AN INTRODUCTION OF TRAJECTORY MODEL INTO HMM-BASED SPEECH SYNTHESIS

Heiga Zen, Keiichi Tokuda, Tadashi Kitamura

Department of Computer Science and Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya 466-8555 Japan

E-mail: {zen,tokuda,kitamura}@ics.nitech.ac.jp

ABSTRACT

In the synthesis part of a hidden Markov model (HMM) based speech synthesis system which we have proposed, a speech parameter vector sequence is generated from a sentence HMM corresponding to an arbitrarily given text by using a speech parameter generation algorithm. However, there is an inconsistency: although the speech parameter vector sequence is generated under the constraints between static and dynamic features, HMM parameters are trained without any constraints between them in the same way as standard HMM training. In the present paper, we introduce a trajectory-HMM, which has been derived from the HMM under the constraints between static and dynamic features, into the training part of the HMM-based speech synthesis system. Experimental results show that the use of trajectory-HMM training improves the quality of the synthesized speech.

1. INTRODUCTION

The increasing availability of large speech databases makes it possible to construct speech synthesis systems, referred to as corpus-based, by applying statistical learning algorithms. These systems, which can be automatically trained, not only generate natural and high quality synthetic speech but also can reproduce voice characteristics of the original speaker.

For constructing such a system, the use of hidden Markov models (HMMs) has become more popular. They have successfully been applied to modeling the sequence of speech spectra in speech recognition systems, and their performance have been improved by techniques utilizing the flexibility of the HMMs: context-dependent modeling, dynamic features, mixture of Gaussian distributions, parameter tying, speaker and environment adaptation. Speech synthesis systems based on the HMMs can be categorized as follows:

1. Transcription and segmentation of database [1]
2. Construction of inventory of speech segments [2, 3]
3. Run-time selection of multiple instances of speech segments [4, 5]
4. Speech synthesis from HMMs themselves [6–9]

In approaches 1–3, by using a waveform concatenation algorithms, e.g., PSOLA algorithm, a high quality synthetic speech could be produced. However, to obtain various voice qualities, a large amount of speech data is necessary, and it is difficult to collect, store, and label such speech data. On the other hand, in approach 4, voice characteristics of synthetic speech can be changed by transforming HMM parameters appropriately. From this point of view, we have proposed parameter generation algorithms for HMM-based speech synthesis [10], and constructed a speech synthesis system [9]. The main feature of the system is the use of dynamic feature: by inclusion of dynamic features in the state output vector, the dynamic features of speech parameter vector sequence generated in synthesis are constrained to be realistic, as defined by the parameters of the HMMs.

However, there is an inconsistency: although the speech parameter vector sequence is generated from the HMMs under the constraints between static and dynamic features, the HMM parameters are trained without any constraints between them in the same way as standard HMM training. In the present paper, we introduce a trajectory-HMM [11, 12], which has been derived from the HMM under the constraints between static and dynamic features, into the training part of the HMM-based speech synthesis system. Experimental results show that the use of trajectory-HMM training improves the quality of the synthesized speech.

The rest of this paper is organized as follows. Section 2 summarizes the overview of the HMM-based speech synthesis system. Section 3 shows the speech parameter generation algorithm and derivation of the trajectory-HMM. Section 4 describes its training algorithm. Result of subjective listening test is shown in Section 5. Concluding remarks and future plans are presented in the final section.

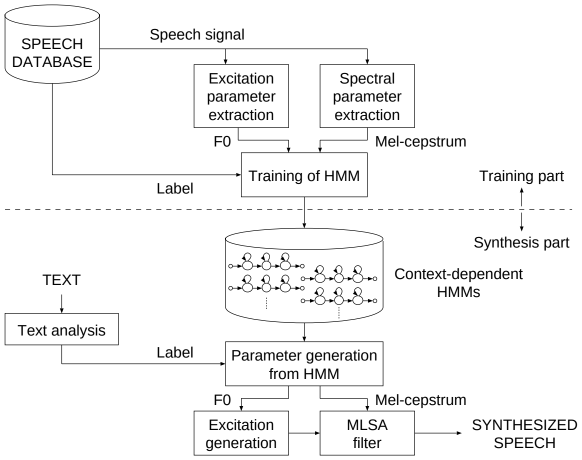


Fig. 1. An overview of the HMM-based speech synthesis system

2. HMM-BASED SPEECH SYNTHESIS SYSTEM

Figure 1 shows the overview of the HMM-based speech synthesis system. In the current system, an output vector of the HMM consists of a spectrum part and an excitation parts. The spectrum part consists of mel-cepstral coefficients including the zeroth coefficients, their delta and delta-delta coefficients. The excitation part consists of a log fundamental frequency (log F_0), its delta and delta-delta coefficients. In the training part, the spectrum part is modeled by continuous distribution HMMs (CD-HMMs) and the excitation part is modeled by multi-space probability distribution HMMs (MSD-HMMs) [13]. HMMs have state duration distributions to model the temporal structure of speech. As a result, the system models not only spectrum parameters but also excitation parameters and durations in a unified framework. To capture acoustic variation associated with contextual factors (e.g., phone identity factors, stress-related factors, locational factors) for spectrum, F_0 pattern and duration, we use context-dependent HMMs. However, as contextual factors increase, their combinations also increase exponentially. To overcome this problem, a decision-tree based context clustering technique [14] is applied to distributions for spectrum, F_0 and state duration in the same manner as in HMM-based speech recognition.

In the synthesis part, first, an arbitrarily given text to be synthesized is converted to a context-based label sequence. Secondly, according to the label sequence, a sentence HMM is constructed by concatenating context-dependent HMMs. State durations of the sentence HMM are determined so as to maximize the output probability of state durations. Then a sequence of mel-cepstral coefficients and log F_0 values including voiced/unvoiced decisions is determined in such a way that its output probability for the HMM is maximized using the speech parameter generation algorithm (case 1 in [10]). Finally, the speech waveform is synthesized directly from the generated mel-cepstral coefficients and log F_0 values by using the MLSA (Mel Log Spectrum Approximation) filter [15].

3. INTRODUCTION OF THE CONSTRAINTS BETWEEN STATIC AND DYNAMIC FEATURES

3.1. Speech parameter generation algorithm

For a given continuous distribution HMM λ, we model the speech parameter vector sequence $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \ldots, \boldsymbol{o}_T^\top\right]^\top$ in such a way that

$$P(\boldsymbol{o} \mid \lambda) = \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda)\, P(\boldsymbol{q} \mid \lambda), \tag{1}$$

is maximized with respect to $\boldsymbol{o}$, where $\boldsymbol{q} = \{q_1, \ldots, q_T\}$ is the state, T is the number of frames in the observation vector sequence. We assume that the speech parameter vector $\boldsymbol{o}_t$ consists of the M-dimensional static feature vector

$$\boldsymbol{c}_t = [c_t(1), c_t(2), \ldots, c_t(M)]^\top \tag{2}$$

and $1, \ldots, (D-1)$-th order dynamic feature vectors, that is

$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta^{(1)}\boldsymbol{c}_t^\top, \ldots, \Delta^{(D-1)}\boldsymbol{c}_t^\top\right]^\top, \tag{3}$$

where $\Delta^{(d)}\boldsymbol{c}_t$ is the d-th dynamic feature vector given by

$$\Delta^{(d)}\boldsymbol{c}_t = \sum_{\tau=-L_-^{(d)}}^{L_+^{(d)}} w^{(d)}(\tau)\boldsymbol{c}_{t+\tau}, \tag{4}$$

$w^{(d)}(\tau)$ is a window coefficient for calculating the d-th dynamic feature. Accordingly, when each state output probability distribution is assumed to be a single Gaussian distribution, $P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda)$ is given by

$$P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda) = \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{o}_t \mid \boldsymbol{\mu}_{q_t}, \boldsymbol{\Sigma}_{q_t}\right) = \mathcal{N}\left(\boldsymbol{o} \mid \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}\right), \tag{5}$$

where $\boldsymbol{\mu}_{q_t}$ and $\boldsymbol{\Sigma}_{q_t}$ are the $DM \times 1$ mean vector and the $DM \times DM$ covariance matrix, respectively, associated with q_t-th state, and

$$
\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{q}} &= \left[\boldsymbol{\mu}_{q_1}^\top, \boldsymbol{\mu}_{q_2}^\top, \ldots, \boldsymbol{\mu}_{q_T}^\top\right]^\top \\
\boldsymbol{\mu}_{q_t} &= \left[\Delta^{(0)}\boldsymbol{\mu}_{q_t}^\top, \Delta^{(1)}\boldsymbol{\mu}_{q_t}^\top, \ldots, \Delta^{(D-1)}\boldsymbol{\mu}_{q_t}^\top\right]^\top \\
\Delta^{(d)}\boldsymbol{\mu}_{q_t} &= \left[\Delta^{(d)}\mu_{q_t}(1), \Delta^{(d)}\mu_{q_t}(2), \ldots, \Delta^{(d)}\mu_{q_t}(M)\right]^\top \\
\boldsymbol{\Sigma}_{\boldsymbol{q}} &= \mathrm{diag}\left[\boldsymbol{\Sigma}_{q_1}, \boldsymbol{\Sigma}_{q_2}, \ldots, \boldsymbol{\Sigma}_{q_T}\right] \\
\boldsymbol{\Sigma}_{q_t} &= \mathrm{diag}\left[\Delta^{(0)}\boldsymbol{\Sigma}_{q_t}, \Delta^{(1)}\boldsymbol{\Sigma}_{q_t}, \ldots, \Delta^{(D-1)}\boldsymbol{\Sigma}_{q_t}\right] \\
\Delta^{(d)}\boldsymbol{\Sigma}_{q_t} &= \mathrm{diag}\left[\Delta^{(d)}\Sigma_{q_t}(1), \Delta^{(d)}\Sigma_{q_t}(2), \ldots, \Delta^{(d)}\Sigma_{q_t}(M)\right].
\end{aligned}
$$

Conditions (4) can be arranged in a matrix form:

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

where

$$
\begin{aligned}
\boldsymbol{c} &= \left[\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \ldots, \boldsymbol{c}_T^\top\right]^\top & (7)\\
\boldsymbol{W} &= \left[\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_T\right]^\top \otimes \boldsymbol{I}_{M\times M} & (8)\\
\boldsymbol{W}_t &= \left[\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \ldots, \boldsymbol{w}_t^{(D-1)}\right] & (9)\\
\boldsymbol{w}_t^{(d)} &= \Big[\underbrace{0, \ldots, 0}_{t-L_-^{(d)}-1}, w^{(d)}(-L_-^{(d)}), \ldots, w^{(d)}(L_+^{(d)}), \underbrace{0, \ldots, 0}_{T-\left(t+L_+^{(d)}\right)}\Big]^\top, & (10)
\end{aligned}
$$

$L_-^{(0)} = L_+^{(0)} = 0$, and $w^{(0)}(0) = 1$. It is obvious that Eq. (5) is maximized when $\boldsymbol{o} = \boldsymbol{\mu}_{\boldsymbol{q}}$, that is, the speech parameter vector sequence becomes a sequence of the mean vectors. This is a result of the independence assumption of state output probabilities of the HMM. To avoid this problem, we use the constraints between static and dynamic features. Under the constraints Eq. (6), maximizing Eq. (5) with respect to $\boldsymbol{o}$ is equivalent to that with respect to $\boldsymbol{c}$. By setting

$$\frac{\partial \log P\left(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{q}, \lambda\right)}{\partial \boldsymbol{c}} = \boldsymbol{0}, \tag{11}$$

we obtain a set of equations

$$\boldsymbol{R}_{\boldsymbol{q}}\boldsymbol{c} = \boldsymbol{r}_{\boldsymbol{q}}, \tag{12}$$

where

$$
\begin{aligned}
\boldsymbol{R}_{\boldsymbol{q}} &= \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W} = \boldsymbol{P}_{\boldsymbol{q}}^{-1} & (13)\\
\boldsymbol{r}_{\boldsymbol{q}} &= \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}. & (14)
\end{aligned}
$$

By solving Eq. (12), we can obtain the static feature vector sequence $\boldsymbol{c}$ maximizing Eq. (5) under the constraints given by Eq. (6).

3.2. Derivation of trajectory-HMM

However, the constraints between static and dynamic features given by Eq. (5) are used only in the synthesis part. These constraints should be used not only in the synthesis part but also in the training part. This inconsistency may degrade the quality of synthetic speech. Recently, we have proposed a statistical model, called trajectory-HMM [11]. It has been derived from the standard HMM under the constraints between static and dynamic features and it can solve above inconsistency.

By introducing the condition Eq. (6) into Eq. (5), the output probability of an observation vector sequence $\boldsymbol{o}$ (one of training data) for an standard HMM λ can be rewritten as a function of $\boldsymbol{c}$ as follows:

$$
\begin{aligned}
P\left(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{q}, \lambda\right) &= \mathcal{N}\left(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}\right) \\
&= K_{\boldsymbol{q}} \cdot \mathcal{N}\left(\boldsymbol{c} \mid \bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}\right), & (15)
\end{aligned}
$$

where $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$, $\boldsymbol{P}_{\boldsymbol{q}}$ and $K_{\boldsymbol{q}}$ are a mean vector corresponding to an utterance, a covariance matrix corresponding to an utterance, and normalization constant independent of $\boldsymbol{c}$, respectively. They are given as follows:

$$
\begin{aligned}
\bar{\boldsymbol{c}}_{\boldsymbol{q}} &= \boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{r}_{\boldsymbol{q}} & (16)\\
\boldsymbol{P}_{\boldsymbol{q}} &= \left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}\right)^{-1} & (17)\\
K_{\boldsymbol{q}} &= \frac{\sqrt{(2\pi)^{MT}\left|\boldsymbol{P}_{\boldsymbol{q}}\right|}}{\sqrt{(2\pi)^{DMT}\left|\boldsymbol{\Sigma}_{\boldsymbol{q}}\right|}} \cdot \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_{\boldsymbol{q}}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}} - \boldsymbol{r}_{\boldsymbol{q}}^\top\boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{r}_{\boldsymbol{q}}\right)\right\}. & (18)
\end{aligned}
$$

Under the condition Eq. (6), we should regard $\boldsymbol{c}$ rather than $\boldsymbol{o}$ as the random variable of the statistical model. From Eq. (15), we may define $P\left(\boldsymbol{c} \mid \boldsymbol{q}, \lambda\right)$ by

$$P\left(\boldsymbol{c} \mid \boldsymbol{q}, \lambda\right) = \mathcal{N}\left(\boldsymbol{c} \mid \bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}\right). \tag{19}$$

As a result, we obtain a new statistical model

$$P\left(\boldsymbol{c} \mid \lambda\right) = \sum_{\text{all } \boldsymbol{q}} P\left(\boldsymbol{c} \mid \boldsymbol{q}, \lambda\right) P\left(\boldsymbol{q} \mid \lambda\right), \tag{20}$$

to which we refer as “trajectory-HMM” [11]. Interestingly, the mean vector sequence $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$ given by Eq. (16) is exactly the same as the speech parameter vector sequence $\boldsymbol{c}$ obtained by solving Eq. (12). Thus, maximization of Eq. (19) can be considered as minimization of the error between training data $\boldsymbol{c}$ and generated speech parameter vector sequence $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$. In the next section, we describe the training algorithm of the trajectory-HMM based on a maximum likelihood criterion.

4. TRAINING ALGORITHM

In common with the HMM training, the EM algorithm may be used for the trajectory-HMM training. The auxiliary function for the trajectory-HMM can be written as

$$Q\left(\lambda, \lambda'\right) = \sum_{\text{all } \boldsymbol{q}} P\left(\boldsymbol{q} \mid \boldsymbol{c}, \lambda\right) \, \log P\left(\boldsymbol{c}, \boldsymbol{q} \mid \lambda'\right), \tag{21}$$

where λ is a set of current parameters and λ' is a new one. It can be shown that by substituting λ' which maximizes Eq. (21) for λ, the likelihood increases unless λ is a critical point of the likelihood. Unfortunately, it is intractable to calculate Eq. (21) because the output probability depends on entire state sequence. To avoid this difficulty, we apply the single state sequence approximation (Viterbi approximation). As a result, the problem is broken down into the following maximization problems:

$$\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{c}, \boldsymbol{q} \mid \lambda) \tag{22}$$

$$\lambda' = \arg\max_{\lambda} P(\boldsymbol{c}, \hat{\boldsymbol{q}} \mid \lambda). \tag{23}$$

4.1. Optimizing the trajectory-HMM parameters

First, we solve the maximization problem of Eq. (23). This problem is equivalent to maximizing

$$\log P(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) = -\frac{1}{2}\Big\{MT\log(2\pi) - \log\left|\boldsymbol{R}_{\boldsymbol{q}}\right| + \boldsymbol{c}^\top \boldsymbol{R}_{\boldsymbol{q}}\boldsymbol{c} + \boldsymbol{r}_{\boldsymbol{q}}^\top \boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{r}_{\boldsymbol{q}} - 2\boldsymbol{r}_{\boldsymbol{q}}\boldsymbol{c}^\top\Big\} \tag{24}$$

with respect to

$$\boldsymbol{m} = \left[\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top \ldots, \boldsymbol{\mu}_N^\top\right]^\top \tag{25}$$

$$\boldsymbol{\phi} = \left[\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1}, \ldots, \boldsymbol{\Sigma}_N^{-1}\right]^\top \tag{26}$$

where N is the total number of trajectory-HMM states.

By setting

$$\frac{\partial \log P(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)}{\partial \boldsymbol{m}} = \boldsymbol{0}, \tag{27}$$

we obtain a set of linear equations

$$\boldsymbol{S}_{\boldsymbol{q}}^\top \boldsymbol{W}\boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{W}^\top \boldsymbol{S}_{\boldsymbol{q}}\boldsymbol{\Phi}^{-1}\boldsymbol{m} = \boldsymbol{S}_{\boldsymbol{q}}^\top \boldsymbol{W}\boldsymbol{c} \tag{28}$$

for determination of $\boldsymbol{m}$ maximizing Eq. (24), where

$$\boldsymbol{\Phi}^{-1} = \operatorname{diag}(\boldsymbol{\phi}) \tag{29}$$

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \boldsymbol{S}_{\boldsymbol{q}}\boldsymbol{m} \tag{30}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} = \operatorname{diag}\left(\boldsymbol{S}_{\boldsymbol{q}}\boldsymbol{\phi}\right) \tag{31}$$

and $\boldsymbol{S}_{\boldsymbol{q}}$ is a $DT \times DMN$ matrix whose elements are 0 or 1 determined according to the state sequence $\boldsymbol{q}$. The dimensionality of Eq. (28) is DMN: although it could be tens of thousands, it is still small enough to solve the set of linear equations using currently available computational resources.

For maximizing Eq. (24) with respect to $\boldsymbol{\phi}$, we apply a steepest descent algorithm using the first derivative

$$\frac{\partial \log P(\boldsymbol{c}, \boldsymbol{q} \mid \lambda)}{\partial \boldsymbol{\phi}} = \frac{1}{2}\boldsymbol{S}_{\boldsymbol{q}}^\top \operatorname{diag}^{-1}\Big(\boldsymbol{W}\boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{W}^\top - \boldsymbol{W}\boldsymbol{c}\boldsymbol{c}^\top\boldsymbol{W}^\top + 2\boldsymbol{\mu}_{\boldsymbol{q}}\boldsymbol{c}^\top\boldsymbol{W}^\top + \boldsymbol{W}\bar{\boldsymbol{c}}_{\boldsymbol{q}}\bar{\boldsymbol{c}}_{\boldsymbol{q}}^\top\boldsymbol{W}^\top - 2\boldsymbol{\mu}_{\boldsymbol{q}}\bar{\boldsymbol{c}}_{\boldsymbol{q}}^\top\boldsymbol{W}^\top\Big) \tag{32}$$

because Eq. (32) is not a quadratic function of $\boldsymbol{\phi}$.

4.2. Obtaining most likely state sequence

Next, we discuss the maximization problem of Eq. (22). Based on the approximation

$$\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{c}, \boldsymbol{q} \mid \lambda) \tag{33}$$

$$= \arg\max_{\boldsymbol{q}} \frac{1}{K_{\boldsymbol{q}}} P(\boldsymbol{o}, \boldsymbol{q} \mid \lambda) \tag{34}$$

$$\approx \arg\max_{\boldsymbol{q}} P(\boldsymbol{o}, \boldsymbol{q} \mid \lambda), \tag{35}$$

we can use the Viterbi algorithm for the HMM. However, this approximation reduces the accuracy of the state alignment. To overcome this problem, an algorithm to obtain sub-optimal state sequence by using time-recursive likelihood computation and the Viterbi search with delayed decision has been derived in [12].

4.3. Training procedure

Training procedure of the trajectory-HMM can be summarized as follows:

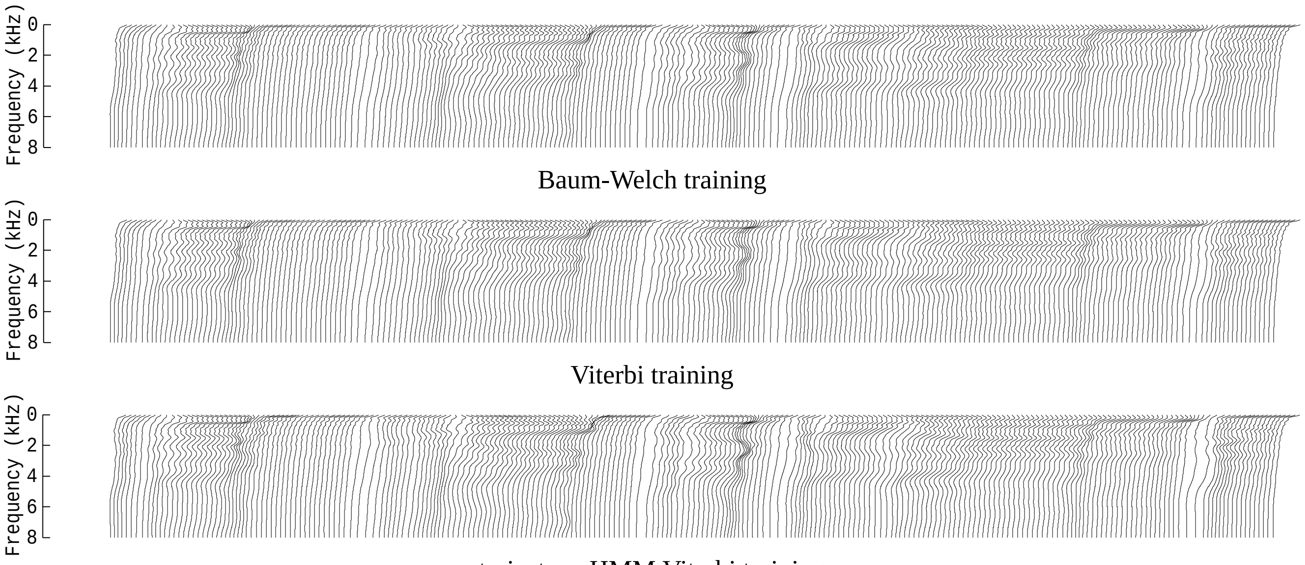
1. Initialize trajectory-HMM parameters by HMM parameters;
2. Select an initial state sequence for each training data by the Viterbi algorithm;
3. Update $\boldsymbol{m}$ by solving Eq. (28) according to the given state sequences;
4. Update $\boldsymbol{\phi}$ by the steepest descent algorithm using Eq. (32) according to the given state sequences;
5. Find state sequences by the algorithm proposed in [12];
6. If the model likelihood for the training data has not converged, go to step 3, otherwise stop iteration;

5. EXPERIMENT

5.1. Experimental conditions

We used first 1096 sentences from CMU ARCTIC database [16] uttered by a male speaker AWB for training. Speech signals were sampled at a rate of 16 kHz and windowed by a 25-ms Blackman window with a 5-ms shift, and mel-cepstral coefficients were obtained by a mel-cepstral analysis technique. Fundamental frequency (F_0) values were extracted by the ESPS get_f0 for at 5-ms intervals. Feature vector consisted of spectrum and F_0 parameter vectors: the spectrum parameter vector consisted of 25 mel-cepstral coefficients including the zeroth coefficient, their delta and delta-delta coefficients, and the F_0 parameter vector consisted of $\log F_0$, its delta and delta-delta. We used 5-state left-to-right with noskip HMM structure.

In this work, the following contextual factors were taken into account:



Baum-Welch training

Viterbi training

trajectory-HMM Viterbi training

Fig. 2. Generated spectra for a sentence fragment “tropic land”

- phoneme:
  - {before preceding, preceding, current, succeeding, after succeeding} phoneme
  - position of current phoneme in current syllable
- syllable:
  - number of phonemes at {preceding, current, succeeding} syllable
  - {stress[1], accent[2]} of {preceding, current, succeeding} syllable
  - position of current syllable in current {word, phrase}
  - number of {preceding, succeeding} {stressed, accented} syllables in current phrase
  - number of syllables {from previous, to next} {stressed, accented} syllable
  - vowel within current syllable
- word:
  - guess at part of speech of {preceding, current, succeeding} word
  - number of syllables in {preceding, current, succeeding} word
  - position of current word in current phrase
  - number of {preceding, succeeding} content words in current phrase
  - number of words {from previous, to next} content word
- phrase:
  - number of syllables in {preceding, current, succeeding} phrase
  - position in major phrase
  - ToBI endtone of current phrase
- utterance:
  - number of {syllables, words, phrases} in current utterance

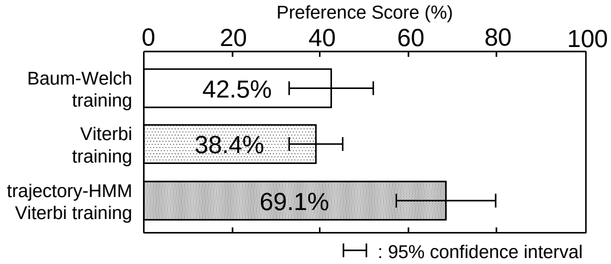


Fig. 3. Preference scores

These contextual factors were extracted using feature extraction functions of the Festival speech synthesis system [17] from the utterance information included in the database. We applied a decision-tree based context clustering technique based on an MDL criterion [18] to distributions for spectrum, F_0, and state duration. For spectrum and F_0, decision trees were constructed for each state position. The resultant trees for spectrum, F_0, and state duration had 978, 1180, and 449 leaves in total, respectively.

5.2. Experimental results

To compare the effects of the training algorithm for the quality of synthetic speech, we constructed 3 acoustic models trained by different algorithms. First, we trained the HMMs by using the Baum-Welch (EM) algorithm (model parameters maximizing Eq. (1) were estimated). Then we constructed the trajectory-HMMs by the algorithm described in Section 3 using the Baum-Welch trained HMMs as ini-

[1] The lexical stress of the syllable as specified from the lexicon entry corresponding to the word related to this syllable.

[2] An intonational accent of the syllable predicted by a CART tree (0 or 1).

tial models (model parameters maximizing Eq. (19) were estimated). To investigate the effect of the Viterbi approximation, we also trained the HMMs by the Viterbi training using the Baum-Welch trained HMMs as initial models (model parameters maximizing Eq. (5) were estimated). Both the trajectory-HMMs and the Viterbi trained HMMs were not iteratively estimated and the model parameters in the F_0 part were not updated. Therefore, synthesized F_0 patterns and durations from these 3 models were exactly the same.

Figure 2 shows the sequences of speech spectra calculated from the generated mel-cepstrum vectors from the Baum-Welch trained HMMs, the Viterbi trained HMMs, and the trajectory-HMMs for a sentence fragment "tropic land" taken from a sentence not included in the training data. It is seen from Fig. 2, formant structure of the sequence of speech spectra generated from the trajectory-HMMs gets clearer than that from other models.

To evaluate the effectiveness of the trajectory-HMM training, subjective listening test was conducted. We compared the quality of the synthesized speech generated from the Baum-Welch trained HMMs, the Viterbi trained HMMs, and the trajectory-HMMs by paired comparison tests. Subjects were 8 persons, and presented a pair of synthesized speech from different models in random order and then asked which speech sounded more natural. For each subject, 20 test sentences were chosen at random from 42 test sentences not contained in the training data sentence set.

Figure 3 shows the preference scores. It can be seen from the figure that the use of the trajectory-HMM training improved the quality of synthetic speech. Although the Viterbi approximation was used both in the Viterbi training and the trajectory-HMM Viterbi training, the quality of synthetic speech generated from the trajectory-HMMs was better than that from the Viterbi trained HMMs. It indicates that this improvement was achieved by the introduction of the trajectory-HMM training, not by the Viterbi approximation.

6. CONCLUSION

In the present paper, we introduced a trajectory-HMM which we had proposed into the training part of the HMM-based speech synthesis system. Experimental results showed that the use of trajectory-HMM training improved the quality of the synthesized speech.

Future work includes applying the trajectory-HMM training not only for the spectrum part but also the F_0 part. The synthesized speech generated by the latest system can be found at [19].

7. ACKNOWLEDGEMENT

Authors would like to thank Prof. Takao Kobayashi, Dr. Takashi Masuko, and Dr. Yoshihiko Nankaku for helpful discussions.

8. REFERENCES

[1] A. Ljolje, J. Hirschberg, and J.P.H. van Santen, "Automatic speech segmentation for concatenative inventory selection," in *Progress in speech synthesis*, J.P.H. van Santen, R.W. Sproat, J.P. Olive, and J. Hirshberg, Eds. Springer-Verlag, 1997.

[2] R.E. Donovan and P.C. Woodland, "Automatic speech synthesizer parameter estimation using HMMs," in *Proc. of ICASSP*, 1995, pp. 640–643.

[3] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe, "Recent improvements on Microsoft's trainable text-to-speech system - Whistler," in *Proc. of ICASSP*, 1997, pp. 959–962.

[4] H. Hon, A. Acero, X. Huang, J. Liu, and M. Plumpe, "Automatic generation of synthesis units for trainable text-to-speech synthesis," in *Proc. of ICASSP*, 1998, pp. 293–306.

[5] R.E. Donovan and E.M. Eide, "The IBM trainable speech synthesis system," in *Proc. of ICSLP*, 1998, vol. 5, pp. 1703–1706.

[6] A. Falaschi, M. Giustiniani, and M. Verola, "A hidden Markov model approach to speech synthesis," in *Proc. of Eurospeech*, 1989, pp. 187–190.

[7] M. Giustiniani and P. Pierucci, "Phonetic ergodic HMM for speech synthesis," in *Proc. of Eurospeech*, 1991, pp. 349–352.

[8] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," in *Proc. of ICASSP*, 1996, pp. 389–392.

[9] 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. of Eurospeech*, 1999, vol. 5, pp. 2347–2350.

[10] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. of ICASSP*, 1995, pp. 660–663.

[11] K. Tokuda, H. Zen, and T. Kitamura, "Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features," in *Proc. of Eurospeech 2003*, 2003, pp. 865–868.

[12] H. Zen, K. Tokuda, and T. Kitamura, "A Viterbi algorithm for a trajectory model derived from HMM with explicit relationship between static and dynamic features," in *Proc. of ICASSP 2004*, 2004, to appear.

[13] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. of ICASSP*, 1999, pp. 229–232.

[14] J.J. Odell, *The use of context in large vocabulary speech recognition*, Ph.D. thesis, Cambridge University, 1995.

[15] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. of ICASSP*, 1983, pp. 93–96.

[16] J. Kominek and A.W. Black, "CMU ARCTIC databases for speech synthesis," Tech. Rep. CMU-LTI-03-177, Carnegie Mellon University, 2003.

[17] A.W. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," `http://www.festvox.org/festival/`.

[18] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. of Eurospeech*, 1997, pp. 99–102.

[19] `http://kt-lab.ics.nitech.ac.jp/˜zen/HTS/`.