ACOUSTIC SEGMENTATION USING SWITCHING STATE KALMAN FILTER

Yanli Zheng and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801

ABSTRACT

Segmenting the acoustic signal in the TIMIT database by a switching state Kalman filter model is reported in this paper. According to the assumption that the high dimensional acoustic feature vector of the LSF (Line Spectrum Frequency) of the speech signal is probably embedded in a low dimensional space, a two dimensional vector is used to represent the continuous state vector in this model. The parameters of the model are initialized by PPCA (probabilistic principle component analysis) and first order vector auto-regression, and are re-estimated by the EM algorithm. We show that this model can be used to classify vowels, nasals, frication and silence by an approximate Viterbi inference.

1. INTRODUCTION

In the conventional speech recognition system, the features of the speech signal, for example MFCC, are modeled by HMM. The dynamics of the speech signal are captured by the movement of the discrete state variable. In this model, the dependency among the frames within each discrete state is ignored.

The application of a stochastic linear system in speech recognition is introduced in [1], where the dimension of the continuous space is the same as the dimension of the observation space. Comparing to the independent-frame HMM, the model in [1] showed a superior performance. A more parsimonious, target directed dynamic model is reported in [2], where a performance comparable to conventional HMM is achieved. A derivation of the EM algorithm for the estimation of extended Kalman filter (EKF) parameters is reported in [3], and a dual estimation method with much better convergence properties is proposed.

Although the most appropriate model structure for speech recognition is still an open question, it is obvious that inter-frame dependence of acoustic observations is caused in part by the continuous movement of the tongue, lips and other articulators. Research with the articulator data [4, 5] indicates that two factors are sufficient to represent the steady state of the articulators of the vowels. In this paper, we demonstrate that the switching Kalman filter model with only a two dimensional vector as the hidden continuous state can group the acoustic data into vowels, nasals, frication and silence effectively.

2. SWITCHING KALMAN FILTER MODEL

The switching state model can be described by the following equations:

$$\begin{aligned}
&x_t = A(s_t)x_{t-1} + v_t(s_t), \\
&v_t(s_t) \sim N(0, Q(s_t)) \\
&y_t = C(s_t)x_t + \mu(s_t) + w_t(s_t), \\
&w_t(s_t) \sim N(0, R(s_t)) \\
&T(i,j) = \Pr(s_t = i \mid s_{t-1} = j), \\
&\mathit{where}\ i, j = 1, 2, \cdots G
\end{aligned}$$

A switching state Kalman filter can be used to model piecewise linear dynamic time series. The equations above assume that the discrete switching state s_t is also a first order Markov process.

2.1. EM learning of switching Kalman filter

In the learning process, we assume that the status of the discrete state is known (i.e. the segmented data is used in the training), so only the continuous states and parameters of the Kalman filter are estimated.

In the EM algorithm, sufficient statistics of the hidden continuous state are calculated by the Kalman smoother in the E step, and the values of the parameters were found by maximizing the lower bound of the likelihood function in the M step [6].

$$\sum_{n=1}^{N}\log p(Y_n^T \mid \theta) \geq \sum_{n=1}^{N}\{\int dX\cdot p(X_n \mid Y_n,\theta_{old})\log p(X_n,Y_n \mid \theta_{new}) - \int dX\cdot p(X_n \mid Y_n,\theta_{old})\log p(X_n \mid Y_n,\theta_{old})\}$$

where N is the number of different training sequences, $Y_n^T = \{y_{n,1}, y_{n,2}, \cdots y_{n,T}\}$, $X_n^T = \{x_{n,1}, x_{n,2}, \cdots x_{n,T}\}$,and $\theta(s_t = j) = \{A_j, C_j, \mu_j, Q_j, R_j\}$,where $j = 1,2,\cdots G$, G is the total number of classes ,

E step: The sufficient statistics are calculated as follows:

$$\hat{x}_{t|T} = E[x_t \mid Y^T]$$

$$V_{t|T} = \mathrm{cov}[x_t x_t^{'} \mid Y^T]$$

$$V_{t,t-1|T} = \mathrm{cov}[x_t x_{t-1}^{'} \mid Y^T]$$

$$< \hat{x}_{t|T}\hat{x}_{t|T}^{'} S_t(j) >= V_{t|T} S_t\ (j) + \hat{x}_{t|T}\hat{x}_{t|T}^{'} S_t(j)$$

$$< \hat{x}_{t|T}\hat{x}_{t-1|T}^{'} S_t(j) >= V_{t,t-1|T} S_t\ (j) + \hat{x}_{t|T}\hat{x}_{t-1|T}^{'} S_t(j)$$

M step: The parameter updating formulas are as follows:

$$A_j = [\sum_{n=1}^{N}\sum_{t=2}^{T_n} < \hat{x}_{t|T}\hat{x}_{t-1|T}{}' S_t(j) >][\sum_{n=1}^{N}\sum_{t=2}^{T_n} < \hat{x}_{t-1|T}\hat{x}'_{t-1|T}\ S_t(j) >]^{-1}$$

$$C_j = [\sum_{n=1}^{N}\sum_{t=2}^{T_n} y_{t,n}\cdot\hat{x}_{t,n|T}^{'} S_t(j)][\sum_{n=1}^{N}\sum_{t=2}^{T_n} < \hat{x}_{t|T}\hat{x}_{t|T}{}' S_t(j) >]^{-1}$$

$$Q_j = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n} < \hat{x}_{t|T}\hat{x}_{|Tt}{}' S_t(j) > - A_j < \hat{x}_{t|T}\hat{x}_{t-1|T}{}' S_t(j) >}{[\sum_{n=1}^{N} T_n - N]}$$

$$R_j = [\sum_{n=1}^{N} T_n(j)]^{-1}[\sum_{n=1}^{N}\sum_{t=1}^{T_n}(y_{t,n}y_{t,n}^{'} S_t(j) - C_j\hat{x}_{t,n|T}^{'} y_{t,n}^{'} S_t(j)]$$

$$\mu_j = [\sum_{n=1}^{N} T_n(j)]^{-1}[\sum_{n=1}^{N}\sum_{t=1}^{T_n} y_{t,n} S_t(j)]$$

2.2. Approximate optimal discrete state sequence inference using Viterbi approximation [7]

Given the trained model, the segmentation problem is defined as to find the optimal sequence of states S^{T*}, where $S^{T*} = \{S_1^*, S_2^*, \cdots, S_T^*\}$, and $S^{T*} = \arg\max_{S^T}\{\log p(S^T \mid Y)\}$.

The exact inference of the optimal state sequence is intractable, since

$$p(Y^T \mid S^T) = \int dX^T p(Y^T \mid X^T, S^T) p(X^T \mid S^T)$$

and $p(X^T \mid S^T)$ is a 2T-dimensional jointly Gaussian distribution.

The approximate Viterbi inference can be stated as follow:

1. Initialization

$$J_1(i) = \frac{1}{2}(y_1 - Cx_1)'(V_1 + R_i)^{-1}(y_1 - Cx_1) + \frac{1}{2}|V_1 + R_i| - \log(\pi_i)$$

$$i = 1,2,\cdots,G$$

2. Recursion

$$J_t(i) = \min_j(J_{t-1}(j) + J_{t,t-1}(i,j))$$

$$J_{t,t-1}(i,j) = \frac{1}{2}(y_t - C_i\hat{x}_{t|t-1}(i,j))'(C_i V_{t|t-1}(i,j) C_i + R_i)^{-1}(y_t - C_i\hat{x}_{t|t-1}(i,j)) + \frac{1}{2}|C_i V_{t|t-1}(i,j) C_i + R_i| - \log(T_{j\to i})$$

$$\varphi_t^*(i) = \arg\min_j(J_{t-1}(j) + J_{t,t-1}(i,j))$$

$$j = 1,2,\cdots,G, \quad i = 1,2,\cdots,G \quad t = 2,3,\cdots,T$$

3. Termination

$$s_T^* = \arg\min_i\ (J_T(i))$$

4. Path backtracking

$$s_t^* = \varphi(s_{t+1}^*) \quad t = T-1,\cdots,1$$

3. EXPERIMENT

We used this switching state Kalman filter to analyze the acoustic signals in the TIMIT database. In this experiment, the four discrete states are used to represent categories of vowels, nasals, frication and silence respectively.

The original waveform data is down-sampled to 8 KHz. The acoustic features are represented by the coefficients of a 10th order LSF (line spectral frequency). The frame rate is 10ms with a window size of 20 ms, and a Hamming window is used. The reason to choose LSF as the feature vector is that the series of LSF coefficients during vowel and glides may be modeled by a linear dynamic model [8]. The dimension of the hidden continuous state is 2, which is supposed to represent the low dimensional dynamics embedded in a high dimensional observation space.

In this experiment, the data of 10 female speakers in the DR1 of the TRAIN data in TIMIT were used to train the model, and 4 other female speakers in the DR1 of the TRAIN data were used as the testing set.

3.1. Initialization and learning the parameters of the model

Although EM algorithm is guaranteed to increase the log likelihood until convergence, it does not guarantee to find the global optimal solution. When starting from a poor initial point, the converged solution may be trapped in some local maximum which is far away from the optimal solution. In our experiment, C_j and R_j are initialized by probabilistic PCA [9], and the state matrix A_j and Q_j are initialized by the first order vector auto-regression, where j=1,2,3,4 represent 4 categories or 4 discrete states, and 1 is the state of vowels, semivowels and glides; 2 is the state of nasals; 3 is the state of frication which includes stop release and fricative segments, and 4 is the state of the silence and stop closure.

After the initialization, the model is learned by the EM algorithm introduced in section 2.1. It was found that the learning algorithm tends to be unstable if no constraint is put on the noise covariance of Q_j and R_j. To avoid instability, Q_j and R_j matrices are kept unchanged during the EM updating. It was found that the learning is converged after only 2 iterations which indicates that the initial parameters are near a local optimal solution. It is instructive to take a look at the state matrix A_j and state noise covariance matrix Q_j found by the learning algorithm.

$$A_1 = \begin{bmatrix} 0.94 & 0.01 \\ 0.02 & 0.96 \end{bmatrix} \quad Q_1 = \begin{bmatrix} 0.12 & 0 \\ 0 & 0.07 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 0.7 & 0.09 \\ 0.03 & 0.62 \end{bmatrix} \quad Q_2 = \begin{bmatrix} 0.43 & 0 \\ 0 & 0.58 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 0.77 & 0.03 \\ -0.02 & 0.7 \end{bmatrix} \quad Q_3 = \begin{bmatrix} 0.35 & 0 \\ 0 & 0.44 \end{bmatrix}$$

$$A_4 = \begin{bmatrix} 0.67 & 0.02 \\ -0.09 & 0.76 \end{bmatrix} \quad Q_4 = \begin{bmatrix} 0.49 & 0 \\ 0 & 0.37 \end{bmatrix}$$

Notice that the diagonal elements of A_1 are significantly larger than other A_j's, and the diagonal elements of Q_1 are significantly smaller than other Q_j's. One explanation of the results is that the acoustic features of vowels are more correlated (change more smoothly), and are better represented by a low dimensional manifold.

3.2. Manner Segmentation using the switching Kalman filter

We used the decoding algorithm introduced in section 2.2 to segment speech data from the TIMIT database. The number of frames used in training and testing sets and the correctness rate are listed in Table 1. As an illustration of the segmentation, a decoding sequence as well the ground truth is plotted in Figure 1. And one sample of the trajectory of the state space is shown in Figure 2.

Table 2 is a confusion matrix, showing the percent misclassifications of each type. Nasals are primarily misclassified as vowels, while frication and silence segments are primarily misclassified as one another.

4. CONCLUSION

In this paper, a switching state Kalman filter model is used to segment acoustic signals from TIMIT into four different subcategories. The preliminary result shows that this model can successfully capture dynamics of the low dimensional manifold embedded in the high dimensional feature space (LSF feature space). The (frame) classification rate of the vowels is about 90%. Our experiment shows that the parameters initialized by PPCA (probabilistic principle component analysis) and first order vector auto-regression tends to be a local optimal solution. And the Viterbi decoding works well with this initial set of parameters.

Table 1. Experiment result

	Training Set			
	Vowels	Nasals	Stops and Fricatives	Silence and Stop Closure
#Frames	10914	1480	6340	5272
Correct %	89	61	51	80
	Testing Set			
#Frames	4579	580	2239	1939
Correct %	92	57	51	71

Table 2. Confusion matrix of the segmentation

	Classification results of the training set			
	Vowels	Nasals	Stops and Fricatives	Silence and Stop Closure
Vowels	0.88	0.03	0.05	0.02
Nasals	0.22	0.61	0.06	0.10
Stops and Fricatives	0.14	0.02	0.51	0.32
Silence and Stop Closure	0.06	0.02	0.13	0.80
	Classification results of the training set			
Vowels	0.92	0.02	0.05	0.01
Nasals	0.37	0.57	0.04	0.03
Stops and Fricatives	0.16	0.01	0.51	0.32
Silence and Stop Closure	0.06	0.04	0.18	0.71

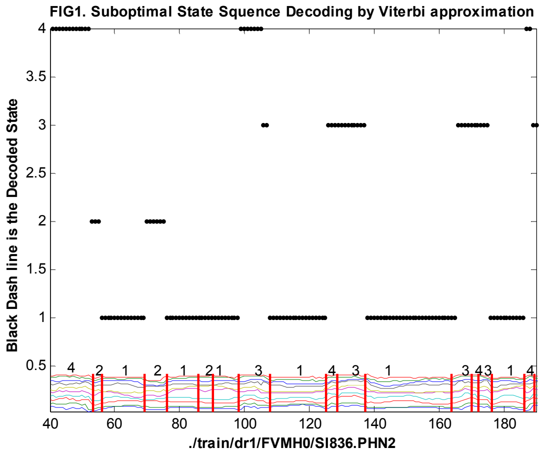

Figure1. The segmentation of the speech signal, where the red vertical line is the true boundary of the segmentation. Signals with amplitude less than 0.5 are LSF coefficients, and the digits just above them are the true class labels of the LSF frames. The black dashed lines are the segmentation resulting from Viterbi decoding with amplitude representing its category. Group 1 is vowels, semivowels and glides; group 2 is nasals; group 3 is stops and fricatives; and group 4 is silence and stop closure.

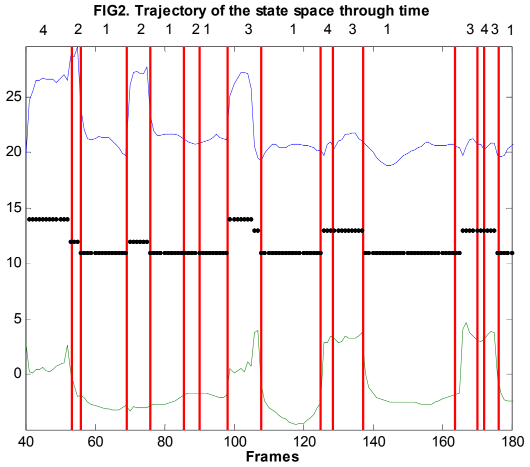

Figure2. One sample of the trajectory of the continuous state, where the red vertical line is the true boundary of the segmentation. the digits just above the box are the true group of the LSF frames, and the horizontal lines in the middle are the segmentation by the Viterbi algorithm. The group of 1 to 4 corresponds to the amplitude of 11 to 14.

5. REFERENCES

[1] V. Digalakis, J.R.Rohlicek, and M. Ostendorf, "ML estimation of stochastic linear system with EM algorithm and its application," *IEEE Trans. Speech AudioProcess.1*, 431-442, (1993).

[2] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," JASA 108(6), 3036-3048, *(2000).*

[3] R. Togneri, J. Ma, L. Deng, "Parameter estimation of a target-directed dynamic system model with switching states," Signal Processing 81, 975-987, (2001).

[4] R. Harshman, P. Ladefoged, and L. Goldstein, "Factor analysis of tongue shapes," JASA 62(3), 693-707, (1977)

[5] Y. Zheng and M. Hasegawa-Johnson, "PARAFAC analysis of the three-dimensional tongue shape," JASA 113(1), (2003 in print)

[6] C.L. Giles and M. Gori, "*Adaptive Processing of Sequences and Data Structures*," Lecture Notes in Artificial Intelligence, 168-197. Berlin: Springer-Verlag.

[7] V. Pavlovic, J. M. Rehg, T.J. Cham and K. Murphy, "A Dynamic Bayesian Network Approach to Figure Tracking Using Learned Dynamical Models," Proc. International Conference on Computer Vision. Kerkyra, Greece (1999) .

[8] M. Hasegawa-Johnson and A. Alwan, "Speech Coding: Fundamentals and Applications," *Wiley Encyclopedia of Telecommunications*, J Proakis editor (in press).

[9] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," Journal of the Royal Statistical Society, Series B **61**(3), 611–622, (1999)