IDIAP RESEARCH REPORT

A Bayesian Switching Linear Dynamical System for Scale-Invariant robust speech extraction

Bertrand Mesot [a,b]   David Barber [c]

IDIAP–RR 07–52

October 2007

submitted for publication

[a] IDIAP Research Institute
[b] École Polytechnique Fédérale de Lausanne (EPFL)
[c] University College London

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 07–52

A Bayesian Switching Linear Dynamical System for Scale-Invariant robust speech extraction

Bertrand Mesot David Barber

October 2007

submitted for publication

Abstract. Most state-of-the-art automatic speech recognition (ASR) systems deal with noise in the environment by extracting *noise robust* features which are subsequently modelled by a Hidden Markov Model (HMM). A limitation of this feature-based approach is that the influence of noise on the features is difficult to model explicitly and the HMM is typically over sensitive, dealing poorly with unexpected and severe noise environments. An alternative is to model the raw signal directly which has the potential advantage of allowing noise to be explicitly modelled. A popular way to model raw speech signals is to use an Autoregressive (AR) process. AR models are however very sensitive to variations in the amplitude of the signal. Our proposed Bayesian Autoregressive Switching Linear Dynamical System (BAR-SLDS) treats the observed noisy signal as a *scaled*, *clean hidden* signal plus noise. The variance of the noise and signal scaling factor are automatically adapted, enabling the robust identification of scale-invariant clean signals in the presence of noise.

1 The Switching AR-HMM

A basic way to model a speech waveform—represented as a sequence of unidimensional samples $v_{1:T}$—is by means of an Autoregressive (AR) process. An AR process models the sample v_t as the sum of a linear combination of the R previous samples and a random, Gaussian distributed, innovation η_t:

$$v_t = \sum_{r=1}^{R} c_r v_{t-r} + \eta_t \quad \text{with} \quad \eta_t \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ represents the normal (Gaussian) distribution with mean μ and variance σ^2, and c_r are the AR coefficients. Since an AR process is too simple to model the strong non-stationarities typically encountered in speech signals, a useful extension is to consider each sample v_t as being generated by one of S different AR processes. The switching between the various AR processes is controlled by $p(s_t \,|\, s_{t-1})$, where s_t is the index of the AR process used at time t. In the Switching AR-HMM (SAR-HMM) proposed in [4], the joint distribution of the sequence of observations $v_{1:T}$ and states $s_{1:T}$ is

$$p(v_{1:T}, s_{1:T}) = \prod_{t=1}^{T} p(v_t \,|\, s_t, v_{t-R:t-1})\, p(s_t \,|\, s_{t-1})$$

where $p(s_1 \,|\, s_0) \equiv p(s_1)$ is a specified prior distribution. If we define $\tilde{\mathbf{v}}_t = \begin{bmatrix} v_{t-1} & \dots & v_{t-R} \end{bmatrix}^{\mathsf{T}} \equiv v_{t-R:t-1}$ and $\mathbf{c} = \begin{bmatrix} c_1 & \dots & c_R \end{bmatrix}^{\mathsf{T}}$, then Eq. 1 defines a Gaussian emission distribution for the current sample v_t:

$$p(v_t \,|\, s_t, \tilde{\mathbf{v}}_t) = \frac{1}{\sqrt{2\pi\sigma^2_{s_t}}} \exp\left\{ -\frac{1}{2\sigma^2_{s_t}} (v_t - \mathbf{c}^{\mathsf{T}}_{s_t} \tilde{\mathbf{v}}_t)^2 \right\}$$

where the mean and variance depend on the current state s_t. It is desirable to prevent the switch state changing too rapidly and the speech signal is therefore considered as the concatenation of a number of fixed-length segments within which the state cannot change. This corresponds to the joint distribution

$$p(v_{1:T}, s_{1:N}) = \prod_{n=1}^{N} p(s_n \,|\, s_{n-1}) \prod_{t=t_n}^{t_{n+1}-1} p(v_t \,|\, s_n, \tilde{\mathbf{v}}_t).$$

Gain Adaptation

Whilst Eq. 1 is invariant under rescaling of the signal $v_{1:T}$ in the zero-noise limit, in noisy environments, the equation does not remain invariant. In particular, if the signal is scaled by a factor α, we would require the innovation variance to scale by a factor α^2. In other words, the 'gain' of the sequence, σ, needs to be appropriately set for each sequence. This problem is generally addressed by performing Gain Adaptation (GA) [4, 6], replacing in Eq. 1, for each segment n and state s, the variance σ^2 by the segment-state variance σ^2_{ns} which maximises the likelihood of the observations in the n-th segment, i.e.,

$$\sigma^2_{ns} = \arg\max_{\sigma^2} p(v_{t_n:t_{n+1}-1} \,|\, \sigma^2).$$

However, increasing the innovation σ^2_{ns} allows the model to produce wilder uncontrolled fluctuations in the signal. Ideally, we may wish to have a model which deals with changes in overall signal level by simply re-scaling the underlying signal, thus controlling the form of the signal more carefully.

2 The Bayesian AR-SLDS

An alternative to adapting the innovation is to consider the observed sample v_t as a *scaled* version of a scale-invariant *hidden* sample w_t plus noise:

$$v_t = b w_t + \eta^{\mathcal{V}}_t \quad \text{with} \quad \eta^{\mathcal{V}}_t \sim \mathcal{N}(0, \sigma^2_{\mathcal{V}}) \tag{2}$$

and to model the 'clean' hidden sample w_t with a switching AR process:

$$w_t = \mathbf{c}_s^{\mathsf{T}} \tilde{\mathbf{w}}_t + \eta_t^{\mathcal{W}}(s) \quad \text{with} \quad \eta_t^{\mathcal{W}}(s) \sim \mathcal{N}\big(0, \sigma_{\mathcal{W}}^2(s)\big). \tag{3}$$

In this manner, no innovation-inflation is required, provided that the observed signal is simply a scaled, noisy version of an underlying AR process. For a given observed sequence, the setting of b and $\sigma_{\mathcal{V}}^2$ is unknown a-priori and needs to be determined. To solve this problem we treat both parameters as random variables and introduce a Normal-Gamma prior[1]

$$b \,|\, \nu, s \sim \mathcal{N}\big(\mu_s, \nu^{-1}\sigma_s^2\big) \quad \text{and} \quad \nu \,|\, s \sim \mathcal{G}(\alpha_s, \beta_s) \tag{4}$$

where $\mathcal{G}(\alpha, \beta)$ is the Gamma distribution with shape α and inverse scale β. Similarly to the SAR-HMM, we consider a segmental approach where the state, scaling factor and noise variance are kept constant over a segment. Using $\vartheta_n = \{b_n, \nu_n\}$, Eqs. 2, 3 and 4 correspond to the distributions $p(v_t \,|\, w_t, \vartheta_n)$, $p(w_t \,|\, w_{t-1}, s_n)$ and

$$p(\vartheta_n \,|\, s_n) = p(b_n \,|\, \nu_n, s_n)\, p(\nu_n \,|\, s_n)$$

respectively. The joint distribution $p(v_{1:T}, w_{1:T}, \vartheta_{1:N}, s_{1:N})$ defined by this model is equal to

$$\prod_{n=1}^{N} p(\vartheta_n \,|\, s_n)\, p(s_n \,|\, s_{n-1}) \prod_{t=t_n}^{t_{n+1}-1} p(v_t \,|\, w_t, \vartheta_n)\, p(w_t \,|\, w_{t-1}, s_n). \tag{5}$$

For fixed ϑ_n the model corresponds to a special case of a Switching Linear Dynamical System (SLDS) where the underlying dynamics is constrained to be autoregressive. We will thus refer to the model defined by (5), which includes a prior over ϑ_n, as the Bayesian AR-SLDS (BAR-SLDS).

2.1 Parameter Optimisation

Given a set of M training sequences[2] $\{v_{1:T}^1, \ldots, v_{1:T}^M\}$, we want to find the parameter setting $\Psi^\star$ which maximises the total log-likelihood of the training sequences, i.e.,

$$\Psi^\star = \arg\max_{\Psi} \sum_{m=1}^{M} \log p(v_{1:T}^m \,|\, \Psi) \tag{6}$$

where Ψ is equal to

$$\bigcup_{s} \Big\{\mathbf{c}_s, \sigma_{\mathcal{W}}^2(s), \mu_s, \sigma_s^2\Big\} \;\cup\; \bigcup_{i,j} \Big\{p(s_n = j \,|\, s_{n-1} = j)\Big\}.$$

The prior distribution $p(s_1)$ is not optimised, but simply set to one for the first state and zero otherwise. Since our aim is to train the model on clean signals and to later test it on noisy data, we do not use a prior on ν during training and set appropriate value for α_s and β_s during testing. The likelihood of a sequence $v_{1:T}$ is

$$p(v_{1:T} \,|\, \Psi) = \sum_{s_{1:N}} \int_{\substack{\vartheta_{1:N} \\ w_{1:T}}} p(v_{1:T}, w_{1:T}, \vartheta_{1:N}, s_{1:N} \,|\, \Psi) \tag{7}$$

The sum/integral in Eq. 7 makes an explicit solution to Eq. 6 difficult to obtain. The usual approach would then be to use the Expectation Maximisation (EM) algorithm. However, the non-linear interaction between w_t and ϑ_n in Eq. 2 renders computing the required EM posterior distributions intractable.

[1] To ease notation, we prefer using the inverse variance $\nu = 1/\sigma_{\mathcal{V}}^2$.

[2] For simplicity, we will assume that they all have the same length.

2.2 Variational Inference

We propose to use a variational approach where the true posterior distribution is approximated by the simpler distribution

$$q(w_{1:T}, \vartheta_{1:N}, s_{1:N}) = q(w_{1:T} \,|\, s_{1:N})\, q(\vartheta_{1:N} \,|\, s_{1:N})\, q(s_{1:N})$$

where the problematic dependency between w_t and ϑ_n has been removed. By considering the Kullback-Leibler (KL) divergence $\mathrm{KL}\big(q(w_{1:T}, \vartheta_{1:N}, s_{1:N}) \,||\, p(w_{1:T}, \vartheta_{1:N}, s_{1:N} \,|\, \Psi)\big)$, we obtain the following lower bound on the log-likelihood

$$\log p(v_{1:T} \,|\, \Psi) \geq -\big\langle \log q(w_{1:T}, \vartheta_{1:N}, s_{1:N})\big\rangle_q + \big\langle \log p(v_{1:T}, w_{1:T}, \vartheta_{1:N}, s_{1:N} \,|\, \Psi)\big\rangle_q \tag{8}$$

Our aim is therefore to find the q distribution for which the lower bound is as close as possible to the true log-likelihood. Differentiating the bound with respect to $q(\vartheta_{1:N} \,|\, s_{1:N})$ yields

$$q(\vartheta_{1:N}, s_{1:N}) \propto p(\vartheta_{1:N} \,|\, s_{1:N})\, p(s_{1:N}) \,\exp\Big\{\big\langle \log p(v_{1:T}, w_{1:T} \,|\, \vartheta_{1:N}, s_{1:N})\big\rangle_{q(w_{1:T} \,|\, s_{1:N})}\Big\}$$

and differentiating with respect to $q(w_{1:T}, s_{1:N})$ yields

$$q(w_{1:T}, s_{1:N}) \propto p(w_{1:t} \,|\, s_{1:N})\, p(s_{1:N}) \,\exp\Big\{\big\langle \log p(v_{1:T} \,|\, w_{1:T}, \vartheta_{1:N}, s_{1:N})\big\rangle_{q(\vartheta_{1:N} \,|\, s_{1:N})}\Big\}. \tag{9}$$

2.2.1 Finding $q(\vartheta_n \,|\, s_n)$

Since we chose conjugate priors, the posterior distribution has the same form as the prior, hence

$$b_n \,|\, \nu_n, s_n \sim \mathcal{N}(\hat{\mu}_{s_n}, \nu_n^{-1}\hat{\sigma}^2_{s_n}) \;\text{ and }\; \nu_n \,|\, s_n \sim \mathcal{G}(\hat{\alpha}_{s_n}, \hat{\beta}_{s_n}).$$

After some algebra we obtain[3]

$$\hat{\sigma}^2_s = \sigma^2_s \left[1 + \sigma^2_s \sum_t \langle w_t^2\rangle\right]^{-1}, \qquad \hat{\mu}_s = \hat{\sigma}^2_s \left[\frac{\mu_s}{\sigma^2_s} + \sum_t v_t \langle w_t\rangle\right]$$

$$\hat{\alpha}_s = \alpha_s + \frac{1}{2}L_n, \qquad \hat{\beta}_s = \beta_s + \frac{1}{2}\sum_t \left[v_t^2 + \frac{\mu_s}{\sigma^2_s} - \frac{\hat{\mu}_s}{\hat{\sigma}^2_s}\right]$$

where the averages are taken with respect to $q(w_t \,|\, s_n)$, $L_n = t_{n+1} - t_n$ and the sums are carried out from t_n to $t_{n+1} - 1$.

2.2.2 Finding $q(w_t \,|\, s_n)$ and $q(s_n)$

The right-hand-side of Eq. 9 can be written as

$$\prod_{n=1}^{N} p(s_n \,|\, s_{n-1}) \prod_{t=t_n}^{t_{n+1}-1} q(v_t \,|\, w_t, s_n)\, p(w_t \,|\, w_{t-1}, s_n) \tag{10}$$

with $\log q(v_t \,|\, w_t, s_n)$ given by[4]

$$\begin{aligned}\big\langle \log p(v_t \,|\, w_t, \vartheta_n, s_n)\big\rangle &= \frac{1}{2}\big\langle \log \nu\big\rangle - \frac{1}{2}\big\langle \nu (v_t - b w_t)^2\big\rangle \\ &= \frac{1}{2}\big\langle \log \nu\big\rangle - \frac{1}{2}\langle \nu\rangle\big(v_t - \langle b\rangle w_t\big)^2 - \frac{1}{2}\underbrace{\big\langle \nu\big(b - \langle b\rangle\big)^2\big\rangle}_{\hat{\sigma}^2_{s_n}} w_t^2\end{aligned}$$

[3] While the prior parameters depends on s only, the posterior's depends on s and n. To ease notation we dropped the n index however.

[4] Irrelevant constant terms are ignored.

where the averages are over $q(\theta_n \mid s_n)$. Since, $\langle b \rangle = \hat{\mu}_{s_n}$ and $\langle \nu \rangle = \hat{\alpha}_{s_n}/\hat{\beta}_{s_n}$, $q(v_t \mid w_t, s_n)$ is proportional to

$$\exp\left\{ -\frac{1}{2} \begin{bmatrix} v_t - \hat{\mu}_{s_n} w_t \\ \hat{\sigma}_{s_n} w_t \end{bmatrix}^\mathsf{T} \begin{bmatrix} \frac{\hat{\beta}_{s_n}}{\hat{\alpha}_{s_n}} & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} v_t - \hat{\mu}_{s_n} w_t \\ \sigma_{s_n} w_t \end{bmatrix} \right\}.$$

This can equivalently be written as a stochastic linear equation defined on an augmented observation $\mathbf{v}_t$ [2],

$$\mathbf{v}_t = \mathbf{B}_{s_n} w_t + \boldsymbol{\eta}_t(s_n) \quad \text{with} \quad \boldsymbol{\eta}_t(s_n) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{s_n})$$

where

$$\mathbf{v}_t = \begin{bmatrix} v_t \\ 0 \end{bmatrix}, \quad \mathbf{B}_{s_n} = \begin{bmatrix} \hat{\mu}_{s_n} \\ \hat{\sigma}_{s_n} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{s_n} = \begin{bmatrix} \frac{\hat{\beta}_{s_n}}{\hat{\alpha}_{s_n}} & 0 \\ 0 & 1 \end{bmatrix}.$$

By replacing v_t by $\mathbf{v}_t$ in (10) we see that (10) corresponds to a SLDS for which the posteriors $q(w_t \mid s_n)$ and $q(s_n)$ can be computed using any of the numerous available algorithms found in the literature; see [1] for a review and comparison. For the experiments presented in this article, we used the Expectation Correction (EC) algorithm [1] which provides a fast and accurate procedure for computing the desired posteriors. We also used EC to find a first estimate of $q(w_t \mid s_n)$ and $q(s_n)$ by running the algorithm on a SLDS where the parameters where set to their mean value. Variational inference was then carried out by iteratively applying the forumlae of Sections 2.2.1 and 2.2.2.

2.3 Parameter Updating

Update formulae for the parameters in Ψ can be obtained by means of the Variational Bayesian EM Algorithm [3]. This corresponds to maximising the lower bound given by Eq. 8 with respect to Ψ. Differentiating the lower bound with respect to Ψ and setting the result equal to zero yields

$$\mathbf{c}_s = \left\langle \tilde{\mathbf{w}}_t \tilde{\mathbf{w}}_t^T \right\rangle^{-1} \left\langle w_t \tilde{\mathbf{w}}_t \right\rangle, \qquad \mu_s = \langle b \rangle$$

$$\sigma^2_{\mathcal{W}}(s) = \frac{1}{\langle L_n \rangle} \left\langle (w_t - \mathbf{c}_s^\mathsf{T} \tilde{\mathbf{w}}_t)^2 \right\rangle, \qquad \sigma_s^2 = \frac{1}{\langle L_n \rangle} \left\langle \nu (b - \mu_s)^2 \right\rangle$$

where the averages must be interpreted as

$$\langle x \rangle = \sum_{n=1}^{N} q(s_n = s) \sum_{t=t_n}^{t_{n+1}-1} \langle x \rangle_{q(w_t \mid s_n)\, q(\theta_n \mid s_n)}.$$

The updated formula for the transition distribution is

$$p(s_n = j \mid s_{n-1} = j) = \frac{\sum_{n>1} q(s_{n-1} = i, s_n = j)}{\sum_{n>1} q(s_{n-1} = i)}.$$

3 Results

To test the potential benefit of the proposed scale-invariant model, we examined the reconstructions of scaled noisy signals provided by the BAR-SLDS compared with the more standard gain-adaptation procedure. Clean, non-noisy utterances of the digit ‘one’, were taken from TI-DIGITS [5], downsampled to 8 kHz. An AR-SLDS model was trained on these data using the formulae of Section 2.3[5]. As a demonstration, a single digit clean utterance of a ‘one’ was taken, from which a scaled noisy version

[5] Since the model was trained on clean data, no prior was used on ν. The training was stopped after convergence of the lower bound given by Eq. 8. The model was composed of 10 states and were using 10 AR coefficients and a left-right transition matrix. The segment length was of 140 samples. This corresponds to 1.75 ms at a sampling frequency of 8 kHz.

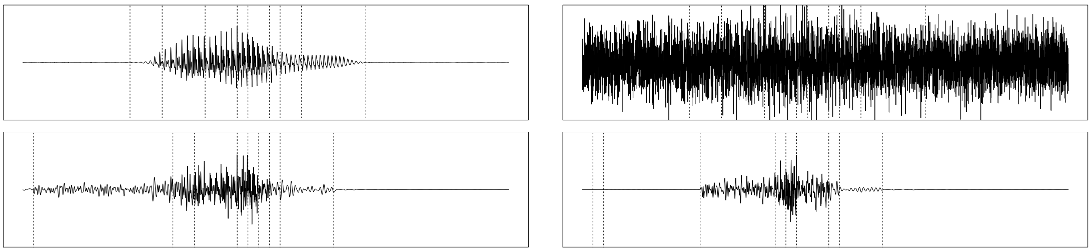

Figure 1: Comparison of signal reconstruction. Top: original (left) and corrupted (right) waveform of a ‘one’. Bottom: most likely reconstruction as given by the gain-adapted AR-SLDS (left) and the BAR-SLDS (right). The dashed lines indicate the most likely state segmentation. The state segmentation of the clean signal is displayed on the noisy signal as well.

Noise Var.	SNR (dB)	GA AR-SLDS	BAR-SLDS
clean	—	97.0%	87.0%
10^{-5}	19.7	94.8%	83.3%
10^{-4}	10.6	84.0%	78.3%
10^{-3}	0.7	61.2%	64.0%

Table 1: Comparison of the recognition accuracy of the Gain-Adapted AR-SLDS and the Bayesian AR-SLDS for various levels of noise. The second column gives the approximate Signal to Noise Ratio (SNR).

of the signal was then formed and corrupted by additive Gaussian noise at SNR 0. Given this scaled-noisy signal, the posterior $q(w_t, s_n)$ can be used to reconstruct the most likely (ML) clean speech signal. Fig. 1 shows the ML reconstructed clean signal given by the gain-adapted AR-SLDS and the BAR-SLDS. The BAR-SLDS does not allow the innovation to change, resulting in less variability in the underlying signal and a cleaner denoising, particularly at the edges where the signal level is low. On the other hand, the gain-adapted AR-SLDS provides a reasonable reconstruction but, as a result of the extra innovation required to explain the change in signal-level, allows the reconstructed signal too much freedom, particularly in the low signal level areas, as anticipated.

An interesting comparison is the recognition performance of the BAR-SLDS compared to the gain-adapted AR-SLDS. We repeated the above training procedure, fitting an AR-SLDS model to each of the 11 digits in the TI-DIGITS database. For a given test utterance $v_{1:T}$, recognition was performed by picking the digit model for which the likelihood of the corresponding augmented observation $\mathbf{v}_{1:T}$ was the highest. To evaluate the accuracy of the BAR-SLDS in the presence of noise, we corrupted the original clean test utterances with additive stationary Gaussian noise. To give the model the opportunity to remove noise we specified a prior on ν with a mean of 1 and a large variance. Table 1 compares the recognition accuracy of the proposed Bayesian AR-SLDS with the Gain-Adapted AR-SLDS proposed in [6]. Although there is a slight improvement at SNR 0, the BAR-SLDS is otherwise less accurate than its gain-adapted counterpart. This drop in performance can be explained by the fact that the BAR-SLDS does not, as yet, adapt the innovation variance, using only the scale to allow for changes in the signal. A natural extension of the BAR-SLDS model is therefore to allow the innovation to be adapted, as well as the scale. Such a model should have the benefit that the innovation adaptation will be required only in those cases that cannot be well explained by simpler rescalings of the underlying clean signal.

4 Conclusion & Future Work

We proposed a Bayesian approach to deal with variations in the signal amplitude in AR models. As expected, the approach results in cleaner reconstructions than approaches which simply adapt innovation variance. Whilst our proposed solution is quite natural, the model throws up some significant technical challenges. Our technique is, to our knowledge, the first variational approximation of the Bayesian SLDS which retains dependencies between switch and continuous latent states by exploiting state-of-the-art inference procedures. Such technical advances will hopefully lead to the wider application of SLDS style models in signal processing areas. The presented model is part of our continuing programme of development of models for dealing with noisy signals. In the future, we will consider priors on the innovation variance and, possibly, on the AR coefficients. Another useful extension would be to use an AR noise model which would allow complex non-stationary noise sources to be considered.

Acknowledgements

This work was supported by the Swiss NSF through the MULTI project and the Swiss OFES through the PASCAL Network.

References

[1] D. Barber. Expectation correction for smoothed inference in switching linear dynamical systems. *Journal of Machine Learning Research*, 7:2515–2540, November 2006.

[2] D. Barber and S. Chiappa. Unified Inference for Variational Bayesian Linear Gaussian State-Space Models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2006.

[3] M. J. Beal and Z. Ghahramani. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In Oxford University Press, editor, *Bayesian Statistics 7*, pages 453–464, 2003.

[4] Y. Ephraim and W. J. J. Roberts. Revisiting autoregressive hidden Markov modeling of speech signals. *IEEE Signal Processing Letters*, 12(2):166–169, February 2005.

[5] R.G. Leonard. A database for speaker independent digit recognition. In *Proceedings of ICASSP84*, volume 3, 1984.

[6] B. Mesot and D. Barber. Switching linear dynamical systems for noise robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(6):1850–1858, August 2007.