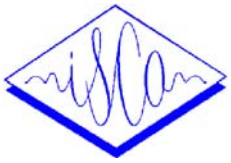

4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

Optimal Filtering and Smoothing for Speech Recognition Using a Stochastic Target Model

G. Ramsay and L. Deng

Department of Electrical & Computer Engineering
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1.

ABSTRACT

This paper presents a stochastic target model of speech production, where articulator motion in the vocal tract is represented by the state of a Markov-modulated linear dynamical system, driven by a piecewise-deterministic control trajectory, and observed through a non-linear function representing the articulatory-acoustic mapping. Optimal filtering and smoothing algorithms for estimating the hidden states of the model from acoustic measurements are derived using a measure-change technique, and require solution of recursive integral equations. A sub-optimal approximation is developed, and illustrated using examples taken from real speech.

1. Introduction

The modelling of temporal and contextual variation observed in the acoustic signal is a central concern in speech recognition, and is not handled well by traditional HMM methodologies. Since much of the observed variability in speech can be attributed to articulatory dynamics, it is reasonable to assume that incorporation of an explicit model of articulatory phenomena should eventually lead to improved performance, and to a more compact and precise parameterization of many sources of contextual variation.

In a recent paper [1], we proposed a stochastic target model of speech production, where articulator motion in the vocal tract is represented by the state of a linear dynamical system observed through a non-linear function representing the articulatory-acoustic mapping. The system parameters are modulated by a Markov chain representing phonological sequences, and the system is driven by a control input consisting of a piecewise-deterministic trajectory modelling dynamic targets in the articulatory space.

The contribution of the present paper is the derivation of the optimal non-linear filtering and smoothing recursions needed to estimate the hidden states of the model from acoustic measurements. Results from a full articulatory recognition task are not yet available, but data-fitting experiments on cepstral data from real speech are provided to illustrate a number of phenomena affecting performance.

2. Model Formulation

Assume an underlying complete probability space $(\Omega, \mathcal{F}, P)$.

Let $S = \{S_m : m \in \mathbb{N}\}$ be a finite-state homogeneous Markov chain generating phonological sequences, taking values in a finite alphabet $\mathcal{S} = \{s_i : i = 1 \ldots N_S\}$. Define a transition matrix $\Pi = [\pi(i,j) : i,j \in \mathcal{S}]$, where $\pi(i,j) = P(S_{m+1} = j | S_m = i)$, $\pi(i,j) \geq 0$, $\sum_{j \in \mathcal{S}} \pi(i,j) = 1$, and assign an initial distribution $\pi_1 : \mathcal{S} \to [0,1]$.

Let $T = \{T_m : m \in \mathbb{N}\}$ be a process representing the durations of the states in S, where each T_m is drawn from a Poisson distribution with mean $\mu_\tau(S_m)$ chosen from a family of duration parameters $\mathcal{T} = \{\mu_\tau(i) \in \mathbb{R} : i \in \mathcal{S}\}$ Define also $\tau = \{\tau_m : m \in \mathbb{N}\}$, where $\tau_1 = 1$, $\tau_m = \tau_{m-1} + T_m$ $(\forall m > 0)$.

To simplify notation, construct a Markovian representation $Z = \{Z_n : n \in \mathbb{N}\}$ of the semi-Markov chain (S,T), where Z_n takes values in a countable state-space $\mathcal{Z} = \{z_t^i : i \in \mathcal{S}, t \in \mathbb{N}\}$ and define mappings $f_S : \mathcal{Z} \to \mathcal{S}$, $f_T : \mathcal{Z} \to \mathbb{N}$ by $f_S(z_t^i) = i$, $f_T(z_t^i) = t$. Let $\overline{\Pi}$ and $\overline{\pi}_1$ be the transition matrix and initial distribution for Z such that the processes $(f_S(Z_{\tau_m}), f_T(Z_{\tau_{m+1}-1}))$ and (S_m, T_m) are indistinguishable.

Each phonological sequence produced by (S,T) induces a statistical distribution of possible target trajectories on an articulatory space $\mathcal{X} = \mathbb{R}^p$, representing abstract motor commands. The trajectories are assumed to be *piecewise-deterministic*, in that the shape of any section of trajectory corresponding to a particular state is defined by parameters drawn from an associated probability distribution when the state is entered. The subsequent evolution of the trajectory is completely determined by these parameters until the next phonological transition occurs. The shape of each parameter distribution models systematic static/dynamic compensatory effects in realizing a particular phonological unit.

A previous paper presented a model for piecewise-constant trajectories, where the distributions modelled variation in absolute spatial targets. Here we present an extension that allows for piecewise-polynomial trajectories, where distributions describe the offset (position), slope (velocity), etc., with optional continuity constraints across segment boundaries.

Let $U = \{U_n : n \in \mathbb{N}\}$ be a process representing target trajectories in $\mathcal{X}$. Assume that the sample paths of U consist of polynomial segments $U_{[\tau_m,\tau_{m+1})}$ of order Q, with coefficients drawn from a family of distributions $\psi = \{\psi(i,j) : i \in \mathcal{S}, j = 1 \ldots Q\}$ according to the current Markov state S_m. Suppose for convenience that $\psi(i,j) \sim N(\mu_u(i,j) \in \mathbb{R}^p, \Sigma_u(i,j) \in \mathbb{R}^{p\times p})$, defined by a set of target parameters $\mathcal{U} = \{(\mu_u(i,j), \Sigma_u(i,j)) : i \in \mathcal{S}, j = 1 \ldots Q\}$.

To obtain a Markovian representation for U, define processes $U^i = \{U^i_n : n \in \mathbb{N}\} : i = 0 \ldots Q$, $\Psi^i = \{\Psi^i_n : n \in \mathbb{N}\} : i = 1 \ldots Q$, such that U^i records the i'th difference of the target process $U \equiv U^0$, and Ψ^i records changes in the i'th polynomial coefficient. These are generated by the recursion

$$\begin{aligned} \Psi^i_n &= I_{B_i}(Z_n)[\mu_u(f_S(Z_n), i) + \Sigma_u^{\frac{1}{2}}(f_S(Z_n), i)V^i_n] + \Sigma_\Delta^{\frac{1}{2}} V^i_n, \\ U^i_n &= I_{A_i^c}(Z_n)[U^i_{n-1} + U^{i+1}_{n-1}] + \Psi^{i+1}_n, \end{aligned}$$

where $V^i = \{V^i_n : n \in \mathbb{N}\}$ are zero-mean, unit variance Gaussian i.i.d. processes, Σ_Δ is a small positive-definite noise covariance, and $U^Q_n, U^i_0 \equiv 0$. Here $I_{A_i}, I_{B_i} : \mathcal{Z} \to \{0,1\}$ are indicator functions for sets $A_i, B_i \subset \{z \in \mathcal{Z} : f_T(z) = 1\}$ which determine the behaviour of the target process at segment boundaries. States in A_i^c preserve continuity of the i'th difference when a transition occurs, whereas states in B_i introduce a random jump in U^i_n sampled from $\psi(f_S(Z_n), i)$.

Let $X = \{X_n : n \in \mathbb{N}\}$ describe the state of an articulatory model evolving on $\mathcal{X}$, and let $Y = \{Y_n : n \in \mathbb{N}\}$ represent indirect observations of X in a measurement space $\mathcal{Y} = \mathbb{R}^q$. The following model for (X,Y) is assumed,

$$\begin{aligned} X_n &= \sum_{j=1}^{d-1} A(f_S(Z_n), j) X_{n-j} + A(f_S(Z_n), d) U_{n-1} + V_n, \\ Y_n &= h(X_n) + W_n, \end{aligned}$$

with initial state $X_0 \sim N(\mu_0 \in \mathbb{R}^p, \Sigma_0 \in \mathbb{R}^{p\times p})$. System matrices are chosen from a family of system parameters $\mathcal{A} = \{A(i,j) \in \mathbb{R}^{p\times p} : i \in \mathcal{S}, j = 1 \ldots d\}$, $\sum_j A(i,j) = I$. Gaussian i.i.d. noise processes $V = \{V_n : n \in \mathbb{N}\}$, $W = \{W_n : n \in \mathbb{N}\}$ represent unmodelled disturbances in $\mathcal{X}$ and $\mathcal{Y}$, independent of Z, U, X_0, with $V_n \sim N(0, \Sigma_v \in \mathbb{R}^{p\times p})$ $W_n \sim N(0, \Sigma_w \in \mathbb{R}^{q\times q})$. The non-linear function $h : \mathcal{X} \to \mathcal{Y}$ represents the articulatory-acoustic mapping, and is assumed known. This completes the model description.

The use of a semi-Markov process for modelling the underlying phonological sequence is standard [2], and follows a number of recent proposals for "segment-based" HMMs. Models with trend functions have been developed previously [3][4]; we note in particular the model described by Russell [5][6], where prior distributions are placed on the trajectory parameters, and the "target-state" model proposed by Digalakis et al. [7][8], both of which are similar to the U process described above, which is adapted from [9][10]. Several of the above proposals have been described as "target models"; however, in all of these models, the "targets" are essentially reference templates with some residual variation. The stochastic target model outlined here is closer to proposals in [11][12], in that the target trajectory represents an underlying goal towards which the hidden articulatory state relaxes asymptotically, and the dynamics are not reset at state boundaries, allowing undershoot effects and inter-state interaction to occur. It provides a useful separation between different levels of phonetic representation, parameterized independently, and an important distinction is made between sources of systematic linguistic variability (S,T,U) and unmodelled surface noise (V,W). The remainder of this paper develops the filtering/smoothing recursions needed to calculate estimates of S, T, U, X from sample paths of Y.

3. State Estimation

The general state estimation problem concerns the calculation of $E\{\phi_k | \mathcal{F}^Y_n\}$ for an arbitrary integrable $\mathcal{F}^X_k$-measurable random variable ϕ_k, where the filtrations $\mathcal{F}^X = \{\mathcal{F}^X_n : n \in \mathbb{N}\}$, $\mathcal{F}^Y = \{\mathcal{F}^Y_n : n \in \mathbb{N}\}$, $\mathcal{G} = \{\mathcal{G}_n : n \in \mathbb{N}\}$ are defined by $\mathcal{F}^X_n = \sigma(Z_k, U^i_k, X_k : k \le n)$, $\mathcal{F}^Y_n = \sigma(Y_k : k \le n)$, $\mathcal{G}_n = \mathcal{F}^X_n V \mathcal{F}^Y_n$. Adopting the reference-probability approach proposed by Elliott (e.g. [13]), construct processes $\lambda = \{\lambda_n : n \in \mathbb{N}\}$, $\Lambda = \{\Lambda_n : n \in \mathbb{N}\}$ according to

$$\begin{aligned} \lambda_n &\triangleq \exp(\frac{1}{2} h(X_n)^T \Sigma_w^{-1} h(X_n) - h(X_n)^T \Sigma_w^{-1} Y_n), \\ \Lambda_n &\triangleq \Pi_{k\le n} \lambda_k. \end{aligned}$$

Λ is a uniformly-integrable $\mathcal{G}$-martingale under P; define a new measure $\overline{P}$ on $(\Omega, \mathcal{G}_\infty)$ by setting the restriction of the Radon-Nikodym derivative $d\overline{P}/dP$ to $\mathcal{G}_n$ equal to Λ_n. The measures P and $\overline{P}$ are mutually absolutely continuous, and invoking a discrete-time version of the Girsanov theorem, it can be shown that, under $\overline{P}$, $\mathcal{F}^X_\infty$ and $\mathcal{F}^Y_\infty$ are independent σ-algebras, Y is an i.i.d. Gaussian process with distribution $N(0, \Sigma_w)$, and the restrictions of P and $\overline{P}$ to $(\Omega, \mathcal{F}^X_\infty)$ coincide. A reverse measure change defined by similar processes $\overline{\lambda}$, $\overline{\Lambda}$, with $\overline{\lambda}_n = 1/\lambda_n$ recovers the original probability measure P. Furthermore, conditional expectations under P and $\overline{P}$ are linked by the Kallianpur-Striebel formula,

$$E\{\phi_k | \mathcal{F}^Y_n\} = \frac{\overline{E}\{\phi_k \overline{\Lambda}_n | \mathcal{F}^Y_n\}}{\overline{E}\{\overline{\Lambda}_n | \mathcal{F}^Y_n\}} \triangleq \frac{\sigma_n(\phi_k)}{\sigma_n(1)} : k \le n.$$

State estimation therefore reduces to calculation of $\sigma_n(\phi_k)$, since $\sigma_n(1)$ is a normalizing factor. This is easier than the original problem since Y is independent under $\overline{P}$. Of particular interest are the forward, backward, and smoothed unnormalized joint conditional densities α_k, β_k, and γ_k of $Z_k, U^0_k \ldots U^{Q-1}_k, X_k \ldots X_{k-d+2}$, defined such that for all integrable Borel functions $\phi_k(Z_k, U^0_k \ldots U^{Q-1}_k, X_k \ldots X_{k-d+2})$

$$\begin{aligned} \sigma_n(\phi_k) &= \sum_{\mathcal{Z}} \int \phi_k \ \gamma_k \ du_0 \ldots du_{Q-1} \ dx_1 \ldots dx_{d-1}, \\ \sigma_k(\phi_k) &= \sum_{\mathcal{Z}} \int \phi_k \ \alpha_k \ du_0 \ldots du_{Q-1} \ dx_1 \ldots dx_{d-1}, \\ \beta_k &= \overline{E}\{(\Lambda_n/\Lambda_{k-1}) | \mathcal{F}^X_{k-1} V \mathcal{F}^Y_n\}. \end{aligned}$$

Observe that, using independence properties under $\overline{P}$,

$$\overline{E}\{\phi_k\ \overline{\Lambda}_n|\mathcal{F}_n^Y\} = \overline{E}\{\phi_k\ \beta_{k+1}\ \overline{\Lambda}_k|\mathcal{F}_n^Y\}$$
$$\overline{E}\{\phi_k\ \overline{\Lambda}_k|\mathcal{F}_k^Y\} = \overline{E}\{\overline{E}\{\phi_k\ \overline{\lambda}_k|\mathcal{F}_{k-1}^X V\mathcal{F}_k^Y\}\overline{\Lambda}_{k-1}|\mathcal{F}_k^Y\}$$
$$\beta_k = \overline{E}\{\overline{\lambda}_k\overline{E}\{(\overline{\Lambda}_n/\overline{\Lambda}_k)|\mathcal{F}_k^X V\mathcal{F}_n^Y\}|\mathcal{F}_{k-1}^X V\mathcal{F}_n^Y\}$$

Re-writing each quantity as a function of variables at time $k-1$ or $k+1$ using the model equations, we can apply the independence of the Y process under $\overline{P}$ to express the conditional expectations as integrals w.r.t. an appropriate density, where the observation y_k enters only as a constant parameter. To accomplish this, define for $i = 1 \ldots Q-1$

$$K_i(z,u,v_1,v_2) = (I_{B_i}(z)\Sigma_u^{\frac{1}{2}}(f_S(z),i) + \Sigma_\Delta^{\frac{1}{2}})^{-1}$$
$$[u - I_{A_i^c}(z)(v_1+v_2) - I_{B_i}(z)\mu_u(f_S(z),i)],$$
$$K_0(z,x_0\ldots x_{d-1},v) = \Sigma_v^{-\frac{1}{2}}[x_0 - A(f_S(z),1)x_1$$
$$\ldots - A(f_S(z),d-1)x_{d-1} - A(f_S(z),d)v)].$$

Substituting the new variables in the expectations, integrating over z,$u_0\ldots u_{Q-1}$,x, defining $\Gamma(x) = \exp(x^Tx)/(2\pi)^{\frac{p}{2}}$, and comparing the result with the original definitions of α_k, β_k, γ_k demonstrates that the conditional densities satisfy

$$\alpha_k(z,u_0\ldots u_{Q-1},x_1\ldots x_{d-1}) = \sum_{\zeta\in\mathcal{Z}}\overline{\pi}(\zeta,z)\int\overline{\lambda}_k(y_k,x_1)$$
$$\Gamma(K_0(z,x_1,\ldots,x_{d-1},\xi,v_0))\prod_{i=1}^{Q-1}\Gamma(K_i(z,u_i,v_i,v_{i+1}))$$
$$\alpha_{k-1}(\zeta,v_0\ldots v_{Q-1},x_2\ldots x_{d-1},\xi)\ dv_0\ldots dv_{Q-1}\ d\xi,$$
$$\beta_k(z,u_0\ldots u_{Q-1},x_1\ldots x_{d-1}) = \sum_{\zeta\in\mathcal{Z}}\overline{\pi}(z,\zeta)\int\overline{\lambda}_k(y_k,\xi)$$
$$\Gamma(K_0(\zeta,\xi,x_1,\ldots,x_{d-1},u_0))\prod_{i=1}^{Q-1}\Gamma(K_i(\zeta,v_i,u_i,u_{i+1}))$$
$$\beta_{k+1}(\zeta,v_0\ldots v_{Q-1},\xi,x_1\ldots x_{d-2})\ dv_0\ldots dv_{Q-1}\ d\xi,$$

$$\gamma_k(z,u_0\ldots u_{Q-1},x_1\ldots x_{d-1}) = \alpha_k\ \beta_{k+1}.$$

where α_0 is the prior density and $\beta_n \equiv 1$. The resulting pair of integral equations define forward-backward recursions for the unnormalized conditional densities α_k, β_k, γ_k. In the case where $h(\cdot)$ is linear, the integral can be calculated explicitly and the conditional density propagates as a mixture of Gaussians with an exponentially-growing number of components, one for each possible state path through the Markov chain. In general, the integral must be evaluated numerically, or approximations introduced. If the measurement function $h(\cdot)$ is expanded locally at each time step for each component in the current mixture as a Taylor series centred on the previous conditional mean, a linearized approximation can be developed which performs reasonably in practice. After considerable manipulation, it is possible to show that this is in fact equivalent to constructing a tree of extended Kalman smoothers matched to each sequence of linearized systems defined by the underlying Markov chain. Several proposals of this kind have already been investigated extensively in the target-tracking literature [14], and the algorithm for each path can be written in the Joseph form of the Bryson-Frazier two-filter smoother. The standard details are omitted here.

4. Simulation Results

To illustrate the behaviour of the model and the state-estimation algorithms, the simulation results presented here have been carried out on a small corpus of acoustic data, and the articulatory interpretation of the hidden states is ignored at present, setting the observation function $h(.)$ to be the identity mapping. The aim was to examine the data-fitting capability of the model, and to determine the importance of the different parameters during estimation; computational burden precludes a full evaluation at present.

Training and test data, each consisting of 100 hand-labelled words from a single speaker in the speaker-dependent portion of the DARPA Resource Management database, were parameterized using 12 mel-frequency cepstral coefficients, calculated every 10ms using a 15ms Hamming window. Word models were constructed from phone transcriptions, sharing a set of 47 single-state phone models; each phone was represented by a second-order linear system with piecewise-linear target trajectories, with and without continuity constraints across boundaries. The models were initialized by training the target trajectories alone, and then re-trained using 10 iterations of the EM algorithm; Σ_v and Σ_w were fixed throughout. Word labels were supplied during training, but without phone boundaries. During testing, the hidden X and U trajectories were estimated and compared with the original data.

Figure 1 shows a typical result for the word "Tuscaloosa's"; Figure 2 shows the same data with continuity imposed on the U trajectory, but not its slope. The effect of allowing a distribution of targets is clear from the two different trajectories modelling the /s/ phone. Imposing continuity constraints tends to make the U process model the data directly; when U is free to vary across boundaries it behaves like a true target, the X process models the dynamics, and the influence of the system matrices is more pronounced.

Figure 3 illustrates the effect of varying Σ_v for the word "Arkansas's", and demonstrates the importance of controlling this parameter carefully. As Σ_v is increased, the model interprets more of the signal variation as modelling error, and does not attempt to fit it using the target process U. When Σ_v and Σ_w are trained, the effect of these parameters can often swamp the behaviour of the phonetic model. By adjusting the noise covariances, the model can be made to ignore a certain amount of global "non-linguistic" variation, sharpening the target distributions. A key feature of the model is that it provides a number of mechanisms for fitting different sources of variation, each of which can be controlled and examined during training and recognition.

5. Conclusions

The stochastic target model presented in [1] has been extended to include piecewise-deterministic control trajectories with continuity constraints. The optimal non-linear filtering and smoothing recursions required to estimate the hidden states of the model from acoustic data have been derived, and a sub-optimal approximation technique has been outlined. Preliminary data-fitting experiments indicate that the model is capable of accounting for coarticulation across and within phone boundaries, and makes use of a number of different mechanisms for modelling possible sources of variability. Future work will compare the performance of the model on articulatory and acoustic recognition tasks.

6. REFERENCES

1. Ramsay G. and Deng L. Maximum-likelihood estimation for articulatory speech recognition using a stochastic target model. In *Proc. Eurospeech-95*, pages 1401–1404, 1995.
2. Russell M.J. and Moore R. Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. In *Proc. ICASSP-85*, pages 2376–2379, 1988.
3. Deng L., Aksmanovic M., Sun D., and Wu J. Speech recognition using hidden Markov models with polynomial regression functions as non-stationary states. *IEEE Trans. on Speech and Audio Proc.*, 2(4):507–520, 1994.
4. Afify M., Gong Y., and Haton J.-P. Stochastic trajectory models for speech recognition : an extension to modelling time correlation. In *Proc. Eurospeech-95*, pages 515–518, 1995.
5. Russell M. A segmental HMM for speech pattern matching. In *Proc. ICASSP-93*, pages 499–502, 1993.
6. Holmes W.J. and Russell M.J. Speech recognition using a linear dynamic segmental HMM. In *Proc. Eurospeech-95*, pages 1611–1614, 1995.
7. V. Digalakis, J.R. Rohlicek, and M. Ostendorf. ML estimation of a stochastic linear dynamical system with the EM algorithm and its application to speech recognition. *IEEE Transactions on speech and audio processing*, 1(4):431–442, 1993.
8. Ross K. A dynamical system model for recognizing intonation patterns. In *Proc. Eurospeech-95*, pages 993–996, 1995.
9. Shumway R.H. and Stoffer D.S. Dynamic linear models with switching. *Journal of the American Statistical Association*, 86(415):763–769, 1991.
10. Gordon K. and Smith A.F.M. Modelling and monitoring discontinuous changes in time series. In Spall J., editor, *Bayesian Analysis of Time Series and Dynamic Models*, pages 359–391. Marcel Dekker: New York, 1988.
11. Shirai K. and Honda M. Estimation of articulatory motion. In *Dynamic Aspects of Speech Production*. University of Tokyo Press, 1976.
12. Fujisaki H. Functional models of articulatory and phonatory dynamics. In *Articulatory Modelling and Phonetics*, pages 49–64. G.A.L.F., 1977.
13. Aggoun L., Elliott R.J., and Moore J.B. A measure change derivation of continuous state Baum-Welch estimators. *Journal of Mathematical Systems, Estimation, and Control*, 5(3):1–12, 1995.
14. Bar-Shalom Y. and Fortmann T.E. *Tracking and Data Association*. Academic Press, 1988.

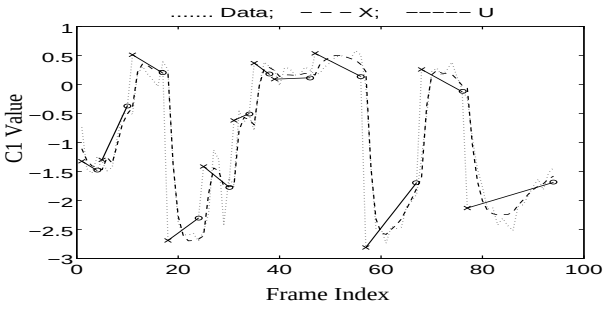

Figure 1: "Tuscaloosa's" : without continuity constraint.

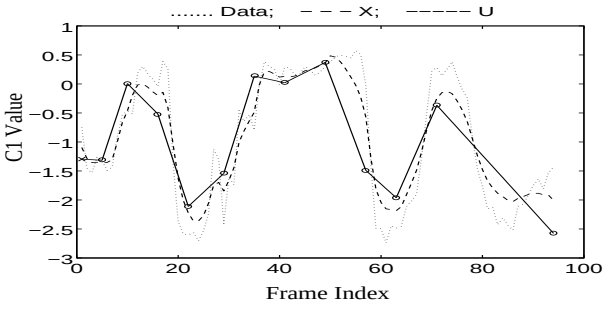

Figure 2: "Tuscaloosa's" : with continuity constraint.

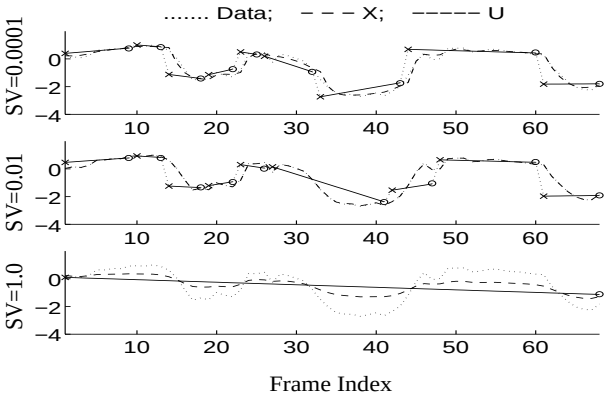

Figure 3: Effect of Σ_v on state estimation.