Minimum Kullback–Leibler Divergence Parameter Generation for HMM-Based Speech Synthesis

Zhen-Hua Ling, *Member, IEEE*, and Li-Rong Dai

***Abstract*—This paper presents a parameter generation method for hidden Markov model (HMM)-based statistical parametric speech synthesis that uses a similarity measure for probability distributions. In contrast to conventional maximum output probability parameter generation (MOPPG), the method we propose derives a parameter generation criterion from the distribution characteristics of the generated acoustic features. Kullback–Leibler (KL) divergence between the sentence HMM used for parameter generation and the HMM estimated from the generated features is calculated by upper bound approximation. During parameter generation, this KL divergence is minimized either by optimizing the generated acoustic parameters directly or by applying a linear transform to the MOPPG outputs. Our experiments show both these approaches are effective for alleviating over-smoothing in the generated spectral features and for improving the naturalness of synthetic speech. Compared with the direct optimization approach, which is susceptible to over-fitting, the feature transform approach gives better performance. In order to reduce the computational complexity of transform estimation, an offline training method is further developed to estimate a global transform under the minimum KL divergence criterion for the training set. Experimental results show that this global transform is as effective as the transform estimated for each sentence at synthesis stage.**

***Index Terms*—Hidden Markov model (HMM), Kullback–Leibler (KL) divergence, parameter generation, speech synthesis.**

I. Introduction

HIDDEN Markov model (HMM)-based statistical parametric speech synthesis has become a mainstream speech synthesis method in recent years [1], [2]. In this method, the spectrum, F0 and segment durations are modeled simultaneously within a unified HMM framework [1]. At synthesis time, these features are predicted from the sentence HMM, which is decided by the results of text analysis. Maximum output probability parameter generation (MOPPG)[1] incorporating dynamic features [3] is currently the most popular parameter generation method. 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 the synthetic speech may be degraded by the parametric synthesizer itself, inaccuracy of acoustic modelling, and the over-smoothing effect of parameter generation [6].

Among these three factors that degrade the quality of synthetic speech, this paper focuses on the over-smoothing problem, which is closely related to the parameter generation method used at synthesis time. In the conventional MOPPG algorithm, the acoustic parameters are predicted so as to maximize their output probabilities from the sentence HMM given the text analysis results of the input sentence [3]. Although the MOPPG outputs evolve from piecewise-constant mean sequences into smooth trajectories by incorporating the constraints between static and dynamic features, they still tend to distribute near the means of the HMM state probability density functions (pdfs). These means are estimated by averaging observations with similar context descriptions in the training set. This averaging process improves the robustness of parameter generation. However, the detailed characteristics of the speech parameters are lost, especially for the spectral parameters. The generated spectral envelopes are over-smoothed, which leads to a muffled voice quality in the synthetic speech.

Many methods have been proposed to overcome this over-smoothing problem, such as post-filtering after parameter generation [5], [7], using real speech parameters or segments to generate the speech waveform [8], [9], or sampling trajectories from the predictive distribution [10], [11], and so on. Some of these methods are incorporated into the parameter generation criterion directly, e.g., the global variance (GV) method [12]. In this method, a global statistical model is trained for the variances of the spectral parameters within each training sentence. During synthesis, the spectral parameters are generated by maximizing the weighted product of the probability functions of the sentence HMM and the GV model. This method has been shown to be effective at alleviating the over-smoothing of generated spectral envelopes by increasing their variance, and thus improving the naturalness of synthetic speech significantly [12]. The GV method has been extended from spectral parameters to the log power spectrum [13] and has been integrated into the generation error measurement of minimum generation error (MGE) training to decrease computational complexity at synthesis time

Manuscript received July 29, 2011; revised October 27, 2011, December 18, 2011; accepted December 28, 2011. Date of publication January 02, 2012; date of current version March 14, 2012. This work was supported by the National Nature Science Foundation of China under Grant 60905010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chung-Hsien Wu.

The authors are with iFLYTEK Speech Lab, University of Science and Technology of China, Hefei 230027, China (e-mail: zhling@ustc.edu; lrdai@ustc.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2182511

[1]This algorithm is also named maximum-likelihood parameter generation (MLPG) in the literature. In order to clarify the technical difference between "likelihood" (which interprets the probability distribution as a function of the model parameters given a fixed outcome) and "probability" (which interprets the probability distribution as a function of the outcome given fixed model parameters), the term "output probability" is adopted in this paper to replace "likelihood" for describing the parameter generation criterion.

[14]. A parameter generation method based on segment-wise representation has also been proposed [15] to cope with the over-smoothing problem. In this method, the output probability function in the parameter generation criterion is calculated using the means of the generated state segments instead of the parameters of individual frames, which allows the parameters to move far away from the distribution center of each state, thus alleviating the over-smoothing effect. This segment-wise parameter generation method can achieve similar mean opinion score (MOS) results in terms of naturalness as the GV approach [15].

We see that both the GV method and the segment-wise parameter generation method guide the generation of acoustic features by examining the similarity between the distribution parameters of acoustic models trained on natural speech and the same distribution parameters derived from the generated acoustic features. The distribution parameter is the sentence-level variance in the GV method [12] and a state-level mean in the segment-wise method [15]. The effectiveness of these two methods suggests that such distribution similarity measures can help the parameter generation criterion overcome the over-smoothing problem because it relaxes the constraint inherent in the MOPPG algorithm that the acoustic features for each frame should be close to the distribution center.

However, both these methods have shortcomings in how they incorporate a distribution similarity measure into the parameter generation criterion. First, they both consider only one limited parameter of the generated acoustic feature distribution, and lack accurate calculation of the divergence between the distributions of natural and generated features. Second, the distribution similarity measurement used in both methods must be combined with other components, such as the probability function of the sentence HMM [12] or the norm constraint of the static features [15], in order to construct the final parameter generation criterion, which introduces the issue of weight tuning for how best to combine these components.

To address these shortcomings, an explicit distribution similarity measure-based parameter generation algorithm is proposed in this paper. We expect the sentence HMM estimated from the generated acoustic features to be as close as possible to the HMM used for parameter generation. Kullback–Leibler (KL) divergence [16] is a popular distribution divergence measure and is adopted in this paper to calculate the distance between these two HMMs. This criterion is integrated into the parameter generation algorithm in two separate approaches. In the first approach, it is applied to optimize the generated acoustic parameters directly. In the second approach, a linear transform is estimated under this criterion and applied to the MOPPG output for each sentence. An offline transform estimation method is further developed to reduce the computational complexity of the linear transform method at synthesis time.

This paper is organized as follows. The framework of HMM-based parametric speech synthesis and the conventional MOPPG algorithm is briefly reviewed in Section II. Section III describes our proposed methods in detail. Experimental results are introduced in Section IV and Section V concludes this paper.

II. HMM-Based Parametric Speech Synthesis

A. Model Training

Fig. 1 shows a diagram of standard HMM-based speech synthesis systems. It consists of a training stage and the synthesis stage. During training, the F0 and spectral parameters of D dimensions are extracted from the waveforms contained in the training set. Then a set of context-dependent HMMs Λ are estimated to maximize the likelihood function $P(\boldsymbol{o} \mid \Lambda)$ for these features. Here $\boldsymbol{o} = [\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \ldots, \boldsymbol{o}_T^\top]^\top$ is the observation feature sequence, $(\cdot)^\top$ denotes the matrix transpose and T is the length of the sequence. The observation feature vector $\boldsymbol{o}_t \in \mathcal{R}^{3D}$ for the tth frame typically consists of static acoustic parameters $\boldsymbol{c}_t \in \mathcal{R}^D$ and their delta and acceleration components as

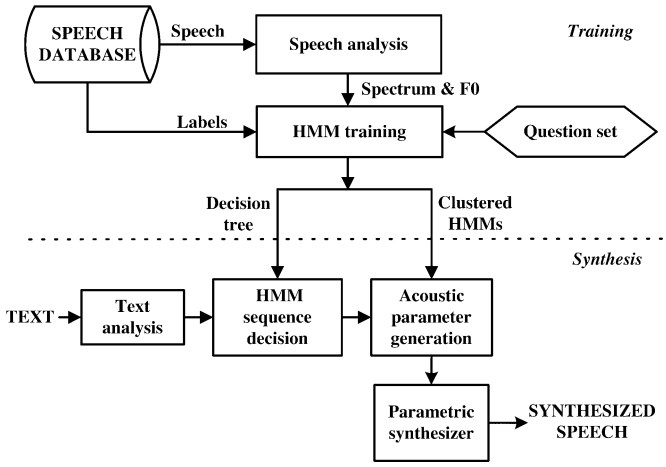

Fig. 1. Diagram of a typical HMM-based parametric speech synthesis system.

$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \triangle \boldsymbol{c}_t^\top, \triangle^2 \boldsymbol{c}_t^\top\right]^\top \tag{1}$$

where

$$\Delta \boldsymbol{c}_t = 0.5\boldsymbol{c}_{t+1} - 0.5\boldsymbol{c}_{t-1} \quad \forall t \in [2, T-1] \tag{2}$$

$$\Delta \boldsymbol{c}_1 = \Delta \boldsymbol{c}_2, \Delta \boldsymbol{c}_T = \Delta \boldsymbol{c}_{T-1} \tag{3}$$

and

$$\Delta^2 \boldsymbol{c}_t = \boldsymbol{c}_{t+1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t-1} \quad \forall t \in [2, T-1] \tag{4}$$

$$\Delta^2 \boldsymbol{c}_1 = \Delta^2 \boldsymbol{c}_2, \quad \Delta^2 \boldsymbol{c}_T = \Delta^2 \boldsymbol{c}_{T-1}. \tag{5}$$

Therefore, the complete feature sequence $\boldsymbol{o}$ can be considered to be a linear transform of the static feature sequence $\boldsymbol{c} = [\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \ldots, \boldsymbol{c}_T^\top]^\top$ such that

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c} \tag{6}$$

where $\boldsymbol{W} \in \mathcal{R}^{3TD \times TD}$ is determined by the delta and acceleration calculation functions in (2)–(5) [3]. Because F0 is only defined for voiced speech frames, a multi-space probability distribution (MSD) [17] is applied to incorporate a distribution for F0 into the probabilistic framework of the HMM. A decision-tree-based model clustering technique is adopted to deal with the data-sparsity problem and to estimate the parameters of models whose context description is missing in the training set. The tree construction is guided by the minimum description length (MDL) criterion [18] after initial training of context-dependent HMMs. Next, a state alignment is conducted using the trained HMMs to train context-dependent state duration probabilities [1] for state duration prediction. A single-mixture Gaussian distribution is used to model the duration probability for each state. A decision-tree-based model clustering technique is similarly applied to these duration distributions.

B. Maximum Output Probability Parameter Generation

During synthesis, the MOPPG algorithm is used to generate acoustic parameters for waveform reconstruction [3]. The result of front-end linguistic analysis on the input text is used to determine the sentence HMM λ. The state sequence $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ is predicted using the trained state duration probabilities [1]. Then, the speech feature sequence is generated by maximizing $P(\boldsymbol{o} \mid \lambda, \boldsymbol{q})$. Considering the constraints between static and dynamic features as in (6), the parameter generation criterion can be rewritten as

$$\boldsymbol{c}^* = \arg\max_{\boldsymbol{c}} P(\boldsymbol{W}\boldsymbol{c} \mid \lambda, \boldsymbol{q}) \tag{7}$$

where $\boldsymbol{c}^*$ is the output of MOPPG. By setting

$$\frac{\partial P(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{q}, \lambda)}{\partial \boldsymbol{c}} = \boldsymbol{0} \tag{8}$$

we obtain

$$\boldsymbol{c}^* = \left(\boldsymbol{W}^\top \boldsymbol{U}_{\boldsymbol{q}}^{-1} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{U}_{\boldsymbol{q}}^{-1} \boldsymbol{m}_{\boldsymbol{q}} \tag{9}$$

where $\boldsymbol{m}_{\boldsymbol{q}} = [\boldsymbol{\mu}_{q_1}^\top, \ldots, \boldsymbol{\mu}_{q_T}^\top]^\top$ and $\boldsymbol{U}_{\boldsymbol{q}} = \mathrm{diag}(\boldsymbol{\Sigma}_{q_1}, \ldots, \boldsymbol{\Sigma}_{q_T})$ are the mean vector and covariance matrix of the sentence as decided by the state sequence $\boldsymbol{q}$ [3].

III. Minimum Kullback–Leibler Divergence Parameter Generation

The maximum output probability criterion used in the MOPPG algorithm restricts the generated acoustic features to distributing close to the means of the HMM states, which adds to the over-smoothing problem and degrades the naturalness of the synthetic speech. To overcome this problem, we propose a distribution similarity measure-based parameter generation method. In this method, the sentence HMM derived from the trained HMM using the input text is defined as the *target model* for parameter generation. Another HMM is estimated from the generated acoustic features of each sentence, which is termed the *generated model*. Rather than examine the generated acoustic features frame by frame as in the MOPPG algorithm, we expect the stochastic characteristics of the generated acoustic features to be as similar as possible to those of natural recordings given specific context information. This means the distance between the *target model* which describes natural features and the *generated model* which represents the generated features should be minimized. Kullback–Leibler (KL) divergence [16] is adopted here to calculated the divergence between these two HMMs. The details of this proposed method will be introduced next.

A. Estimation of the Generated HMM

For each sentence to be synthesized, $\lambda = (\{a_{ij}\}, \{b_i\}, \{\pi_i\})$ and $\bar{\lambda} = (\{\bar{a}_{ij}\}, \{\bar{b}_i\}, \{\bar{\pi}_i\})$ represent the *target HMM* and the *generated HMM*, respectively, where π_i and $\bar{\pi}_i$ are initial state distributions of λ and $\bar{\lambda}$; a_{ij} and $\bar{a}_{ij}$ denote state transition probability from state i to state j in λ and $\bar{\lambda}$. $b_i = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{m}_i, \boldsymbol{\Sigma}_i)$ and $\bar{b}_i = \mathcal{N}(\boldsymbol{o}_t; \bar{\boldsymbol{m}}_i, \bar{\boldsymbol{\Sigma}}_i)$ are the pdf of state i for models λ and $\bar{\lambda}$, where $\mathcal{N}(;\boldsymbol{m}, \boldsymbol{\Sigma})$ is a Gaussian distribution with a mean vector $\boldsymbol{m}$ and a covariance matrix $\boldsymbol{\Sigma}$; the mean vector and covariance matrix consist of static, delta and acceleration components as[2]

$$\boldsymbol{m}_i = [m_{i1}, m_{i2}, m_{i3}]^\top, \boldsymbol{\Sigma}_i = \mathrm{diag}\left(\sigma_{i1}^2, \sigma_{i2}^2, \sigma_{i3}^2\right) \tag{10}$$

$$\bar{\boldsymbol{m}}_i = [\bar{m}_{i1}, \bar{m}_{i2}, \bar{m}_{i3}]^\top, \bar{\boldsymbol{\Sigma}}_i = \mathrm{diag}\left(\bar{\sigma}_{i1}^2, \bar{\sigma}_{i2}^2, \bar{\sigma}_{i3}^2\right). \tag{11}$$

The parameters of the target HMM λ are determined by the trained HMM set and the context information derived from the input text. Meanwhile, the parameters of the generated HMM $\bar{\lambda}$ need to be estimated from the generated acoustic features $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$. Because of the small amount of training data available within a single sentence, the state transition probabilities a_{ij} are copied over to use as the values of $\bar{a}_{ij}$, and the distribution parameters of $\bar{b}_i$ are estimated under the *maximum a posteriori* (MAP) [19] criterion to improve the robustness of parameter estimation. Using MAP estimation, we have [19]

$$\bar{m}_{id} = \alpha \frac{\sum_{t=1}^{T} \gamma_i(t) \boldsymbol{o}_{t,d}}{\sum_{t=1}^{T} \gamma_i(t)} + (1-\alpha) m_{id}^p \tag{12}$$

$$\bar{\sigma}_{id}^2 = \alpha \frac{\sum_{t=1}^{T} \gamma_i(t) \boldsymbol{o}_{t,d}^2}{\sum_{t=1}^{T} \gamma_i(t)} + (1-\alpha)\left(\sigma_{id}^{p\,2} + m_{id}^{p\,2}\right) - \bar{m}_{id}^2 \tag{13}$$

where $d \in \{1,2,3\}$; $\boldsymbol{o}_{t,d}$ means the generated acoustic feature at time t and for dimension d, $\gamma_i(t)$ is the occupancy probability of state i at time t, m_{id}^p, and $\sigma_{id}^{p\,2}$ denote the prior distribution parameters of state i and dimension d

$$\alpha = \frac{t_{i+1} - t_i}{t_{i+1} - t_i + \beta} \tag{14}$$

is the adaptation coefficient controlling the balance between the prior distribution and the estimation given by the observed training data; β is set manually.

Because the state pdf $\bar{b}_i$ is context-dependent and rich context features including detailed phonetic and prosodic descriptions [2] are adopted here for context-dependent model training, the chances of finding identical state pdfs within a sentence are very low. For the sake of simplifying the derivation, we assume that each state appears only once in a sentence and use the hard segmentation results given by state duration prediction to replace the calculation of $\gamma_i(t)$. Therefore, (12) and (13) can be rewritten as

$$\bar{m}_{id} = \alpha \frac{\boldsymbol{p}_i^\top \boldsymbol{s}_{id}}{t_{i+1} - t_i} + (1-\alpha) m_{id}^p, \tag{15}$$

$$\bar{\sigma}_{id}^2 = \alpha \frac{\boldsymbol{s}_{id}^\top \boldsymbol{s}_{id}}{t_{i+1} - t_i} + (1-\alpha)\left(\sigma_{id}^{p\,2} + m_{id}^{p\,2}\right) - \bar{m}_{id}^2 \tag{16}$$

where $t_i \in \{1, 2, \ldots\}$ denotes the beginning frame index of state i which is determined by the predicted state sequence $\boldsymbol{q}$; $\boldsymbol{p}_i$ is an all-one vector of length $t_{i+1} - t_i$; $\boldsymbol{s}_{id}$ is the acoustic feature

[2]Because the covariance matrices of state pdfs are commonly set to be diagonal in practical implementation of HMM-based speech synthesis systems, the parameter generation for each dimension is independent. Therefore, the number of static feature dimensions D is set to 1 in this section to simplify the notation.

vector of state i and dimension d. Here, $\boldsymbol{s}_{id}$ is extracted from the feature sequence $\boldsymbol{o}$ as

$$\boldsymbol{s}_{id} = \boldsymbol{K}_{id}\boldsymbol{o} = \boldsymbol{K}_{id}\boldsymbol{W}\boldsymbol{c} \tag{17}$$

where $\boldsymbol{K}_{id} \in \mathcal{R}^{(t_{i+1}-t_i)\times 3T}$ is the matrix for feature sequence segmentation and its nth row is defined as

$$\{\boldsymbol{K}_{id}\}_{n,k} = \begin{cases} 1 & k = (n + t_i - 2) * 3 + d, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

In (12) and (13), m_{id}^p and ${\sigma_{id}^p}^2$ are introduced to represent the prior distribution of generated acoustic features for a given context input. It is inappropriate to derive these prior distributions from the context-dependent HMM set Λ trained in Section II-A because the distribution characteristics of the generated acoustic features may be different from those of the natural features. Furthermore, these prior distribution parameters should be estimated before the acoustic features for the input sentence are generated and observed. Therefore, the acoustic features generated by MOPPG for the text of all sentences in the training set are used here to estimate these prior distribution parameters. First, a prior HMM set Λ^p is trained on these features under the maximum-likelihood criterion and the decision trees for model clustering are set to be the same as the ones of HMM set Λ trained in Section II-A. Then, m_{id}^p and ${\sigma_{id}^p}^2$ can be derived from Λ^p using the context information of the sentence for synthesis.

B. KL Divergence Between Target and Generated Models

Once the target HMM λ and the generated HMM $\bar{\lambda}$ are given, the KL divergence in symmetrical form between them is defined as

$$D(\lambda \parallel \bar{\lambda}) = D_{\mathrm{KL}}(\lambda \parallel \bar{\lambda}) + D_{\mathrm{KL}}(\bar{\lambda} \parallel \lambda) \tag{19}$$

where

$$D_{\mathrm{KL}}(\lambda \parallel \bar{\lambda}) = \int_{\mathcal{R}^{3DT}} P(\boldsymbol{x} \mid \lambda) \ln \frac{P(\boldsymbol{x} \mid \lambda)}{P(\boldsymbol{x} \mid \bar{\lambda})} d\boldsymbol{x}. \tag{20}$$

is the directional KL divergence using λ as the reference model. However, there is no closed form solution for calculating the KL divergence between two HMMs. Thus, the upper bound of the KL divergence between two left-to-right HMMs [20] is adopted as an approximation in our method as follows:

$$D(\lambda \parallel \bar{\lambda}) \leqslant \sum_{i=1}^{S} \left[\frac{D_{\mathrm{KL}}(b_i \parallel \bar{b}_i)}{1 - a_{ii}} + \frac{D_{\mathrm{KL}}(\bar{b}_i \parallel b_i)}{1 - \bar{a}_{ii}} + \frac{(a_{ii} - \bar{a}_{ii}) \log(a_{ii}/\bar{a}_{ii})}{(1 - a_{ii})(1 - \bar{a}_{ii})} \right] \doteq D_{\mathrm{crt}} \tag{21}$$

where S is the number of states in the sentence HMM and

$$D_{\mathrm{KL}}(b_i \parallel \bar{b}_i) = \sum_{d=1}^{3} \frac{1}{2} \left[\ln \frac{\bar{\sigma}_{id}^2}{\sigma_{id}^2} - 1 + \frac{\sigma_{id}^2 + (\bar{m}_{id} - m_{id})^2}{\bar{\sigma}_{id}^2} \right] \tag{22}$$

is the KL divergence between two Gaussian distributions for the ith HMM state. Substituting (22) into (21) and considering $\bar{a}_{ii} = a_{ii}$, D_{crt} can be rewritten as

$$D_{\mathrm{crt}} = \sum_{i=1}^{S} \sum_{d=1}^{3} \frac{1}{2(1 - a_{ii})} \left[-2 + \frac{\sigma_{id}^2 + (\bar{m}_{id} - m_{id})^2}{\bar{\sigma}_{id}^2} + \frac{\bar{\sigma}_{id}^2 + (m_{id} - \bar{m}_{id})^2}{\sigma_{id}^2} \right]. \tag{23}$$

During parameter generation, the minimum KL divergence criterion is implemented by minimizing this upper bound D_{crt}.

Note our method adopts the symmetrical form of KL divergence in (19) instead of the directional form in (20), where the target HMM λ is used as the reference model. In addition to the fact that the directional KL divergence is asymmetrical and thus not a distance metric, another important reason for this is that the directional KL divergence between two Gaussian distributions in (22) is unequally influenced by deviations of $\bar{\sigma}_{id}^2$ in different directions. For example, assuming $\bar{m}_{id} = m_{id}$ in (22), the condition that $\bar{\sigma}_{id}^2 = 10\sigma_{id}^2$ can lead to much smaller $D_{\mathrm{KL}}(b_i \parallel \bar{b}_i)$ than the condition that $\bar{\sigma}_{id}^2 = 0.1\sigma_{id}^2$. This means the criterion using directional KL divergence could lead to overly large variances in the generated acoustic features. Therefore, the symmetrical form of KL divergence is adopted to avoid this problem in our proposed method.

C. Minimum KL Divergence Parameter Generation

The minimum KL divergence criterion is adopted to optimize the generated acoustic parameters directly so that

$$\boldsymbol{c}^* = \arg\min_{\boldsymbol{c}} D_{\mathrm{crt}}. \tag{24}$$

In order to determine $\boldsymbol{c}^*$, we iteratively update $\boldsymbol{c}$ using a steepest decent algorithm similar to that used in [12]

$$\boldsymbol{c}^{(k+1)} = \boldsymbol{c}^{(k)} - \epsilon \cdot \left. \frac{\partial D_{\mathrm{crt}}}{\partial \boldsymbol{c}} \right|_{\boldsymbol{c}=\boldsymbol{c}^{(k)}} \tag{25}$$

where k denotes the iteration number and ϵ is the step size.

Substituting (14)–(17) into (23), we obtain

$$\frac{\partial D_{\mathrm{crt}}}{\partial \boldsymbol{c}} = \sum_{i=1}^{S} \sum_{d=1}^{3} \frac{1}{2(1 - a_{ii})} \times \left[2(\bar{m}_{id} - m_{id}) \left(\frac{1}{\sigma_{id}^2} + \frac{1}{\bar{\sigma}_{id}^2} \right) \cdot \frac{\partial \bar{m}_{id}}{\partial \boldsymbol{c}} + \left(\frac{1}{\sigma_{id}^2} - \frac{\sigma_{id}^2 - (\bar{m}_{id} - m_{id})^2}{\bar{\sigma}_{id}^4} \right) \frac{\partial \bar{\sigma}_{id}^2}{\partial \boldsymbol{c}} \right] \tag{26}$$

where

$$\frac{\partial \bar{m}_{id}}{\partial \boldsymbol{c}} = \frac{\alpha}{t_{i+1} - t_i} \boldsymbol{W}^\top \boldsymbol{K}_{id}^\top \boldsymbol{p}_i, \tag{27}$$

$$\frac{\partial \bar{\sigma}_{id}^2}{\partial \boldsymbol{c}} = \frac{\alpha}{t_{i+1}-t_i}\boldsymbol{W}^\top \boldsymbol{K}_{id}^\top \left(\boldsymbol{I} - \frac{\boldsymbol{p}_i \boldsymbol{p}_i^\top}{t_{i+1}-t_i}\right)\boldsymbol{K}_{id}\boldsymbol{W}\boldsymbol{c} - 2\bar{m}_{id}\frac{\partial \bar{m}_{id}}{\partial \boldsymbol{c}} \tag{28}$$

$\boldsymbol{I} \in R^{(t_{i+1}-t_i)\times(t_{i+1}-t_i)}$ is an identity matrix. For the first iteration, $\boldsymbol{c}^{(0)}$ is initialized using the output of the MOPPG method. The iterative updating stops either when D_{crt} decreases by an amount smaller than a given threshold or the number of iterations reaches a preset maximum value.

D. Minimum KL Divergence Feature Transform

When estimating each state pdf $\bar{b}_i$ of the generated HMM $\bar{\lambda}$ using (15)–(16), the number of frames within each state segment $\boldsymbol{s}_{id}$ is very limited. Although the MAP method has been adopted to deal with the data-sparsity problem, $\bar{b}_i$ still tends to have much smaller variance than b_i of the target HMM λ, which is estimated after decision-tree-based model clustering as introduced in Section II-A. Thus, the variance of the generated parameters could become overly large when optimized directly under the minimum KL divergence criterion in (24). This potential risk of over-fitting could introduce unanticipated noise and discontinuities into the generated feature trajectories. Some experimental results related to this issue will be shown in the next section.

Meanwhile, an alternative approach is proposed here to avoid this problem. Here, a linear transform is estimated for the MOPPG outputs for each sentence under the minimum KL divergence criterion. The flowchart is shown in Fig. 2. Because the same linear transform is applied to all frames within a sentence, this method is able to compensate the over-smoothing effect of MOPPG and preserve the temporal continuity of transformed feature trajectories at the same time. The transform matrix is set to be diagonal with an extra bias vector. We can use a scalar to represent each frame's static acoustic feature for the sake of simplifying the notation. For the tth frame, the static acoustic feature generated by MOPPG $c_t^{(\mathrm{MP})}$ is transformed such that

$$c_t = l \cdot c_t^{(\mathrm{MP})} + h \tag{29}$$

where l and h denote the diagonal transform matrix and bias vector of a single dimension. The distribution parameters $\bar{m}_{id}$ and $\bar{\sigma}_{id}^2$ estimated from MOPPG outputs using (15)–(16) are simultaneously transformed to $\hat{m}_{id}$ and $\hat{\sigma}_{id}^2$ such that

$$\hat{m}_{id} = l \cdot \bar{m}_{id} + h \cdot \delta_d \tag{30}$$

$$\hat{\sigma}_{id}^2 = l^2 \cdot \bar{\sigma}_{id}^2 \tag{31}$$

where

$$\delta_d = \begin{cases} 1, & d = 1, \\ 0, & d = 2,3 \end{cases} \tag{32}$$

is decided by the delta and acceleration calculation functions in (2)–(5). The optimal linear transform parameters are determined under the minimum KL divergence criterion as

$$\{l^*, h^*\} = \arg\min_{\{l,h\}} D_{\mathrm{crt}} \tag{33}$$

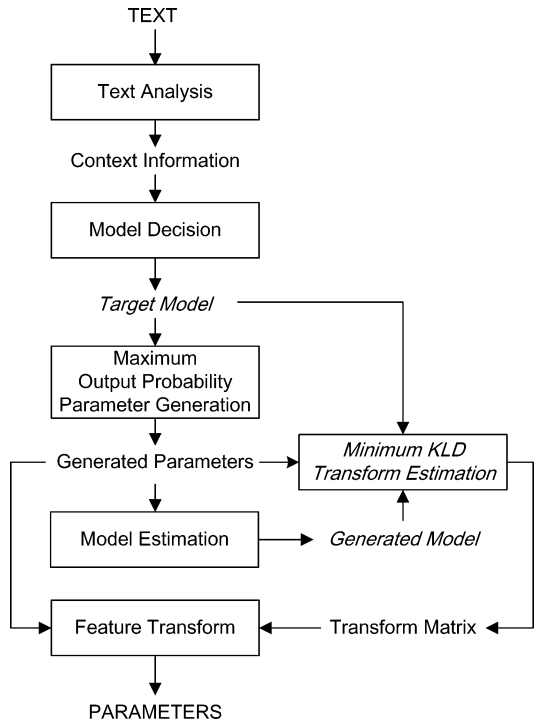

Fig. 2. Flowchart for the minimum KL divergence feature transform method.

where $\hat{m}_{id}$ and $\hat{\sigma}_{id}^2$ are used to replace $\bar{m}_{id}$ and $\bar{\sigma}_{id}^2$ in (23) to calculate D_{crt}. Equation (33) can also be solved using the steepest descent algorithm as

$$\tau^{(k+1)} = \tau^{(k)} - \epsilon \cdot \left.\frac{\partial D_{\mathrm{crt}}}{\partial \tau}\right|_{\tau=\tau^{(k)}} \tag{34}$$

where τ stands for either l or h. Substituting (30)–(31) into (23), we have

$$\frac{\partial D_{\mathrm{crt}}}{\partial \tau} = \sum_{i=1}^{S}\sum_{d=1}^{3}\frac{1}{2(1-a_{ii})} \times \left[2(\hat{m}_{id}-m_{id})\left(\frac{1}{\sigma_{id}^2}+\frac{1}{\hat{\sigma}_{id}^2}\right)\cdot\frac{\partial \hat{m}_{id}}{\partial \tau} + \left(\frac{1}{\sigma_{id}^2} - \frac{\sigma_{id}^2-(\hat{m}_{id}-m_{id})^2}{\hat{\sigma}_{id}^4}\right)\frac{\partial \hat{\sigma}_{id}^2}{\partial \tau}\right] \tag{35}$$

where

$$\frac{\partial \hat{m}_{id}}{\partial l} = \bar{m}_{id}, \quad \frac{\partial \hat{m}_{id}}{\partial h} = \delta_d \tag{36}$$

$$\frac{\partial \hat{\sigma}_{id}^2}{\partial l} = 2l \cdot \bar{\sigma}_{id}^2, \quad \frac{\partial \hat{\sigma}_{id}^2}{\partial h} = 0. \tag{37}$$

$l^{(0)} = 1$ and $h^{(0)} = 0$ are used for the first iteration and the iterative updating stops when the descent of D_{crt} is smaller than a given threshold value or the number of iterations reaches a preset maximum value.

E. Global Transform Estimation Using Training Database

In the minimum KL divergence linear transform method described in Section III-D, it is necessary to estimate the transform

parameters for each sentence at the synthesis stage, but unfortunately the steepest descent updating in (34) incurs high computational cost. Therefore, we develop a method to estimate the linear transform globally at the training stage as

$$\{l^*, h^*\} = \arg\min_{\{l.h\}} \sum_k D_{\text{crt}}^{(k)} \tag{38}$$

where k denotes the kth sentence in the training database and $D_{\text{crt}}^{(k)}$ is calculated in the same way as (23). Equation (38) is solved using a steepest descent algorithm similar to (34). At synthesis time, this linear transform is applied to the MOPPG outputs for each sentence to obtain the final generated acoustic parameters. This method can be considered to represent a kind of post-filtering of the MOPPG outputs, with the post-filter being in the form of a linear transform that is estimated under the minimum KL divergence criterion at the training stage.

IV. Experiments

A. Experimental Conditions

A 1-hour Chinese speech database produced by a professional female speaker was used in our experiments. It consisted of 1050 sentences together with the segmental and prosodic labels. 1000 sentences were selected for training and the remaining 50 sentences were used as a test set. The waveforms were recorded in 16-kHz/16-bit format. In addition to logarithmized F0, 41-order mel-cepstrum (including 0th coefficient) were derived from the spectral envelope by STRAIGHT [21] analysis at 5-ms frame shift. For the spectral and F0 features, a 5-state left-to-right HMM structure with no skips (not hidden semi-Markov model) was adopted to train context-dependent phone models, whose covariance matrices were set to be diagonal. Single-mixture Gaussian distributions were used to model the state duration probabilities. Decision-tree-based distribution clustering [18] was applied in the context-dependent model training to avoid the data-sparsity problem. Here, the question set for tree splitting was designed specifically to match the characteristics of Chinese. A modified version of the HTS toolkit [22] based on HTS-1.1b was used to train the system.

Five parameter generation methods were compared in our experiments. A description of each of these methods is listed in Table I. Here, the *GV* method employed a single-Gaussian distribution with a diagonal covariance matrix for the sentence-level variances of the static features. *KLD_GEN*, *KLD_FT*, and *GLB_FT* followed the methods proposed in Sections III-C, D, and E, respectively, where β in the MAP estimation of (14) was empirically set to 50. The stopping threshold of a minimum decrease of D_{crt} was set to 1 and the maximum iteration number was set to 100 for the iterative updating of (25) and (34). In our experiments, we focused on the effects of these parameter generation methods on spectral features. For each sentence in the test set, five stimuli were synthesized using these five methods to generate spectral parameters, but using the same MOPPG method to generate F0 contours.[3] Fig. 3 shows the convergence of the *KLD_GEN* method for one test sentence. We found that convergence is almost always achieved in (25) and (34) within 100 iterations. In terms of computational complexity, the *KLD_GEN* and *KLD_FT* methods are comparable to the *GV* method and much higher than *MOPPG* because iterative optimization for each dimension of the static acoustic features is required. The computational cost of the *GLB_FT* method at synthesis stage is close to that of the *MOPPG* method because its transform matrix is estimated globally before synthesis.

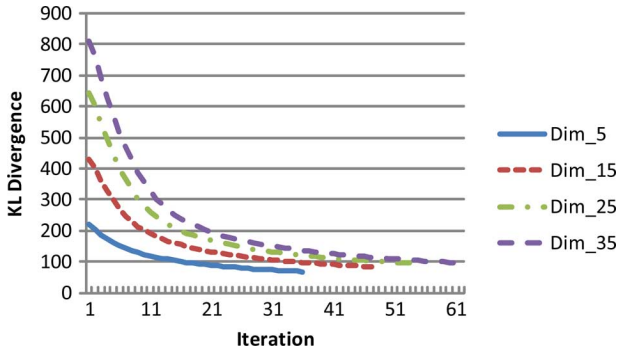

Fig. 3. Convergence of KL divergence in the iterative updating of *KLD_GEN* method for one test sentence. The KL divergences of the 5th, 15th, 25th, and 35th mel-cepstrum dimensions are shown as examples.

TABLE I
Description of Parameter Generation Methods Used in Our Experiments

Method	Description
MOPPG	Conventional maximum output probability parameter generation algorithm
GV	Maximum output probability parameter generation with global variance
KLD_GEN	Minimum KL divergence parameter generation
KLD_FT	Minimum KL divergence feature transform
GLB_FT	Minimum KL divergence feature transform using global transform estimated on training set

B. Objective Evaluation

The difference between the target HMM and the generated HMM using *MOPPG* was studied. An example is shown in Fig. 4, where generated spectral parameters (including the static, delta, and acceleration components of the 0th mel-cepstrum coefficient) are illustrated together with the corresponding mean and standard deviation sequences of the target and generated HMMs. The generated HMM was estimated according to (15)–(16). Comparing the two columns in Fig. 4, we see that the mean sequences of the target HMM and the generated HMM are quite similar. However, the state pdfs of the generated HMM have much smaller variances than the target HMM, especially for the delta and acceleration components.

In order to study such distribution differences mathematically, the KL divergences per state between the target HMM

[3]Some examples of the synthetic speech using the five methods can be found at http://staff.ustc.edu.cn/~zhling/MKLDParaGen/demo.html.

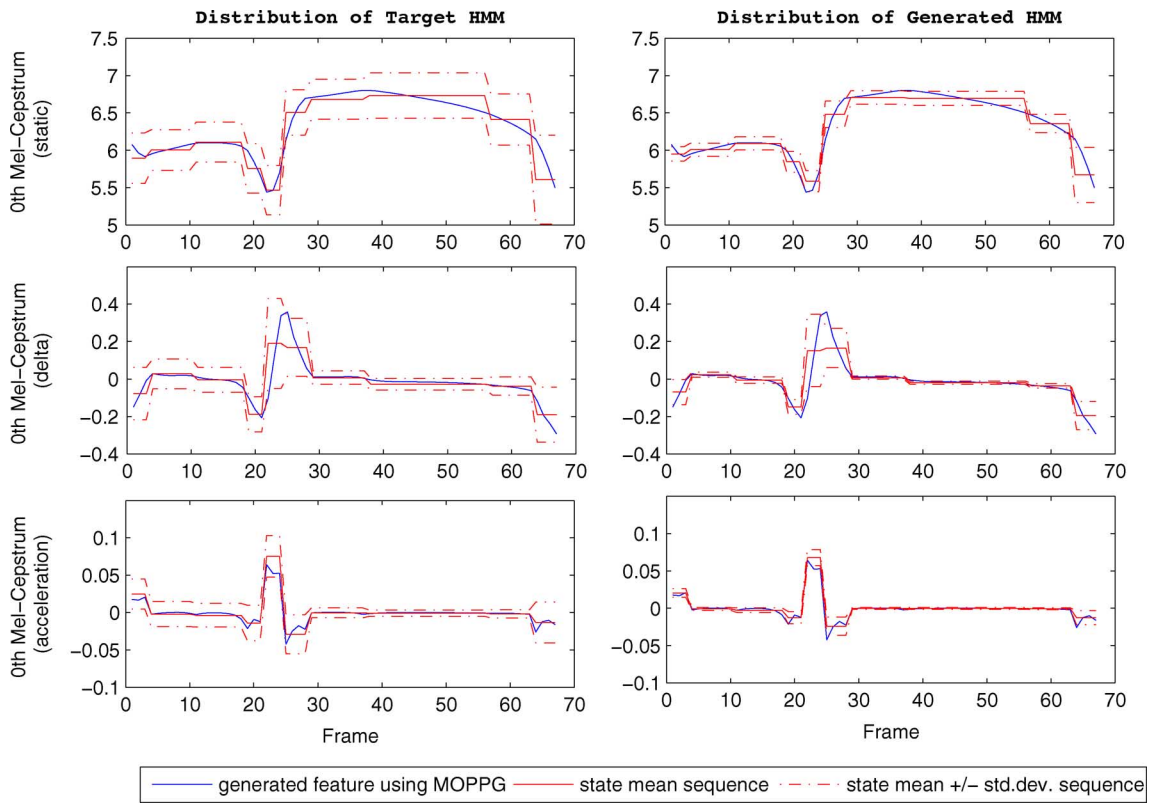

Fig. 4. Example for the state mean and standard deviation parameters of the target HMM (solid and dashed red lines in left column) and the generated HMM (solid and dashed red lines in right column). The spectral parameters (the solid blue lines in both columns) are generated from the target HMM using *MOPPG* and are used to estimated the generated HMM according to (15)–(16).

and the generated HMM were calculated according to (23) for each sentence in the test set using the natural recordings of the test sentences, the *MOPPG* outputs and the *GV* outputs, respectively. The average KL divergences for each mel-cepstrum dimension are shown in Fig. 5. From this figure, we can see that the HMM estimated from natural parameters has much smaller KL divergence with respect to the target HMM than the models estimated from the parameters generated by *MOPPG* and *GV*. For the *MOPPG* method, the generated models for the dynamic features have higher KL divergence than those for the static features. This is consistent with the findings from Fig. 4. The KL divergence also increases with mel-cepstrum coefficient index. Furthermore, the *GV* method reduces the KL divergence between the target HMM and the generated HMM, even though it does not optimize the HMM divergence directly. The static dimensions exhibit greater reduction in KL divergence than the dynamic dimensions because our GV model only handles the sentence-level variances of the static features [12].

We also calculated the average KL divergence between the target HMM and the generated HMMs using the *KLD_GEN* and *KLD_FT* methods for the test set. The results are shown in Fig. 6. It is found that both the *KLD_GEN* and the *KLD_FT* methods can generate spectral features that have a far more similar distribution with respect to the target HMM than the *MOPPG* and *GV* methods shown in Fig. 5. The KL divergence given by *KLD_GEN* is smaller than that of *KLD_FT* for most mel-cepstrum dimensions because the minimum KL divergence criterion is applied to guide the generation of the acoustic parameters directly in the *KLD_GEN* method. There are still some dimensions where *KLD_GEN* has higher KL divergence than *KLD_FT*. This may be caused by the local-optimization property of the steepest descent algorithm. For some mel-cepstrum dimensions, the KL divergence calculated from the parameters generated by *KLD_GEN* are much smaller than that calculated from natural parameters. This implies there may exist a problem of over-fitting when optimizing the generated acoustic features directly under the minimum KL divergence criterion, as discussed in Section III-D.

The mel-cepstral distortion (MCD) on the test set between the natural spectral parameters and the parameters generated using the five methods listed in Table I were calculated. To simplify the calculation of MCD, the spectral parameters were generated using state durations derived from state alignment performed on the natural speech. The results are shown in Fig. 7. We can see both the *GV* method and our proposed minimum KL divergence parameter generation methods increase mel-cepstral distortion in comparison with the conventional *MOPPG* method. Similar findings have been previously described in [23], i.e., that increasing the GV of the generated parameters usually causes an increase of MCD.

C. Subjective Evaluation

Twenty sentences in the test set were selected for subjective evaluation. Their synthetic results using the *MOPPG*, *GV*, *KLD_GEN*, and *KLD_FT* methods were evaluated by five Chinese-native listeners. The listeners were required to give a score from 1 (very unnatural) to 5 (very natural) for each synthesized

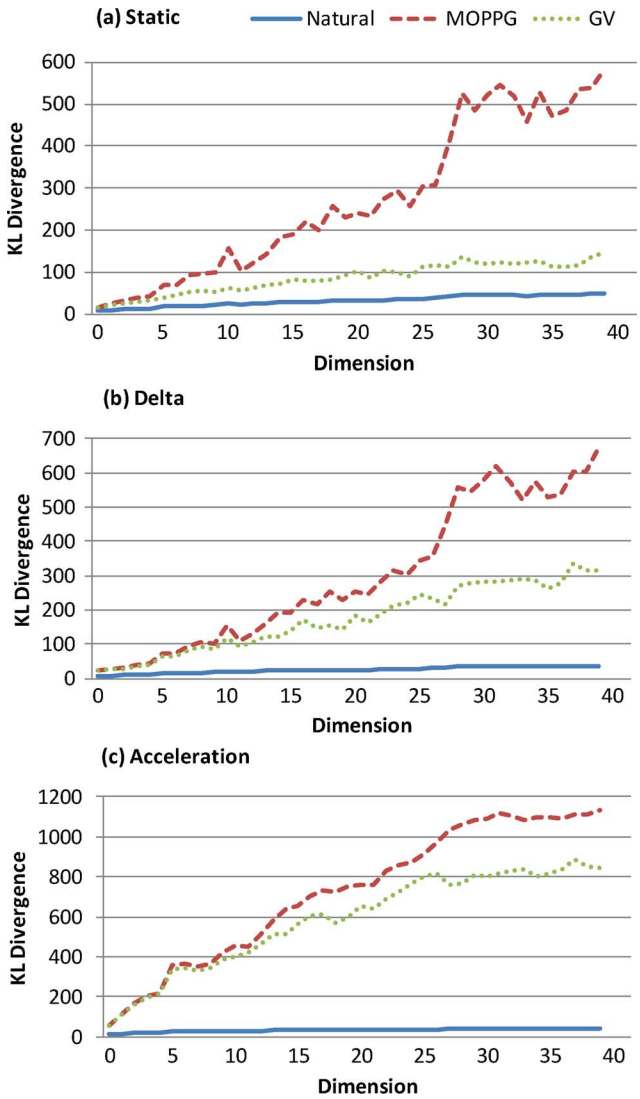

Fig. 5. Average KL divergence of each mel-cepstrum dimension between the target HMM and the generated HMMs in the test set. The generated HMMs are estimated using the natural recordings of the test sentences, the *MOPPG* outputs and the *GV* outputs, respectively.

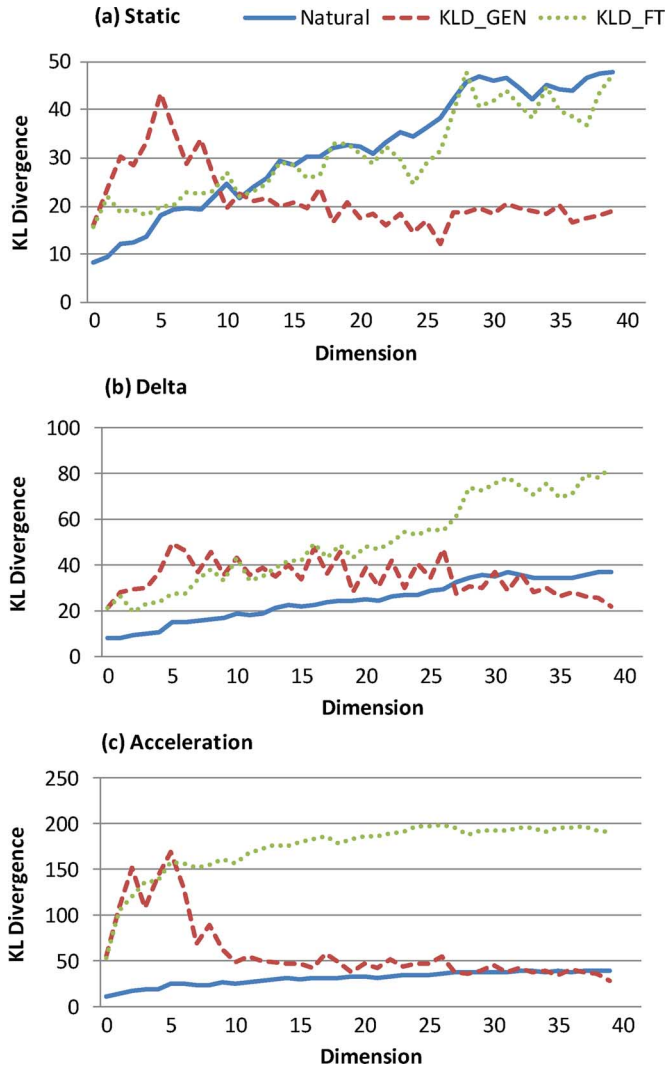

Fig. 6. Average KL divergence of each mel-cepstrum dimension between the target HMM and the generated HMMs in the test set. The generated HMMs are estimated using the natural recordings of the test sentences, the *KLD_GEN* outputs and the *KLD_FT* outputs, respectively.

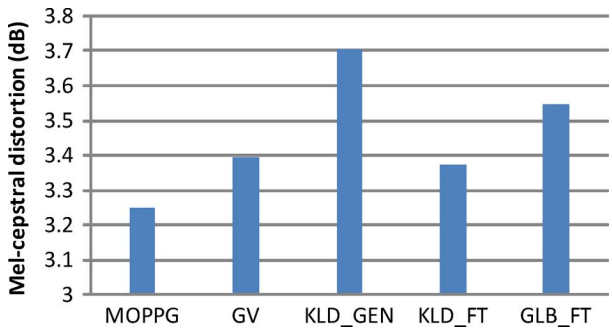

Fig. 7. Mel-cepstral distortions (dB) on test set between the natural parameters and the parameters generated using the five methods listed in Table I.

utterance. The mean opinion scores (MOS) with 95% confidence interval for the four methods are shown in Fig. 8.

Compared with the conventional *MOPPG* method, both *KLD_GEN* and *KLD_FT* methods improve the naturalness of the synthetic speech significantly when they are applied to the generation of mel-cepstra. This is inconsistent with the MCD evaluation results shown in Fig. 7 and proves the effectiveness of our proposed minimum KL divergence parameter generation criterion. We also find that the naturalness of *KLD_GEN* is not as good as *KLD_FT*. Fig. 9 gives an example comparing the spectral parameters generated using *MOPPG*, *KLD_GEN*, and *KLD_FT*. We can see that both the *KLD_GEN* and *KLD_FT* methods increase the variance of the generated acoustic features. However, the trajectory generated by *KLD_GEN* contains much more noise than the other two methods. This noise is caused by the over-fitting problem of the direct optimization approach as discussed in Section III-D. They lead to discontinuity in the generated parameters and degrade the naturalness of the synthetic speech in this evaluation. On the other hand, the *KLD_FT* method is able to alleviate the over-smoothing problem of *MOPPG* while preserving the temporal continuity of the generated feature trajectories. Furthermore, the performance of the *KLD_FT* method is as good as the *GV* method. The advantage of *KLD_FT* is that it requires no extra models aside from the context-dependent phoneme HMM and it is not

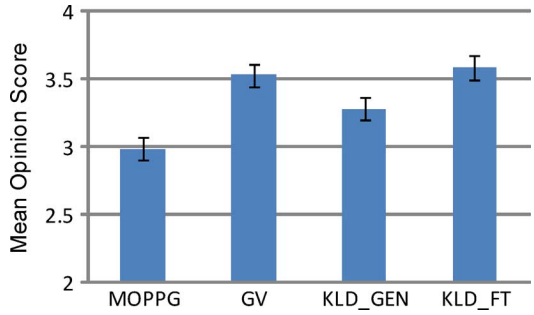

Fig. 8. Mean opinion scores (MOS) with 95% confidence interval for the *MOPPGGV*, *KLD_GEN*, and *KLD_FT* methods listed in Table I.

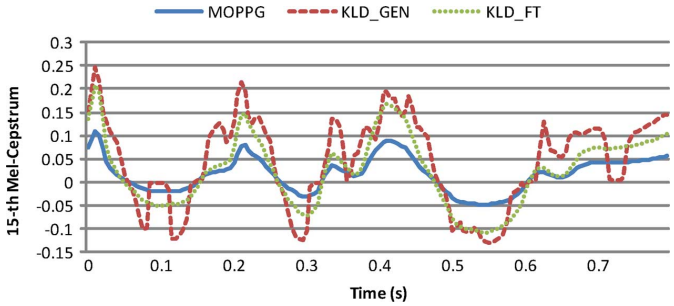

Fig. 9. Sample trajectory of spectral parameters (the 15th mel-cepstrum) generated using the *MOPPG*, *KLD_GEN*, and *KLD_FT* methods.

necessary to tune the weights for integrating multilevel acoustic models with the *GV* method.

Another preference test was conducted to compare the performance of the *GV*, *KLD_FT*, and GLB_FT methods directly. Fifteen sentences in the test set were selected and synthesized using these three methods, respectively. Five Chinese-native listeners took part in the test. Table II shows the preference scores between every pair of the three methods and the p-values given by t-test. We see that there is no significant difference between the *GV* and *KLD_FT* methods at the 5% significance level, and the GLB_FT method is significantly better than *GV* and *KLD_FT*. The superiority of GLB_FT over *KLD_FT* is attributed to the more robust estimation of the linear transform using richer context information of the whole training set. Combining the results shown in Fig. 8 and Table II, we find the naturalness of speech synthesized using the conventional MOPPG algorithm can be improved significantly by applying the global linear transform estimated under the minimum KL divergence criterion as a simple post-filtering operation. Fig. 10 shows the values of the linear transform estimated by GLB_FT in our experiment. We can see that the variances of all mel-cepstrum dimensions are enlarged after the linear transform because the estimated transform factors are always larger than one. There is a trend that the higher dimensions of the mel-cepstrum coefficients get larger transform factors than the lower dimensions. This is consistent with Fig. 5 where the KL divergence of MOPPG outputs increases with mel-cepstrum coefficient index, and the linear transforms in Fig. 10 are estimated to reduce such distribution divergences.

D. Discussion

In our experiments, the value of β in (14) for the MAP estimation is set to 50 empirically. In order to evaluate whether

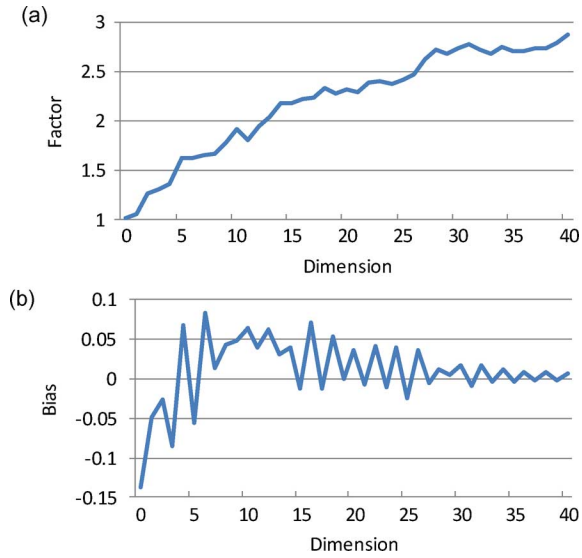

Fig. 10. Global linear transform estimated using the GLB_FT method, including (a) the transform factor and (b) the bias for each mel-cepstrum dimension.

TABLE II
SUBJECTIVE PREFERENCE SCORES (%) AMONG SPEECH SYNTHESIZED USING THE *GV*, *KLD_FT*, AND GLB_FT METHODS, WHERE N/P DENOTES "NO PREFERENCE" AND p MEANS THE p-VALUE OF t-TEST BETWEEN TWO METHODS

GV	*KLD_FT*	*GLB_FT*	N/P	p
29.3	18.7	–	52.0	0.1842
12.0	–	34.7	53.3	0.0034
–	24.0	44.0	32.0	0.0348

TABLE III
MEL-CEPSTRAL DISTORTIONS (dB) FOR THE TEST SENTENCES GENERATED USING THE *KLD_FT* METHOD WITH DIFFERENT VALUES OF β

	$\beta = 10$	$\beta = 50$	$\beta = 100$
$\beta = 10$	–	0.083	0.099
$\beta = 50$	0.083	–	0.035
$\beta = 100$	0.099	0.035	–

the system performance is sensitive to the setting of β, we generated the sentences in the test set using the *KLD_FT* method with $\beta = 10, 50$, and 100, respectively. We found that the subjective perception of these three groups of synthetic speech are very close. Furthermore, the mel-cepstral distortions among the parameters generated using the three β configurations were calculated. The results are shown in Table III. It can be seen that the differences are very small, which means the proposed method is insensitive to β in the range of (10–100).

Based on all the experimental results presented, we assert our proposed criterion of calculating KL divergence between the target and generated HMMs is indeed able to reproduce the qualities of naturalness in the generated acoustic features to a certain extent. For example, the spectral features derived from the natural recordings exhibit low KL divergence, as shown in Fig. 5. The KL divergences of *GV*, *KLD_GEN*, and *KLD_FT*

are much lower than that of the *MOPPG* method, as shown in Figs. 5 and 6, which is consistent with the subjective evaluation results in Fig. 8. However, in some instances, the proposed KL divergence criterion does not work well, e.g., the *KLD_FT* method has higher KL divergence but better naturalness than the *KLD_GEN* method. Some experimental results related to the KL divergence of the generated acoustic features were also presented in [10] and [11], where a method of sampling acoustic feature sequences from a trajectory HMM was studied. It was reported that the models estimated using the samples drawn from a trajectory HMM can converge well to the target model in terms of KL divergence [10]. The spectrum of the sampled trajectories appeared similar to the natural speech qualitatively, but results obtained in a subjective evaluation were not as strong as for the conventional MOPPG method [11]. Our experimental results, together with the work of [10] and [11], indicate the minimum KL divergence parameter generation method is currently still far from ideal and the estimation of both target and generated models could be improved. The trajectory HMM provides a better candidate to represent the target model than the standard HMM since it has better predictive distributions [11]. Compared with the target model, to achieve reliable estimation of the generated model is more difficult. The process of estimating the generated model in our proposed method is affected by the data-sparsity problem. In [10], the generated model was estimated using repeatedly sampled trajectories and its KL divergence decreased as the number of drawn samples increased.

V. Conclusion

We have proposed a minimum KL divergence parameter generation method for HMM-based statistical parametric speech synthesis in this paper. It aims to alleviate the *over-smoothing* problem caused by the conventional maximum output probability parameter generation (MOPPG) algorithm and to improve the quality of synthetic speech. In our approach, the distribution parameters of the generated acoustic features are first estimated using the *maximum a posteriori* (MAP) method. Then, the KL divergence between the *target HMM*, which is used for parameter generation, and the *generated HMM*, which is estimated from the generated acoustic features, is derived as a measure to guide parameter generation. Two approaches, namely direct optimization of feature trajectories and a linear transform of MOPPG outputs, have been proposed in order to integrate the minimum KL divergence criterion into the parameter generation procedure. In our experiments, both these approaches have been shown to significantly improve the naturalness of speech synthesized using MOPPG and mel-cepstrum features. For the linear transform approach, our experimental results show that it is not necessary to estimate the transform matrix for each input sentence at synthesis time, which is computationally expensive. Instead, this transform matrix can be estimated on the training set according to the minimum KL divergence criterion, which makes the postfiltering of MOPPG outputs simple and efficient. Finally, to improve the estimation of the target and generated HMMs and to overcome the over-fitting problem of the direct optimization approach will be the focus of our future work.

Acknowledgment

The authors would like to thank Dr. S. Wei of iFLYTEK Research for useful discussions on the parameter generation methods and Dr. K. Richmond of CSTR, University of Edinburgh, for proofreading the manuscript and making helpful suggestions. The authors would also like to thank the associate editor and the anonymous reviewers for their insightful and helpful comments.

References

[1] 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. Eurospeech*, 1999, pp. 2347–2350.

[2] K. Tokuda, H. Zen, and A. W. Black, , S. Narayanan and A. Alwan, Eds., "HMM-based approach to multilingual speech synthesis," in *Text to Speech Synthesis: New Paradigms and Advances*. Upper Saddle River, NJ: Prentice-Hall, 2004.

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

[4] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

[5] 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.

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

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. Eurospeech*, 2001, pp. 2263–2266.

[8] J. Yu, M. Zhang, J.-H. Tao, and X. Wang, "A novel HMM-based TTS system using both continuous HMMs and discrete HMMs," in *Proc. ICASSP*, 2007, pp. 709–712.

[9] Z.-H. Ling and R.-H. Wang, "HMM-based unit selection using frame sized speech segments," in *Proc. Interspeech*, 2006, pp. 2034–2037.

[10] K. Tokuda, H. Zen, and T. Kitamura, "Reformulating the HMM as a trajectory," model Tech. Rep. of IEICE, 2004.

[11] M. Shannon, H. Zen, and W. Byrne, "The effect of using normalized models in statistical speech synthesis," in *Proc. Interspeech*, 2011, pp. 121–124.

[12] 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.

[13] Z.-H. Ling, Y. Hu, and L.-R. Dai, "Global variance modeling on the log power spectrum of LSPs for HMM-based speech synthesis," in *Proc. Interspeech*, 2010, pp. 825–828.

[14] 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, pp. 4621–4624.

[15] T. Tiomkin, D. Malah, and S. Shechtman, "Statistical text-to-speech synthesis based on segment-wise representation with a norm constraint," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1077–1082, Jul. 2010.

[16] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79–86, 1951.

[17] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM (invited paper)," *IEICE Trans. Inf. Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[18] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, no. 2, pp. 79–86, 2000.

[19] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[20] Y. Zhao, C.-S. Zhang, F. K. Soong, C. Min, and X. Xiao, "Measuring attribute dissimilarity with HMM KL-divergence for speech synthesis,," in *Proc. 6th ISCA Speech Synth. Workshop*, 2007, pp. 206–210.

[21] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[22] [Online]. Available: http://hts.sp.nitech.ac.jp/

[23] T. Toda, , P. Dymarski, Ed., "Modeling of speech parameter sequence considering global variance for HMM-based speech synthesis," in *Hidden Markov Models, Theory and Applications*. Rijeka, Croatia: InTech, 2011.

Zhen-Hua Ling (M'10) received the B.E. degree in electronic information engineering and the M.S. and Ph.D. degrees in signal and information processing from University of Science and Technology of China, Hefei, in 2002, 2005, and 2008 respectively.

From October 2007 to March 2008, he was a Marie Curie Fellow at the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, U.K., From July 2008 to February 2011, he was a joint Postdoctoral Researcher at the University of Science and Technology of China and iFLYTEK Co., Ltd., China. He is currently an Associate Professor at the University of Science and Technology of China. His research interests include speech synthesis, voice conversion, speech analysis, and speech coding.

Dr. Ling is a member of ISCA. He was awarded an IEEE Signal Processing Society Young Author Best Paper Award in 2010.

Li-Rong Dai was born in China in 1962. He received the B.S. degree in electrical engineering from Xidian University, Xi'an, China, in 1983 and the M.S. degree from Hefei University of Technology, Hefei, China, in 1986, and the Ph.D. degree in signal and information processing from the University of Science and Technology of China, Hefei, in 1997.

He joined University of Science and Technology of China in 1993. He is currently a Professor of the School of Information Science and Technology, University of Science and Technology of China. His current research interests include speech synthesis, speaker and language recognition, speech recognition, digital signal processing, voice search technology, machine learning, and pattern recognition. He has published more than 50 papers in these areas.