Global Variance Modeling on the Log Power Spectrum of LSPs for HMM-based Speech Synthesis

Zhen-Hua Ling, Yu Hu, Li-Rong Dai
iFLYTEK Speech Lab, University of Science and Technology of China, P.R.China
zhling@ustc.edu, yuhu@iflytek.com, lrdai@ustc.edu.cn

Abstract

This paper presents a method to model the global variance (GV) of log power spectrums derived from the line spectral pairs (LSPs) in a sentence for HMM-based parametric speech synthesis. Different from the conventional GV method where the observations for GV model training are the variances of spectral parameters for each training sentence, our proposed method directly models the temporal variances of each frequency point in the spectral envelope reconstructed using LSPs. At synthesis stage, the likelihood function of trained GV model is integrated into the maximum likelihood parameter generation algorithm to alleviate the over-smoothing effect on the generated spectral structures. Experiment results show that the proposed method can outperform the conventional GV method when LSPs are used as the spectral parameters and improve the naturalness of synthetic speech significantly.

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

1. Introduction

The hidden Markov model (HMM)-based parametric speech synthesis method has been proposed and made significant progress in recent years [1–3]. In this method, the spectrum, F0 and segment durations are modeled simultaneously within a unified HMM framework [1]. At synthesis time, these features are directly predicted from the HMMs by means of a maximum-likelihood parameter generation algorithm which incorporates dynamic features [2]. The predicted parameter trajectories are then sent to a parametric synthesizer to reconstruct the speech waveform. This method is able to synthesize highly intelligible and smooth speech sounds [4, 5].

However, the quality of its synthetic speech suffers from the using of parametric synthesizer and the average effect of statistical modeling. The spectral envelopes generated from the acoustic HMMs are over-smoothed which leads to the muffled voice of synthetic speech. Many methods have been proposed to overcome the over-smoothing effect for the HMM-based parametric speech synthesis, such as post-filtering [5, 6], using rich context models [7], or combing unit selection strategy with statistical models [8]. A global variance (GV) method was also proposed to solve this problem [9]. In this method, a statistical model was trained for the temporal variances of spectral parameters in each training sentence. During synthesis, the optimal spectral parameters were generated to maximize not only the likelihood of the acoustic HMMs but also the likelihood of the GV model. This method has been proved to be effective to alleviate the over-smoothing of the generated spectral envelopes and improve the naturalness of synthetic speech when mel-cepstrums are used as the spectral parameters [9]. Besides, the global variance has also been introduced into the generation error measurement of minimum generation error (MGE) training to decrease the computational cost at synthesis time [10].

On the other hand, comparing with the spectral parameters, which are used to present the spectral envelope at each frame with fewer parameter dimensions, the spectral envelope itself is more physically meaningful and more directly related with the subjective perception on the speech quality. Therefore, we extend the conventional GV method to explicitly model temporal variances of the spectral envelopes derived from spectral parameters in this paper. Line spectral pairs (LSPs), which are commonly used in the HMM-based parametric speech synthesis systems [5, 7], are adopted as the spectral parameters in our implementation. At training stage, a GV model is trained for the spectral envelopes derived from the LSPs of all training sentences. This model is then integrated into the parameter generation criterion to guide the generation of LSPs at synthesis time.

This paper is organized as follows. Section 2 reviews the conventional maximum likelihood parameter generation algorithm with GV models. Section 3 describes our proposed method in detail. Section 4 and 5 are the experiments and conclusions.

2. Parameter Generation with GV Model

2.1. Maximum likelihood parameter generation

In HMM-based parametric speech synthesis, the maximum likelihood parameter generation algorithm considering dynamic features is used to generated acoustic parameters for waveform reconstruction [2]. After the sentence HMM λ is determined by the results of front-end linguistic analysis on input text and the state sequence q is predicted using the trained state duration probabilities [1], the maximum likelihood parameter generation algorithm is applied to generate the speech feature vector sequence a = [a_1, a_2, ..., a_T]^T to maximize P(a|λ, q). In order to keep the smoothness of the generated speech parameters, the dynamic features including the velocity and acceleration coefficients are used as constraints, and the speech feature vector of each frame can be written as

\[ a_t = [c_t, \Delta c_t, \Delta^2 c_t]^T. \]  

Therefore, the complete feature sequence a can be presented as a linear transform of the static feature sequence c = [c_T, c_{T+1}, ..., c_T]^T as a = Wc, where W is determined by the velocity and acceleration calculation functions [2]. Thus, the parameter generation criterion can be rewritten as

\[ c^* = \arg \max_c P(Wc|\lambda, q). \]
By setting \( \partial P(W|q, \Lambda)/\partial c = 0 \), we obtain
\[
e^* = \left( W^T U^{-1} W \right)^{-1} W^T U^{-1} m, \tag{3}
\]
where \( m = [\mu_1^T, \ldots, \mu_T^T]^T \) and \( U = \text{diag}(\Sigma_1, \ldots, \Sigma_T) \) are the mean vector and covariance matrix of the sentence decided by the state sequence \( q \) \cite{2}.

2.2. Parameter generation considering GV model

The GV vector \( v(c) \) of the static feature trajectory in a sentence is calculated as \cite{9}
\[
v(c) = [v(c)_1, v(c)_2, \ldots, v(c)_D]^T, \tag{4}
\]
\[
v(c)_d = \frac{\sum_{t=1}^{T} (c_{t,d} - \bar{c}_d)^2}{T}, \tag{5}
\]
\[
\bar{c}_d = \frac{\sum_{t=1}^{T} c_{t,d}}{T}, \tag{6}
\]
where \( D \) is the number of dimensions of the static feature vector \( c_t \) and \( c_t = [c_{t,1}, c_{t,2}, \ldots, c_{t,D}]^T \). At training stage, a GV model \( \lambda_c \) is trained for the GV vectors of all training sentences using a single Gaussian distribution \( \mathcal{N}(\mu_c, \Sigma_c) \). In the parameter generation method considering GV \cite{9}, the optimal static feature sequence \( c \) is determined by maximizing not only the likelihood of the static and dynamic feature vectors but also the likelihood of the GV vector. Thus, the criterion is defined as
\[
L = \log \left\{ p(W|\lambda, q)^w \cdot p(v(c)|\lambda_0) \right\}, \tag{7}
\]
where \( w \) denotes the weight to balance the two likelihood functions, which is commonly set to the ratio of dimensions between vector \( v(c) \) and \( Wc \), i.e., \( 1/(3T) \). Because (7) cannot be optimally directed, an iterative updating method is applied to solve the optimal static feature sequence \( c^* \) that maximize the \( L \) function \cite{9}.

3. GV Modeling on Log Power Spectrum

3.1. GV of log power spectrum using LSPs

In the conventional parameter generation algorithm considering GV model, the GV vector is calculated for the spectral parameters of each training sentence. However, to alleviate the over-smoothness of the spectral envelope should be the ultimate aim of the spectral GV modeling. Therefore, we propose a method to calculate and model the GV of log power spectrum directly in this paper. The linear spectral pairs (LSPs) are adopted as the spectral parameters in our implementation. Assume the log power spectrum sequence for a sentence calculated from the static LSP feature sequence \( c^* \) is \( s = [s_1, s_2, \ldots, s_K]^T \), where \( s_t = [s_{t,1}, s_{t,2}, \ldots, s_{t,K}]^T \) and \( K \) is the number of sampling points within frequency range \([0, \pi]\). Based on the definition of LSPs \cite{11}, the log power spectrum \( s_{t,k} \) of frequency point \( \omega_k \) at frame \( t \) can be calculated as
\[
s_{t,k} = -10 \cdot \log_{10} \left[ \frac{1}{4} \left( P_k^2 + Q_k^2 \right) \right] + \frac{20}{\ln 10} c_{t,D}, \tag{8}
\]
where
\[
P_k^2 = 4 \cos^2 \frac{\omega_k}{2} \prod_{l=1}^{(D-1)/2} 4 \left( \cos \omega_k - \cos c_{t,2l-1} \right)^2, \tag{9}
\]
\[
Q_k^2 = 4 \sin^2 \frac{\omega_k}{2} \prod_{l=1}^{(D-1)/2} 4 \left( \cos \omega_k - \cos c_{t,2l} \right)^2, \tag{10}
\]
\[
\omega_k = \frac{k\pi}{K}, \tag{11}
\]
The GV of the log power spectrum sequence \( s \) is calculated by
\[
v(s) = [v(s)_1, v(s)_2, \ldots, v(s)_K]^T, \tag{12}
\]
\[
v(s)_k = \frac{\sum_{t=1}^{T} (s_{t,k} - \bar{s}_k)^2}{T}, \tag{13}
\]
\[
\bar{s}_k = \frac{\sum_{t=1}^{T} s_{t,k}}{T}. \tag{14}
\]

3.2. Parameter generation with GV on log power spectrum

A single Gaussian distribution \( \mathcal{N}(\mu_c, \Sigma_c) \) is trained to get the GV model \( \lambda_c \) for the GV vectors \( v(s) \) calculated on the log power spectrum of LSPs for all training sentences. This model is integrated into the maximum likelihood parameter generation algorithm to guide the prediction of LSP parameters. The criterion of parameter generation can be written as
\[
e^* = \arg \max_c L, \tag{15}
\]
\[
L = \log \left\{ p(W|\lambda, q)^w \cdot p(v(s)|\lambda_0) \right\}, \tag{16}
\]
where the weight \( w \) is set to \( K/(3T) \) to balance the difference of feature dimension numbers between the two models.

In order to determine \( e^* \), we iteratively update \( e \) using the steepest descent algorithm similar to \cite{9}
\[
e^{(i+1)} = e^{(i)} + \alpha \cdot \frac{\partial L}{\partial e} \bigg|_{e=e^{(i)}}, \tag{17}
\]
where \( i \) denotes the number of iterations and \( \alpha \) is the step size. Based on the definition of function \( L \) in (16), the gradient in (17) can be calculated as
\[
\frac{\partial L}{\partial e} = w \left( -W^T U^{-1} Wc + W^T U^{-1} m \right) + [v'_1, v'_2, \ldots, v'_T]^T, \tag{18}
\]
where
\[
v'_t = [v'_t(1), v'_t(2), \ldots, v'_t(D)]^T, \tag{19}
\]
\[
v'_t(d) = \left[ \frac{\partial v_t(s)}{\partial c_{t,d}}, \ldots, \frac{\partial v_t(s)}{\partial c_{t,d}} \right] \cdot \Sigma_s^{-1} \cdot (v(s) - \mu_s). \tag{20}
\]
According to the calculation of global variance in (12)-(14), we have
\[
\frac{\partial v_t(s)}{\partial c_{t,d}} = \frac{2}{T} \cdot (s_{t,k} - \bar{s}_k) \cdot \frac{\partial s_{t,k}}{\partial c_{t,d}} \cdot \left( 1 - \frac{1}{T} \right). \tag{21}
\]
Further, the \( \partial s_{t,k}/\partial c_{t,d} \) can be derived from (8) as
- When \( d = 1, 2, \ldots, D - 1 \),
  \[
  \frac{\partial s_{t,k}}{\partial c_{t,d}} = -\frac{20}{\ln 10} \frac{P_k^2}{P_k^2 + Q_k^2} \cdot \sin c_{t,d}, \tag{22}
  \]
where
\[
X_k^2 = \begin{cases} 
  P_k^2 & d \text{ is odd}, \\
  Q_k^2 & d \text{ is even}.
\end{cases} \tag{23}
\]
Therefore, to optimize the static spectral feature sequence in (7) was realized by optimizing each spectral parameter dimension separately in our implementation.

$$\frac{\partial s_{t,k}}{\partial c_{i,j}} = \frac{20}{\ln 10}$$

The static feature sequence is updated iteratively as (17) until the increasing of criterion function $L$ is smaller than a given threshold $\epsilon$.

4. Experiments

4.1. Experimental Conditions

A 5-hour Chinese speech database broadcasted by a professional female speaker was used in our experiments. It consisted of 4,706 sentences together with the segmental and prosodic labels. The waveforms were recorded in 16kHz/16bit format. Besides the logarithmized F0, 40-order LSPs and an extra gain dimension were derived from the spectral envelope by STRAIGHT [12] analysis at 5ms frame shift. For the spectral and F0 features, 5-state left-to-right ergodic HMM structure was adopted in our system. Single-mixture Gaussian distribution was used to model the state phone duration probabilities. Decision-tree-based model clustering [13] was applied in the context-dependent model training to avoid the data-sparsity problem. Here, the question set for tree splitting was designed considering the characteristics of Chinese.

Three systems were compared in our experiment:

- **Baseline**: The parameters were generated without GV model as introduced in Section 2.1.
- **GV LSP**: A GV model was trained for the LSP parameters and integrated into the parameter generation criterion as introduced in Section 2.2. The steepest decent algorithm [9] was applied to maximize the criterion function (7).
- **GV SPG**: The proposed GV method for the log power spectrum of LSPs was adopted as introduced in Section 3. The number of frequency sampling points was set to $K = 512$.

In the iterative updating procedure of parameter generation for system GV LSP and GV SPG, the step size $\alpha$ and the convergence threshold $\epsilon$ were set to 0.001 and 0.01 respectively. Examples for the convergence curve of the criterion function used in the parameter generation of these two systems are illustrated in Fig. 1 and 2. From these two figures, we see that the criterion functions converge and the log likelihoods of the GV models increase significantly by iteratively updating the static spectral feature sequence.

4.2. Subjective Evaluation

20 sentences out of the training set were selected and synthesized by the three systems respectively. These synthesized speech were evaluated by 5 Chinese-native listeners. The listeners were required to give a score from 1 (very unnatural) to 5 (very natural) for each synthesized speech. The mean opinion scores (MOS) with 95% confidence interval for the three systems are shown in Fig. 3. Comparing the performance of Baseline and GV LSP in Fig. 3, we find that when LSPs are used as the spectral features, the conventional GV method on spectral parameters can not improve the naturalness of synthetic speech significantly. This is different from the experimental results reported in [9] where the mel-cepstrums were used as the spectral features. We explain this difference as the intrinsic properties of LSPs. The LSPs have much stronger correlation among dimensions than mel-cepstrums. The shape of spectral envelope can not be affected explicitly by a single dimension of LSPs, but depends on the closeness of two adjacent LSP dimensions. Therefore, when the diagonal covariance matrix was used in our models, to enhance the variance of each LSP dimension separately may not help alleviate the over-smoothness of generated spectral envelopes significantly. On the other hand, the proposed GV method on the power spectrum of LSPs can improve the naturalness of synthetic speech significantly because it directly models the variance of spectral envelope. In the iterative updating of parameter generation as (17)-(24), the correlation between the
different dimensions of LSPs are also taken into account because the calculation of \(\partial s_{t,k}/\partial c_{t,d}\) depends on not only \(c_{t,d}\) but also the static spectral parameters of other dimensions at frame \(t\).

Fig. 4 shows examples for the spectrograms of synthetic speech using the three systems. From this figure, we see that the spectrograms of GV_LSP is slightly better than that of Baseline. While, system GV_SPG can generate speech with more enhanced formant structures than system Baseline and GV_LSP.

5. Conclusions

In this paper, we have proposed an GV modeling method for the log power spectra derived from LSPs in the HMM-based parametric speech synthesis. This GV model has been combined with the acoustic HMMs to guide the prediction of spectral parameters under maximum likelihood criterion. Our experimental results have shown that the conventional GV method can not achieve satisfactory performance when LSPs are used as the spectral parameters. Whereas, the proposed GV method on the log power spectrum can alleviate the over-smoothing effect of generated spectral structures and improve the naturalness of synthetic speech significantly. To compare the proposed method with the conventional GV modeling on mel-cepstrums and to evaluate the effectiveness of proposed method using the spectral parameters other than LSPs will be the tasks of our future work.

6. Acknowledgements

This work was partially supported by the China Postdoctoral Science Foundation and the National Nature Science Foundation of China (Grant No. 60905010).

7. References