mean of the GV pdf is fixed and all generated parameter sequences has GVs close to the target through the synthesis process. In this study, we analyze and discuss this problem from the aspects of multiple objective measures. In addition, we propose a simple but objectively better alternative technique based on affine transformation of the speech parameters as a form of post-processing. The estimation of the transformation coefficient is based on minimum GV error criterion for the training data, which is similar to the MGE criterion including GV distortion in [7].

2. Conventional spectral enhancement using GV compensation
2.1. Parameter generation considering GV
The \( d \)-th dimension of the global variance (GV) vector \( v \) defined for the parameter sequence of an utterance is given by

\[
v(d) = \frac{1}{T} \sum_{t=1}^{T} \left( c_t(d) - \overline{c(d)} \right)^2
\]

(1)

\[
\overline{c(d)} = \frac{1}{T} \sum_{t=1}^{T} c_t(d)
\]

(2)

where \( c_t(d) \) is the \( d \)-th dimension of speech parameter vector \( c_t \) of \( c \) at the \( t \)-th frame. In [4], the pdf of GVs is modeled by a single Gaussian pdf. In the parameter generation, when a state sequence \( q \) is given, the optimum parameter sequence \( \hat{c} \) is determined so as to maximize an objective function consisting of the HMM and GV log pdfs as follows:

\[
\mathcal{L}_q = \omega \log P(\alpha q, \lambda) + \log P(\nu | \lambda_{GV})
\]

(3)

where \( \alpha = [\alpha_1^T, \ldots, \alpha_t^T, \ldots, \alpha_T^T]^T \), and \( \alpha_t \) is an observation vector including static and dynamic features calculated from \( c \) at \( t \)-th frame [15]. \( \lambda \) and \( \lambda_{GV} \) are model parameter sets of the
HMMs and the GV pdf, respectively. The constant \( \omega \) denotes the GV weight for controlling the balance between the two pdfs. \( \omega \) is usually set to the ratio of the dimensions of vectors \( v \) and \( a \), i.e., \( \omega = 1 / (3T) \) [4]. In contrast with a standard parameter generation algorithm [15], there is no closed form solution for maximizing Eq. (3), and a gradient method is used.

### 2.2. Fast GV compensation method

From the perspective of the computational cost, the GV-based parameter generation described in Sect. 2.1 is more expensive than the standard parameter generation without the GV model. Alternatively, we can use a simpler method where the parameters generated by the standard generation algorithm is converted so as to maximize the GV likelihood, which is used as the initialization for the gradient method in the GV-based parameter generation [4]. The conversion formula is utterance-adaptive and is given by an affine transform as

\[
\mathbf{c}'(d) = \sqrt{\frac{\mu_v(d)}{\sigma_v(d)}} \left( \mathbf{c}(d) - \mathbf{c_\text{G}}} \right) + \mathbf{c_\text{G}} \\
\]

where \( \mu_v(d) \) is the \( d \)-th dimension of the mean vector of \( \lambda_v \). Recently, another approach was proposed using the Lagrange multiplier where the constraint of the target GV is explicitly taken into account in the parameter generation and voice conversion processes [16, 17].

### 2.3. Problem of the conventional approach

By using the GV compensation in the spectral parameter generation, the spectral peaks are emphasized and the subjective quality of the synthetic speech is improved compared to the synthetic speech without the GV compensation. However, the simple modeling of GVs using a single Gaussian pdf causes an unnatural property of GVs of the generated parameter sequences. Specifically, the variance of GVs of synthetic speech samples becomes much smaller than that of the original speech samples because the same GV pdf with a fixed GV mean is used for all utterances to be synthesized. Consequently, the GVs of generated parameter sequences are very close to the mean parameter of the GV pdf, which is much different from the property of the natural speech.

Figure 1 shows an example of this problem. In the figure, the pairs of GVs of the fifth mel-cepstral coefficient of the original and generated parameter sequences are plotted. In the generated parameters, GVPG and NOGV represent the parameter generation algorithms with and without the GV model, respectively. The mean value of the GV pdf estimated using the training data is also shown as MEAN. From the result, we can see that the GVs of the parameter sequences generated by GVPG are distributed almost around the GV mean. In contrast, the GVs of NOGV has a certain correlation to those of the original parameter sequences though there is a large bias. Since the GV constraint in the parameter generation is very strong, the correlation of the NOGV substantially decreases after applying the GV-based variance compensation.

### 3. GV compensation using global affine transformation

From the example and discussion in Sect. 2.3, we came up with an idea that we can consider an affine transformation that compensates the GVs of the parameter sequences while keeping the correlation to the GVs of the original parameter sequences. Specifically, when we assume that the mean value \( \mathbf{c}(d) \) does not change through the conversion process as in the case of Sect. 2.2, the affine transformation is given by the following simple form as

\[
\mathbf{c}'(d) = a_d \left( \mathbf{c}(d) - \mathbf{c_\text{G}} \right) + \mathbf{c_\text{G}} \\
= a_d \mathbf{c}(d) + (1 - a_d) \mathbf{c_\text{G}} \\
\]

where \( a_d \) is a global transformation parameter for the \( d \)-th dimension. Eq. (6) can be viewed as the interpolation between \( \mathbf{c}(d) \) and \( \mathbf{c_\text{G}} \). To increase the variance, \( a_d \) should be set to \( a_d > 1 \), and in that case the conversion becomes extrapolation. Different from the conversion method in Sect. 2.2, the transformation parameter is fixed for all utterances to be synthesized, which leads to keep the GV correlation in the conversion. We optimize the transformation parameter \( a_d \) for the training data of \( N \) utterances using a minimum GV error criterion as follow:

\[
a_{d_d}^\text{opt} = \arg \min_{a_d} \sum_{n=1}^{N} \left( v_g^{(n)}(d) - v_o^{(n)}(d) \right)^2 \\
= \arg \min_{a_d} \sum_{n=1}^{N} \left( a_d v_g^{(n)}(d) - v_o^{(n)}(d) \right)^2 \\
\]

where \( v_g^{(n)}(d) \) and \( v_o^{(n)}(d) \) are the GVs of the generated parameter sequence of the \( n \)-th utterance in the \( d \)-th dimension and before and after the affine transformation, and \( v_o^{(n)}(d) \) is the GV of the original parameter sequence of the \( n \)-th utterance in the \( d \)-th dimension. Similar approaches using a global parameter set trained from the training data were introduced in [16, 18] to avoid the utterance-by-utterance parameter optimization.

### 4. Analysis of spectral enhancement using GV compensation

#### 4.1. Experimental conditions

We used speech data of a male professional narrator MHT taken from the ATR Japanese speech database set B [19]. The speaker

![Figure 1: Relation of GVs of the 5th mel-cepstrum coefficients between the original and generated parameters with (GVPG) and without (NOGV) GV-based parameter generation. Mean value of the GV pdf is also shown as MEAN.](image-url)
The conventional parameter generation algorithm considering GV [4]
GVAT The fast GV compensation method using utterance-adaptive affine transformation described in Sect. 2.2
GVGT The proposed GV compensation using global affine transformation optimized for the training data with a minimum GV error criterion

4.2. Mean and variance characteristics of GV

First, we confirmed the over-smoothing effect on mel-cepstrum by examining the mean values of GV for respective mel-cepstral coefficients of test samples. Figure 2 shows the result for NOGV, GVPG, GVGT. From the figure, there were clear decreases of GV means especially in higher dimensions when the GV compensation was not applied. The reduction of GV means are well recovered by using both the conventional and proposed GV compensation techniques. Then, we examined the variances of GV for respective mel-cepstral coefficients of test samples. Figure 3 shows the result. As in the case with the GV means, there were decreases of GV variances when the standard parameter generation was used. However, different from the case of the GV means, GV variances were much smaller than those of the original speech when using the GV-based parameter generation. This is because the target GV mean was fixed in the parameter generation of all the test samples. In contrast to GVPG, the proposed conversion effectively recovered not only the GV means but also the GV variances, and the variance characteristics became closer to those of the original speech.

4.3. Spectral and GV distortions

Next, we examined how close the GV of the generated parameter sequences were to those of the original parameter sequences. Table 1 shows the average Euclidean distances of GV vectors between original and generated parameter sequences for mel-cepstrum for NOGV, GVPG, GVAT, and GVGT on the test set. From the result, we see that the GV distortions were decreased by using the GV compensation techniques. When comparing the three techniques, we confirmed that the proposed technique is more effective than the conventional ones in the GV distortion reduction. We also examined the spectral distortion between the original and generated parameters on the test set. As the objective measure, we used the mel-cepstral distance which is often used for the objective evaluation in parametric speech synthesis. Table 2 shows the result. As is known in the previous study [14], we confirmed that all the GV compensation techniques increased the mel-cepstral distance. However, it is also seen that the proposed technique suppressed the increase of the distortion compared to the conventional techniques.
4.4. Correlation of GVs

Finally, we compared the conventional and proposed techniques in terms of the correlation of the GVs between the original and generated parameter sequences. Figure 4 shows the average correlation coefficients of GVs between original and generated parameter sequences in each dimension of mel-cepstrum. It is noted that the correlation coefficients of NOGV and GVGT were identical because the correlation property was kept through the affine transformation. From the figure, we see that fixing the target GV parameter in the conventional GV-based parameter generation significantly degraded the correlation in each dimension, and the correlation coefficients were around zero in higher dimensions. In contrast, the GVs of the original parameter sequences were well predicted using the standard parameter generation from context-dependent HSMMs, and the performance is kept by using the proposed technique. To intuitively confirm the effectiveness of the proposed technique, we also show an example of the distribution of GVs as a scatter plot. Figure 5 shows the relation of GVs of the first mel-cepstrum coefficients between original and generated parameters using different GV compensation techniques. From the figure, it is seen that the values of GVGT are diagonally distributed and the proposed technique recovers both the bias and the tilt caused by the standard parameter generation algorithm.

5. Discussions

In the previous section, we analyzed the performance of the conventional GV-based parameter generation by comparing with the proposed GV compensation technique based on global affine transformation using objective measures. The experimental results have shown that the proposed technique outperformed the conventional technique in spite of its simple implementation. However, when the author carefully checked the subjective quality of the synthetic speech, the quality of the samples with both techniques was quite similar, and there was no prominent improvement of naturalness and similarity. The similar result was reported in the previous study [16] where the listeners could not perceive a significant improvement of naturalness in the MOS test among different implementations of the GV compensation. From these results, the relation between the objective measures and subjective quality seems to be not so simple.

However, our goal in this study is not to outperform the conventional GV compensation techniques including more sophisticated or complicated implementations but to provide an alternative that is simple and computationally not expensive. In this perspective, the proposed technique has an advantage that the technique can be easily applied to any other types of parameter speech synthesis, e.g., techniques based on deep neural networks [23] and Gaussian process regression [24], with a small increase of computational cost, which is sometimes difficult in several techniques with complicated implementations of the GV compensation. In addition, it should be noted that there is no degradation from the GV-based parameter generation in the objective measures.

6. Conclusions

In this paper, we analyzed the conventional parameter generation algorithm considering GV and showed the problem on the basis of the objective measures. We also proposed a simple but effective GV compensation technique based on a global affine transformation for each dimension of the spectral parameter vectors. From the experimental results, we have shown that the proposed technique outperformed the GV-based parameter generation in multiple objective measures such as variance characteristics, spectral and GV distortions, and GV correlations. These results indicate that the proposed technique can be used as an alternative of the conventional technique, which saves the computational cost in the synthesis process. We will evaluate the subjective quality with a formal listening test. In addition, the GV compensation in the model space including both spectral and prosodic features will be discussed in the future work, which is more desirable to realize a TTS system with lower computational costs.

7. Acknowledgments

The authors wish to thank Prof. Takao Kobayashi of Tokyo Institute of Technology for the use of label data with phone segmentation and accent information. Part of this work was supported by JSPS Grant-in-Aid for Scientific Research 23700195.
8. References


