INTERSPEECH 2015

Modulation Spectrum-Constrained Trajectory Training Algorithm for HMM-Based Speech Synthesis

Shinnosuke Takamichi[1,2], *Tomoki Toda*[1], *Alan W. Black*[2], *Satoshi Nakamura*[1]

[1] Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
[2] Language Technologies Institute, Carnegie Mellon University (CMU)

shinnosuke-t@is.naist.jp

Abstract

This paper presents a novel training algorithm for Hidden Markov Model (HMM)-based speech synthesis. One of the biggest issues causing significant quality degradation in synthetic speech is the over-smoothing effect often observed in generated speech parameter trajectories. Recently, we have found that a Modulation Spectrum (MS) of the generated speech parameters is sensitively correlated with the over-smoothing effect, and have proposed the parameter generation algorithm considering the MS. The over-smoothing effect is effectively alleviated by the proposed parameter generation algorithm. On the other hand, it loses the computationally-efficient generation processing of the conventional generation algorithm. In this paper, the MS is integrated into the training stage instead of the parameter generation stage in a similar manner as our previous work on Gaussian Mixture Model (GMM)-based spectral parameter trajectory conversion. The trajectory HMM is trained with a novel objective function consisting of both the conventional trajectory HMM likelihood and a newly implemented MS likelihood. This training framework is further extended to the F_0 component. The experimental results demonstrate that the proposed algorithm yields improvements in synthetic speech quality while preserving a capability of the computationally-efficient generation processing.

Index Terms: HMM-based speech synthesis, over-smoothing, global variance, modulation spectrum, trajectory training

1. Introduction

Statistical parametric speech synthesis based on Hidden Markov Models (HMMs) [1] is an effective framework for generating diverse types of synthetic speech. Speech parameters, i.e., spectral and excitation features and HMM-state duration are simultaneously modeled with context-dependent HMMs in a unified framework [2]. In synthesis, the speech parameter trajectories are generated by maximizing the likelihood of the HMMs [3]. This approach allows us not only to apply several techniques for flexibly controlling synthetic speech [4, 5, 6] to various speech-based systems [7, 8], but also to build the speech synthesizer without complicated tuning compared to sample-based [9] or deep neural nets-based [10] speech synthesis. The further merit of HMM-based speech synthesis is the computationally-efficient speech parameter generation [3]. This generation algorithm is very helpful to deploy the speech-based systems that need the fast speech synthesis, e.g., speech-to-speech translation system [11].

One of the critical problems in HMM-based speech synthesis is that the parameter trajectories generated from the HMMs are often over-smoothed. This phenomenon causes significant degradation of the perceptual quality and makes synthetic speech sound muffled [12]. To address this over-smoothing problem, we have found Modulation Spectrum (MS) [13, 14, 15] as a feature well quantify the over-smoothing effect. The MS is defined as the power spectrum of the speech parameter trajectories, and is regarded as an extension of the Global Variance (GV) [16]. The MS of the generated trajectories is often lower than that of natural speech parameter trajectories. [17] integrated a metric on the MS into the parameter generation algorithm to keep the MS close to natural one, and they reported the improvements in synthetic speech quality. However, as [18, 19, 20] reported in the parameter generation considering the GV, the parameter generation algorithm considering the MS also loses the conventional computationally-efficient generation ability because the objective function in synthesis does not solved in a closed form.

As a method to recover such features while adopting the computationally-efficient generation algorithm, [21] have proposed a metric to integrate the GV into the training stage instead of the synthesis stage. They reformulated trajectory HMMs [22] imposing the constraint between the static and dynamic features for spectral and F_0 components. By training the trajectory HMMs with the GV constraint, the computationally-efficient generation algorithm is straightforwardly adopted, but quality benefits by the GV metric is observed in synthetic speech. We can expect that same reformulation with the MS will give us the further gain in synthetic speech quality.

This paper proposes the MS-constrained trajectory training algorithm to HMM-based speech synthesis in the same manner as our previous work [23] on Gaussian Mixture Model (GMM)-based voice conversion [24]. The trajectory HMM is trained with a novel objective function consisting of both the conventional trajectory HMM likelihood and a newly implemented MS likelihood. This training framework is further extended to the F_0 component. The proposed training algorithm is compared to the basic training [2], the conventional trajectory training [22], and GV-constrained trajectory training [21] in term of synthetic speech quality. The result demonstrates the proposed training algorithm achieves the best synthetic speech quality compared to these training algorithms.

2. Basic Framework [2]

2.1. Training Algorithm [1]

In HMM-based speech synthesis, A HMM parameter set $\boldsymbol{\lambda}$ is estimated using the contextual factor sequence $\boldsymbol{X}$ of input text and the speech feature sequence $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ of T frames as follows:

$$\boldsymbol{\lambda} = \operatorname{argmax} L_{\mathrm{basic}} = \operatorname{argmax} P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\lambda}\right). \tag{1}$$

The output probability density function of HMM-state index q is given as:

$$P\left(\boldsymbol{Y}_t \middle| \boldsymbol{X}, q, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right), \tag{2}$$

where $\boldsymbol{Y}_t$ is given by $3D$-dimensional joint static and dynamic feature vectors, $\left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top, \Delta\Delta\boldsymbol{y}_t^\top\right]^\top$, where $\boldsymbol{y}_t = \left[y_t(1), \cdots, y_t(d), \cdots, y_t(D)\right]^\top$ is represented as a D-dimensional vector at frame t, and d is a dimensional index. $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes Gaussian distribution of a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The HMM parameter set $\boldsymbol{\lambda}$ consists Q HMM-states where each HMM-state has the individual mean vector $\boldsymbol{\mu}_q$ and covariance matrix $\boldsymbol{\Sigma}_q$.

2.2. Parameter Generation Algorithm [3]

Given the contextual factor sequence $\boldsymbol{X}$, the generated parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ is analytically determined by maximizing the output probability of the speech feature vector sequence target $\boldsymbol{Y}$ given $\boldsymbol{X}$ under a constraint $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ as follows:

$$\begin{aligned} \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} &= \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{y} | \boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) &(3)\\ &= \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{r}_{\hat{\boldsymbol{q}}} = \left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}}, &(4)\end{aligned}$$

where $\boldsymbol{W}$ is a $3DT$-by-DT weight matrix to calculate the dynamic features [3]. $\hat{\boldsymbol{q}} = \left[\hat{q}_1, \cdots, \hat{q}_t, \cdots, \hat{q}_T\right]$ is the sub-optimum state sequence determined by maximizing state duration probability distribution function $P\left(\boldsymbol{q}|\boldsymbol{X}, \boldsymbol{\lambda}\right)$, where q_t is a sub-optimum HMM-state index at frame t. The mean vector $\boldsymbol{E}_{\hat{\boldsymbol{q}}} = \left[\boldsymbol{\mu}_{\hat{q}_1}^\top, \cdots, \boldsymbol{\mu}_{\hat{q}_t}^\top, \cdots, \boldsymbol{\mu}_{\hat{q}_T}^\top\right]^\top$ and the covariance matrix $\boldsymbol{D}_{\hat{\boldsymbol{q}}} = \operatorname{diag}_{3D}\left[\boldsymbol{\Sigma}_{\hat{q}_1}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_t}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_T}\right]$ are calculated using the corresponding HMM-state, where the notation diag_{3D} denotes the construction of a block diagonal matrix that has the $3D$-by-$3D$ diagonal elements.

3. Trajectory Training [22]

Trajectory HMM is reformulated by imposing the constraint between the static and dynamic features. The objective function for the trajectory training is written as:

$$L_{\mathrm{trj}} = P\left(\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}, \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\right). \tag{5}$$

The mean vector $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ is given by Eq. (4) and the inter-frame correlation is effectively modeled by the temporal covariance matrix $\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}$. In training, the HMM parameters are updated by maximizing L_{trj}.

3.1. Estimation of Model Parameters

Here, let $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1^\top, \cdots, \boldsymbol{\mu}_q^\top, \cdots, \boldsymbol{\mu}_Q^\top\right]^\top$ and $\left[\boldsymbol{\Sigma}_1^{-1}, \cdots, \boldsymbol{\Sigma}_q^{-1}, \cdots, \boldsymbol{\Sigma}_Q^{-1}\right]^\top$ be the joint parameters of $\boldsymbol{\mu}_q$ and $\boldsymbol{\Sigma}_q^{-1}$ over all HMM-states, respectively. The mean vector $\boldsymbol{E}_{\hat{\boldsymbol{q}}}$ and the precision matrix $\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}$ are represented as:

$$\begin{aligned} \boldsymbol{E}_{\hat{\boldsymbol{q}}} &= \boldsymbol{S}_{\hat{\boldsymbol{q}}}\boldsymbol{\mu}, &(6)\\ \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} &= \operatorname{diag}_{3D}\left[\boldsymbol{S}_{\hat{\boldsymbol{q}}}\boldsymbol{\Sigma}^{-1}\right], &(7)\end{aligned}$$

where $\boldsymbol{S}_{\hat{\boldsymbol{q}}} = \left[\boldsymbol{S}_{\hat{q}_1}, \cdots, \boldsymbol{S}_{\hat{q}_T}\right]^\top \otimes \boldsymbol{I}_{3D}$ is a $3DT$-by-$3DQ$ matrix, $\boldsymbol{S}_{\hat{q}_t}$ is an M-dimensional vector of which the q-th component is 1 when $q = \hat{q}_t$ and otherwise are 0 as shown in Fig. 1, and $\boldsymbol{I}_{3D}$ indicates the $3D$-by-$3D$ identity matrix.

To optimize these model parameters for the objective function, we employ the steepest descent algorithm. as follows:

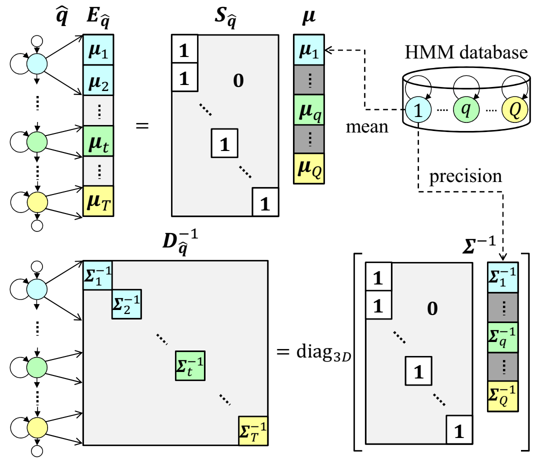

Figure 1: Graphic representation between variables used in this paper.

$$\boldsymbol{\mu}^{(i+1)} = \boldsymbol{\mu}^{(i)} + \alpha \left.\frac{\partial \log L_{\mathrm{trj}}}{\partial \boldsymbol{\mu}}\right|_{\boldsymbol{\mu}=\boldsymbol{\mu}^{(i)}}, \tag{8}$$

where α is a learning rate, and i is an iteration index. $\boldsymbol{\Sigma}^{-1}$ are also optimized in the same manner. The gradients are given by:

$$\begin{aligned} \frac{\partial \log L_{\mathrm{trj}}}{\partial \boldsymbol{\mu}} &= \boldsymbol{S}_{\hat{\boldsymbol{q}}}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right), &(9)\\ \frac{\partial \log L_{\mathrm{trj}}}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{1}{2}\boldsymbol{S}_{\hat{\boldsymbol{q}}}^\top \operatorname{diag}_{3D}^{-1}\left[\boldsymbol{W}\left(\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1} + \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}^\top - \boldsymbol{y}\boldsymbol{y}^\top\right)\right.\\ &\quad \left. - \boldsymbol{E}_{\hat{\boldsymbol{q}}}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} - \boldsymbol{y}\right)^\top \boldsymbol{W}^\top - \boldsymbol{W}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} - \boldsymbol{y}\right)\boldsymbol{E}_{\hat{\boldsymbol{q}}}^\top\right], &(10)\end{aligned}$$

4. Global Variance (GV)-Constrained Trajectory Training [21]

4.1. Global Variance (GV) [16]

The GV $\boldsymbol{v}(\boldsymbol{y}) = \left[v(1), \cdots, v(D)\right]^\top$ is defined as the second order moment of the trajectory $\boldsymbol{y}$, and its d-th component is given as:

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right)^2. \tag{11}$$

4.2. Objective Function for GV-Constrained Training

$\{\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\}$ is updated by maximizing the following objective function L_{gvtrj} consisting of the trajectory HMM and GV likelihoods:

$$L_{\mathrm{gvtrj}} = P\left(\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{v}}\right)^{\omega_{\mathrm{v}}T}, \tag{12}$$

$$P\left(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{v}}\right) = \mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{v}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right), \boldsymbol{\Sigma}_{\mathrm{v}}\right), \tag{13}$$

where ω_{v} is a weight of the GV likelihood, $\boldsymbol{\Sigma}_{\mathrm{v}}$ is a covariance matrix of the GV, and $\boldsymbol{\lambda}_{\mathrm{v}}$ is a model parameter set of the GV. This algorithm updates the model parameters to make the GV of the generated parameter sequence close to natural one.

5. Modulation Spectrum (MS)-Constrained Trajectory Training

5.1. Modulation Spectrum [13]

The MS is defined as the power spectrum of the parameter sequence; i.e., temporal fuctuation of the parameter sequence is decomposed into individual modulation frequency components and their power values are represented as the MS. In this paper, the MS $\boldsymbol{s}(\boldsymbol{y})$ of the parameter sequence $\boldsymbol{y}$ is defined as:

$$\boldsymbol{s}(\boldsymbol{y}) = \left[\boldsymbol{s}(1)^\top, \cdots, \boldsymbol{s}(d)^\top, \cdots, \boldsymbol{s}(D)^\top\right]^\top, \quad (14)$$

$$\boldsymbol{s}(d) = \left[s_d(0), \cdots, s_d(f), \cdots, s_d\left(D'_{\mathrm{s}} - 1\right)\right]^\top \quad (15)$$

$$s_d(f) = R^2_{d,f} + I^2_{d,f} \quad (16)$$

$$= \left(\sum_{t=1}^{T} y_t(d)\cos kt\right)^2 + \left(\sum_{t=1}^{T} y_t(d)\sin kt\right)^2 \quad (17)$$

where $2D_{\mathrm{s}}$ is a length of Discrete Fourier Transform (DFT), $k = -\pi f/D_{\mathrm{s}}$ is a modulation frequency, f is a modulation frequency index, and D'_{s} is the number of MS dimension in each feature dimension, where $D'_{\mathrm{s}} < D_{\mathrm{s}}$. We can control the highest modulation frequency considered in this criterion by adjusting the ratio of D_{s} to D'_{s}. In this paper, the MS is calculated utterance by utterance.

5.2. Objective Function for MS-Constrained Trajectory Training

We integrate the MS compensation into the trajectory training. The objective function consists of both the trajectory likelihood and the MS likelihood as follows:

$$L_{\mathrm{mstrj}} = P\left(\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{s}}\right)^{\omega_{\mathrm{s}} T/D'_{\mathrm{s}}}, \quad (18)$$

$$P\left(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{s}}\right) = \mathcal{N}\left(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{s}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right), \boldsymbol{\Sigma}_{\mathrm{s}}\right), \quad (19)$$

where $\boldsymbol{\lambda}_{\mathrm{s}}$ is a model parameter set of the MS, and $\boldsymbol{\Sigma}_{\mathrm{s}}$ is a $D'_{\mathrm{s}}D$-by-$D'_{\mathrm{s}}D$ covariance matrix, and ω_{s} is a weight of the MS likelihood. The trajectory likelihood and the MS likelihood are normalized by the ratio of the number of feature dimensions when $\omega_{\mathrm{s}} = 1$. $\boldsymbol{\Sigma}_{\mathrm{s}}^{-1}$ is represented as $\left[\boldsymbol{p}_{\mathrm{s}}^{(1)}, \cdots, \boldsymbol{p}_{\mathrm{s}}^{(d)}, \cdots, \boldsymbol{p}_{\mathrm{s}}^{(D)}\right]$, where $\boldsymbol{p}_{\mathrm{s}}^{(d)}$ is $D'_{\mathrm{s}}D$-by-D matrix of which columns correspond to $\boldsymbol{s}(d)$. The MS likelihood works as a penalty term to alleviate the reduction of the temporal fluctuation of the generated parameter sequence. $\boldsymbol{\Sigma}_{\mathrm{s}}$ is in advance estimated using training data.

5.3. Estimation of Model Parameters

The model parameters are estimated in the same way as for GV-constrained trajectory training. Let L_{ms} be the MS likelihood $\mathcal{N}\left(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{s}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right), \boldsymbol{\Sigma}_{\mathrm{s}}\right)$. The logarithm function of L_{mstrj} is given by:

$$\log L_{\mathrm{mstrj}} = \log L_{\mathrm{trj}} + \omega_{\mathrm{s}} \frac{T}{D'_{\mathrm{s}}} \log L_{\mathrm{ms}}, \quad (20)$$

and the gradients of $\log L_{\mathrm{ms}}$ are given as:

$$\frac{\partial \log L_{\mathrm{ms}}}{\partial \boldsymbol{\mu}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{W} \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{s}_{\hat{\boldsymbol{q}}}, \quad (21)$$

$$\frac{\partial \log L_{\mathrm{ms}}}{\partial \boldsymbol{\Sigma}^{-1}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}^\top \mathrm{diag}_{3D}^{-1}\left[\boldsymbol{W}\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{s}_{\hat{\boldsymbol{q}}}\left(\boldsymbol{E}_{\hat{\boldsymbol{q}}} - \boldsymbol{W}\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right)\right], \quad (22)$$

where

$$\boldsymbol{s}_{\hat{\boldsymbol{q}}} = \left[\boldsymbol{s}_1'^\top, \cdots, \boldsymbol{s}_t'^\top, \cdots, \boldsymbol{s}_T'^\top\right]^\top, \quad (23)$$

$$\boldsymbol{s}'_t = \left[s_t(1), \cdots, s_t(d), \cdots, s_t(D)\right]^\top, \quad (24)$$

$$s_t(d) = 2\boldsymbol{f}_t(d)\boldsymbol{p}_{\mathrm{s}}^{(d)\top}\left(\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}(\hat{\boldsymbol{y}})\right), \quad (25)$$

$$\boldsymbol{f}_t(d) = \left[f_{t,d}(0), \cdots, f_{t,d}(f), \cdots, f_{t,d}\left(D'_{\mathrm{s}} - 1\right)\right]^\top, \quad (26)$$

$$f_{t,d}(f) = \hat{R}_{d,f}\cos kt + \hat{I}_{d,f}\sin kt, \quad (27)$$

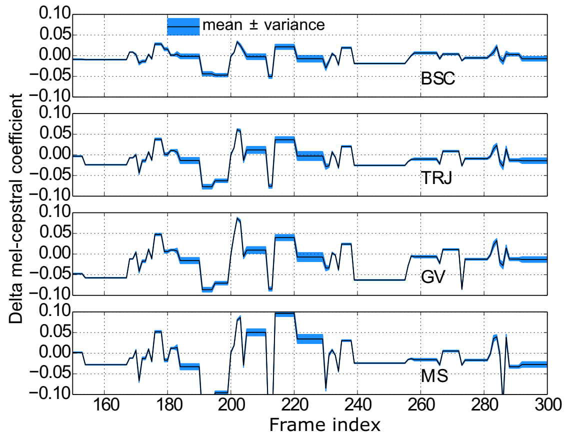

Figure 2: An example of the output probabilities of HMMs trained by several training algorithms.

where $\hat{R}_{d,f}$ and $\hat{I}_{d,f}$ are calculated using the d-th dimensional components of $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$.

5.4. Discussion

As reported in [23, 21], it is unnecessary to consider the MS in parameter generation because the HMM parameters are optimized to make the MS of the generated parameter sequence close to the natural one. Consequently, the basic computationally-efficient parameter generation algorithm is employed. This also enables to avoid the large footprint discussed in the parameter generation algorithm considering the MS [17].

Multi-Space probability Distribution (MSD)-HMM [25] is unsuitable for the implementation of the proposed algorithm for F_0 contour because the MS modeling of the non-continuous sequence is inaccurate [17]. To solve this problem, this paper adopted continuous F_0 modeling [26]. Moreover, 0-mean MS modeling is also adopted [17], which means that the MS is calculated from the F_0 contour that the utterance-level F_0 is subtracted.

Fig. 2 draws the output probabilities at each frame. We can see that the variance of trajectory training ("TRJ") is slightly larger than that of basic training ("BSC"), and the mean of GV-constrained ("GV") or MS-constrained ("MS") trajectory training is significantly changed compared to "TRJ." It is observed that the mean of "MS" tends to transit greatly from the neighbor HMM-state[1].

6. Experimental Evaluation

6.1. Experimental Condition

We trained a context-dependent phoneme Hidden Semi-Markov Model (HSMM) [27] for an English male speaker "RMS" from the CMU ARCTIC database [28]. We used 593 sentences from subset A for training and 100 sentences from subset B for evaluation. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 0th-through-24th mel-cepstral coefficients were extracted as a spectral parameter and log-scaled F_0 and 5 band-aperiodicity [29, 30] were extracted as excitation pa-

[1] Note that the frames that have same statistics correspond to the same HMM-state.

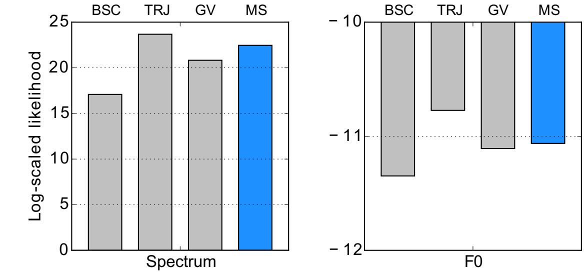

Figure 3: Trajectory likelihoods for the natural spectral parameter sequence and continuous F_0 contour.

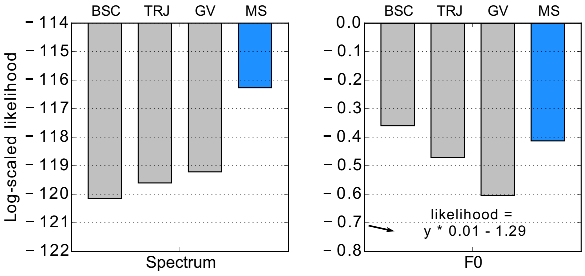

Figure 4: MS likelihoods for the natural spectral parameter sequence and continuous F_0 contour.

rameters. The STRAIGHT analysis-synthesis system [31] was employed for parameter extraction and waveform generation. The feature vector consisted of spectral and excitation parameters and their delta and delta-delta features. 5-state left-to-right HSMMs were used. The DFT length to calculate the MS was set to 2048. Diagonal covariance matrices were used in the HSMM, the GV probability density function, and the MS probability density function. The likelihood weight ω_{v} and ω_{s} were set to 0.5 [21] and 1.0, respectively. D'_s for spectrum and F_0 components were set to $D_s/2(= 50$ Hz) [23] and $D_s/10(= 10$ Hz) [13], respectively.

We compared the following training algorithms:

BSC: basic training (L_{basic}) [2]
TRJ: trajectory training (L_{trj}) [22]
GV: GV-constrained trajectory training (L_{gvtrj}) [21]
MS: proposed MS-constrained trajectory training (L_{mstrj})

These training algorithms were applied to spectral and F_0 components, and the "BSC" training algorithm was applied to aperiodic component. First, "BSC" training is performed to estimate the HMM parameter set $\boldsymbol{\lambda}$, then, "TRJ" training algorithm updates the parameters. "GV" and "MS" training algorithm updates after "TRJ" training. Note that the voiced/unvoiced regions of F_0 contour never changes in all training algorithms.

The trajectory likelihood and the MS likelihood for the natural parameter trajectories of the evaluation data were firstly calculated to analyze the effect of the proposed algorithm. Then, the speech quality of the synthetic speech are evaluated in the perceptual evaluation.

6.2. Objective Evaluation

Fig. 3 and Fig. 4 illustrate the trajectory likelihood L_{trj} and the MS likelihood L_{ms} for the natural parameter trajectories of the evaluation data, respectively. Note that the trajectory likelihood is normalized by the total number of frames T, and the MS likelihood is similarly normalized by the number of dimension D'_{s}. We can see that the proposed MS-constrained trajectory training ("MS") dramatically improves the MS likelihood compared to other training algorithms for the spectral component. On the other hand, the MS likelihood of "MS" is lower than that of the basic training algorithm ("BSC") for F_0 component. To discuss this result, we calculated the proposed objective function divided by T, and the values are -12.64 and -12.35 for "BSC" and "MS," respectively. Therefore, the proposed objective function is certainly improved by the proposed algorithm. Therefore, we can confirm the implementation of the proposed algorithm for both spectral and F_0 components.

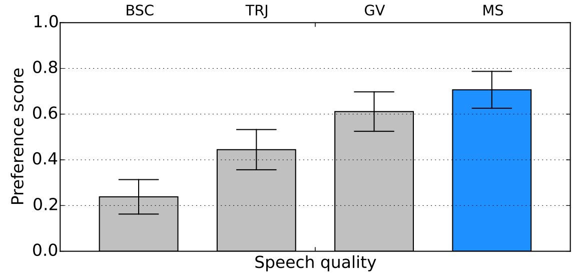

Figure 5: Preference scores on synthetic speech quality with 95% confidence interval.

6.3. Subjective Evaluation

A preference AB test was conducted by 6 listeners. We presented every pair of generated speech of 4 training algorithms in a random order, and we forced listeners to select speech sample that sounds better quality.

Fig. 5 illustrates the result. We can see that the proposed algorithm "MS" achieves the best score. This result demonstrates that the proposed MS-constrained trajectory training for spectral and F_0 components makes it possible to improve the synthetic speech quality while preserving the computationally-efficient generation ability.

7. Conclusion

This paper proposed a novel training algorithm for HMM-based speech synthesis in order to produce the high-quality speech while preserving the computationally-efficient generation algorithm. The Modulation Spectrum (MS) have been integrated into the trajectory HMM training for both spectral and F_0 components. The experimental results yielded the quality improvement in synthetic speech. As a future work, we combine both the proposed algorithm and the rich context modeling [32].

Acknowledgements: Part of this work was supported by JSPS KAKENHI Grant Number 26280060 and Grant-in-Aid for JSPS Fellows Grant Number 26 · 10354, and part of this work was executed under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation."

8. References

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[2] 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*, Budapest, Hungary, Apr. 1999, pp. 2347–2350.

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

[4] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 4, pp. 199–206, 2000.

[5] J. Yamagishi and T. Kobayashi., "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans., Inf. and Syst.*, vol. E90-D, no. 2, pp. 533–543, 2007.

[6] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans., Inf. and Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[7] K. Shirota, K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Integration of speaker and pitch adaptive training for HMM-based singing voice synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 2578–2582.

[8] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal diabilities: Voice banking and reconstruction," *Acoust. Sci. technol.*, vol. 33, pp. 1–5, 2012.

[9] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[10] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 3872–3876.

[11] T. Kano, S. Takamichi, S. Sakti, T. T. G. Neubig, and S. Nakamura, "Generalizing continuous-space translation of paralinguistic information," in *Proc. INTERSPEECH*, Lyon, France, Aug 2013, pp. 2614–2618.

[12] S. King and V. Karaiskos, "The blizzard challenge 2011," in *Proc. Blizzard Challenge workshop*, Turin, Italy, Sept. 2011.

[13] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify modulation spectrum in HMM-based speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 290–294.

[14] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-based post-filter for GMM-based voice conversion," in *Proc. APSIPA ASC*, Siem Reap, Cambodia, Dec. 2014.

[15] ——, "Modified modulation spectrum-based post-filter for HMM-based speech synthesis," in *Proc. GlobalSIP*, Atlanta, United States, Dec. 2014, pp. 710–714.

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

[17] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015.

[18] T. Nose and A. Ito, "Analysis of spectral enhancement using global variance in HMM-based speech synthesis," in *Proc. INTERSPEECH*, MAX Atria, Singapore, May 2014, pp. 2917–2921.

[19] H. Hwang, Y. Tsao, H. Wang, Y. Wang, and S. Chen, "Incorporating global variance in the training phase of GMM-based voice conversion," in *Proc. APSIPA*, Kaohsiung, Taiwan, Oct. 2013, pp. 1–6.

[20] M. Shannon and W. Byrne, "Fast, low-artifact speech synthesis considering global variance," in *Proc. ICASSP*, Vancouver, Canada, May. 2013, pp. 7869–7873.

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

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

[23] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015.

[24] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

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

[26] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio, Speech and Language*, vol. 19, no. 5, pp. 1071–1079, 2011.

[27] H. Zen, K. Tokuda, T. K. T. Masuko, and T. Kitamura, "Hidden semi-Markov model based speech synthesis system," *IEICE Trans., Inf. and Syst., E90-D*, no. 5, pp. 825–834, 2007.

[28] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases for speech synthesis research," in *Tech. Rep. CMU-LTI-03-177*, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, U.S.A., 2003.

[29] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *MAVEBA 2001*, Firentze, Italy, Sept. 2001, pp. 1–6.

[30] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006, pp. 2266–2269.

[31] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[32] S. Takamichi, T. Toda, Y. Shiga, S. Sakti, G. Neubig, and S. Nakamura, "Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 239–250, 2014.