Ways to Implement Global Variance in Statistical Speech Synthesis

Hanna Silén, Elina Helander, Jani Nurminen, Moncef Gabbouj

Department of Signal Processing, Tampere University of Technology, Finland
hanna.silen@tut.fi, elina.helander@tut.fi, jani.nurminen@tut.fi, moncef.gabbouj@tut.fi

Abstract

Hidden Markov model-based speech synthesis is prone to over-smoothing of spectral parameter trajectories. The maximum-likelihood parameter generation favors smooth tracks and the utterance-level variance of each parameter trajectory is significantly reduced compared to the original recordings. This results in muffled speech. To retain the natural variance, statistical global variance modeling has been used in parameter generation. The modeling increases the utterance-level variance in synthesis, but it is computationally demanding: there is no closed-form solution and an iterative approach is used. In this paper, we analyze the performance of two simple alternative approaches for retaining the natural variance of spectral parameters in synthesis, namely variance scaling and histogram equalization. Both methods apply analytically solvable parameter generation and impose the natural variance afterwards as an efficient post-processing step. Subjective evaluations carried out on English data confirm that the achieved synthesis quality is higher compared to simple post-filtering and similar to the standard global variance modeling.
Index Terms: statistical speech synthesis, global variance, variance scaling, histogram equalization

1. Introduction

In statistical speech synthesis that is most often used as a synonym for hidden Markov model (HMM) based speech synthesis [1], speech is generated from statistical models that have been learned from a speech database. The footprint of a statistical speech synthesis system is usually small and the generated speech is rather free from artifacts. However, the approach typically suffers from over-smoothing resulting in slightly muffled-sounding speech.

In this paper we investigate the enhancement of over-smoothed spectral parameters, namely mel-cepstral coefficients (MCCs) [2]. The over-smoothing takes place both in time and frequency-domain. In frequency-domain, fine details of the speech are lost and formants are broadened. As a simple post-processing step, post-filtering common in speech coding [3, 4] can be used to emphasize formants after parameter generation.

In time-domain, a synthesized feature trajectory has much less variation compared to a recorded trajectory as illustrated in Fig. 1. To alleviate the problem, the concept of global variance (GV) has been introduced [5]. It was first used in the context of voice conversion [6] but it has become a well-established technique for reducing over-smoothing also in HMM-based speech synthesis [7]. The aim is to retain (to an extent) the original utterance-level variation of a speech feature trajectory. The parameter generation algorithm maximizes a likelihood that is a combination of HMM and GV likelihoods. The GV likelihood can be regarded as a penalty for reduction of the variance of the generated parameter trajectory.

The above-mentioned post-filtering and parameter generation with GV modeling are the most common ways to reduce over-smoothing in statistical speech synthesis. The GV-based approach allows precise control on the individual spectral features and has been found to improve synthesis similarity to the target speaker compared to post-filtering in voice conversion [6] and the speech naturalness in HMM-based speech synthesis [5]. The standard maximum-likelihood (ML) parameter generation [8] problem without GV can be approximated in a way that has a closed-form solution, but for the joint optimization of HMM and GV likelihoods, there typically is no closed-form solution. The use of GV is hence computationally more complex which can be considered an obstacle e.g. for small handheld devices. Alternative methods for improving natural variation in HMM-based speech synthesis have been proposed e.g. in [9] and [10].

In this paper, we introduce two alternative methods for efficient implementation of utterance-level GV. The synthesis performance of the methods is similar to the parameter generation with GV, but the techniques are less computationally demanding and are applied after parameter generation as a post-processing step. The techniques are 1) *variance scaling* and 2) *histogram equalization.* Both approaches employ the efficient closed-form parameter generation of [8]. In VS, the generated feature sequence is scaled with a GV (mean variance of training data utterances) but unlike in the traditional GV parameter generation, the variance is imposed as a post-processing step. This kind of closed-form ML parameter generation with variance scaling is typically also used in the traditional GV parameter generation as an initialization step of the iterative estimation. HEQ has been used earlier to reduce the over-smoothing effect of the GMM-based F_0 conversion in [11] and in this paper it is used to equalize the histogram of each spectral feature according to the average training data histograms. Experimental results show that the standard parameter generation with GV can be replaced with any of these two techniques.

This paper is organized as follows. In Section 2 we briefly review the parameter generation algorithm with and without GV. Section 3 describes two alternative methods, variance scaling and histogram equalization, to alleviate the over-smoothing problem. Evaluation results are reported in Section 4. Section 5 provides discussion and Section 6 concludes the paper.

2. Speech parameter generation

Synthetic speech parameters are created from pre-trained HMMs with means, variances, and transition probabilities learned from training data using one of the parameter generation algorithms of [8]. Typically, ML estimation of speech parameters is employed and in the simplest case it can be completed analytically. To improve naturalness, joint optimization of HMM and global variance (GV) likelihoods [5, 6] can be used. These are briefly discussed in the following.

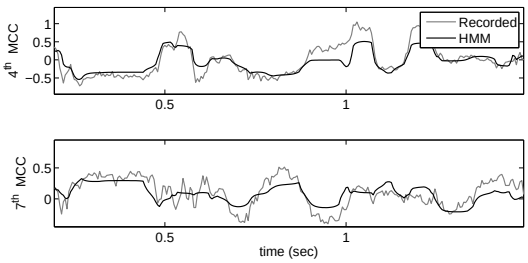

Figure 1: *Over-smoothing effect in HMM-based speech synthesis. Utterance-level MCC trajectories of 4th and 7th coefficient for US English female speaker (CMU Arctic speaker* `slt`*).*

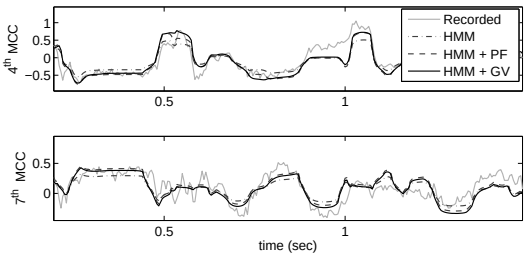

Figure 2: *Enhancement of over-smoothing from post-filtering and global variance modeling for the utterance of Fig. 1.*

2.1. ML parameter generation

As a result of the training phase, a set of context-dependent HMMs is formed with each model consisting of a distribution (or a mixture of them) with means and variances as well as transition probabilities learned from the training data.

We denote the set of HMMs by λ and the observation sequence to be estimated by $\mathbf{O} = [\mathbf{o}_1^T, \mathbf{o}_2^T, \ldots, \mathbf{o}_N^T]^T$, N being the number of frames in the utterance. The observation vectors $\mathbf{o}_n$ are column vectors with M static coefficients augmented with their dynamic coefficients. The ML estimate for the synthetic speech parameter sequence $\mathbf{O}$ is [8]:

$$\mathbf{O}^* = \arg\max_{\mathbf{O}} P(\mathbf{O}|\lambda) = \arg\max_{\mathbf{O}} \sum_{\text{all } \mathbf{q}} P(\mathbf{O}, \mathbf{q}|\lambda), \quad (1)$$

under the condition $\mathbf{O} = \mathbf{WC}$, where $\mathbf{C}$ contains only the static coefficients of $\mathbf{O}$ and matrix $\mathbf{W}$ a linear transformation between $\mathbf{C}$ and $\mathbf{O}$. The solution of (1) can be approximated by a separate estimation of an optimal state sequence and an optimal observation matrix given the state sequence. The closed-form solution is then found efficiently by Cholesky decomposition.

2.2. ML parameter generation with GV modeling

To alleviate the flattening effect of HMM training and ML parameter generation, joint optimization of HMM and GV likelihoods is typically used [5]. GV model (λ_v) is learned from the training data and it models the utterance-level variance of each static feature. Typically λ_v is a single-Gaussian model; either a single model for every utterance or a context-dependent model taking into account for instance the length of an utterance.

Instead of (1), the estimation problem is of the form:

$$\mathbf{O}^* = \arg\max_{\mathbf{O}} \sum_{\text{all } \mathbf{q}} P(\mathbf{O}, \mathbf{q}|\lambda)^{\omega} P(\mathbf{v}(\mathbf{c})|\lambda_v), \quad (2)$$

with a single-Gaussian GV model:

$$P(\mathbf{v}(\mathbf{c})|\lambda_v) = \mathcal{N}(\mathbf{v}(\mathbf{c}); \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v) \quad (3)$$

with mean and variance $\boldsymbol{\mu}_v$ and $\boldsymbol{\Sigma}_v$. The GV term can be seen as a penalty for the over-smoothing. Factor ω is a constant scaling term adjusting the balance between the two likelihoods.

The use of GV alleviates the flattening of spectral trajectories inherent in (1). This is illustrated in Fig. 2, where example trajectories for the utterance of Fig. 1 are synthesized with and without GV (HMM + GV, HMM). For comparison, post-filtered trajectories are given as well (HMM + PF). However, since the GV term in (2) makes the problem analytically unsolvable, an iterative approach – or trajectory training with GV to make the related parameter generation problem analytically solvable [12] – is needed in order to find the optimal feature sequence $\mathbf{O}^*$.

3. Efficient implementation of GV

The alternative methods proposed here impose the natural utterance-level variance as an efficient post-processing step. Instead of (2), we use the analytically-solvable approximation of the standard ML parameter generation of (1). Hence the parameter generation phase can be completed analytically without a need for iterative computing. The generated parameter trajectories are generally smooth and provide a stable basis for the modifications. In the first and the simplest approach, variance scaling (VS), we scale the generated trajectories to a new utterance-level global variance learned from the training data. In the second approach, histogram equalization (HEQ), each trajectory is transformed using its histogram and a target histogram from the training data. They are described in the following.

3.1. Variance scaling

The initial speech trajectories generated by the closed-form approximation of (1) are scaled to a new utterance-level variance learned from the training data. This approach is very simple and intuitive and provides rather similar results to (2) but is computationally significantly less demanding.

We denote the mth spectral feature trajectory generated by the approximation of (1) by $\mathbf{c}_m = [c_m(1), \ldots, c_m(N)]^T$. If each feature dimension is independent from other dimensions, as can be assumed in the case of MCCs, each dimension can be scaled separately. The resulting variance-scaled feature value $c'_m(n)$ is then:

$$c'_m(n) = \frac{\sigma_m^{gv}}{\sigma_m} [c_m(n) - \mu_m] + \mu_m, \quad (4)$$

where $n = 1, 2, \ldots, N$. Here μ_m and $(\sigma_m)^2$ are the mean and variance of the mth trajectory before VS and $(\sigma_m^{gv})^2$ is the utterance-level target GV learned from the training data.

3.2. Histogram equalization

Histogram equalization (HEQ) provides a more detailed conversion of the spectral trajectories generated by the closed-form approximation of (1). HEQ is a widely used tool in image processing and it provides a simple, nonlinear and non-parametric conversion of the parameter tracks. Since there is a closed-form solution for the initial parameter tracks, computational complexity is rather low.

Following the HEQ in [11], the range from $\min(\tilde{\mathbf{c}}_m)$ to $\max(\tilde{\mathbf{c}}_m)$ of zero-mean normalized feature values:

$$\tilde{\mathbf{c}}_m = \mathbf{c}_m - \mu_m, \quad (5)$$

where μ_m denotes the original utterance-level feature mean, is divided into L uniformly spaced bins so that:

$$\min(\tilde{\mathbf{c}}_m) = a_m^{(1)} < a_m^{(2)} < \ldots < a_m^{(L)} = \max(\tilde{\mathbf{c}}_m). \quad (6)$$

The histogram $h_{\tilde{\mathbf{c}}_m}$ of $\tilde{\mathbf{c}}_m$ is then defined as:

$$h_{\tilde{\mathbf{c}}_m}(i) = \frac{N_m^{(i)}}{N}, \quad (7)$$

where $N_m^{(i)}$ is the number of frames assigned to the bin $a_m^{(i)}$, $i = 1, \ldots L$, and N is the total number of frames.

The target histogram, here the average histogram of all training utterances, is formed separately in a similar way. All feature tracks are normalized to zero mean and the range of the training feature tracks is then divided uniformly into L bins such that $b_m^{(1)} < b_m^{(2)} < \ldots < b_m^{(L)}$ as in (6). The histogram value at each bin $b_m^{(i)}$ could be computed similarly to (7).

The equalization procedure transforms the ML-generated synthesized feature tracks by using the ratio of the bin distances in the synthesized and natural (average) histograms as well as the actual bin locations. As a result, the dynamic range of the feature values is increased (or decreased) according to the natural histogram. The HEQ operation is here formulated as [11]:

$$c'_m(n) = \frac{b_m^{(i+1)} - b_m^{(i)}}{a_m^{(i+1)} - a_m^{(i)}} \left[\tilde{c}_m(n) - a_m^{(i)}\right] + b_m^{(i)} + \mu_m, \quad (8)$$

where $a_m^{(i)}$ is the location of the histogram bin closest to $\tilde{c}_m(n)$ and $b_m^{(i)}$ the corresponding bin in the target histogram formed using the training data. The last term cancels the effect of mean-normalization in (5).

4. Experiments

The GV implementations for spectral enhancement proposed in the previous section were evaluated with a listening test. The evaluation setup and the results are described in the following. Synthesis samples are available at `http://www.cs.tut.fi/sgn/arg/silen/is2012/GV`.

4.1. Evaluation systems

The following four systems were considered in the evaluations.

Proposed systems:

- `VS`: Variance scaling of Section 3.1 (post-processing),
- `HEQ`: Histogram equalization of Section 3.2 (post-processing).

Both `VS` and `HEQ` are performed as a post-processing step after the closed-form parameter generation of Section 2.1. For `HEQ`, we set $L = 50$ and discarded the highest and lowest 1 % of the feature values considered to be outliers when forming the target histogram. No post-filtering was used in `VS` or `HEQ`.

Reference systems:

- `GV`: Global variance modeling jointly in parameter generation as in Section 2.2,
- `PF`: Parameter generation of Section 2.1 with post-filtering (post-processing).

Table 1: *Average preference percentage (95% confidence intervals) for the CMU Arctic data (speakers `rms` and `slt`).*

	No preference (%)	Reference preferred (%)	Proposed preferred (%)
`PF` vs. `HEQ`	10.6 ± 4.8	20.6 ± 6.3	68.8 ± 7.3
`PF` vs. `VS`	36.9 ± 7.6	18.8 ± 6.1	44.4 ± 7.8
`GV` vs. `HEQ`	21.3 ± 6.4	43.1 ± 7.8	35.6 ± 7.5
`GV` vs. `VS`	65.0 ± 7.5	18.1 ± 6.0	16.9 ± 5.9

System `GV` is computationally more demanding providing high quality while `PF` is a more light-weight approach using closed-form parameter generation (as also in `VS` and `HEQ`) and simple filtering to enhance the speech formants. In all four test setups, excitation parameters (F_0 and band-aperiodicity) were taken from the parameter generation of 2.1 without GV.

4.2. Evaluation data

Two US English sets from the publicly available CMU Arctic database (`http://www.festvox.org`) were used: female voice `slt` and male voice `rms`. For both speakers, the first half of the data (set A with 593 utterances) was used for HMM training and the second half of the data (set B with 539 utterances) was used for development and testing.

For both speakers, and all four systems, 10 different samples were synthesized. Both proposed systems (`VS` and `HEQ`) were evaluated against both reference systems (`PF` and `GV`), resulting in altogether 80 sample pairs in the evaluation. The order of the pairs and samples was randomized. Eight listeners (non-native but all fluent in English) participated in the test. For each sample pair, the listeners were asked which of the samples had better quality or if the quality was equal.

Speech waveforms with 16 kHz sampling rate were parameterized using STRAIGHT [13] into a spectral envelope encoded as MCCs of order 39, log-F_0, and voice aperiodicity of five frequency bands. For HMM-training and speech parameter generation (both for standard and GV parameter generation), HTS version 2.2 [7] was used. In the modeling, 5-state, left-to-right hidden semi-Markov models (HSMMs) with no state skips allowed were used. Synthesis durations were taken from the original recordings aligned using the trained HSMMs. Silences were discarded in the post-processing and GV modeling. For post-filtering, the standard post-filter factor 1.4 was used.

4.3. Listening test results

The results of the listening test are given in Table 1. The results indicate that both of the proposed methods (`VS` and `HEQ`) are able to outperform the pure post-filtering approach (`PF`). Especially `HEQ` seems to outperform `PF` significantly. Compared to the more complex parameter generation with GV (`GV`), `VS` is rated equal in quality while in the comparison of `HEQ` and `GV`, preference varies from sentence to sentence.

A deeper analysis of the speaker-wise results reveals that in the test `GV` vs. `HEQ`, the preference also is somewhat speaker-dependent: for `slt`, `GV` is preferred more often while for `rms`, `HEQ` gets a higher preference percentage. Furthermore, the preference of `HEQ` was also found to be listener-dependent with some listeners showing more preference towards `HEQ`.

Figures 3 and 4 show an example of utterance-level MCC trajectories (the 4th and 7th MCC) and histograms (the 4th MCC) in synthesis for the speaker `slt`. The utterance is the same as in figures 1 and 2. The figure with histograms shows

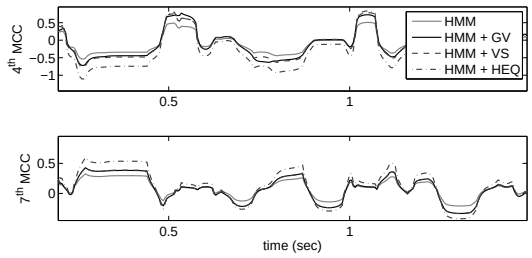

Figure 3: *MCC trajectories for the proposed and traditional GV parameter generation for the utterance of Fig. 1.*

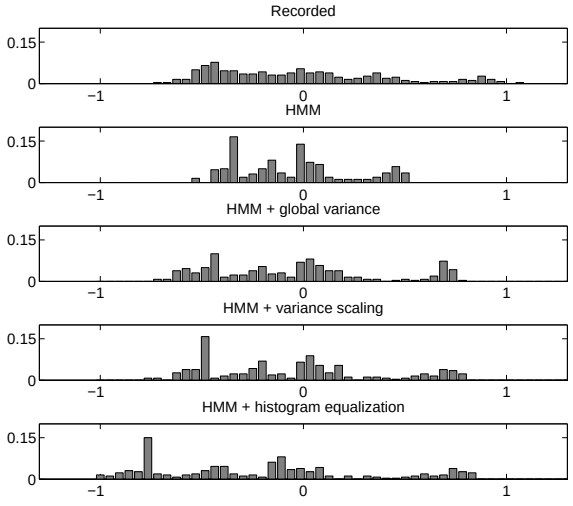

Figure 4: *Histogram of the 4th MCC for the utterance of Fig. 1 as recorded and generated using methods of sections 2–3.*

that compared to the original recorded utterance, the dynamic range of HMM-generated utterance is compressed significantly. Both VS and HEQ as well as GV are able to widen the dynamic range compared to pure HMM-generated utterance.

5. Discussion

In addition to the statistical speech synthesis, parameter generation with GV has been successfully applied in voice conversion in order to improve the perceived quality. It is presumable that the use of VS and HEQ can also be used to improve the voice conversion quality. Furthermore, they can be used with any voice conversion algorithm and not only the ones that utilize the parameter generation algorithm. The audio sample page given in Section 4 also contains samples from voice conversion.

The natural speech histograms were calculated from the whole training database but a more detailed selection of an utterance histogram based on the synthesized sentence could be applied. The GV was currently calculated over the whole training database context-independently, since according to our experiments, the context-dependent tree clustering of GV available in the current HTS version resulted in a tree with only two leaf nodes, the only clustering question concerning the length of the utterance. However, more detailed GV models could also improve the VS technique. Furthermore, in this paper, the over-smoothing reduction operations (GV, PF, and HEQ) were applied to each MCC separately, but more care should be taken if applied to line spectral frequency coefficients.

The speaker similarity was not evaluated in the experiments. However, initial listenings suggest that HEQ might produce higher similarity to the original recordings compared to GV and VS. Between GV and VS, no noticeable differences were found.

6. Conclusions

In this paper, we have described two alternative methods for alleviating the over-smoothing effect in HMM-based speech synthesis. The traditional parameter generation with GV is computationally complex and we have proposed to replace it with either simple variance scaling or histogram equalization. Variance scaling provides highly similar results to GV but histogram equalization produces slightly different results. Some listeners preferred the HEQ version and some the GV/VS version. The preference is also dependent on the speaker used in the synthesis. In the final analysis, the experimental results suggest that both proposed techniques can be used to replace the parameter generation with global variance in order to reduce the computational complexity without sacrificing the synthesis quality.

7. References

[1] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in *Proc. of 2002 IEEE Workshop on Speech Synthesis*, 2002.

[2] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. of ICASSP*, 1992.

[3] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "CELP coding based on mel-cepstral analysis," in *Proc. of ICASSP*, 1995.

[4] A. M. Kondoz, *Digital speech coding for low bit rate communication systems*. Wiley and Sons, 2004.

[5] 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, pp. 816–824, May 2007.

[6] ——, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. of ICASSP*, 2005.

[7] K. Tokuda, K. Oura, K. Hashimoto, S. Shiota, H. Zen, J. Yamagishi, T. Toda, T. Nose, S. Sako, and A. W. Black, "HMM-based speech synthesis system (HTS)," 2011. [Online]. Available: http://hts.sp.nitech.ac.jp/

[8] K. Tokuda, T. Kobayashi, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of ICASSP*, 2000.

[9] S. Tiomkin and D. Malah, "Statistical text-to-speech synthesis with improved dynamics," in *Proc. of Interspeech*, 2008.

[10] A. Sorin, S. Shechtman, and V. Pollet, "Uniform speech parameterization for multi-form segment synthesis," in *Proc. of Interspeech*, 2011.

[11] Z.-Z. Wu, T. Kinnunen, E. Chng, and H. Li, "Text-independent F0 transformation with non-parallel data for voice conversion," in *Proc. of Interspeech*, 2010.

[12] T. Toda and S. Young, "Trajectory training considering global variance for HMM-based speech synthesis," in *Proc. of ICASSP*, 2009.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, 1999.