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

Shinnosuke Takamichi\textsuperscript{1,2}, Tomoki Toda\textsuperscript{1}, Alan W. Black\textsuperscript{2}, Satoshi Nakamura\textsuperscript{1}

\textsuperscript{1} Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
\textsuperscript{2} Language Technologies Institute, Carnegie Mellon University (CMU)

\texttt{shinnosuke-t@is.naist.jp}

\section*{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.

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

\section{1. Introduction}

Statistical parametric speech synthesis based on Hidden Markov Models (HMMs) \cite{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 \cite{2}. In synthesis, the speech parameter trajectories are generated by maximizing the likelihood of the HMMs \cite{3}. This approach allows us not only to apply several techniques for flexibly controlling synthetic speech \cite{4, 5, 6} to various speech-based systems \cite{7, 8}, but also to build the speech synthesizer without complicated tuning compared to sample-based \cite{9} or deep neural nets-based \cite{10} speech synthesis. The further merit of HMM-based speech synthesis is the computationally-efficient speech parameter generation \cite{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 \cite{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 \cite{12}. To address this over-smoothing problem, we have found Modulation Spectrum (MS) \cite{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) \cite{16}. The MS of the generated trajectories is often lower than that of natural speech parameter trajectories. \cite{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 \cite{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, \cite{21} have proposed a metric to integrate the GV into the training stage instead of the synthesis stage. They reformulated trajectory HMMs \cite{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 \cite{23} on Gaussian Mixture Model (GMM)-based voice conversion \cite{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 \cite{2}, the conventional trajectory training \cite{22}, and GV-constrained trajectory training \cite{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.

\section{2. Basic Framework \cite{2}}

\subsection{2.1. Training Algorithm \cite{1}}

In HMM-based speech synthesis, A HMM parameter set $\lambda$ is estimated using the contextual factor sequence $X$ of input text and the speech feature sequence $Y = [Y_1^T, \ldots, Y_T^T]$ of $T$ frames as follows:

\begin{equation}
\lambda = \arg \max \ L_{\text{basic}} = \arg \max \ P(Y | X, \lambda). \quad (1)
\end{equation}

Copyright © 2015 ISCA
The output probability density function of HMM-state index \( q \) is given as:
\[
P(Y_t | X_q, \lambda) = N(y_t; \mu_q, \Sigma_q),
\]
where \( Y_t \) is given by \( 3D \)-dimensional joint static and dynamic feature vectors, \([y_t^1, \Delta y_t^1, \Delta \Delta y_t^1]^\top\), where \( y_t = [y_t(1), \ldots, y_t(d), \ldots, y_t(D)]^\top \) is represented as a \( D \)-dimensional vector at frame \( t \), and \( d \) is a dimensional index, \( N(\cdot; \mu, \Sigma) \) denotes Gaussian distribution of a mean vector \( \mu \) and a covariance matrix \( \Sigma \). The HMM parameter set \( \lambda \) consists \( Q \) HMM-states where each HMM-state has the individual mean vector \( \mu_q \) and covariance matrix \( \Sigma_q \).

2.2. Parameter Generation Algorithm [3]

Given the contextual factor sequence \( X \), the generated parameter sequence \( \tilde{y}_q = [\tilde{y}_q^1, \ldots, \tilde{y}_q^d, \ldots, \tilde{y}_q^D] \) is analytically determined by maximizing the output probability of the speech feature vector sequence target \( Y \) given \( X \) under a constraint \( Y = W X \) as follows:
\[
\tilde{y}_q = \arg \max_y P(W y|X, \tilde{q}, \lambda)
\]
\[
= R_q^{-1} q = (W^\top D_q^{-1} W)^{-1} W^\top D_q^{-1} E_q,
\]
where \( W \) is a \( 3DT \)-by-\( DT \) weight matrix to calculate the dynamic features [3]. \( \tilde{q} = [\tilde{q}_1, \ldots, \tilde{q}_i, \ldots, \tilde{q}_T] \) is the sub-optimum state sequence determined by maximizing state duration probability distribution function \( P(q_t X, \lambda) \), where \( q_t \) is a sub-optimum HMM-state at frame \( t \). The mean vector \( E_q = [\mu_q^1, \ldots, \mu_q^i, \ldots, \mu_q^D] \) and the covariance matrix \( D_q = \text{diag}_{3D} [\Sigma_q^1, \ldots, \Sigma_q^i, \ldots, \Sigma_q^D] \) are calculated using the corresponding HMM-state, where the notation \( \text{diag}_{3D} \) 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:
\[
L_{trj} = P(y_t | X_q, \tilde{q}, \lambda) = \mathcal{N}(y_t; \tilde{y}_q, R_q^{-1})
\]
(5)
The mean vector \( \tilde{y}_q \) is given by Eq. (4) and the inter-frame correlation is effectively modeled by the temporal covariance matrix \( R_q^{-1} \). In training, the HMM parameters are updated by maximizing \( L_{trj} \).

3.1. Estimation of Model Parameters

Here, let \( \mu = [\mu^1_1, \ldots, \mu^1_q, \ldots, \mu^q_1, \ldots, \mu^q_q]^\top \) and \([\Sigma^1, \ldots, \Sigma^i, \ldots, \Sigma^j, \ldots, \Sigma^T]^\top\) be the joint parameters of \( \mu_q \) and \( \Sigma_q \) over all HMM-states, respectively. The mean vector \( E_q \) and the precision matrix \( D_q^{-1} \) are represented as:
\[
E_q = S_q \mu,
\]
\[
D_q^{-1} = \text{diag}_{3D} [S_q \Sigma^{-1}]
\]
(7)
where \( S_q = [S_q^1, \ldots, S_q^q] \) is a \( 3D \)-by-\( 3DQ \) matrix, \( S_q^q \) is an \( M \)-dimensional vector of which the \( q \)-th component is 1 when \( q = \hat{q}_i \) and otherwise are 0 as shown in Fig. 1, and \( I_{3D} \) indicates the \( 3D \)-by-\( 3D \) identity matrix.

To optimize these model parameters for the objective function, we employ the steepest descent algorithm, as follows:
\[
\mu^{(i+1)} = \mu^{(i)} + \alpha \frac{\partial \log L_{trj}}{\partial \mu} |_{\mu=\mu^{(i)}}
\]
(8)
where \( \alpha \) is a learning rate, and \( i \) is an iteration index. \( \Sigma^{-1} \) are also optimized in the same manner. The gradients are given by:
\[
\frac{\partial \log L_{trj}}{\partial \mu} = S_q^\top D_q^{-1} W (y - \tilde{y}_q)
\]
(9)
\[
\frac{\partial \log L_{trj}}{\partial \Sigma^{-1}} = \frac{1}{2} S_q^\top \text{diag}_{3D} [W (R_q^{-1} + \tilde{y}_q \tilde{y}_q^\top - y y^\top) - E_q (\tilde{y}_q - y)^\top W^\top - W (\tilde{y}_q - y) E_q^\top].
\]
(10)


4.1. Global Variance (GV) [16]

The GV \( v(y) = [v(1), \ldots, v(D)]^\top \) is defined as the second order moment of the trajectory \( y \), and its \( d \)-th component is given as:
\[
v(d) = \frac{1}{T} \sum_{t=1}^{T} y_t(d) - \frac{1}{T} \sum_{t=1}^{T} y_t(d)
\]
(11)

4.2. Objective Function for GV-Constrained Training

\( \{\mu, \Sigma^{-1}\} \) is updated by maximizing the following objective function \( L_{gvtrj} \) consisting of the trajectory HMM and GV likelihoods:
\[
L_{gvtrj} = P(y | X, \tilde{q}, \lambda) P(v(y) | X, \tilde{q}, \lambda, \lambda_v)\omega_v^\top,
\]
(12)
\[
P(v(y) | X, \tilde{q}, \lambda, \lambda_v) = \mathcal{N}(v(y); v(\tilde{q}_v), \Sigma_v)
\]
(13)
where \( \omega_v \) is a weight of the GV likelihood, \( \Sigma_v \) is a covariance matrix of the GV, and \( \lambda_v \) 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 fluctuation 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 $s(y)$ of the parameter sequence $y$ is defined as:

$$ s(y) = [s(1)^T, \ldots, s(d)^T, \ldots, s(D)^T]^T, $$

$$ s(d) = [sa(0), \ldots, sa(f), \ldots, sa(D') - 1)^T $$

$$ s_d(f) = R_d^2(f) + I_d(f) $$

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

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

$$ L_{\text{ms}} = P(y|X, \hat{\tilde{q}}, \lambda) P(s(y)|X, \hat{\tilde{q}}, \lambda, \lambda_\Sigma) $$

where $N(s(y); \hat{\tilde{q}}, \Sigma)$ is a Multi-Space probability Distribution (MSD)-HMM [25].

$$ s(d) = \Sigma s, $$

$$ \Sigma_{a} = \Sigma_{a}' = \Sigma_{a}D' $$

where $\lambda_\Sigma$ is a model parameter set of the MS, and $\Sigma_a$ is a $D'D$-by-$D'D$ covariance matrix, and $\omega$ 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 $\omega_t = 1$. $\Sigma_a^{-1}$ is represented as $\left[p_1^{(s)}, \ldots, p_d^{(s)}, \ldots, p_s^{(D)}\right]$, where $p_1^{(s)}$ is $D'D$-by-$D'$. The MS likelihood works as a penalty term to alleviate the reduction of the temporal fluctuation of the generated parameter sequence. $\Sigma_a$ 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 $L_{\text{ms}}$ be the MS likelihood $N(s(y); \hat{\tilde{q}}, \Sigma_a)$. The logarithm function of $L_{\text{ms}}$ is given by:

$$ \log L_{\text{ms}} = \log L_{\text{ms}} + \omega_0 \frac{T}{D_a} \log L_{\text{ms}}, $$

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

$$ \frac{\partial \log L_{\text{ms}}}{\partial \mu} = S_q D_q^{-1} W R_q^{-1} s_q, $$

$$ \frac{\partial \log L_{\text{ms}}}{\partial \Sigma^{-1}} = -S_q \text{diag}(D) W R_q^{-1} s_q (E_q - W \hat{\tilde{q}}), $$

where $s_q = [s_1^T, \ldots, s_1^T, \ldots, s_s^T]^T$, $s_1 = [s_1(1), \ldots, s_1(d), \ldots, s_1(D)]^T$, $s_1(d) = 2f_{1d}(p^{(s)}(y) - s(y))$, $f_{1d}(d) = f_{1d}(0), \ldots, f_{1d}(f), \ldots, f_{1d}(D') - 1)^T$, $f_{1d}(f) = \hat{R}_{d, f} \cos kt + \hat{I}_{d, f} \sin kt$, $\hat{R}_{d, fi} \cos kt + \hat{I}_{d, fi} \sin kt$.

### 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 (“TR”) 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 “TR”. It is observed that the mean of “MS” tends to transit greatly from the neighbor HMM-state.

### 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 dataset [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 parameters.

---

1Note that the frames that have same statistics correspond to the same HMM-state.
parameters. 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, while preserving the computationally-efficient generation algorithm. The Modulation Spectrum (MS) have been integrated into the trajectory HMM training for both spectral and F0 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


