SWITCHING ACAUSAL FILTERS FOR SPEECH MODELING

Yasuhiro Minami and Hirokazu Kameoka
NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
minami@cslab.kecl.ntt.co.jp

ABSTRACT

This paper shows a unified model of dynamical systems in speech processing that includes speech recognition and pitch modeling. For this purpose, we propose the use of switching acausal filters (SAFs), which exchange multiple acausal filters. These filters are defined by identical linear dynamical systems that exchange the roles of observation value and system input. This paper describes the formulation of recognition, training, and feature generation methods for SAFs, which can be applied to several previously proposed speech models. As an example, we show that an HMM with dynamic features and our F0 control method can be modeled by the proposed formulation. An HMM synthesis method can also be modeled using the formulations. From these results, we demonstrate the unification capability of SAFs.

Index Terms—Acausal filter, HMM, delta features, delta-delta features, Kalman filter.

1. INTRODUCTION

Various models for speech recognition have been developed recently with the aim of capturing speech dynamics [1][2][3][4][5][6]. However, the theoretical backgrounds of these methods are so widely varied that they seem to be completely different from each other. It is thus important to compare these models systematically in order to determine how effectively the models capture the nature of speech dynamics.

The work in [7] classified various models for speech recognition from the viewpoint of a segmental model. However, that paper did not classify HMMs with the delta and delta-delta features as a segmental model. In speech recognition, HMMs with delta features are thought to comprise two independent processes: calculation of delta and delta-delta features and transition of hidden states. Our assumption is that since delta and delta-delta features as well as state transitions characterize the speech dynamics, these two processes should be unified in a single dynamical system of segmental models. We have found that the speech synthesis method based on HMMs with delta and delta-delta features [8] can be formulated precisely as a specified fixed-lag Kalman filter [9], which can be used for estimating distributions of hidden states in linear dynamical systems. Therefore, a linear dynamical system should be the key framework for such unification, and, moreover, we believe that the recognition, training, and synthesis algorithms for HMMs with delta and delta-delta features can be unified based on linear dynamical systems.

In this paper, we propose a unified framework, switching acausal filters (SAFs), that describes HMMs with dynamic features using linear dynamical systems. SAFs exchange the roles of system inputs and observations in linear dynamical systems. Although this innovation seems small, it leads us to a new vista of speech signal processing, since SAFs offer the possibility of handling a large variety of speech processing tasks. The switching acausal filters are inspired by the switching AR models described in [10], which switch AR models in every speech frame. While switching AR models express speech dynamics in state equations in linear dynamical system, switching acausal filters express speech dynamics in observation equations. In this paper, we first show the general structure of SAFs and the formulation of algorithms for not only recognition and training but also synthesis.

We show that this formulation can represent speech synthesis using HMMs with delta and delta-delta features [11] as well as speech recognition using these HMMs. In addition, we express a previous parameter estimation algorithm for a F0 control model [12] using SAFs.

2. SWITICHING ACAUSAL FILTERS

The basic linear dynamical systems used here are first explained in this section, and then the switching acausal filters are described.

2.1 Definition of a linear dynamical system

The general formulation of a linear dynamical system is represented as

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{G}\mathbf{w}_t + \mathbf{u}_t , \qquad (1)$$

$$\mathbf{z}_t = \mathbf{C}\mathbf{x}_t + \mathbf{v}_t , \qquad (2)$$

where $\mathbf{F}$, $\mathbf{C}$, and $\mathbf{G}$ are the linear transform matrices, and $\mathbf{x}_t$, $\mathbf{w}_t$, and $\mathbf{v}_t$ are random vector variables whose distributions are Gaussian. Equations (1) and (2) are called state and observation equations, respectively. The distributions of $\mathbf{w}_t$ and $\mathbf{v}_t$ are independent Gaussian noise sources:

$$\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{Q_t}) \text{ and} \qquad (3)$$

$$\mathbf{v}_t \sim N(\mathbf{0}, \mathbf{R}_t) , \qquad (4)$$

where $\mathbf{Q_t}$ and $\mathbf{R}_t$ are covariance matrices. At time *t*, $\mathbf{x_t}$ is the variable, $\mathbf{u}_t$ is the system input, and $\mathbf{z}_t$ is the observation. A Kalman filter and smoother address the general problem of trying to estimate hidden state sequence $\mathbf{x}_t$, given observations [13].

2.2 Definition of switching acausal filters

Switching acausal filters stochastically switch the acausal filters at each time step. These acausal filters are created by linear dynamical systems. We set the parameters as

$$\mathbf{x}_t = [x_{t+n}, x_{t+n-1}, ..., x_t, ..., x_{t-n}]' , \tag{5}$$

$$\mathbf{F} = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ 1 & \ddots & 0 & 0 & 0 \\ 0 & \ddots & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}, \tag{6}$$

$$\mathbf{u}_t = [y_{t+n}, 0, ...0]' \text{, and} \tag{7}$$

$$\mathbf{z}_t = \mathbf{m}_{q_t} . \tag{8}$$

We exchange the roles of $\mathbf{u}_t$ and $\mathbf{z}_t$ in linear dynamical systems. This means that while $\mathbf{u}_t$ and $\mathbf{z}_t$ are generally defined as a system input and an observation, here, conversely, $\mathbf{u}_t$ and $\mathbf{z}_t$ are defined as an observation and a system input. This modification is original and an important characteristic of SAFs.

Introducing this change, $\mathbf{u}_t$ is set to a vector whose first column is observed signal y_{t+n}, where *T+1* is the length of the observed signal and $y_{T+\tau}$ is set to 0 if $\tau > 0$. 2*n*+1 is the window length for $\mathbf{x}_t$. **F** is a shift operation matrix whose roles denote identical equations, except for the row 0. $\mathbf{u}_t$ and **F** perform this equation:

$$\begin{aligned} \mathbf{x}_t &= [x_{t+n}, x_{t+n-1}, ..., x_t, ..., x_{t-n}]' \\ &= [y_{t+n}, y_{t+n-1}, ..., y_t, ..., y_{t-n}]'. \end{aligned} \tag{9}$$

The initial vector of $\mathbf{x}_t$ is set to

$$\begin{aligned} \mathbf{x}_0 &= [x_n, x_{n-1}, ..., x_0, ..., x_{-n}]' \\ &= [y_n, y_{n-1}, ..., y_0, 0, ..., 0]'. \end{aligned} \tag{10}$$

Since for recognition and training we want to set the fixed vector to $\mathbf{x}_t$ as shown in Eq. (9), we set $\mathbf{w}_t = \mathbf{0}$, which means $\mathbf{G}\mathbf{w}_t = \mathbf{0}$.

Finally we introduce switch state q_t , which specifies $\mathbf{z}_t$, and $\mathbf{R}_t$ matrices to use at time *t*. We set $\mathbf{z}_t = \mathbf{m}_{q_t}$, where $\mathbf{m}_{q_t}$ is the fixed vector in the states. This is another important characteristic of SAFs, and it is completely different from general linear dynamical systems.

When we set $\mathbf{R}_t = \mathbf{R}_{q_t}$, random variable $\mathbf{v}_{q_t}$ is yielded by the following distribution:

$$\mathbf{v}_{q_t} \sim N(\mathbf{0}, \mathbf{R}_{q_t}) . \tag{11}$$

Using these settings, observation Eq. (2) can be rewritten as

$$\begin{aligned} \mathbf{m}_{q_t} &= \mathbf{C}\mathbf{x}_t + \mathbf{v}_{q_t} \\ &= \mathbf{C}[x_{t+n}, x_{t+n-1}, ..., x_t, ..., x_{t-n}]' + \mathbf{v}_{q_t} , \end{aligned} \tag{12}$$

$$\mathbf{C} = [\mathbf{c}_1, ..\mathbf{c}_i, ..., \mathbf{c}_I]' , \tag{13}$$

$$\mathbf{c}_i = [c_{i,t+n}, c_{i,t+n-1}, ..., c_{i,t}, ..., c_{i,t-n}]' . \tag{14}$$

These equations show a set of acausal filters. $\mathbf{c}_i$ is a vector of the coefficients of each filter. Since $\mathbf{m}_{q_t}$ and $\mathbf{R}_{q_t}$ are dependent on the states, we can select the characteristics of the filters by changing state q_t at time *t*.

3. BASIC ALGORITHM FOR SAFs

We define three important algorithms: recognition, training parameters, and generation parameters. In conventional speech recognition methods, recognition and training algorithms are often defined. However, generation is rarely defined. This is another important characteristic of SAFs.

3.1 Recognition

The joint distribution of $\mathbf{x}_{0:T}$ and $q_{0:T}$ of SAFs is as follows:

$$\Pr(\mathbf{x}_{0:T}, q_{0:T}) = \frac{1}{norm(\mathbf{C}, q_{0:T})} \prod_{t=0}^{T} \Pr(\mathbf{C}\mathbf{x}_t \mid q_t) \prod_{t=1}^{T} \Pr(q_t \mid q_{t-1}) \Pr(q_0), \text{(15)},$$

where $\Pr(\mathbf{C}\mathbf{x}_t \mid q_t)$ is the following Gaussian probability density:

$$\Pr(\mathbf{C}\mathbf{x}_t \mid q_t) = N(\mathbf{C}\mathbf{x}_t; \mathbf{m}_{q_t}, \mathbf{R}_{q_t}). \tag{16}$$

Note that since random variable $\mathbf{x}_t$ is multiplied by **C** in Eq. (12), a normalizing factor

$$norm(\mathbf{C}, q_{0:T}) = \int \prod_{t=0}^{T} \Pr(\mathbf{C}\mathbf{x}_t \mid q_t) d\mathbf{C}\mathbf{x}_0 ... d\mathbf{C}\mathbf{x}_T \tag{17}$$

is required. If $\Pr(q_0)$, $\Pr(q_t \mid q_{t-1})$, and $N(\mathbf{C}\mathbf{x}_t; \mathbf{m}_{q_t}, \mathbf{R}_{q_t})$ are given, recognition can be performed by calculating Eq. (15) for all possible state sequences except for $norm(\mathbf{C}, q_{0:T})$. If no constraint exists, among the vectors of $\mathbf{C}\mathbf{x}_0 ... \mathbf{C}\mathbf{x}_T$, $norm(\mathbf{C}, q_{0:T})$ should be 1.0. This assumption is made to match the assumption that each dynamic feature in HMMs is not constrained by the other features. Since we use this assumption here, we can obtain this approximation:

$$\begin{aligned} &\Pr(\mathbf{x}_{0:T}, q_{0:T}) \approx \Pr(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1, ...\mathbf{C}\mathbf{x}_T, q_{0:T}) \\ &\prod_{t=0}^{T} \Pr(\mathbf{C}\mathbf{x}_t \mid q_t) \prod_{t=1}^{T} \Pr(q_t \mid q_{t-1}) \Pr(q_0). \end{aligned} \tag{18}$$

(Note that to obtain $norm(\mathbf{C}, q_{0:T})$, we have to calculate the following:

$$norm(\mathbf{C}, q_{0:T}) = \int \prod_{t=0}^{T} \Pr(\mathbf{C}\mathbf{x}_t \mid q_t) dy_0 ... dy_T . \tag{19}$$

Since $\mathbf{C}\mathbf{x}_0, ..., \mathbf{C}\mathbf{x}_T$ have the constraint described in Eq. (12), integration can be performed with respect to the free parameters in $\mathbf{C}\mathbf{x}_0, ..., \mathbf{C}\mathbf{x}_T$, which are $y_0, ..., y_T$.)

3.2 Training parameters

All of the parameters, including $\mathbf{m}_j$, which is the observation parameter in a linear dynamical system, can be trained by the approximated EM algorithm. Specifically, the EM training procedure is used to train this set of parameters: $\theta = \{ \mathbf{R}_j, \mathbf{m}_j, \mathbf{C}, \Pr(q_t = j \mid q_{t-1} = i), \Pr(q_0 = i) \mid \forall i, j \}$.

We assume that one sequence of training data is given. From Eq. (15), the complete log likelihood for the sequence is

$$L = \log(\Pr(\mathbf{x}_{0:T}, q_{0:T})) = \sum_{t=0}^{T} \log \Pr(\mathbf{C}\mathbf{x}_t \mid q_t) + \sum_{t=1}^{T} \log \Pr(q_t \mid q_{t-1}) + \log \Pr(q_0) - \log(norm(\mathbf{C}, q_{0:T})). \quad (20)$$

In EM we iteratively maximize the expected value of the average complete data log likelihood:

$$\hat{L} = E_{P(\mathbf{x}_{0:T}, q_{0:T})}[L] = \sum_{q_0} .... \sum_{q_T} \Pr(\mathbf{x}_{0:T}, q_{0:T}; \theta^{old}) \log(\Pr(\mathbf{x}_{0:T}, q_{0:T}; \theta)). \quad (21)$$

Here, θ is the expected parameter set, and θ^{old} is the current parameter set. To maximize the likelihood, we take the derivatives with respect to each parameter in θ and set them to 0. We introduce the same approximation, setting the term $norm(\mathbf{C}, q_{0:T})$ to 1.0, to thus obtain the equation:

$$\begin{aligned} \tilde{L} &= \sum_{q_0} .... \sum_{q_T} \Pr(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1, ...\mathbf{C}\mathbf{x}_T, q_{0:T}; \theta^{old}) \\ &\cdot \log(\Pr(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1, ...\mathbf{C}\mathbf{x}_T, q_{0:T}; \theta)) \\ &= \Pr(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1, ...\mathbf{C}\mathbf{x}_T) \\ &[\sum_{t=0}^{T} \sum_{q_t} \left(\sum_{\{q_\tau, \tau \neq t\}} \Pr(q_{0:T} \mid \mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1, ...\mathbf{C}\mathbf{x}_T; \theta^{old}) \right) \\ &\cdot \log \Pr(\mathbf{C}\mathbf{x}_t \mid q_t)] \\ &+ \sum_{q_0} .... \sum_{q_T} \Pr(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1, ...\mathbf{C}\mathbf{x}_T, q_{0:T}; \theta^{old}) \\ &\left[\sum_{t=0}^{T} \log \Pr(q_t \mid q_{t-1}) + \log \Pr(q_0) \right]. \end{aligned} \quad (22)$$

To obtain $\mathbf{m}_j$, we take the derivative with respect to $\mathbf{m}_j$ and set it to 0 as

$$\frac{\partial \tilde{L}}{\partial \mathbf{m}_j} = 0 \quad (23)$$

Consequently we obtain

$$\frac{\partial}{\partial \mathbf{m}_j} \sum_{t=0}^{T} \sum_{q_t = j} W_t^{\,j} \left[\log \Pr(\mathbf{C}\mathbf{x}_t \mid q_t) + ... \right] = 0, \quad (24)$$

$$\frac{\partial}{\partial \mathbf{m}_j} \sum_{t=0}^{T} \sum_{q_t = j} W_t^{\,j} \left[-\frac{1}{2} \left[(\mathbf{C}\mathbf{x}_t - \mathbf{m}_{q_t})' \mathbf{R}_{q_t}{}^{-1} (\mathbf{C}\mathbf{x}_t - \mathbf{m}_{q_t}) \right] \right] = 0, \quad (25)$$

and $$\mathbf{m}_j = \frac{1}{\left(\sum_{t=0}^{T} W_t^{\,j} \right)} \sum_{t=0}^{T} W_t^{\,j} \mathbf{C}\mathbf{x}_t . \quad (26)$$

Here, $W_t^{\,j}$ is the weight value computed during the forward-backward step described in the Appendix.

In a similar way, we take the derivative with respect to $\mathbf{R}_j$ and set it to 0 as

$$\frac{\partial \tilde{L}}{\partial \mathbf{R}_j^{-1}} = 0. \quad (27)$$

We obtain

$$\frac{\partial}{\partial \mathbf{R}_j^{-1}} \sum_{t=0}^{T} \sum_{q_t = j} W_t^{\,j} \left[\log \Pr(\mathbf{C}\mathbf{x}_t \mid, q_t) + ... \right] = 0, \quad (28)$$

$\mathbf{R}_j$ can be obtained as

$$\mathbf{R}_j = \frac{1}{\left(\sum_{t=0}^{T} W_t^{\,j} \right)} \sum_{t=0}^{T} W_t^{\,j} (\mathbf{C}\mathbf{x}_t - \mathbf{m}_j)(\mathbf{C}\mathbf{x}_t - \mathbf{m}_j)'. \quad (29)$$

To obtain $\mathbf{C}$, we take the derivative of the average likelihood with respect to $\mathbf{C}$ and set it to 0 as

$$\frac{\partial \tilde{L}}{\partial \mathbf{C}} = 0 \, . \quad (30)$$

Consequently, we obtain

$$\frac{\partial}{\partial \mathbf{C}} \sum_{t=0}^{T} \sum_{q_t = j} W_t^{\,j} \left[-\frac{1}{2} \left[(\mathbf{C}\mathbf{x}_t - \mathbf{m}_{q_t})' \mathbf{R}_{q_t}{}^{-1} (\mathbf{C}\mathbf{x}_t - \mathbf{m}_{q_t}) \right] \right] = 0 \text{,(31)}$$

$$\sum_{t=0}^{T} \sum_{q_t = j} W_t^{\,j} \mathbf{R}_{q_t}{}^{-1} \mathbf{C}\mathbf{x}_t \mathbf{x}_t' - \sum_{t=0}^{T} \sum_{q_t = j} W_t^{\,j} \mathbf{R}_{q_t}{}^{-1} m_j \mathbf{x}_t' = 0, \text{ and } \quad (32)$$

$$\mathbf{C} = \sum_{t=0}^{T} \sum_{q_t = j} W_t^{\,j} \mathbf{m}_j \mathbf{x}_t' \left[\sum_{t=0}^{T} \sum_{q_t = j} W_t^{\,j} \mathbf{x}_t \mathbf{x}_t' \right]^{-1} . \quad (33)$$

Transition probability can be obtained by taking the derivative of the average likelihood with respect to $a(i, j) = \Pr(q_t = j \mid q_{t-1} = i)$ as

$$\frac{\partial \tilde{L}}{\partial a(i, j)} = 0. \quad (34)$$

Using the constraint $\sum_j a(i, j) = 1.0$, we thus obtain

$$a(i, j) = \frac{\sum_{t=1}^{T} \Pr(q_{t-1} = i, q_t = j \mid \mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1, ...\mathbf{C}\mathbf{x}_T)}{\sum_{t=0}^{T-1} W_t^i} . \quad (35)$$

Initial transition probability can be obtained by taking the derivative of the average likelihood with respect to $\pi(j) = \Pr(q_0 = j)$. (36)

Finally, we obtain

$$\pi(j) = W_0^j \, . \quad (37)$$

These parameter estimation algorithms only consider one training speech data sequence. However, it is easy to extend them to handle multiple speech data sequences by summing up waited parameters.

3.3 Generating features

In Sections 3.1 and 3.2, we treat $\mathbf{x}_t$ as a fixed value. If we treat it as a random variable, we can generate signals from trained switching acausal filters. These generated signals provide the variables by which the likelihood is maximized with trained parameters.

Here, suppose that state sequence $q_{0:T}$ is given. To generate $\mathbf{x}_t$ under this condition, we release $\mathbf{w}_t$ from 0 and set

$$\mathbf{Q}_t = \zeta \,, \tag{38}$$

where ζ is a sufficiently large real number.

We also set $\mathbf{u}_t$, $\mathbf{x}_0$, and $\mathbf{G}$ as

$$\mathbf{u}_t = [0, 0, ...0]' \tag{39}$$

$$\mathbf{x}_0 = [0, 0, ..., 0]' \tag{40}$$

$$\mathbf{G} = [1, 0, ..., 0]' \,. \tag{41}$$

This means that $\mathbf{x}_t$ is released from the constraint of input signal y_t . We can construct a linear dynamical system by setting the values defined here and in 3.2 to Eqs. (1) and (2), and a Kalman smoother using the linear dynamical system is operated to generate speech signals. The Kalman smoother's objective function, that is, a Hamiltonian function [13], is

$$\min_{\mathbf{x}_0, \mathbf{w}_0, ..\mathbf{w}_T} \left[(\mathbf{m}_{q_0} - \mathbf{C}\mathbf{x}_0)' \mathbf{R}_{q_0}{}^{-1} (\mathbf{m}_{q_0} - \mathbf{C}\mathbf{x}_0) + \sum_{i=1}^{T} (\mathbf{m}_{q_t} - \mathbf{C}\mathbf{x}_t)' \mathbf{R}_{q_t}{}^{-1} (\mathbf{m}_{q_t} - \mathbf{C}\mathbf{x}_t) + \sum_{i=1}^{T} \mathbf{w}_t{}' \mathbf{Q}_t{}^{-1} \mathbf{w}_t \right]. \tag{42}$$

Here, we have a constraint:

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{G}\mathbf{w}_t + \mathbf{u}_t \,. \tag{43}$$

Using Eq. (43), we can modify Eq. (42) to

$$\min_{\mathbf{x}_0, \mathbf{w}_0, ..\mathbf{w}_T} \left[\sum_{i=0}^{T} (\mathbf{m}_{q_t} - \mathbf{C}\mathbf{x}_t)' \mathbf{R}_{q_t}{}^{-1} (\mathbf{m}_{q_t} - \mathbf{C}\mathbf{x}_t) + \sum_{i=1}^{T} \frac{\mathbf{w}_t{}' \mathbf{w}_t}{\zeta} \right]. \tag{44}$$

If ζ is large enough, the second term of Eq. (44) can be ignored.

From Eqs. (39), (41) and (43) we obtain

$$x_{t+n} = \mathbf{w}_t = w_t \,. \tag{45}$$

Consequently, we obtain the objective function:

$$\min_{x_{-n}, ..., x_{T+n}} \left[\sum_{i=0}^{T} (\mathbf{m}_{q_t} - \mathbf{C}\mathbf{x}_t)' \mathbf{R}_{q_t}{}^{-1} (\mathbf{m}_{q_t} - \mathbf{C}\mathbf{x}_t) \right]. \tag{46}$$

The Kalman filter generates $\mathbf{x}_{0:T}$ that minimizes Eq. (46).

4. RERATION TO HMM WITH DELTA AND DELTA-DELTA FEATURES

An HMM with delta and delta-delta features can be modeled by a specified switching acausal filter. For simplicity, we only describe the one-dimensional feature case. We also suppose that the distribution of a feature in a state is single Gaussian distribution. Here, we set

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1/5 & 1/10 & 0 & -1/10 & -1/5 \\ 1/14 & -1/28 & -1/14 & -1/28 & 1/14 \end{bmatrix}, \tag{47}$$

where $2n+1$ is 5. $\mathbf{C}$ is an example matrix that calculates static and dynamic features denoted by $y_t, \Delta y_t, \Delta\Delta y_t$ from a time series of speech input y_t as

$$\mathbf{C}\mathbf{x}_t = \mathbf{C}[y_{t+n}, y_{t+n-1}, ..., y_t, ..., y_{t-n}]' = [y_t, \Delta y_t, \Delta\Delta y_t]' \,. \tag{48}$$

We set $\mathbf{m}_j$ to the means of the feature and dynamic features and set $\mathbf{R}_j$ to the variance of the feature and dynamic features as

$$\mathbf{m}_j = \left[m_j, \Delta m_j, \Delta\Delta m_j \right]', \quad \text{and} \tag{49}$$

$$\mathbf{R}_j = \begin{bmatrix} \sigma_j{}^2 & 0 & 0 \\ 0 & \Delta\sigma_j{}^2 & 0 \\ 0 & 0 & \Delta\Delta_j \sigma^2 \end{bmatrix}, \tag{50}$$

where j is the state number.

From Eq. (22), the approximate complete data log likelihood for the sequence is

$$\begin{aligned} \tilde{L} &= \log(\Pr(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1, ..., \mathbf{C}\mathbf{x}_T, q_{0:T})) \\ &= \sum_{t=0}^{T} \log b_{q_t}(\mathbf{C}\mathbf{x}_t) + \sum_{t=1}^{T} \log a(q_{t-1}, q_t) + \log \pi(q_0), \end{aligned} \tag{51}$$

where

$$b_{q_t}(\mathbf{C}\mathbf{x}_t) = N([y_t, \Delta y_t, \Delta\Delta y_t]'; [m_{q_t}, \Delta m_{q_t}, \Delta\Delta m_{q_t}]', \mathbf{R}_{q_t}) \,, \tag{52}$$

$$a(q_{t-1}, q_t) = \Pr(q_t \mid q_{t-1}), \text{ and} \tag{53}$$

$$\pi(q_0) = \Pr(q_0). \tag{54}$$

Eq. (51) is identical to the complete log likelihood of HMMs with delta and delta-delta features. Using the above formulation, we also confirmed that Eqs., (26), (29), (35), and (37) are the same as the training equations of HMMs with delta and delta-delta features (This is reasonable. To do this, we introduce the assumption that there is no constraint among the vectors of $\mathbf{C}\mathbf{x}_0 ... \mathbf{C}\mathbf{x}_T$. This means that there is no constraint among $y_t, \Delta y_t, \Delta\Delta y_t$ here. This is a well-known assumption to make HMMs in speech recognition.).

These formulations can be easily extended to multidimension mixture HMMs without loss of generality.

Still, there is room for debate about the $norm(\mathbf{C}, q_{0:T})$ term, which is ignored in this formulation. Calculating term $norm(\mathbf{C}, q_{0:T})$ is strongly related to the work in [14]. In the HMM case, Eqs. (15) and (19) show the same objective function described in this article.

5. RELATION TO SPEECH SYNTHESIS METHOD USING HMMS

A speech synthesis method using HMMs with delta and delta-delta features described in [11] can be expressed by a Kalman smoother with a specified switching acausal filter. Here, it is assumed that the parameters $\mathbf{m}_j$ and $\mathbf{R}_j$ are obtained using the algorithm in Section 3. In addition, we introduce Eqs. (39)-(41) and (47)−(50). Using these settings, we can obtain the state and observation equations. A Kalman smoother operates using Eqs. (1) and (2) to generate a smoothed state sequence. From Eq. (46), the objective function of the smoothed sequence can be rewritten as

$$\min_{x_{-n},..,x_{T+n}}\left[\sum_{i=0}^{T}([m_{q_t},\Delta m_{q_t},\Delta\Delta m_{q_t}]'-\mathbf{C}\mathbf{x}_t)'\mathbf{R}_{q_t}{}^{-1}([m_{q_t},\Delta m_{q_t},\Delta\Delta m_{q_t}]'-\mathbf{C}\mathbf{x}_t)\right]. \quad (55)$$

This is equivalent to the objective function of the HMM speech synthesis method [11]. Therefore, using a Kalman filter with a switching acausal filter incorporates the HMM speech synthesis method.

6. RELATION TO F0 CONTROL MODEL

A switching acausal filter contains the F0 control model described in [12], which expressed the F0 contour as

$$\alpha\frac{d^2y}{d^2t}+\beta\frac{dy}{dt}+\gamma y=\varphi\,, \quad (56)$$

where y denotes the observed F0 contour and φ denotes the system input. The model approximated first- and second-order derivatives as

$$\left.\frac{d^2y}{d^2t}\right|_t\approx\mathbf{B}\mathbf{y}_t \text{ and} \quad (57)$$

$$\left.\frac{dy}{dt}\right|_t\approx\mathbf{A}\mathbf{y}_t\,, \quad (58)$$

where **A** and **B** are linear functions that approximate the first and second derivatives and $\mathbf{y}_t$ is defined as

$$\mathbf{y}_t=[y_{t+n},y_{t+n-1},...,y_t,...,y_{t-n}]'\,. \quad (59)$$

The F0 control model assumed that φ is a step function whose step values depend on states. Under this assumption, Eq. (56) can be rewritten in a discrete manner as

$$(\alpha\mathbf{B}+\beta\mathbf{A}+\gamma\mathbf{D})\mathbf{y}_\mathbf{t}=m_{q_t}+\varepsilon\,. \quad (60)$$

Here, m_{q_t} is a fixed value in the state q_t, and ε is an error that is independent on the state [12]. The objective of the F0 control model is to determine α, β, and γ using observation data.

This can be solved by using switching acausal filters. First we make state and observation equations for the switching acausal filters corresponding to the F0 control model by setting $\mathbf{C}$, $\mathbf{m}_j$, and $\mathbf{R}_j$ as

$$\mathbf{C}=(\alpha\mathbf{B}+\beta\mathbf{A}+\gamma\mathbf{D}), \quad (61)$$

$$\mathbf{m}_j=m_j\,\text{, and} \quad (62)$$

$$\mathbf{R}_j=\mathbf{R}=\sigma^2\,. \quad (63)$$

Here, σ^2 is the variance of ε.

To maximize the average complete data log likelihood, we take the derivative of Eq. (22) with respect to α, β, and γ as

$$\frac{\partial\tilde{L}}{\partial\alpha}=\frac{\partial\tilde{L}}{\partial\mathbf{C}}\frac{\partial\mathbf{C}'}{\partial\alpha}=0\,, \quad (64)$$

$$\frac{\partial\tilde{L}}{\partial\beta}=\frac{\partial\tilde{L}}{\partial\mathbf{C}}\frac{\partial\mathbf{C}'}{\partial\beta}=0\,\text{, and} \quad (65)$$

$$\frac{\partial\tilde{L}}{\partial\gamma}=\frac{\partial\tilde{L}}{\partial\mathbf{C}}\frac{\partial\mathbf{C}'}{\partial\gamma}=0\,. \quad (66)$$

From Eq. (64), we obtain

$$\begin{aligned}&\frac{\partial}{\partial\mathbf{C}}\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\left[-\frac{1}{2}\left[(\mathbf{C}\mathbf{x}_t-m_{q_t})'\mathbf{R}^{-1}(\mathbf{C}\mathbf{x}_t-m_{q_t})\right]\right]\frac{\partial\mathbf{C}'}{\partial\alpha}\\&=\left[\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}\mathbf{C}\mathbf{x}_t\mathbf{x}_t'-\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}m_j\mathbf{x}_t'\right]\frac{\partial}{\partial\alpha}[\alpha\mathbf{B}+\beta\mathbf{A}+\gamma\mathbf{D}]'=0\end{aligned} \quad (67)$$

$$\text{and}\left[\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}\mathbf{C}\mathbf{x}_t\mathbf{x}_t'-\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}m_j\mathbf{x}_t'\right]\mathbf{B}'=0\,. \quad (68)$$

From Eq. (65), we obtain

$$\begin{aligned}&\frac{\partial\tilde{L}}{\partial\beta}=\frac{\partial}{\partial\mathbf{C}}\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\left[-\frac{1}{2}\left[((\mathbf{C}\mathbf{x}_t-m_{q_t})'\mathbf{R}^{-1}(\mathbf{C}\mathbf{x}_t-m_{q_t})\right],...\right]\frac{\partial\mathbf{C}'}{\partial\beta}=\\&=\left[\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}\mathbf{C}\mathbf{x}_t\mathbf{x}_t'-\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}m_j\mathbf{x}_t'\right]\frac{\partial}{\partial\beta}[\alpha\mathbf{B}+\beta\mathbf{A}+\gamma\mathbf{D}]'\,.\end{aligned} \quad (69)$$

We set 0 to this equation to obtain

$$\left[\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}\mathbf{C}\mathbf{x}_t\mathbf{x}_t'-\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}m_j\mathbf{x}_t'\right]\mathbf{A}'=0\,. \quad (70)$$

From Eq. (66), we obtain

$$\begin{aligned}&\frac{\partial\tilde{L}}{\partial\gamma}=\frac{\partial}{\partial\mathbf{C}}\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\left[-\frac{1}{2}\left[(\mathbf{C}\mathbf{x}_t-m_{q_t})'\mathbf{R}^{-1}(\mathbf{C}\mathbf{x}_t-m_{q_t})\right],...\right]\frac{\partial\mathbf{C}'}{\partial\gamma}=\\&\left[\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}\mathbf{C}\mathbf{x}_t\mathbf{x}_t'-\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}m_j\mathbf{x}_t'\right]\frac{\partial}{\partial\gamma}[\alpha\mathbf{B}+\beta\mathbf{A}+\gamma\mathbf{D}]'\,.\end{aligned} \quad (71)$$

We set 0 to this equation to obtain

$$\left[\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}\mathbf{C}\mathbf{x}_t\mathbf{x}_t'-\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{R}^{-1}m_j\mathbf{x}_t'\right]\mathbf{D}'=0\,. \quad (72)$$

Finally, we obtain the following three equations:

$$\left[\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{C}\mathbf{x}_t\mathbf{x}_t'\mathbf{B}'-\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^jm_j\mathbf{x}_t'\mathbf{B}'\right]=0, \quad (73)$$

$$\left[\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{C}\mathbf{x}_t\mathbf{x}_t'\mathbf{A}'-\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^jm_j\mathbf{x}_t'\mathbf{A}'\right]=0,\text{ and} \quad (74)$$

$$\left[\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{C}\mathbf{x}_t\mathbf{x}_t'\mathbf{D}'-\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^jm_j\mathbf{x}_t'\mathbf{D}'\right]=0. \quad (75)$$

We can rewrite these three equations simply in matrix form as

$$\begin{bmatrix}\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{B}\mathbf{x}_t\mathbf{x}_t{}'\mathbf{B}' & \sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{A}\mathbf{x}_t\mathbf{x}_t{}'\mathbf{B}' & \sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{D}\mathbf{x}_t\mathbf{x}_t{}'\mathbf{B}'\\ \sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{B}\mathbf{x}_t\mathbf{x}_t{}'\mathbf{A}' & \sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{A}\mathbf{x}_t\mathbf{x}_t{}'\mathbf{A}' & \sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{D}\mathbf{x}_t\mathbf{x}_t{}'\mathbf{A}'\\ \sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{B}\mathbf{x}_t\mathbf{x}_t{}'\mathbf{D}' & \sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{A}\mathbf{x}_t\mathbf{x}_t{}'\mathbf{D}' & \sum_{t=0}^{T}\sum_{q_t=j}W_t{}^j\mathbf{D}\mathbf{x}_t\mathbf{x}_t{}'\mathbf{D}'\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\\ \gamma\end{bmatrix}=\begin{bmatrix}\sum_{t=0}^{T}\sum_{q_t=j}W_t{}^jm_j\mathbf{x}_t'\mathbf{B}'\\ \sum_{t=0}^{T}\sum_{q_t=j}W_t{}^jm_j\mathbf{x}_t'\mathbf{A}'\\ \sum_{t=0}^{T}\sum_{q_t=j}W_t{}^jm_j\mathbf{x}_t'\mathbf{D}'\end{bmatrix}. \quad (76)$$

These equations are almost the same as the equation described in [12]. The difference is that the above equations have weight values and summations for *j*. Since the switching acausal filters use an approximation EM algorithm, while the F0 control model uses a Viterbi-training algorithm, the weights and the summations in the above equations are the inference terms obtained by the forward-backward process. In addition, by using switching acausal filters, the transition probabilities between states can be obtained by Eq. (35). These results show that switching acausal filters incorporate and extend the F0 control model.

7. CONCLUSION

This paper proposed switching acausal filters (SAFs) and showed algorithms for the recognition, training and speech generation processes of SAFs. The key characteristic of SAFs is to exchange the roles of system input and observations, which leads us to a new vista of speech recognition. We showed that SAFs contain HMMs with delta and delta-delta features that are commonly used for speech recognition. We also showed that they contain the HMM synthesis method. In addition to these tools, we showed that switching acausal filters incorporate the F0 control model [12].

Although the problems related to these developments have already been solved by many research efforts in different ways, by redefining them using linear dynamical systems, we can now compare our methods with the other methods called segmental models. Our findings reveal that this provides a good perspective for improving speech modeling. Furthermore, our results indicate that SAFs show much promise.

APPENDIX

Forward and backward procedures for calculating inference $W_t^j = \Pr(q_t = j \mid \mathbf{Cx}_0,...,\mathbf{Cx}_T)$ are required. Although there are several formulations to calculate this, we introduce a forward-backward calculation using HMM parameter estimation. Note that introducing the approximation, this formulation is the complete the same as the HMM forward-backward algorithm.

[Forward recursion]

Assuming that $\Pr(q_{t-1} = i, \mathbf{Cx}_0,...,\mathbf{Cx}_{t-1})$ has been obtained, forward recursion can be calculated by

$$\begin{aligned}&\Pr(q_t = j, \mathbf{Cx}_0,...,\mathbf{Cx}_t)\\&= \Pr(\mathbf{Cx}_t \mid q_t = j)\sum_i \Pr(q_{t-1} = i, \mathbf{Cx}_0,...,\mathbf{Cx}_{t-1})\Pr(q_t = j \mid q_{t-1} = i).\end{aligned} \tag{77}$$

The following are the initial conditions:

$$\Pr(q_0 = j) = \pi(j)\,, \tag{78}$$

$$\Pr(\mathbf{Cx}_0 \mid q_0 = j) = N(\mathbf{Cx}_0, \mathbf{m}_j \mathbf{R}_j)\ . \tag{79}$$

[Backward recursion]

Assuming that $\Pr(\mathbf{Cx}_{t+2},...,\mathbf{Cx}_T \mid q_{t+1} = k)$ has already been obtained, backward recursion can be calculated by

$$\begin{aligned}&\Pr(\mathbf{Cx}_{t+1},...,\mathbf{Cx}_T \mid q_t = j)\\&= \sum_k \Pr(q_{t+1} = k \mid q_t = j)\Pr(\mathbf{Cx}_{t+1} \mid q_{t+1} = k)\Pr(\mathbf{Cx}_{t+2},...,\mathbf{Cx}_T \mid q_{t+1} = k)\,.\end{aligned} \tag{80}$$

There are various ways to set initial values. One is to restrict the possible final nodes. Another is to introduce only one final state.

[Inference calculation]

From the results of the forward recursion and backward recursion, we obtain

$$\begin{aligned}\Pr(q_t = i \mid \mathbf{Cx}_0,...,\mathbf{Cx}_T) &= \frac{\Pr(q_t = i, \mathbf{Cx}_0,...,\mathbf{Cx}_T)}{\Pr(\mathbf{Cx}_0,...,\mathbf{Cx}_T)}\\&= \frac{\Pr(q_t = i, \mathbf{Cx}_0,...,\mathbf{Cx}_t)\Pr(\mathbf{Cx}_{t+1},...,\mathbf{Cx}_T \mid q_t = i)}{\Pr(\mathbf{Cx}_0,...,\mathbf{Cx}_T)}.\end{aligned} \tag{81}$$

REFERENCES

[1] J. Picone, S. Pike, R. Regan, T. Kamm, J. Bridle, L. Deng, Z. Ma, H. Richards, and M. Schuster, "Initial Evaluation of Hidden Dynamic Models on Conversational Speech," Proc. ICASSP, pp. 109-112, 1999.

[2] L. Deng, "A Dynamic, Feature-based Approach to the Interface between Phonology and Phonetics for Speech Modeling and Recognition," Speech Communication, 24 (4), pp. 299-323, 1998.

[3] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "ML Estimation of a Stochastic Linear System with the EM Algorithm and Its Application to Speech Recognition," IEEE Trans. Speech Audio Processing, Vol. 1, No. 4, pp. 431-442, 1993.

[4] H. Gish and K. Ng, "Parametric Trajectory Models for Speech Recognition," Proc. ICSLP, pp. 466-469, 1996.

[5] W. J. Holmes and M. J. Russell, "Probabilistic-trajectory Segmental HMMs," Computer Speech and Language, vol. 13, pp. 3-37, 1999.

[6] R. Iyer, H. Gish. M.-H. Siu, G. Zavaliagkos, and S. Matsoukas, "Hidden Markov Models for Trajectory Modeling," Proc ICSLP, 1998.

[7] M. Ostendorf, V. Digalakis, and O. Kimball. "From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," IEEE Trans. Speech and Audio Processing, 4(5), pp. 360-378, 1996.

[8] S. Furui, "Speaker-Independent Isolated Word Recognition using Dynamic Features of Speech Spectrum," IEEE Trans. Acoustics, Speech and Signal Processing, 34(1) , pp. 52-59, 1986.

[9] Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri, "A Theoretical Analysis of Speech Recognition based on Feature Trajectory Models," Proc. ICSLP, pp. 549-552, 2004.

[10] K. P. Murphy, "Switching Kalman Filters," http://www.cs.ubc.ca/~murphyk/

[11] K. Tokuda, T. Kobayashi, and S. Imai, "Speech Parameter Generation from HMM using Dynamic Features," Proc. ICASSP, pp. 660-663, 1995.

[12] Y. Ohishi, H. Kameoka, K. Kashino, and K. Takeda, "Parameter Estimation Method of F0 Control Model for Singing Voices," Interspeech, pp.139-142, 2008.

[13] T. Kailath, A. H. Sayed, and B. Hassibi, "Linear Estimation," Prentice Hall, 2000.

[14] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a Trajectory Model by Imposing Explicit Relationships Between Static and Dynamic Feature Vector Sequences," Computer Speech and Language, 21(1), pp. 153-173, 2007.