TRAJECTORY TRAINING CONSIDERING GLOBAL VARIANCE FOR HMM-BASED SPEECH SYNTHESIS

Tomoki Toda[†], *Steve Young*[‡]

[†] Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
[‡] Department of Engineering, University of Cambridge, UK
tomoki@is.naist.jp, sjy@eng.cam.ac.uk

ABSTRACT

This paper presents a novel method for training hidden Markov models (HMMs) for use in HMM-based speech synthesis. The primary goal of HMM parameter optimization is to ensure that parameters generated from the trained models exhibit similar properties to natural speech. In this paper, two major problems in conventional training are addressed: 1) the inconsistency between the training and synthesis optimization criterion; and 2) the over-smoothing caused by the statistical modeling process. The proposed method integrates the global variance (GV) criterion into a trajectory training method to give a unified framework for both training and synthesis which provides both a consistent optimization criterion and a closed form solution for parameter generation. The experimental results demonstrate that the proposed method yields a significant improvement in the naturalness of synthetic speech.

Index Terms— speech synthesis, hidden Markov models, training criterion, trajectory likelihood, global variance

1. INTRODUCTION

The hidden Markov model (HMM) is an effective framework for modeling the acoustics of speech and its introduction has enabled significant progress in speech and language technologies. Recently, HMM-based speech synthesis [1] has attracted attention as a corpus-based approach to Text-to-Speech (TTS) which has the potential for realizing very flexible TTS systems.

A basic framework of HMM-based speech synthesis consists of training and synthesis processes. In the training process, speech parameters such as spectral envelope and F_0 are extracted from speech waveforms and then their time sequences are modeled by context-dependent HMMs. A joint vector of static and dynamic features is usually used as an observation vector to model the dynamic characteristics of speech acoustics. In the synthesis process, a smoothly varying speech parameter trajectory is generated by maximizing the likelihood of a composite sentence HMM subject to a constraint between static and dynamic features with respect to not the observation vector sequence but the static feature vector sequence [2]. Finally a vocoding technique is employed for generating a speech waveform from the generated speech parameters. Although HMM-based speech synthesis has many attractive features such as completely data-driven voice building, flexible voice quality control, speaker adaptation, small footprint, and so on, it has the significant drawback that the quality of the synthetic speech is noticeably degraded compared to the original spoken audio.

The main weakness of the basic framework for HMM-based speech synthesis is the inconsistency between the training and synthesis criteria, i.e., likelihoods for the joint static and dynamic feature vectors in the training process compared to likelihoods for only the static feature vectors in the synthesis process. Consequently, the trained model parameters are not optimum for parameter generation. To address this problem, Zen *et al*. [3] have proposed a training method based on the trajectory HMM, which is derived by imposing an explicit relationship between static and dynamic features on the traditional HMM. This method allows the utilization of a unified criterion, i.e., trajectory likelihood, in both training and synthesis processes. In a similar spirit, Wu and Wang [4] have proposed minimum generation error (MGE) training. This method optimizes the HMM parameters so that an error between the generated and natural parameters is minimized.

In a different approach to improving the synthetic speech quality, Toda and Tokuda [5] have introduced a new criterion on a higher-order moment called the global variance (GV), which is the variance of the static feature vectors calculated over a time sequence (e.g., over an utterance) in the parameter generation process. The static feature vectors generated from the HMMs in the traditional generation process are usually over-smoothed and this is one of the main factors causing the muffled effect in HMM-synthesized speech. Since the GV is inversely correlated with these smoothing effects, a metric on the GV of the generated parameters effectively works as a penalty term in the parameter generation process. It has been reported that synthetic speech quality can be significantly improved by generating the parameter trajectory while keeping its GV close to the natural one.

In an attempt to apply the idea of considering the GV in the HMM training process, Wu *et al*. [6] have proposed MGE training which considers the error in the GV between natural and generated parameters as well as the generation error as mentioned above. The HMM parameters are optimized so that both the generated parameters and their GV are similar to the natural ones. It is clear that this idea can also be applied to the ML-based training process, i.e., the trajectory training, in which not only frame-by-frame generation errors but also a correlation of the errors over a time sequence is considered.

In this paper, we propose a trajectory training method which incorporates the GV. The HMM parameters are optimized in the sense of maximum likelihood subject to a constraint on the GV of the generated parameter trajectory. Consequently, we obtain a unified framework which consistently uses the same criterion in both training and synthesis. A similar method has been proposed in [7], but its effectiveness has never been reported. We show that our proposed method yields significant improvements to the naturalness of synthetic speech. Moreover, trajectory training has hitherto only been applied to the spectral component of the speech signal [3]. Here we extend the framework to include multi-space probability distribution HMMs (MSD-HMMs) [8] to enable trajectory training to be applied to the F_0 component as well as the spectral component.

The authors are grateful to Prof. Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

2. BASIC FRAMEWORK

2.1. HMM Training

Let us assume a D-dimensional static feature vector $\boldsymbol{c}_t = [c_t(1), \cdots, c_t(D)]^\top$ at frame t. We use a speech parameter vector $\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta^{(1)}\boldsymbol{c}_t^\top, \Delta^{(2)}\boldsymbol{c}_t^\top]^\top$ consisting of not only the static feature vector but also dynamic feature vectors $\Delta^{(1)}\boldsymbol{c}_t$, $\Delta^{(2)}\boldsymbol{c}_t$ as the observation vector. The sequences of vectors $\boldsymbol{o}_t$ and $\boldsymbol{c}_t$ are written as $\boldsymbol{o} = [\boldsymbol{o}_1^\top, \cdots, \boldsymbol{o}_t^\top, \cdots, \boldsymbol{o}_T^\top]^\top$ and $\boldsymbol{c} = [\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_T^\top]^\top$, respectively.

In the traditional HMM, the probability density of $\boldsymbol{o}$ given an HMM state sequence $\boldsymbol{q} = [q_1, \cdots, q_t, \cdots, q_T]$ is written as

$$P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{U}_{\boldsymbol{q}}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{q_t}, \boldsymbol{U}_{q_t}) \quad (1)$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{U})$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{U}$. The mean vector sequence $\boldsymbol{\mu}_{\boldsymbol{q}}$ and the covariance matrix sequence $\boldsymbol{U}_{\boldsymbol{q}}$ are given by

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{q}} &= \left[\boldsymbol{\mu}_{q_1}^\top, \cdots, \boldsymbol{\mu}_{q_t}^\top, \cdots, \boldsymbol{\mu}_{q_T}^\top\right]^\top && (2) \\ \boldsymbol{U}_{\boldsymbol{q}} &= \mathrm{diag}\left[\boldsymbol{U}_{q_1}, \cdots, \boldsymbol{U}_{q_t}, \cdots, \boldsymbol{U}_{q_T}\right]. && (3) \end{aligned}$$

In the training process, the HMM parameter set $\boldsymbol{\lambda}$ is optimized in the sense of maximum likelihood as follows:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) P(\boldsymbol{q}|\boldsymbol{\lambda}). \quad (4)$$

2.2. Parameter Generation [2]

Given the HMM state sequence $\boldsymbol{q}$,[1] the static feature vector sequence is determined by

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) = \overline{\boldsymbol{c}}_{\boldsymbol{q}} \quad \text{subject to} \quad \boldsymbol{o} = \boldsymbol{W}\boldsymbol{c} \quad (5)$$

where $\boldsymbol{W}$ is a window matrix extending the static feature vector sequence to the observation vector sequence consisting of static and dynamic features. The ML estimate $\overline{\boldsymbol{c}}_{\boldsymbol{q}}$ is given by

$$\begin{aligned} \overline{\boldsymbol{c}}_{\boldsymbol{q}} &= \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{r}_{\boldsymbol{q}} && (6) \\ \boldsymbol{P}_{\boldsymbol{q}}^{-1} &= \boldsymbol{W}^\top \boldsymbol{U}_{\boldsymbol{q}}^{-1} \boldsymbol{W} && (7) \\ \boldsymbol{r}_{\boldsymbol{q}} &= \boldsymbol{W}^\top \boldsymbol{U}_{\boldsymbol{q}}^{-1} \boldsymbol{\mu}_{\boldsymbol{q}}. && (8) \end{aligned}$$

3. IMPLEMENTING TRAJECTORY TRAINING FOR BOTH SPECTRAL AND F_0 COMPONENTS

3.1. Trajectory HMM [3]

The traditional HMM is reformulated as a trajectory HMM by imposing an explicit relationship between static and dynamic features, which is given by $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$. The probability density of $\boldsymbol{c}$ in the trajectory HMM given the state sequence $\boldsymbol{q}$ is then written as

$$P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{c}; \overline{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}) = \frac{1}{Z_{\boldsymbol{q}}} P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) \quad (9)$$

where

$$Z_{\boldsymbol{q}} = \frac{\sqrt{(2\pi)^{DT}|\boldsymbol{P}_{\boldsymbol{q}}|}}{\sqrt{(2\pi)^{3DT}|\boldsymbol{U}_{\boldsymbol{q}}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}}^{-1} \boldsymbol{\mu}_{\boldsymbol{q}} - \boldsymbol{r}_{\boldsymbol{q}}^\top \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{r}_{\boldsymbol{q}})\right). \quad (10)$$

The mean vector $\overline{\boldsymbol{c}}_{\boldsymbol{q}}$ varies within states and inter-frame correlation is modeled by the temporal covariance matrix $\boldsymbol{P}_{\boldsymbol{q}}$ that is generally full even if using the same number of model parameters as in the traditional HMM. Note that the mean vector of the trajectory HMM is equivalent to the ML estimate of the generated static feature sequence shown by Eq. (6) because the parameter generation process in Eq. (5) is reformed as the maximization process of the trajectory likelihood $P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\lambda})$.

3.2. Estimation of Model Parameters [3]

The HMM parameter set $\boldsymbol{\lambda}$ is estimated by maximizing the trajectory likelihood $\mathcal{L}_{\boldsymbol{q}}(= P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\lambda}))$ given the HMM state sequence $\boldsymbol{q}$.[2] The mean vectors and diagonal covariance matrices at all HMM states (from 1 to N), which are given by

$$\begin{aligned} \boldsymbol{m} &= \left[\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top, \cdots, \boldsymbol{\mu}_N^\top\right]^\top && (11) \\ \boldsymbol{\Sigma}^{-1} &= \left[\boldsymbol{U}_1^{-1}, \boldsymbol{U}_2^{-1}, \cdots, \boldsymbol{U}_N^{-1}\right]^\top, && (12) \end{aligned}$$

are simultaneously updated. The mean vectors $\boldsymbol{m}$ are iteratively updated using the following gradient,

$$\frac{\partial \mathcal{L}_{\boldsymbol{q}}}{\partial \boldsymbol{m}} = \boldsymbol{A}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}}^{-1} \boldsymbol{W}(\boldsymbol{c} - \overline{\boldsymbol{c}}_{\boldsymbol{q}}) \quad (13)$$

where $\boldsymbol{A}_{\boldsymbol{q}}$ is a $3MT \times 3MN$ matrix whose elements are 0 or 1 determined according to the state sequence $\boldsymbol{q}$. The covariance matrices $\boldsymbol{\Sigma}$ are iteratively updated using the following gradient,

$$\begin{aligned} \frac{\partial \mathcal{L}_{\boldsymbol{q}}}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{1}{2} \boldsymbol{A}_{\boldsymbol{q}}^\top \text{on-diag} \Big[& \boldsymbol{W}(\boldsymbol{P}_{\boldsymbol{q}} + \overline{\boldsymbol{c}}_{\boldsymbol{q}} \overline{\boldsymbol{c}}_{\boldsymbol{q}}^\top - \boldsymbol{c}\boldsymbol{c}^\top)\boldsymbol{W}^\top \\ & -2\boldsymbol{\mu}_{\boldsymbol{q}}(\overline{\boldsymbol{c}}_{\boldsymbol{q}} - \boldsymbol{c})^\top \boldsymbol{W}^\top \Big] \quad (14) \end{aligned}$$

where on-diag[$\cdot$] denotes the extraction of only diagonal elements from a square matrix. Note that a closed form solution also exists for the estimation of the mean vectors but a very large set of linear equations needs to be solved.

3.3. Trajectory Training for F_0 modeling

The MSD-HMM for the F_0 component usually models each of the static and dynamic features separately with different streams because the dynamic features at voiced/unvoiced boundaries are not easily calculated. The probability density of $\boldsymbol{o}$ given $\boldsymbol{q}$ is written as

$$\begin{aligned} & P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) = \\ & \prod_{n=0}^{2} \prod_{t \in U^{(n)}} (1 - w_{q_t}^{(n)}) \prod_{t \in V^{(n)}} w_{q_t}^{(n)} \mathcal{N}(\Delta^{(n)} c_t; \mu_{q_t}^{(n)}, U_{q_t}^{(n)}) \quad (15) \end{aligned}$$

where $\Delta^{(0)} c_t = c_t$, and $w_{q_t}^{(n)}$ is a weight for a continuous space on which static or dynamic features are modeled. The unvoiced and voiced frames for each stream are denoted as $t \in U^{(n)}$ and $t \in V^{(n)}$, respectively.

To derive the trajectory MSD-HMM in the similar manner as mentioned above, we first approximate Eq. (15) with a single stream as follows:

$$\begin{aligned} P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) &= \prod_{t \in U^{(0)}} (1 - w_{q_t}^{(0)}) \prod_{t \in V^{(0)}} w_{q_t}^{(0)} \mathcal{N}(\boldsymbol{c}_t; \boldsymbol{\mu}_{q_t}, \boldsymbol{U}_{q_t}) \\ &= P(\boldsymbol{o}_V|\boldsymbol{q}_V, \boldsymbol{\lambda}) \prod_{t \in U^{(0)}} (1 - w_{q_t}^{(0)}) \prod_{t \in V^{(0)}} w_{q_t}^{(0)} \quad (16) \end{aligned}$$

[1] We usually use the suboptimum HMM state sequence determined by maximizing only a likelihood of the duration model.

[2] This paper uses the suboptimum HMM state sequence determined by the Viterbi algorithm and the traditional likelihood, i.e., $P(\boldsymbol{q}|\boldsymbol{\lambda})P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda})$.

where $\boldsymbol{o}_V$ is a time sequence of static and dynamic features and $\boldsymbol{q}_V$ is a state sequence of only the voiced frames. In order to ignore the dynamic features at voiced/unvoiced boundary frames, precisions for the dynamic features ${U_{q_t}^{(1)}}^{-1}$ and ${U_{q_t}^{(2)}}^{-1}$ are set to zeros at the boundaries. Note that the same modification is usually performed in the speech parameter generation for the F_0 component. The MSD-HMM is reformulated as the trajectory MSD-HMM by imposing $\boldsymbol{o}_V$ = $\boldsymbol{W}_V \boldsymbol{c}_V$ on $P(\boldsymbol{o}_V|\boldsymbol{q}_V, \boldsymbol{\lambda})$. Consequently, the probability density of $\boldsymbol{c}$ given $\boldsymbol{q}$ is given by

$$P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\lambda}) = P(\boldsymbol{c}_V|\boldsymbol{q}_V, \boldsymbol{\lambda}) \prod_{t \in U^{(0)}} (1 - w_{q_t}^{(0)}) \prod_{t \in V^{(0)}} w_{q_t}^{(0)}. \qquad (17)$$

4. GV-CONSTRAINED TRAJECTORY TRAINING

In order to integrate parameter generation considering the GV [5] into an ML-based training framework, we propose using trajectory training subject to a constraint on the GV. This approach is quite different from the method described in [7] since it uses a different definition of the GV probability density function, it updates both means and covariances, and it is applied to both spectral and F_0 components.

4.1. Objective Function

The proposed objective function $\mathcal{L}'_{\boldsymbol{q}}$ is given by

$$\mathcal{L}'_{\boldsymbol{q}} \quad = \quad P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\lambda}) P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_v)^{\omega T} \qquad (18)$$

where $\boldsymbol{v}(\boldsymbol{c}) = [v(1), \cdots, v(D)]^\top$ is a GV vector of the static feature vector sequence $\boldsymbol{c}$, which is calculated by

$$v(d) \quad = \quad \frac{1}{T} \sum_{t=1}^{T} (c_t(d) - \langle c(d) \rangle)^2 \qquad (19)$$

$$\langle c(d) \rangle \quad = \quad \frac{1}{T} \sum_{\tau=1}^{T} c_\tau(d). \qquad (20)$$

The GV is calculated utterance by utterance and the probability density of the GV is modeled by

$$P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_v) \quad = \quad \mathcal{N}(\boldsymbol{v}(\boldsymbol{c}); \boldsymbol{v}(\overline{\boldsymbol{c}}_{\boldsymbol{q}}), \boldsymbol{\Sigma}_v). \qquad (21)$$

Note that the mean vector of the GV probability density is defined as the GV of the mean vector of the trajectory HMM, which is equivalent to the GV of the generated parameters from the HMM shown by Eq. (6). Hence, the GV likelihood $P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_v)$ works as a penalty term to make the GV of the generated parameters close to that of the natural ones. The balance between the two likelihoods $P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\lambda})$ and $P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_v)$ is controlled by the GV weight ω.

4.2. Estimation of model parameters

Given the HMM state sequence $\boldsymbol{q}$, the GV weight ω, and the GV covariance matrix $\boldsymbol{\Sigma}_v$, the HMM parameter set is estimated by maximizing the proposed objective function $\mathcal{L}'_{\boldsymbol{q}}$. The mean vectors and the covariance matrices are iteratively updated using the following gradients,

$$\frac{\partial \mathcal{L}'_{\boldsymbol{q}}}{\partial \boldsymbol{m}} \quad = \quad \boldsymbol{A}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}}^{-1} \boldsymbol{W} \left(\boldsymbol{c} - \overline{\boldsymbol{c}}_{\boldsymbol{q}} + \omega \boldsymbol{P}_{\boldsymbol{q}} \overline{\boldsymbol{x}}_{\boldsymbol{q}} \right) \qquad (22)$$

$$\frac{\partial \mathcal{L}'_{\boldsymbol{q}}}{\partial \boldsymbol{\Sigma}^{-1}} \quad = \quad \frac{1}{2} \boldsymbol{A}_{\boldsymbol{q}}^\top \text{on-diag} \Big[\boldsymbol{W} (\boldsymbol{P}_{\boldsymbol{q}} + \overline{\boldsymbol{c}}_{\boldsymbol{q}} \overline{\boldsymbol{c}}_{\boldsymbol{q}}^\top - \boldsymbol{c}\boldsymbol{c}^\top) \boldsymbol{W}^\top$$
$$-2 \boldsymbol{\mu}_{\boldsymbol{q}} (\overline{\boldsymbol{c}}_{\boldsymbol{q}} - \boldsymbol{c}_{\boldsymbol{q}})^\top \boldsymbol{W}^\top + 2\omega \boldsymbol{W} \boldsymbol{P}_{\boldsymbol{q}} \overline{\boldsymbol{x}}_{\boldsymbol{q}} (\boldsymbol{\mu}_{\boldsymbol{q}} - \boldsymbol{W} \overline{\boldsymbol{c}}_{\boldsymbol{q}})^\top \Big] (23)$$

where

$$\overline{\boldsymbol{x}}_{\boldsymbol{q}}^{(d)} = -2 \left(\overline{\boldsymbol{c}}_{\boldsymbol{q}}^{(d)} - \left\langle \overline{\boldsymbol{c}}_{\boldsymbol{q}}^{(d)} \right\rangle \right) (\boldsymbol{v}(\overline{\boldsymbol{c}}_{\boldsymbol{q}}) - \boldsymbol{v}(\boldsymbol{c}))^\top \boldsymbol{p}_v^{(d)}. \qquad (24)$$

4.3. Parameter generation

It is not necessary to consider the GV in parameter generation because the HMM parameters are optimized so that the GV of the generated trajectory is close to the natural one. Consequently, the basic parameter generation algorithm shown by Eq. (5) is employed. Note that the basic algorithm is computationally much more efficient compared to the parameter generation algorithm considering the GV [5] that needs an iterative process.

If the proposed objective function $\mathcal{L}'_{\boldsymbol{q}}$ is used in the parameter generation, the static feature vector sequence is determined by

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \mathcal{N}(\boldsymbol{c}; \overline{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}) \mathcal{N}(\boldsymbol{v}(\boldsymbol{c}); \boldsymbol{v}(\overline{\boldsymbol{c}}_{\boldsymbol{q}}), \boldsymbol{\Sigma}_v) = \overline{\boldsymbol{c}}_{\boldsymbol{q}}. \qquad (25)$$

Note that this estimate is equivalent to the ML estimate by the basic algorithm. Therefore, the proposed framework can also be regarded as a unified framework using the same objective function in both the training process and the synthesis process.

5. EXPERIMENTAL EVALUATIONS

5.1. Experimental conditions

For the evaluation, voices were built for each of 4 English speakers (2 males: bdl and rms, and 2 females: clb and slt) in the CMU ARCTIC database [10]. For each speaker, we used sub-set A consisting of about 600 sentences as training data and the remaining sub-set B consisting of about 500 sentences for evaluation. Context-dependent labels were automatically generated from texts using a text analyzer derived from Flite [11].

The 0th through 24th mel-cepstral coefficients were used as spectral parameters and log-scaled F_0 plus aperiodic components for the excitation. STRAIGHT [12] was employed for the analysis-synthesis method. Each speech parameter vector included the static features and their delta and delta-deltas. The frame shift was set to 5 ms.

Context-dependent HMMs were trained for each of the spectral, F_0, and aperiodic components using a decision-tree based context clustering technique. We also trained context-dependent duration models for modeling the state duration probabilities. After initializing using the basic training process, trajectory training was performed for the spectral and F_0 components. Finally, trajectory training was performed subject to a constraint on the GV for the both components. The covariance matrix of the GV probability density function was trained using the GVs calculated from individual utterances in the training data. We set the GV weight ω to 0.5. Trajectory training was not applied to the aperiodic components.

For synthesis, the speech parameter sequences were generated from sentence HMMs for given input contexts. The basic speech parameter generation algorithm shown by Eq. (5) was employed and then, a speech waveform was synthesized by filtering the mixed excitation, which was designed by the generated excitation parameters, based on the generated spectral parameters.

5.2. Objective Evaluation

Table 1 shows the log-scaled trajectory, GV, and proposed likelihoods for mel-cepstrum and log-scaled F_0 in the training data and the evaluation data, respectively. The trajectory training causes

Table 1. Log-scaled trajectory likelihood shown by Eq. (9), GV likelihood shown by Eq. (21), and proposed likelihood shown by Eq. (18) (when $\omega = 1.0$) of each model

a) Training data

Likelihoods for mel-cep	Trajectory	GV	Proposed
Basic HMM (Basic)	16.97	-69.22	-52.25
Trajectory HMM (Trj)	28.52	-34.22	-5.70
GV-Trajectory HMM (GV-Trj)	28.11	92.43	120.54
Likelihoods for log F_0	**Trajectory**	**GV**	**Proposed**
Basic HMM (Basic)	2.09	1.59	3.68
Trajectory HMM (Trj)	2.26	1.87	4.13
GV-Trajectory HMM (GV-Trj)	2.25	2.55	4.80

b) Evaluation data

Likelihoods for mel-cep	Trajectory	GV	Proposed
Basic HMM (Basic)	14.49	-69.13	-54.64
Trajectory HMM (Trj)	27.28	-34.87	-7.59
GV-Trajectory HMM (GV-Trj)	26.87	80.34	107.21
Likelihoods for log F_0	**Trajectory**	**GV**	**Proposed**
Basic HMM (Basic)	1.77	1.27	3.04
Trajectory HMM (Trj)	2.02	1.54	3.56
GV-trajectory HMM (GV-Trj)	2.02	2.17	4.19

significant improvements in the trajectory likelihoods because the HMM parameters are optimized so as to directly maximize the trajectory likelihoods. It is interesting to note that the trajectory training also causes improvements in the GV likelihood although the improvement gains are not large. These results suggest that the trajectory training yields better generated parameter trajectories than the basic training. The GV likelihoods are dramatically improved by the GV-constrained trajectory training. Note that this training method does not cause significant reductions to the trajectory likelihoods. This can be observed for both the mel-cepstrum and F_0 components and in both the training and the evaluation data.

Overall these results suggest that the proposed training method does lead to parameter trajectories which more closely resemble the various feature characteristics of real speech.

5.3. Subjective Evaluation

An opinion test was conducted on the naturalness of the synthetic speech to demonstrate the effectiveness of the proposed method. Eight kinds of voices were evaluated: seven combinations of basic/trajectory/GV-constrained trajectory training methods for F_0/mel-cepstrum components and one analysis-synthesis, as shown in **Figure 1**. Ten listeners participated in the test. Each listener evaluated 32 samples consisting of four sentences for each speaker, i.e., 128 samples in total. These sentences were randomly selected for each speaker and each listener from the evaluation data.[3]

Figure 1 shows the result of the test. The trajectory training yields significant quality improvements compared to the basic training and the GV-constrained trajectory training provides further improvements. It can be seen that the largest improvements are obtained by applying these methods to the mel-cepstrum components. This tendency is similar to that observed when considering the GV in the parameter generation process [5].

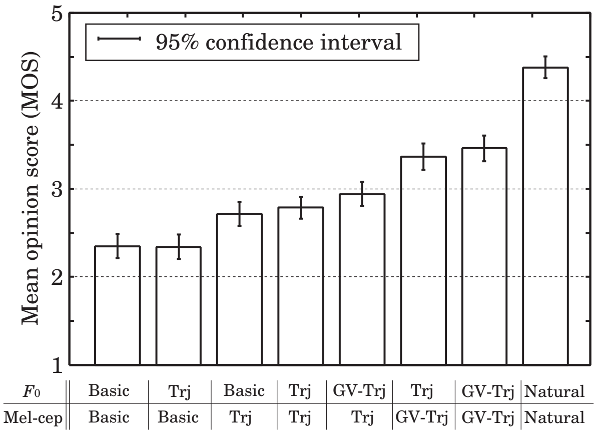

Fig. 1. Result of opinion test on naturalness.

6. CONCLUSIONS

This paper has described a new trajectory training method under a constraint on a global variance (GV) for HMM-based speech synthesis. The proposed method provides a unified framework for training and synthesizing speech using a common criterion, it yields very significant improvements in naturalness, and it allows a more efficient parameter generation process considering the GV based on a closed form solution. Our next step is to investigate whether the proposed method causes significant quality improvements in synthetic speech compared with the conventional GV-based parameter generation.

7. 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. *Proc. EUROSPEECH*, pp. 2347–2350, Budapest, Hungary, Sep. 1999.

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

[3] H. Zen, K. Tokuda, and T. Kitamura. Reformulating the HMM as a trajetory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language*, Vol. 21, pp. 153-173, 2007.

[4] Y.-J. Wu and R.H. Wang. Minimum generation error training for HMM-based speech synthesis. *Proc. ICASSP*, pp. 89–92, Toulouse, France, May 2006.

[5] T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions*, Vol. E90-D, No. 5, pp. 816–824, May 2007.

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

[7] K. Nakamura. *Model training considering global variance for HMM-based speech synthesis*. Master Thesis (in Japanese), Nagoya Institute of Technology, 2007.

[8] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Multi-space probability distribution HMM. *IEICE Trans. Inf. and Syst.*, Vol. E85-D, No. 3, pp. 455–464, 2002.

[9] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel-generalized cepstral analysis – a unified approach to speech spectral estimation. *Proc. ICSLP*, pp. 1043–1045, Yokohama, Japan, Sep. 1994.

[10] J. Kominek and A. W. Black. CMU ARCTIC databases for speech synthesis. *Technical Report*, CMU-LTI-03-177, Language Technologies Institute, Carnegie Mellon University, 2003.

[11] http://www.speech.cs.cmu.edu/flite/index.html

[12] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, Vol. 27, No. 3–4, pp. 187–207, 1999.

[3] Several samples are available from http://spalab.naist.jp/˜tomoki/ICASSP/GV-TrjHMM/index.html