FAST, LOW-ARTIFACT SPEECH SYNTHESIS CONSIDERING GLOBAL VARIANCE

Matt Shannon, William Byrne

Cambridge University Engineering Department, U.K.

ABSTRACT

Speech parameter generation considering global variance (GV generation) is widely acknowledged to dramatically improve the quality of synthetic speech generated by HMM-based systems. However it is slower and has higher latency than the standard speech parameter generation algorithm. In addition it is known to produce artifacts, though existing approaches to prevent artifacts are effective.

We present a simple new theoretical analysis of speech parameter generation considering global variance based on Lagrange multipliers. This analysis sheds light on one source of artifacts and suggests a way to reduce their occurrence. It also suggests an approximation to exact GV generation that allows fast, low latency synthesis. In a subjective evaluation our fast approximation shows no degradation in naturalness compared to conventional GV generation.

***Index Terms*—** Speech synthesis, speech parameter generation considering global variance, artifact, low latency.

1. INTRODUCTION

Speech parameter generation considering global variance (GV generation) [1] dramatically improves the quality of speech generated by statistical speech synthesis systems and is now a standard part of HMM-based systems which aim for high-quality synthesis.

However the conventional algorithm for GV generation uses trajectory-level gradient descent [1], making it slow and introducing latency. In contrast the standard speech parameter generation algorithm (case 1 in [2]) is fast and has an approximate time-recursive variant [3] which has low latency and is still reasonably fast. In practice systems which aim for low latency synthesis often use post-filtering [4, 5, 6] rather than GV generation for this reason.

GV generation is also known to sometimes introduce *artifacts* into the synthesized speech [7, 8, 9, 10, 11]. Here by an artifact we mean a short distortion in the audio, such as a click, pop or short high-pitched whine. Existing implementations of GV generation, such as that found in the *HMM-based speech synthesis system (HTS)* [12], reduce artifacts by carefully tuning the convergence criterion during gradient descent, providing a form of *early stopping* [8, 12].

We present a simple mathematical analysis of GV generation based on Lagrange multipliers. This analysis naturally leads to a new partially analytic algorithm for doing exact GV generation. In addition the analysis provides some insight into one source of artifacts, and we present a simple way to vastly reduce the occurrence of artifacts while maintaining quality.

The analysis also suggests a way to approximate GV generation. This approximation can be implemented as a simple one-shot adjustment to the static parameters of the model after training and before synthesis. The standard speech parameter generation algorithm (or its time-recursive variant) can then be used for fast (or low latency and reasonably fast) approximate GV generation. We will see that in practice this approximation performs very well.

Previous work attempting to reduce artifacts includes using full-covariance models and a full-covariance GV distribution [7], carefully tuning the stopping criterion (for context-dependent GV distributions) [8], and using 'GV-constrained' trajectory HMM training [13]. The present work provides a new interpretation of one source of artifacts, and a new technique to reduce their occurrence.

Previous attempts to make GV generation faster at synthesis time have mainly focussed on incorporating aspects of the GV utility function into training [14, 15, 13, 16]. However unlike the present work these approaches result in a significant increase in the complexity of training. In addition it has been shown that simple linear and non-linear scaling of the mean trajectory performs better than post-filtering and almost as well as GV generation [17]. Our method more directly approximates conventional GV generation. A direct comparison between the two approaches would be informative.

This work was supported in part by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

2. BACKGROUND

In a typical statistical parametric speech synthesis system speech audio is represented as a sequence of *acoustic feature vectors* (or *speech parameters*) [9]. One component of this feature vector sequence is referred to as a *trajectory*. A trajectory is therefore a sequence $c = c_{1:T}$ of T real numbers.

For a given *hidden state sequence* encoding information about the text together with timings, and a given component of the feature vector, typical models use a Gaussian distribution over the trajectory c, and so the log pdf (up to a constant) is a quadratic function

$$A(c) \triangleq -\tfrac{1}{2} c^{\mathsf{T}} P c + b^{\mathsf{T}} c \tag{1}$$

where the precision matrix P and the b-value b of the Gaussian depend on the state sequence [18, 19]. The most likely trajectory $\arg\max_c A(c)$ is the mean trajectory $\mu = P^{-1} b$. The standard speech parameter generation algorithm (case 1 in [2]) computes the mean trajectory efficiently by exploiting the fact that for typical models P is band-diagonal [2].

2.1. Speech parameter generation considering global variance

The *global variance (GV)* $v(c)$ of a trajectory c is given by

$$v(c) \triangleq \frac{1}{T} \sum_t c_t^2 - \left(\frac{1}{T} \sum_t c_t \right)^2 = \frac{1}{T} c^{\mathsf{T}} J c \tag{2}$$

where $J \triangleq I - \frac{1}{T} \mathbb{1}\mathbb{1}^{\mathsf{T}}$, I is the identity matrix, and $\mathbb{1}$ is a vector of ones of length T. Note that $v(c)$ is also a quadratic function.

In speech parameter generation considering global variance [1] we optimize a modified utility function

$$G(c) \triangleq A(c) + \omega \log \mathcal{N}(v(c); \mu_{\mathrm{GV}}, \sigma^2_{\mathrm{GV}}) \tag{3}$$

where μ_{GV} and σ^2_{GV} are the empirical mean and variance of $v(c)$ over the training corpus and ω controls the trade-off between producing a likely trajectory and a trajectory with likely global variance. Typically $\omega = 1/(3T)$ [1].

The conventional algorithm is to optimize $G(c)$ using a form of gradient descent [1]. As mentioned in §1 the stopping criterion is typically carefully tuned, providing a form of early stopping which helps to avoid artifacts. We refer to GV generation run until convergence with no early stopping as *exact* GV generation.

It has been noted that in practice the effect of this utility function is typically to set the global variance of the generated trajectory to be almost exactly equal to μ_{GV} [11, 20].

2.2. Static model parameters

For models such as the standard HMM synthesis framework [9] and the trajectory HMM [18] the precision matrix P and b-value b are of the form $P = \text{diag}(\tau^0) + P^\Delta$ and $b = b^0 + b^\Delta$ where τ^0 is a vector where each entry τ^0_t is the *static precision model parameter* for the hidden state at time t, b^0 is the *static b-value model parameter* for the hidden state at time t, and P^Δ and b^Δ both depend on the *delta* and *delta-delta* model parameters [19]. Here b^0_t is related to the more conventional static mean model parameter μ^0_t by $b^0_t = \tau^0_t \mu^0_t$.

Thus the static precision parameters τ^0 contribute to the diagonal of P and the static b-value parameters b^0 contribute to b.

3. USING LAGRANGE MULTIPLIERS

In this section we present a theoretical analysis of GV generation based on Lagrange multipliers.

3.1. Optimal trajectory with given global variance

We first look at the sub-problem of finding the trajectory c which maximizes $A(c)$ subject to the constraint that its global variance $v(c)$ must equal some particular target value v. This sub-problem may be solved by introducing a *Lagrange multiplier* λ. Define

$$L(c;\lambda) \triangleq A(c) + \tfrac{1}{2}\lambda T v(c) = -\tfrac{1}{2} c^\mathsf{T}(P - \lambda J)c + b^\mathsf{T} c. \quad (4)$$

As long as $P - \lambda J$ is positive definite, the unique global optimum of the quadratic function $L(c;\lambda)$ is given by the trajectory

$$\hat{c}(\lambda) \triangleq (P - \lambda J)^{-1} b. \quad (5)$$

Given λ, if $c \neq \hat{c}(\lambda)$ is a trajectory with the same global variance as $\hat{c}(\lambda)$ then $A(\hat{c}(\lambda)) = L(\hat{c}(\lambda);\lambda) - \frac{1}{2}\lambda T v(\hat{c}(\lambda)) > L(c;\lambda) - \frac{1}{2}\lambda T v(\hat{c}(\lambda)) = A(c)$. Thus $\hat{c}(\lambda)$ is the (unique) optimal trajectory with global variance $v(\hat{c}(\lambda))$. A less terse proof is given in [21].

Subject to fairly mild conditions on P and b, $v(\hat{c}(\lambda)) \to 0$ as $\lambda \to -\infty$ and $v(\hat{c}(\lambda)) \to \infty$ as $\lambda \to \lambda_{\text{J}}$ where λ_{J} is the smallest λ for which $P - \lambda J$ fails to be positive definite [21]. By differentiating $v(\hat{c}(\lambda))$ it is easy to show that this is a strictly increasing function of λ on $(-\infty, \lambda_{\text{J}})$. Therefore for any $v > 0$ there exists a unique λ such that the trajectory $\hat{c}(\lambda)$ has global variance v. Thus the solution to the above sub-problem is always of the form (5) for some $\lambda < \lambda_{\text{J}}$.

We can make (5) simpler to compute by applying the matrix inversion lemma to obtain

$$\hat{c}(\lambda) = (P - \lambda I)^{-1}(b - \nu(\lambda)\mathbb{1}) \quad (6)$$

$$\text{where } \nu(\lambda) = \frac{\lambda b^\mathsf{T}(P - \lambda I)^{-1}\mathbb{1}}{T + \lambda \mathbb{1}^\mathsf{T}(P - \lambda I)^{-1}\mathbb{1}} \quad (7)$$

as long as $P - \lambda I$ is invertible. $P - \lambda I$ is invertible for all $\lambda < \lambda_{\text{J}}$ except $\lambda = \lambda_{\text{I}}$, where λ_{I} is the smallest eigenvalue of P [21]. Since $P - \lambda I$ is band-diagonal, quantities of the form $(P - \lambda I)^{-1}x$ can be computed efficiently using a banded LU decomposition. This allows $\hat{c}(\lambda)$ to be computed in $O(T)$ time.

3.2. Partially analytic GV generation

The trajectory generated by (exact) GV generation is of the form (5) for some λ. To see this, let v be the global variance of the trajectory obtained by maximizing $G(c)$. From §3.1 there is some $\lambda < \lambda_{\text{J}}$ such that $v(\hat{c}(\lambda)) = v$. But $\hat{c}(\lambda)$ is the unique optimum for $A(c)$, and so $G(c)$, amongst trajectories with global variance v. Thus $c = \hat{c}(\lambda)$.

This leads naturally to a new algorithm for GV generation where the one-dimensional optimization of $\lambda \mapsto G(\hat{c}(\lambda))$ is performed numerically with $\hat{c}(\lambda)$ for each λ computed analytically. This *partially analytic* algorithm may be faster than the conventional gradient descent-based algorithm, but we do not investigate its speed here.

3.3. Interpretation in terms of static parameters

Due to the way the static parameters τ^0 contribute to the diagonal of P and the way the static parameters b^0 contribute to b (§2.2) we may view (6) as the standard speech parameter generation algorithm on a modified model where we replace τ^0_t by $\tau^0_t - \lambda$ and b^0_t by $b^0_t - \nu(\lambda)$. Note that λ is *utterance-specific* in the sense that the value of λ selecting during parameter generation may vary from utterance to utterance. This interpretation will be used in §4.1 and §5.

3.4. Global mean squared deviation (GMSD)

In this section we introduce *GMSD generation.* This is similar to GV generation but has a simpler form of $\nu(\lambda)$ which will prove useful in §5. Define the *global mean squared deviation (GMSD)* of a trajectory c around a given value u as $s_u(c) \triangleq \frac{1}{T}\sum_t (c_t - u)^2$. By a similar argument to that used in §3.1 it can be shown that the optimal trajectory with given GMSD is of the form (6) with $\nu(\lambda) = u\lambda$ for some (unique) $\lambda < \lambda_{\text{I}}$ [21]. We set u to be the mean value of c_t over all frames of the training corpus.

4. ARTIFACTS

As discussed in §1 GV generation sometimes introduces artifacts. In this section we show how the analytic results presented in §3 shed light on one source of these artifacts, and we present a modification that reduces their occurrence.

It has been suggested previously that one of the causes of artifacts is the fact that GV generation effectively uses μ_{GV} as the target GV for all utterances (§2.1). This corpus average GV may be too large for certain utterances, particularly short ones, taking the statistical model "out of its comfort zone" and leading to artifacts [10, 11]. We investigate one version of this hypothesis and find it to be a relatively minor source of artifacts for our experimental systems.

4.1. One effect which may lead to artifacts

The analytic results presented in §3.3 show that GV generation may be interpreted as standard parameter generation after subtracting utterance-specific values λ and $\nu(\lambda)$ from the static parameter sequences τ^0 and b^0 respectively. However whereas for the standard case we have $\tau^0_t > 0$, the modified parameter $\tau^0_t - \lambda$ may be less than zero. What impact might this have on the generated trajectory?

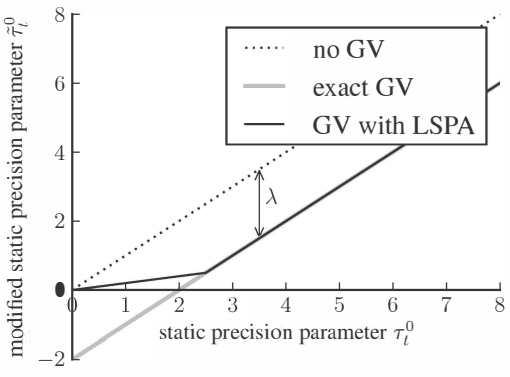

Fig. 1. Conventional exact GV generation (exact GV) effectively subtracts a constant λ from the static precision parameter τ_t^0. Local static parameter adjustment (GV with LSPA) alters this to ensure the new value is never negative. Here $\xi = 0.2$ and $\lambda = 2$ in (8).

In general when using the standard parameter generation algorithm we can roughly think of the static, delta and delta-delta model parameters (§2.2) as encoding soft constraints on what the trajectory value, velocity and acceleration should be while in a given state. The standard parameter generation algorithm trades off these conflicting constraints to produce a compromise trajectory. The precision parameter τ_t^0 controls how important the corresponding soft constraint is: the larger τ_t^0, the more harshly a potential trajectory is penalized for deviating from the static mean value b_t^0/τ_t^0 at time t. In this view $\tau_t^0 < 0$ corresponds to the counterintuitive constraint that the trajectory value at time t should *deviate* from the static mean b_t^0/τ_t^0 as much as possible. The more negative τ_t^0 the more important it is for the trajectory to deviate substantially from this static mean value.

Thus if $\tau_t^0 - \lambda$ in §3.3 becomes too negative we may get local 'excursions' towards large positive or negative values in the trajectory, which may cause audible artifacts in the synthesized audio.

4.2. Local static parameter adjustment (LSPA)

Exact GV generation effectively uses a modified static precision parameter $\tilde{\tau}_t^0$ set to $\tau_t^0 - \lambda$, which may be negative. For *local static parameter adjustment (LSPA)* we instead set

$$\tilde{\tau}_t^0(\lambda) \triangleq \max(\tau_t^0 - \lambda, \xi\tau_t^0). \tag{8}$$

The form of this function is shown in Figure 1. We set $\xi = 0.2$ based on small-scale preliminary experiments.

LSPA may also be viewed in terms of an *adjustment weight* $w_t(\lambda) \triangleq \min(1, \frac{1}{\lambda}(1-\xi)\tau_t^0)$. This definition is chosen so that $\tilde{\tau}_t^0(\lambda) = \tau_t^0 - \lambda w_t(\lambda)$. Note that $0 \leq w_t(\lambda) \leq 1$, with $w_t(\lambda) = 1$ for frames where no adjustment is made.

We generalize (6) to the LSPA case by setting

$$\hat{c}(\lambda) \triangleq (P - \lambda\,\mathrm{diag}(w(\lambda)))^{-1}(b - \nu(\lambda)w(\lambda)). \tag{9}$$

We recover (6) in the case $w(\lambda) = \mathbb{1}$. For LSPA GMSD generation we set $\nu(\lambda) = u\lambda$ as before. For LSPA GV generation we set

$$\nu(\lambda) = \frac{\lambda b^\mathsf{T}(P - \lambda\,\mathrm{diag}(w(\lambda)))^{-1}w(\lambda)}{\mathbb{1}^\mathsf{T}w(\lambda) + \lambda w(\lambda)^\mathsf{T}(P - \lambda\,\mathrm{diag}(w(\lambda)))^{-1}w(\lambda)}. \tag{10}$$

We recover (7) in the case $w(\lambda) = \mathbb{1}$.

LSPA is *local* is the sense that it only makes an adjustment to the static precision parameter in regions where $\tau_t^0 - \lambda$ is negative or close to negative. If the effect mentioned in §4.1 is large then we might hope that LSPA would reduce artifacts while maintaining the other advantages of GV generation. We will see experimentally that this is indeed the case.

5. FIXED LAGRANGE MULTIPLIER

We have seen (§3.3) that trajectories generated by exact GV generation are of the form (6) where λ is an utterance-specific Lagrange multiplier. A natural question is whether quality degrades substantially if we instead use a fixed value of λ for all utterances.

For simplicity we use GMSD generation instead of GV generation when using a fixed value of the Lagrange multiplier λ. We refer to the use of fixed λ together with LSPA as *fixed LSPA*.

Fixing λ has the advantage of allowing very fast generation, since the subtraction of $\lambda w_t(\lambda)$ from the static precision parameters τ_t^0 and the subtraction of $u\lambda w_t(\lambda)$ from the static b-value parameters b_t^0 can be performed off-line as a simple *model* adjustment, and then the standard speech parameter generation algorithm, or its time-recursive variant, used at synthesis time.

We propose two methods to train the fixed value of λ used for each component of the feature vector.

Firstly for each utterance r in the training set (and for a given component of the feature vector) we can find the λ^r such the generated trajectory $c^r(\lambda^r)$ has GMSD equal to that of the corresponding natural trajectory c_{nat}^r. We can then set λ to be the median, or slightly more generally some percentile, of the training set λ^r values.

Secondly we can choose to use the value of λ which minimizes the mean squared error in the GMSD of the generated trajectories over the training set $\frac{1}{R}\sum_{r=1}^{R}[s_u(c^r(\lambda)) - s_u(c_{\text{nat}}^r)]^2$. We refer to this second approach as MSE GMSD training.

6. EXPERIMENTS

We performed two sets of experiments. Firstly we used a range of generation methods with a trajectory HMM system to evaluate the effectiveness of LSPA at reducing artifacts. Secondly we used a range of generation methods with a standard system to evaluate the extent to which fixing the value of the Lagrange multiplier λ causes a degradation in naturalness.

6.1. Systems

The systems were trained on the CMU ARCTIC corpus [22] for the single speaker 'slt' (approximately 1 hour) with 50 held-out utterances. The training regime for the standard system was adapted from the HTS speaker dependent training demo [12]. The trajectory HMM system took the trained standard system as a starting point, and re-estimated the spectral leaf parameters based on a fixed alignment. All other details of the experimental set-up were as in [23].

6.2. Listening test

A Blizzard Challenge-style [24] listening test was conducted over several weeks. It was completed by 24 native English speakers. The listening test consisted of three parts containing 48 utterances each.

In the first part of the listening test the listeners evaluated the naturalness of the methods in §6.3 on a scale of 1 to 5. The second part was similar but using the methods in §6.4. In the third part the listeners heard the methods in §6.3 and for each utterance

were asked to judge whether it contained an artifact, described as "a short distortion in the audio, e.g. a blip, a click, a pop, or a short high-pitched whine, but NOT a short pause in the incorrect position". They were told that they should expect roughly 1 in 10 utterances to contain an artifact, and were presented with two examples, generated by method E1, of utterances containing an artifact. The third part was conducted after the first two so that perceived artifacts would not influence naturalness judgements.

6.3. LSPA evaluation

A number of different generation methods were used with the trajectory HMM system in order to investigate the effectiveness of LSPA for reducing artifacts:

- H uses conventional HTS GV generation with early stopping
- E1 uses exact GV generation. This was implemented using the partially analytic solution in §3.2, but we checked that running conventional HTS gradient descent for millions of iterations gives almost indistinguishable trajectories.
- E2 uses exact GV generation, but using a target GV for each utterance. This target GV is set to the expected GV $\mathbb{E}v(c)$ where $c \sim \mathcal{N}(\mu, P^{-1})$, where $\mu = P^{-1}b$ is the mean trajectory. This may be computed as $\mathbb{E}v(c) = \frac{1}{T}\text{tr}[P^{-1}] - \frac{1}{T^2}\mathbb{1}^\mathsf{T}P^{-1}\mathbb{1} + v(\mu)$ where tr is trace.
- A1 uses LSPA GV generation, using the expected GV as the target GV for each utterance
- A2 uses LSPA GMSD generation, using the expected GMSD as the target GMSD for each utterance

Method E2 allows us to investigate the previously suggested hypothesis that artifacts are partly caused by μ_{GV} being an inappropriate GV value for some utterances (§4). In contrast to E1, E2 allows the model to decide how much global variance it expects for each utterance. Note that, unlike for the unnormalized standard HMM, for the trajectory HMM the corpus mean of the expected GV is very close to the corpus mean of the natural GV, so any decrease in the number of artifacts should not be due to using less GV overall.

We used the trajectory HMM system for assessing LSPA since exact GV generation with this model produced more artifacts than with the standard system, providing a more robust test of LSPA's effectiveness at reducing artifacts.

For methods (N, H, E1, E2, A1, A2) the proportion of utterances judged to contain an artifact were (5%, 15%, 66%, 60%, 18%, 17%) respectively, and the mean opinion scores were (4.6, 2.6, 2.5, 2.7, 2.7, 2.6) respectively. In a Mann-Whitney U test none of the synthetic methods had significantly different naturalness to any other.

These results show that optimizing the global variance utility function introduces many artifacts (E1). This is a weakness in the global variance utility function. As expected early stopping is very effective at reducing artifacts (E1 versus H).

Comparing E1 and E2 we can see that few of the artifacts introduced by GV generation appear to be due to the statistical model being asked to generate trajectories with global variance values it views as unreasonable. Comparing E2 and A1 we can see that the effect outlined in §4.1 appears to be a substantial source of artifacts.

LSPA appears to be almost as effective as early stopping at preventing artifacts and equally as natural (H versus A1 and A2).

The authors perceived almost no 'GV-like' artifacts for systems H, A1 or A2, so listeners may be identifying artifacts not due to GV generation. Perhaps surprisingly the presence of artifacts had very little effect on naturalness judgements, with E1 and E2 rated as natural as H, A1 and A2 despite having many more artifacts.

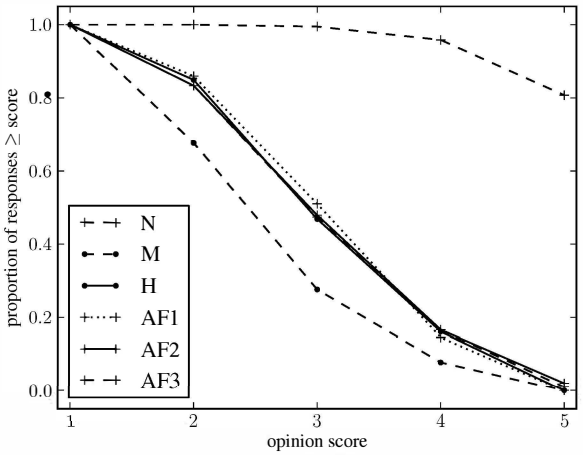

Fig. 2. Complementary cumulative plot showing results for the fixed LSPA evaluation. For an opinion score s, the ordinate gives the proportion of participant responses that were s or greater. For any given opinion score larger ordinate values are better.

6.4. Fixed LSPA evaluation

A number of different generation methods were used with the standard HMM system in order to investigate the effectiveness of fixed LSPA for fast GV generation:

- M uses the standard speech parameter generation algorithm
- H uses conventional HTS GV generation with early stopping
- AF1, AF2 and AF3 use fixed LSPA GMSD generation. The fixed value λ is set using the percentile method at 50% for AF1, the percentile method at 85% for AF2, and the MSE GMSD method for AF3 (see §5).

Preliminary listening by the authors suggested no 'GV-like' artifacts were present for any of the above systems so only naturalness was formally evaluated.

The mean opinion scores for methods (N, M, H, AF1, AF2, AF3) were (4.8, 2.0, 2.5, 2.5, 2.5, 2.5) respectively. The naturalness opinion score results are summarized in Figure 2. We use a *complementary cumulative plot* [23] since it contains more information than a box plot. In a Mann-Whitney U test M was significantly different to all other methods, but none of the GV-based methods (H, AF1, AF2, AF3) was significantly different to any other.

As expected conventional GV generation gives substantial gains over the standard parameter generation algorithm (M versus H).

Fixed LSPA GMSD generation appears to give at least as good naturalness as conventional HTS GV generation (H versus AF1/2/3).

The three methods for choosing the fixed value of λ appear to give very similar results (AF1, AF2 and AF3).

7. CONCLUSION

We have presented a simple new theoretical analysis of GV generation based on Lagrange multipliers. We have shown how this analysis sheds light on one source of the artifacts sometimes introduced by GV generation, and presented a new method (LSPA) that greatly reduces artifacts. We have seen that using a fixed Lagrange multiplier (fixed LSPA) provides a fast approximation to GV generation which showed no degradation in naturalness compared to conventional GV generation in our experiments.

8. REFERENCES

[1] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

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

[3] K. Koishida, K. Tokuda, T. Masuko, and T. Kobayashi, "Vector quantization of speech spectral parameters using statistics of static and dynamic features," *IEICE Trans. Inf. Syst.*, vol. E84-D, no. 10, pp. 1427–1434, 2001.

[4] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "CELP coding based on mel-cepstral analysis," in *Proc. ICASSP 1995*, 1995, pp. 33–36.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis," *IEICE Trans. Inf. Syst. (Japanese edition)*, vol. J87-D-II, no. 8, pp. 1565–1571, 2004.

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

[7] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," in *Proc. Blizzard Challenge Workshop 2006*, 2006.

[8] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge Workshop 2008*, 2008.

[9] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[10] T. Toda, "Modeling of speech parameter sequence considering global variance for HMM-based speech synthesis," in *Hidden Markov models, theory and applications*, P. Dymarski, Ed. In-Tech, 2011.

[11] S. King, "An introduction to statistical parametric speech synthesis," *Sadhana*, vol. 36, no. 5, pp. 837–852, 2011.

[12] HTS working group, "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/, accessed 30 November 2012.

[13] T. Toda and S. Young, "Trajectory training considering global variance for HMM-based speech synthesis," in *Proc. ICASSP 2009*, 2009, pp. 4025–4028.

[14] J. Latorre, K. Iwano, and S. Furui, "Combining Gaussian mixture model with global variance term to improve the quality of an HMM-based polyglot speech synthesizer," in *Proc. ICASSP 2007*, pp. 1241–1244.

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

[16] H. Zen, M. J. F. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE Trans. Audio Speech Language Process.*, vol. 20, no. 3, pp. 794–805, 2012.

[17] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proc. Interspeech 2012*, 2012.

[18] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech and Language*, vol. 21, no. 1, pp. 153–173, 2007.

[19] M. Shannon and W. Byrne, "A formulation of the autoregressive HMM for speech synthesis," Department of Engineering, University of Cambridge, UK, Technical Report CUED/F-INFENG/TR.629, 2009, http://mi.eng.cam.ac.uk/~sms46/papers/shannon2009fah.pdf.

[20] H. Zen, M. J. F. Gales, Y. Nankaku, and K. Tokuda, "Statistical parametric speech synthesis based on product of experts," in *Proc. ICASSP 2010*, 2010, pp. 4242–4245.

[21] M. Shannon and W. Byrne, "Partially analytic speech parameter generation considering global variance," Department of Engineering, University of Cambridge, UK, Technical Report CUED/F-INFENG/TR.682, 2013, http://mi.eng.cam.ac.uk/~sms46/papers/shannon2013partially.pdf.

[22] J. Kominek and A. W. Black, "The CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, Technical Report CMU-LTI-03-177, 2003.

[23] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Trans. Audio Speech Language Process.*, vol. 21, no. 3, pp. 587–597, 2013.

[24] A. W. Black and K. Tokuda, "The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. Interspeech 2005*, 2005, pp. 77–80.