社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－SLP－54（9）
2004／12／20

カルマンフィルタにより生成されたトラジェクトリに基づく音声認識

南　泰浩

ＮＴＴコミュニケーション科学基礎研究所　〒619-0237 京都府相楽郡精華町光台 2-4
E-mail:　minami@cslab.kecl.ntt.co.jp

あらまし　我々はHMMからトラジェクトリを生成し、音声認識を行うHMM-トラジェクトリ法を提案した。しかし、この手法は混合分布への拡張が難しいという問題があった。本報告では、HMM-トラジェクトリ手法の混合分布への拡張を考慮にいれた新たなトラジェクトリ生成手法を提案する。この手法では、混合分布型HMMに対応する複数の線形動的システムを用意し、線形動的システムの観測方程式、状態方程式を決定する。この観測方程式、状態方程式をセレクティングカルマンフィルタにより動作させ、適切なトラジェクトリの生成を行う。本報告では、このセレクティングカルマンフィルタによって、適切なトラジェクトリが生成できることを実験により示す。

キーワード　HMM, カルマンフィルタ,スイッチングカルマンフィルタ,トラジェクトリ

Speech recognition method based on trajectories generated by Kalman filters

Yasuhiro MINAMI

NTT communication science laboratories, NTT corporation

2-4 Hikaridai, Seika-cho, Souraku-gun, Kyoto, Japan 619-0237
E-mail:　minami@cslab.kecl.ntt.co.jp

Abstract We previously proposed a speech recognition method using trajectory models that unfortunately had some difficulty treating HMMs with Gaussian mixture distributions. Therefore, this paper proposes a new trajectory generation method that extends the use of HMM-trajectory with Gaussian mixture HMMs, generates linear dynamic system models from the HMMs, and then performs switching Kalman filters to generate smooth trajectory with them. Experiments show that the method generates smooth trajectories.

Keyword　HMM, Kalman filter, switching Kalman filter, trajectory

1. Introduction

Since HMMs model the acoustic feature vector sequence as a piecewise stationary process, the probability of a given acoustic feature is independent of the sequence of acoustic features preceding and following the current feature. This means that HMMs cannot treat the time-dependent characteristics of speech within that state, which is a widely recognized drawback of speech recognition using HMMs.

We previously proposed a HMM-trajectory method that employs smoothed speech feature trajectories generated directly from HMM statistics instead of HMM mean sequences [1][2][3]. The technique generates a smooth feature vector trajectory by maximizing the HMM likelihood while simultaneously considering the relationship between the static and the dynamic features (delta features and delta-delta features). This procedure is a type of filter that smoothes the HMM mean sequence. Our method has a significant advantage: the relationship between the static and dynamic features, which are ignored in the conventional speech recognition phase, can be used to improve recognition results. In our previous paper, speaker-independent word recognition results showed that the proposed method was effective when a single Gaussian distribution in each HMM state was used [1]. We extended our method that treats Gaussian mixture HMMs by selecting the sequence of Gaussian mixture distributions that gives the best likelihood with the Viterbi algorithm [2]. Although recognition experiments showed the effectiveness of the proposed method, the method proved too simple to be applied to the more complicated Gaussian mixture models.

This report describes a possibility that extends our method to use Gaussian mixture distributions. We

found that a smoother HMM-trajectory method can be written by linear dynamical system (LDS) models [3]. Gaussian mixture HMMs are converted to LDS models using this fact, and then switching Kalman filters [4] generate smoothed trajectory using the linear dynamic system (LDS) models.

In section 2, HMM-trajectory methods are described. The section also describes that generating smoothed trajectory can be expressed by Kalman filter. In section 3, the generating smoothed trajectory using switching Kalman filter is described. In section 4, generating the smoothed trajectory is evaluated using the TIMIT database.

2. HMM-trajectory method

This section describes an overview and the theoretical aspects of the HMM-trajectory method based on [1][3]. Figure 1 shows the flow of the HMM-trajectory method. It is assumed that HMMs have only a single Gaussian distribution in a state. It is also assumed that the recognition result candidates for input speech are obtained in advance and that input speech is segmented by Viterbi algorithms using standard HMMs, based on transcription results. Therefore, state sequences for input speech are available. The HMM-trajectory method uses the sequences to generate smoothed trajectories by first generating mean sequences of features, delta features, and delta-delta features from HMMs. These sequences resemble the step function shown in Figure 1. Sequences are smoothed by maximizing HMM likelihood, while respecting the constraints between static features, delta features, and delta-delta features. This operation is a kind of filter

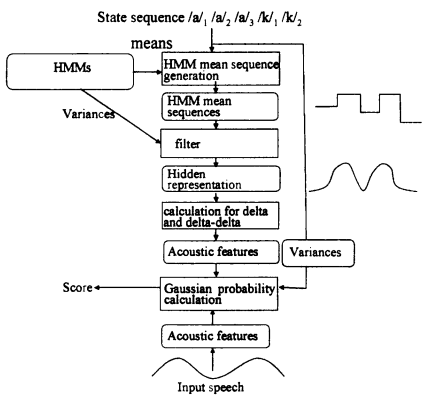

Figure 1. Diagram of HMM-trajectory method.

to obtain a smooth trajectory. The smoothed trajectory is then converted into static features (the trajectory itself), delta features, and delta-delta features (show in Fig. 1). Distributions are defined along with these triple sequences. Distribution variances are calculated using the same training data used to create the original HMMs. The input data are transformed into static features, delta-features, and delta-delta features. Then the sequences of the features are used to calculate the likelihood with the distributions. The above procedure is performed for all of the candidates.

We found that the filter can be written by linear dynamical system (LDS) models [3]. The details are shown below. The set of a static feature, a delta-feature, and a delta-delta feature of the trajectory is denoted by vector $\mathbf{x}_t = [x_t, \Delta x_t, \Delta\Delta x_t]'$, where " ' " denotes the matrix transpose. Suppose that x_t is one dimension vector to make the following formula simple.

If state sequence $S = \{S_1, S_2, ..., S_T\}$ is given, then the total score along with the state sequence is denoted by:

$$P(\mathbf{X} \mid S) = \prod_{t=1}^{T} \frac{1}{(2\pi)^{\frac{3}{2}} |\mathbf{\Sigma}_{S_t}|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(\mathbf{x}_t - \mathbf{m}_{S_t})' \mathbf{\Sigma}_{S_t}^{-1} (\mathbf{x}_t - \mathbf{m}_{S_t})\right]}, \quad (1)$$

ignoring the state transition probabilities, where, $\mathbf{m}_{S_t} = [\mu_{S_t}, \Delta\mu_{S_t}, \Delta\Delta\mu_{S_t}]'$, $\mathbf{\Sigma}_{S_t} = diag[\sigma_{S_t}, \Delta\sigma_{S_t}, \Delta\Delta\sigma_{S_t}]$. $\mathbf{m}_{S_t}$, and $\mathbf{\Sigma}_{S_t}$ are the HMM mean vector and covariance matrices at time t, respectively, and where diag[] denotes a matrix whose elements are diagonal entries. If there are no constraints between $x_t, \Delta x_t$ and $\Delta\Delta x_t$, then the most likely sequence generated by the HMMs should be the HMM mean sequences. However, the following strong constraints exist:

$$\Delta x_t = \sum_{i=-W}^{i=W} i x_{t+i} \Big/ \sum_{i=-W}^{i=W} i^2 \quad , \quad (2)$$

$$\Delta\Delta x_t = \frac{\sum_{i=-W}^{i=W} \{(2W+1)i^2 - (\sum_{j=-W}^{j=W} j^2)\} x_{t+i}}{2\{(\sum_{j=-W}^{j=W} j^4)(2W+1) - (\sum_{j=-W}^{j=W} j^2)^2\}} , \quad (3)$$

where W is the window size used to calculate the delta and delta-delta features. These equations are well-known for calculating the delta and the delta-delta features by using a static feature. When W is set to 1, Eq. (3) is written as:

$$\Delta x_t = \frac{x_{t+1} - x_{t-1}}{2}. \quad (4)$$

With these constraints, the most likely sequence is no longer the mean sequence. We obtain linear equations for all of x_t by taking the logarithm of Eq. (2), replacing all Δx_t and $\Delta\Delta x_t$ in $\mathbf{x}_t$ into x_t, taking the derivative of all of x_t, and setting the result equation to zero. The equations can be solved with an RLS algorithm, which is used in adaptive filters and in the sliding-window concept [5], which only considers the last L frames of x_t. Finally, we obtain recursive equations as follows:

$$\hat{\mathbf{H}}_t = \mathbf{P}^L_{t-1}\boldsymbol{\gamma}_L'(\boldsymbol{\gamma}_L\mathbf{P}^L_{t-1}\boldsymbol{\gamma}_L' + \boldsymbol{\Sigma}_{S_t})^{-1} \quad (5)$$

$$\mathbf{P}^L_{t-1} = \mathbf{J}_L(\mathbf{I} - \hat{\mathbf{H}}_t\boldsymbol{\gamma}_L)\mathbf{P}^L_{t-2}\mathbf{J}_L' + \boldsymbol{\Theta}_t \quad (6)$$

$$\mathbf{X}^L_t = \mathbf{J}_L\mathbf{X}^L_{t-1} + \hat{\mathbf{H}}_t(\mathbf{M}_{s_t} - \boldsymbol{\gamma}_L\mathbf{J}_L\mathbf{X}^L_{t-1}) \quad (7)$$

$$\mathbf{X}^L_t = [x_t, x_{t-1}, \ldots, x_{t-(L-1)}]', \quad (8)$$

where

$$\mathbf{J}_L = \underbrace{\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}}_{L}, \quad (9)$$

where $\mathbf{X}^L_t$ is time-windowed $\mathbf{X}$ defined by $\mathbf{X}^L_t = [x_t, x_{t-1}, \ldots, x_{t-(L-1)}]'$, γ_L is defined by $\gamma_L = [\underbrace{\gamma}_{2W+1}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{L-2W-1}]$, and I is an identify matrix. $\boldsymbol{\Theta}_t$ is defined by $\boldsymbol{\Theta}_t = [\sigma\mathbf{I}, \mathbf{0}, \ldots, \mathbf{0}]$, where σ is a positive large number. γ is an operator to calculate $\mathbf{x}_t$ form sequence of x_t.

The above method that obtained smoothed trajectory was originally proposed for synthesizing speech [6][7]. These equations were modified from their original formulation in [7] to make them similar to RLS formulas because the algorithm's initial value is different from the original formula.

These equations closely resemble recursions in Kalman filters. Sayed and Kaliath proved that Kalman filters are equivalent to filters using RLS algorithms [8]. Thus we discuss whether Eqs. (5) to (7) can be expressed by a Kalman filter formula as follows:

$$\mathbf{X}^L_{t+1} = \mathbf{J}_L\mathbf{X}^L_t + \mathbf{n}_t, \quad (10)$$

$$\mathbf{m}_{S_t} = \boldsymbol{\lambda}_L\mathbf{X}^L_t + \mathbf{w}_t \quad . \quad (11)$$

Processes $\mathbf{n}_t$ and $\mathbf{w}_t$ are assumed to be zero-mean Gaussian noise processes with variances $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Theta}_t$, respectively. From these two equations, we found that the filter in the HMM-trajectory is a Kalman filter.

3. Trajectory generation using switching Kalman Filters

Section 2 describes how a filter can be obtained by Kalman filters. To generate a smoothed trajectory from the mixture HMMs, first multiple linear dynamical system (LDS) models are generated from HMMs. The problem is how to generate a trajectory using multiple Kalman filters. Recently switching Kalman filters have been proposed by Murphy [4]. Switching Kalman filters have linear dynamical system (LDS) models that switch between models to generate smoothed trajectory using the LDS models. Let us briefly explain switching Kalman filters. $\mathbf{z}_t$ denotes a hidden continuous state, $\mathbf{y}_t$ denotes observation, and S_t denotes the discrete state number. Each discrete state has the following state and observation equations:

$$\mathbf{z}_{t+1} = \mathbf{A}\mathbf{z}_t + \mathbf{v}_t \quad (12)$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{z}_t + \mathbf{u}_t \quad . \quad (13)$$

$\mathbf{v}_t \sim N(0, \mathbf{Q}_t)$ and $\mathbf{u}_t \sim N(0, \mathbf{R}_t)$ are independent Gaussian noises. The switching Kalman filters switch these equations at time t. The basic definitions are described below.

$$\mathbf{z}_{t|\tau} = E[\mathbf{Z}_t \mid \mathbf{y}_{1:\tau}] \quad (14)$$

If τ = t, then $\mathbf{z}_{t|\tau}$ is called filtered statistics. If τ < t, then $\mathbf{z}_{t|\tau}$ is called predicted statistics. If τ > t

(total speech length), then $\mathbf{z}_{t|\tau}$ is called smoothed statistics.

$$\mathbf{V}_{t|\tau} = Cov[\mathbf{Z}_t \mid \mathbf{y}_{1:\tau}]. \quad (15)$$

These are the covariance values for $\mathbf{z}_{t|\tau}$.

The basic concept of switching Kalman filters is very simple. Figure 2 shows the operation for each time step of switching Kalman filters. Assume that there are two Kalman filters and their two sets of estimated mean and covariance at time t-1 are $(\mathbf{z}^1_{t-1|t-1}, \mathbf{V}^1_{t-1|t-1})$, $(\mathbf{z}^2_{t-1|t-1}, \mathbf{V}^2_{t-1|t-1})$, which are outputs of filter time t-1. Since switching Kalman filters perform two different filters for each estimated mean and variance at time t-1, four filter outputs are generated. If this procedure were performed to the end of the utterances, large amounts of filter outputs would be generated. To reduce the number of Gaussian distributions, switching Kalman filters combine the four outputs to two outputs in each time step, meaning that the switching Kalman filters keep the number of Gaussian distributions at two in each time.

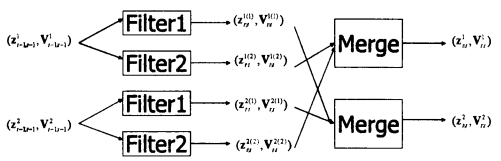

Figure 2. One step operation in switching Kalman filters

Switching Kalman filters consist of three algorithms: filter, smoother, and merge, which are described as follows.

3.1. Filter algorithm

The filter functions of input and output are shown as follows:

$$(\mathbf{z}_{t|t}, \mathbf{V}_{t,|t}, \mathbf{L}_t) = Filter(\mathbf{z}_{t-1|t-1}, \mathbf{V}_{t-1|t-1}, \mathbf{y}_t; \mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}). \quad (16)$$

This function estimates the mean, variance, and likelihood at time t using the estimated mean and variance at time t-1. In this calculation, the filter uses the state and observation equations described in Eqs. (12)(13). The operation, which is the same for ordinary Kalman filter algorithms, is shown as follows:

$$\mathbf{z}_{t|t-1} = \mathbf{A}\mathbf{z}_{t-1|t-1} \quad (17)$$

$$\mathbf{V}_{t-1|t} = \mathbf{A}\mathbf{V}_{t-1|t-1}\mathbf{A}' + \mathbf{Q}_t \quad (18)$$

$$\mathbf{e}_t = \mathbf{y}_t - \mathbf{C}\mathbf{z}_{t|t-1} \quad (19)$$

$$\mathbf{\Psi}_t = \mathbf{C}\mathbf{V}_{t-1|t}\mathbf{C}' + \mathbf{R}_t \quad (20)$$

$$\mathbf{K}_t = \mathbf{V}_{t-1|t}\mathbf{C}'\mathbf{\Psi}_t^{-1}. \quad (21)$$

The likelihood for this filter is calculated by

$$L_t = N(\mathbf{e}_t; 0, \mathbf{\Psi}_t), \quad (22)$$

$$\mathbf{z}_{t|t} = \mathbf{z}_{t|t-1} + \mathbf{K}_t\mathbf{e}_t, \text{ and} \quad (23)$$

$$\mathbf{V}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{C})\mathbf{V}_{t-1|t} = \mathbf{V}_{t-1|t} - \mathbf{K}_t\mathbf{\Psi}_t\mathbf{K}_t' \quad (24).$$

3.2. Smoother algorithm

Smoother functions of input and output are

$$(\mathbf{z}_{t|T}, \mathbf{V}_{t,|T}) = smooth(\mathbf{z}_{t+1|T}, \mathbf{V}_{t+1|T}, \mathbf{z}_{t|t}, \mathbf{V}_{t|t}, \mathbf{V}_{t+1|t+1}; \mathbf{A}, \mathbf{Q}). \quad (25)$$

This function estimates the mean and variance at time t using the estimated mean and variance at time t+1. This function calculates the estimations backward in time. The smoother uses the state and observation equation information described in Eqs. (12)(13). The operation is the same as the ordinary Kalman smoother algorithm and is shown as follows:

$$\mathbf{z}_{t+1|T} = \mathbf{A}\mathbf{z}_{t|t} \quad (26)$$

$$\mathbf{V}_{t+1|t} = \mathbf{A}\,\mathbf{V}_{t|t}\mathbf{A}' + \mathbf{Q} \quad (27)$$

$$\mathbf{J}_t = \mathbf{V}_{t|t}\mathbf{A}'\mathbf{V}_{t+1|t}^{-1} \quad (28)$$

$$\mathbf{z}_{t|T} = \mathbf{z}_{t|t} + \mathbf{J}_t(\mathbf{z}_{t+1|T} - \mathbf{z}_{t+1|t}) \quad (29)$$

$$\mathbf{V}_{t|T} = \mathbf{V}_{t|t} + \mathbf{J}_t(\mathbf{V}_{t+1|T} - \mathbf{V}_{t+1|t})\mathbf{J}_t'. \quad (30)$$

3.3. Merging algorithm

This function combines several Gaussian distributions into a single Gaussian distribution.

Assume a random variable X. μ_X^k is the means of the random variable at state k. $\mathbf{V}_X^k$ is the variance of the random variable. The merge functions of input and output are shown as follows:

$$(\mu_X, \mathbf{V}_X) = merge(\mu_X^k, \mathbf{V}_X^k, P^k) \tag{31}$$

$$\mu_X = \sum_k P^k \mu_X^k \tag{32}$$

$$\mathbf{V}_X = \sum_k P^k \mathbf{V}_X^k + \sum_k P^k (\mu_X^k - \mu_X)(\mu_Y^k - \mu_Y)'. \tag{33}$$

This merger is performed after the filter and smoother algorithms are performed.

3.4. Extension of HMM-trajectory using switching Kalman Filters

Figure 3 shows Switching Kalman filters for HMM-trajectory that correspond to the HMMs that have two states and two mixtures in each state. Each Gaussian distribution in HMM is assigned to one state in the switching Kalman filters. A linear dynamical system is converted from the Gaussian distribution. Therefore each state has a single linear dynamical system model expressed by the state Eq. (10) and observation Eq. (11). To use a Kalman smoother, Eqs. (10) (11) are modified so that the time length of state for $\mathbf{x}_t$ is the same as the window size for the delta and delta-delta features.

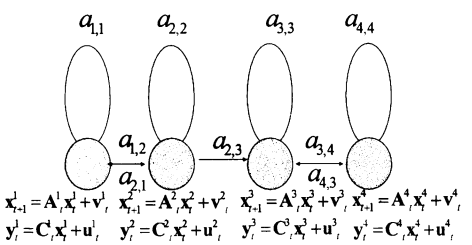

Figure 3. Switching Kalman filters corresponding to HMM having 2 states and 2 mixture Gaussian distributions in a state.

To obtain the nearest trajectory to the input speech, likelihood $L_t = N(\mathbf{e}_t; 0, \Psi_t)$ is changed to

$$L_t = N(\mathbf{z}_t; \mathbf{m}_k, \Sigma_k), \tag{34}$$

where $\mathbf{m}_k$ and Σ_k are a HMM mean vector and a covariance matrix at Kalman filter state k in respectively.

4. Experiments

Smoothed trajectory generated by a switching Kalman filter was compared with the trajectory from the maximum likelihood method. We used a TIMIT database to train 3 HMM states that have 3 mixture components per state. The sampling rate was 16 kHz, the frame shift was 10 msec, and the MFCC order was 12. In addition to MFCC, power was used as a feature parameter. Therefore, the total number of static features was thirteen. After calculating these static features, delta and delta-delta were obtained using Eqs. (2) and (3). HMMs were trained using these features. The state and transition equations shown in (10) and (11) were generated from the HMMs.

Smoothed trajectory was generated for the first test data. Figure 4 shows the result of fifth coefficients of MFCC. Solid lines show the best HMM mean sequence selected by a Viterbi algorithm. Dotted lines show the trajectory generated by maximizing likelihood with the HMM mean sequence. Dotted and dashed lines show the mean trajectory generated by the proposed method.

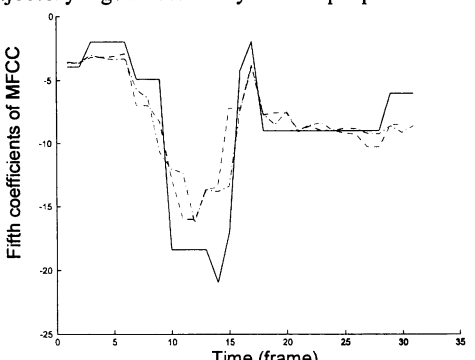

Figure 4. Smoothed trajectory generated by switching Kalman filters. (compared with maximum likelihood method.)

The mean trajectory is the mean of the obtained distributions by switching Kalman filters. The results show that switching Kalman filters with Gaussian mixture HMMs can generate the appropriate trajectory. In addition, we confirmed that the proposed method can generate Gaussina mixture distributions along the mean trajectory,

demonstrating that the proposed method has the possibility to be extended for using Gaussina mixture HMMs.

We tried to perform preliminary small phoneme recognition experiments using the smoothed trajectory generated by switching Kalman filters. Since Eq. (5) requires at least 5 frames to calculate delta and delta delta features, the number of phoneme lengths were 3, 4, and 5 frames. So we could not perform the recognition experiments. We have to extend our method so that it can deal with word or sentence recognition.

5. Conclusion

This report proposes a new trajectory generation method that extends HMM-trajectory by creating linear dynamic systems from Gaussian mixture HMMs. Switching Kalman filters generate trajectory using linear dynamical systems. The generation experiment shows that the filters can be used to generate appropriate trajectory. In the future we will apply this method to large vocabulary speech recognition.

REFERENCES

[1] Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri, "Recognition method with parametric trajectory synthesized using direct relations between static and dynamic feature vector time series," Proc. ICASSP, pp. 957-960, May. 2002.

[2] Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri, "Recognition method with parametric trajectory generated from mixture distribution HMMs," Proc. ICASSP, pp. 124-127, April 2003.

[3] Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri, "A theoretical analysis of speech recognition based on feature trajectory models," Proc. pp. 549-552, Oct. 2004.

[4] K. P. Murphy, "Filtering, Smoothing and the Junction Tree Algorithm," http://www.cs.ubc.ca /~murphyk/papers.html.

[5] C.F.N. Cowan and P.M. Grant, "Adaptive filters," Prentice-Hall, INC., Englewood Cliffs, N. J., 1985.

[6] K. Tokuda, T. Kobayashi, and S. Imai," Speech parameter generation from HMM using dynamic features," Proc. ICASSP, pp. 660-663, May 1995.

[7] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," Proc. ICASSP, pp. 89-392, May 1996.

[8] A. H. Sayed and T. Kailath, "A state-space approach to RLS filtering," IEEE, Signal Process. Mag., vol. 11, pp. 18-60, 1994.