To realize a conversational agent system, we would need HSMMs.

MLLR ADAPTATION FOR
HIDDEN SEMI-MARKOV MODEL BASED SPEECH SYNTHESIS

Junichi Yamagishi, Takashi Masuko†, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 Japan
Email: {junichi.yamagishi,masuko,takao.kobayashi}@ip.titech.ac.jp

Abstract

This paper describes an extension of maximum likelihood linear regression (MLLR) to hidden semi-Markov model (HSMM) and presents an adaptation technique of phoneme/state duration for an HMM-based speech synthesis system using HSMMs. The HSMM-based MLLR technique can realize the simultaneous adaptation of output distributions and state duration distributions. We focus on describing mathematical aspect of the technique and derive an algorithm of MLLR adaptation for HSMMs.

1. Introduction

To realize a conversational agent system, we would need to crystallize speech synthesis system with abilities to change voice characteristics, prosodic features, and emotional expressions arbitrarily. Although the state-of-the-art concatenative speech synthesis system can synthesize natural sounding speech, conversions of voice characteristics, prosodic features, and emotional expressions are still difficult problems even for these systems.

On the other hand, an HMM-based speech synthesis system proposed in [1][2] enables us to change not only spectral features but also prosodic features. In this system, fundamental frequency, and phoneme/state duration are modeled simultaneously in a framework of HMM and converted based on a maximum likelihood linear regression (MLLR) [3] model adaptation framework. Using this technique, we can change voice characteristics of synthetic speech so as to mimic an arbitrary target speaker’s voice with a small amount of speech data uttered by the target speaker. Although we demonstrated that this technique works well and effectively, the implemented method of phoneme/state duration modeling [4] was not mathematically rigorous. In [2][4], state duration probabilities were estimated approximately on the trellis which was obtained in the embedded training of HMM without explicit state duration duration probability for the simplification of implementation. Moreover, the conversion method of phoneme/state duration [2] was based on the statistics from the trellis of HMMs without explicit state duration probability. However, it is possible to incorporate explicit state duration probability into HMMs and reestimate the state duration probability using EM algorithm [5, 6, 7]. This HMM is referred to as “hidden semi-Markov model” (HSM). Using the HSMMs, we can develop a rigorously adapted algorithm of state duration mathematically.

In this paper, we describe an extension of MLLR to HSMM and presents an adaptation algorithm of state duration based on the statistics from HSMMs with explicit state duration probability distributions. The proposing HSMM-based MLLR technique enables us to realize the simultaneous adaptation of output distributions and state duration distributions in a mathematically rigorous way. In this paper, we will focus on describing mathematical aspect of the technique and derive an algorithm of MLLR adaptation for HSMMs. This paper is organized as follows. Section 2 describes a brief review of hidden semi-Markov model. Section 3 presents the HSMM-based MLLR adaptation technique, and detailed descriptions for output distributions and state duration distributions are given in Sect. 3.1 and 3.2, respectively. Section 4 summarizes our findings.

2. Hidden Semi-Markov Model

Before deriving HSMM-based MLLR algorithm, we describe the training method [5, 6, 7] of hidden semi-Markov model. First we explain reestimation process of parameters of HSM. An N-state HSM λ is specified by initial state probability π = {πi}N i=1, state transition probability A = {aij}N i,j=1,i̸=j, output probability distribution B = {bi}N i=1, and state duration probability distribution D = {di}N i=1. Here we assume that bi(·) is Gaussian distribution of speech frame vector α at state i characterized by mean vector µi and diagonal covariance matrix Σi, where n is the dimensionality of the speech frame vector and

\[ b_i(\alpha) = \frac{1}{\sqrt{2\pi |\Sigma_i|}} e^{-\frac{1}{2}(\alpha - \mu_i)^T \Sigma_i^{-1}(\alpha - \mu_i)} \]  

For a given observation sequence O = {α1, · · · , αT} and HSM λ = (A, B, π, D), observation probability of O is written as

\[ P(O|\lambda) = \sum_{i=1}^{N} \sum_{d=1}^{d_i} \gamma_t^d(i) \quad \forall \ t \in [1, T] \]  

where t is frame number of the observation sequence at time t, T is total number of frames of the observation sequence and

\[ d = \frac{1}{\sqrt{2\pi \sigma_i^2}} e^{-\frac{(d-m_i)^2}{2\sigma_i^2}}. \]
\( \gamma_t^d(i) \) is given by

\[
\gamma_t^d(i) = \sum_{j=1}^{N} \alpha_{t-d}(j)a_{ji}p_i(d) \prod_{s=t-d+1}^{t} b_i(o_s) \beta_t(i). \tag{4}
\]

In this equation, \( \alpha_t(i) \) and \( \beta_t(i) \) are forward and backward parameters defined by

\[
\alpha_t(i) = \sum_{d=1}^{T} \sum_{j=1}^{N} \alpha_{t-d}(j)a_{ji}p_i(d) \prod_{s=t-d+1}^{t} b_i(o_s) \tag{5}
\]

\[
\beta_t(i) = \sum_{d=1}^{T-t} \sum_{j=1}^{N} a_{ij}p_j(d) \prod_{s=t+1}^{t+d} b_j(o_s) \beta_{t+d}(j) \tag{6}
\]

where \( a_{ji} \) is the transition probability from state \( j \) to \( i \), \( \alpha_0(i) = \pi_i \), and \( \beta_T(i) = 1 \).

Equation (3) is the objective function in maximum likelihood estimation. However, it is not easy to maximize the objective function directly. Therefore it is convenient to use an auxiliary function which has no hidden parameters and increases the objective function if the auxiliary function is maximized. The auxiliary function \( Q(\lambda', \lambda) \) of current parameters \( \lambda' \) and new parameter \( \lambda \) is defined as follows:

\[
Q(\lambda', \lambda) = \sum_{q \in \Theta} \sum_{l \in \mathcal{T}} P(q, l|O, \lambda') \log P(O, q, l|\lambda) \tag{7}
\]

where \( q = \{q_1, q_2, \ldots, q_T\} \) is a possible state sequence, \( \Theta \) is a set of all possible state sequences, \( \mathcal{T} = \{t_1, t_2, \ldots, t_K\} \) is a possible sequence of state duration for the observation sequence \( O \) and the state sequence \( q \), and \( \mathcal{T} \) is a set of all possible state duration sequences. It is noted that the state sequence \( q \) goes through \( K \) states and \( \sum_{k=1}^{K} t_k = T \).

For given observation sequence \( O \) and model \( X' \), we can derive reestimation formulae of parameters of \( \lambda \) which maximize the auxiliary function. The parameter set \( \lambda = (A, \pi, \sigma, D) \) can be reestimated as

\[
\hat{\mu}_i = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_{it}^d \mu_i}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_{it}^d} \tag{8}
\]

\[
\Sigma_i = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_{it}^d (\mu_i - \mu_i)(\mu_i - \mu_i)^\top}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_{it}^d} \tag{9}
\]

\[
\bar{\alpha}_{ij} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \alpha_{t-d}(i)a_{ji}p_i(d) \prod_{s=t-d+1}^{t} b_i(o_s) \gamma_t^d(j)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j)} \tag{10}
\]

\[
\bar{\pi}_i = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \alpha_{t-d}(i)a_{ji}p_i(d) \prod_{s=t-d+1}^{t} b_i(o_s) \beta_t(i)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \alpha_{t-d}(i) \beta_t(i)} \tag{11}
\]

\[
\bar{a}_{ij} = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \alpha_{t-d}(i)a_{ji}p_i(d) \prod_{s=t-d+1}^{t} b_i(o_s) \beta_t(j)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \alpha_{t-d}(i) \beta_t(j)} \tag{12}
\]

\[
\bar{\pi}_i = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \alpha_{t-d}(i)a_{ji}p_i(d) \prod_{s=t-d+1}^{t} b_i(o_s) \beta_t(j)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \alpha_{t-d}(i) \beta_t(j)} \tag{13}
\]

Note that it is not necessary to reestimate transition and initial state probabilities if left-to-right models without skip paths are used.

### 3. Hidden Semi-Markov Model-based MLLR Adaptation

We consider here a technique for performing maximum likelihood linear regression (MLLR) adaptation [3] for output distributions and state duration distributions of hidden semi-Markov model. HSMM-based MLLR estimates two kinds of regression matrices, one is for output distribution and, the other is for state duration distribution. The problem of HSMM-based MLLR can be written as follows:

\[
W_{\text{max}} = \arg \max_{W} P(O|\lambda, W) \tag{14}
\]

\[
X_{\text{max}} = \arg \max_{X} P(O|\lambda, X). \tag{15}
\]

Here we reestimate these regression matrices which maximize the auxiliary function and iteratively maximize the observation likelihood using EM algorithm.

#### 3.1. Adaptation for Output Distribution

First we derive HSMM-based MLLR algorithm for output distribution. We make assumption that each output distribution is a single Gaussian distribution given by (1).

We assume that the adaptation of the mean vector is achieved by affine transformation. That is, the adapted mean vector \( \hat{\mu}_i \) at state \( i \) is obtained by applying a transformation/regression matrix \( W_i \) to the extended mean vector \( \xi_i \):

\[
\hat{\mu}_i = W_i \xi_i \tag{16}
\]

where \( W_i \) is \( n \times (n+1) \) matrix which transforms the mean vector of output distribution, and \( \xi_i \) is \( (n+1) \) dimensional vector defined as

\[
\xi_i = [1, \mu_i^\top]^\top. \tag{17}
\]

Using (16), the adapted output distribution becomes

\[
b_i(o) = \frac{1}{\sqrt{2\pi}^n |\Sigma_i|} e^{-\frac{1}{2}(o - W_i \xi_i)^\top \Sigma_i^{-1}(o - W_i \xi_i)} \tag{18}
\]

In the auxiliary function (7), only the output distributions \( b_i() \) are affected in the reestimation processes for regression matrix \( W_i \). Hence the auxiliary function (7) can be rewritten as

\[
Q_b(\lambda', \lambda) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{d=1}^{t} P(q_t = i, l_t = d|O, \lambda', W_i) \cdot \log \left( \prod_{s=t-d+1}^{t} b_i(o_s) \right) \tag{19}
\]
Differentiating (19) with respect to $W_i$, we have
\[
\frac{\partial Q_p(X', \lambda)}{\partial W_i} = \sum_{t=1}^{T} \sum_{d=1}^{R} P(q_t = i, l_t = d|O, X', W_i) \frac{\partial}{\partial W_i} \log \left( \prod_{s=t-d+1}^{t} b_i(o_s) \right).
\] (20)

Note that $P(q_t = i, l_t = d|O, X', W_i) = \gamma^d_t(i)$ and the partial differentiation of product of output distribution becomes as follows:
\[
\frac{\partial}{\partial W_i} \log \left( \prod_{s=t-d+1}^{t} b_i(o_s) \right) = \sum_{s=t-d+1}^{t} \Sigma^{-1}_i (o_s - W_i \xi) \xi^\top_i.
\] (21)

Then substituting (20) for (21) and equating the result to zero yield
\[
\sum_{t=1}^{T} \sum_{d=1}^{R} \gamma^d_t(i) \Sigma^{-1}_i \sum_{s=t-d+1}^{t} o_s \xi^\top_i = \sum_{t=1}^{T} \sum_{d=1}^{R} \Sigma^{-1}_i \sum_{s=t-d+1}^{t} o_s \Sigma^{-1}_i (o_s - W_i \xi) \xi^\top_i.
\] (22)

In general, it is impossible to estimate the MLLR regression matrices for each distribution because the amount of adaptation data of a target speaker is small. Therefore, MLLR makes use of tree structures to group the distributions in the model, and to tie the regression matrices at each group. Tying of each regression matrices makes it possible to adapt distributions which have no adaptation data. When the regression matrix $W_z$ is tied across $R$ distributions, (22) becomes
\[
\sum_{t=1}^{T} \sum_{d=1}^{R} \sum_{r=1}^{t} \gamma^d_t(r) \Sigma^{-1}_r \sum_{s=t-d+1}^{t} o_s \xi^\top_i = \sum_{t=1}^{T} \sum_{d=1}^{R} \Sigma^{-1}_r \sum_{s=t-d+1}^{t} o_s \Sigma^{-1}_r (o_s - W_z \xi) \xi^\top_i.
\] (23)

Except for the cases $R = 1$ or $\xi = \xi_2 = \cdots = \xi_R$, we can solve this equation with respect to $W_z$. The computing procedure of $W_z$ is almost the same as HMM-based MLLR [3].

### 3.2. Adaptation for State Duration Distribution

Next we derive MLLR algorithm for state duration distribution. We assume that each state duration distribution is given by a single Gaussian distribution (2).

We further assume that the adaptation of the mean is achieved by linear regression. That is, the adapted mean $\hat{m}_i$ is obtained by applying a regression matrix $X_i$ to the extended mean vector $\phi_i$,
\[
\hat{m}_i = X_i \phi_i.
\] (24)

where $X_i$ is a 2-dimensional vector which transforms the mean of state duration distribution at state $i$, and $\phi_i$ is defined as
\[
\phi_i = [1, m_i]^\top.
\] (25)

Using (24), the adapted state duration distribution becomes
\[
p_i(d) = \frac{1}{\sqrt{2\pi} \sigma^2_i} e^{-\frac{(d - X_i \phi_i)^2}{2\sigma^2_i}}.
\] (26)

In the auxiliary function (7), only the state duration distributions $p_i(\cdot)$ are affected in the reestimation processes for regression matrix $X_i$. Hence the auxiliary function (7) can be rewritten as
\[
Q_p(X', \lambda) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{d=1}^{R} P(q_t = i, l_t = d|O, X', X_i) \cdot \log p_i(d)
\] (27)

Differentiating (27) with respect to $X_i$, we obtain
\[
\frac{\partial Q_p(X', \lambda)}{\partial X_i} = \sum_{t=1}^{T} \sum_{d=1}^{R} P(q_t = i, l_t = d|O, X', X_i) \cdot \frac{\partial}{\partial X_i} \log p_i(d).
\] (28)

The partial differentiation of state duration distribution becomes
\[
\frac{\partial}{\partial X_i} \log p_i(d) = \frac{1}{\sigma_i} \left( d \phi_i^\top - X_i \phi_i \phi_i^\top \right).
\] (29)

Then substituting (28) for (29) and equating the result to zero yield
\[
\sum_{t=1}^{T} \sum_{d=1}^{R} \gamma^d_t(i) \frac{d}{\sigma_i} \phi_i^\top = \sum_{t=1}^{T} \sum_{d=1}^{R} \gamma^d_t(i) X_i \phi_i \phi_i^\top.
\] (30)

To adapt distributions which have no adaptation data, the regression matrix $X_z$ is tied across $R$ distributions. Thus, equation (30) becomes
\[
\sum_{t=1}^{T} \sum_{d=1}^{R} \sum_{r=1}^{t} \gamma^d_t(r) \frac{d}{\sigma_i} \phi_i^\top = \sum_{t=1}^{T} \sum_{d=1}^{R} \sum_{r=1}^{t} \gamma^d_t(r) X_z \phi_i \phi_i^\top.
\] (31)

Except for the cases $R = 1$ or $\phi_1 = \phi_2 = \cdots = \phi_R$, we can solve this equation with respect to $X_z$.
\[
X_z = \left( \sum_{t=1}^{T} \sum_{d=1}^{R} \gamma^d_t(r) \frac{d}{\sigma_i} \phi_i^\top \right)^{-1} \left( \sum_{t=1}^{T} \sum_{d=1}^{R} \sum_{r=1}^{t} \gamma^d_t(r) \phi_i \phi_i^\top \right).
\] (32)

### 3.3. Implementation Problem

Incorporating state duration probability greatly increases computational costs. To keep the computational costs in a reasonable range, we truncate the duration probability $p_i(d)$ at a given maximum duration value $D$. Using this maximum duration value, (23) and (32) become as follows:
\[
\sum_{t=1}^{T} \sum_{r=1}^{R} \sum_{d=1}^{\min(t, D)} \gamma^d_t(r) \Sigma^{-1}_r \sum_{s=t-d+1}^{t} o_s \xi^\top_i = \sum_{t=1}^{T} \sum_{r=1}^{R} \sum_{d=1}^{\min(t, D)} \gamma^d_t(r) d \Sigma^{-1}_r W_z \xi \xi^\top_i.
\] (33)
and

\[
X_z = \left( \sum_{t=1}^{T} \sum_{r=1}^{R} \sum_{d=1}^{\min(t,D)} \frac{\gamma^d(r)}{\sigma^2} \phi_r^T \right) \cdot \left( \sum_{t=1}^{T} \sum_{r=1}^{R} \sum_{d=1}^{\min(t,D)} \frac{\gamma^d(r)}{\sigma^2} \phi_r \phi_r^T \right)^{-1}
\]

(34)

where \( D \) is maximum duration value within any state. Equation (33) is equivalent to HMM-based MLLR [3] if \( D = 1 \) and \( p_i(1) = 1 \) are given.

3.4. Discussions

In the conventional method [2], the adapted regression matrix for state duration distributions was not reestimated in a strict way. Therefore it is not clear whether the objective function converges to a critical points. On the other hand, it is straightforward to prove that the objective function \( P(O|\lambda, W_i, X_i) \) converges to a critical points using the proposed reestimation formulae. Furthermore, the proposed adaptation technique for output distributions includes the HMM-based MLLR of [3]. And also we can apply the proposed adaptation technique for output distributions to multi-space probability distribution [8].

An issue of the proposed adaptation technique for state duration is that there is a possibility of being adapted to a negative mean value of the state duration distribution because we assume normal state duration distribution. Since the state duration distribution should be defined in the positive area essentially, we need to assume a distribution defined in the positive area, such as lognormal, gamma, or Poisson distribution. The extension to these distributions is our future work.

4. Conclusions

This paper describes an extension of MLLR to hidden semi-Markov model and presents an adaptation technique of phoneme/state duration. HSMM-based MLLR algorithm can perform the simultaneous adaptation of output distributions and state duration distributions. Although we have focused on describing mathematical aspect of the technique and deriving the algorithm of MLLR adaptation for HSMMs in this paper, an HSMM-based speech synthesis system has already developed [9] and thus the proposed technique can be easily implemented.

5. References


