A BAYESIAN ALTERNATIVE TO GAIN ADAPTATION IN AUTOREGRESSIVE HIDDEN MARKOV MODELS

Bertrand Mesot and David Barber

IDIAP Research Institute

ABSTRACT

Models dealing directly with the raw acoustic speech signal are an alternative to conventional feature-based HMMs. A popular way to model the raw speech signal is by means of an autoregressive (AR) process. Being too simple to cope with the nonlinearity of the speech signal, the AR process is generally embedded into a more elaborate model, such as the switching autoregressive HMM (SAR-HMM). A fundamental issue faced by models based on AR processes is that they are very sensitive to variations in the amplitude of the signal. One way to overcome this limitation is to use *Gain Adaptation* to adjust the amplitude by maximising the likelihood of the observed signal. However, adjusting model parameters by maximising *test* likelihoods is fundamentally outside the framework of standard statistical approaches to machine learning, since this may lead to overfitting when the models are sufficiently flexible. We propose a statistically principled alternative based on an exact Bayesian procedure in which priors are explicitly defined on the parameters of the AR process. Explicitly, we present the Bayesian SAR-HMM and compare the performance of this model against the standard Gain-Adapted SAR-HMM on a single digit recognition task, showing the effectiveness of the approach and suggesting thereby a principled and straightforward solution to the issue of Gain Adaptation.

Index Terms— Autoregressive processes, Gain control, Bayes procedures, Speech recognition

1. INTRODUCTION

Models dealing directly with the raw acoustic speech signal are an alternative to conventional feature-based Hidden Markov Models (HMMs). One of the most popular examples is the *Autoregressive (AR) Process* which models a sample y_t of a speech signal—represented as a sequence of samples $y_{1:T}$—as a linear combination of the R previous samples plus a Gaussian distributed innovation η

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

where σ^2 is the variance of the innovation and c_r are the AR coefficients. However, an AR process is too simple to model the strong non-stationarities typically encountered in speech signals. A possible way to deal with non-stationarity is to select at each time step t a setting of the AR parameters from a discrete set of possible parameter values, with the switching between the parameters controlled by a Markov Model. This approach is at the root of the *AR Hidden Markov Model* (AR-HMM) proposed by Poritz [1] and its modern-day counterpart the *Switching AR-HMM* (SAR-HMM), proposed by Ephraim and Roberts [2]. At the heart of the above models lies a standard AR process. However, a fundamental limitation of such AR models is that the innovation variance σ^2 does not scale properly with the signal. In particular, if the signal is scaled by a factor α, we would expect the innovation variance to scale by a factor α^2 as well. In other words, the 'gain' of the sequence, σ, needs to be set for each sequence, and has a strong impact on the likelihood of an observed sequence. Finding, therefore, a solution to the gain problem is a key step in the successfull application of such fundamental models as AR processes to acoustic signal analysis. A straightforward approach is to *gain normalise* the signal such that it always has unit variance. An alternate and more effective solution [2, 3] is to replace σ^2 in Eq. 1 by the variance which maximises the likelihood of the speech signal $y_{1:T}$

$$\sigma^2_{\mathrm{ML}} = \arg\max_{\sigma^2} p(y_{1:T} \,|\, \sigma^2). \qquad (2)$$

This approach, called *Gain Adaptation* (GA), has been successfully used for isolated digit recognition with AR-HMMs in clean and noisy environments [2, 3, 4]. Whilst useful in practice, GA does not fit into the usual machine learning framework since, formally, model parameters may only be set on the basis of *training* data. Otherwise, in flexible models, setting model parameters on the basis of test data may lead to overfitting. We therefore consider a statistically principled alternative Bayesian approach to GA which consists in specifying a prior probability distribution on the model parameters. This approach has two potential benefits over standard GA: (i) the variations of the gain can be explicitly controlled, and (ii) the AR coefficients are allowed to change, which may be useful to model inter and intra speaker variations for example.

In this paper we present the Bayesian SAR-HMM which generalises the standard acoustic level SAR-HMM, concurrently dealing with the issues of GA and parameter uncer-

tainty in a computationally efficient and principled manner.

2. THE SAR-HMM

The standard SAR-HMM [2, 3, 4] has a discrete switch variable which can be in S different states, each state representing a particular setting of the AR coefficients c_r and innovation variance σ^2 used in Eq. 1. From a probabilistic viewpoint, the model defines a joint distribution over the sequences of observed samples $y_{1:T}$ and switch states $s_{1:T}$ of the form

$$p(y_{1:T}, s_{1:T}) = \prod_{t=1}^{T} p(y_t \,|\, y_{t-R:t-1}, s_t)\, p(s_t \,|\, s_{t-1}) \qquad (3)$$

where $p(y_t \,|\, y_{t-R:t-1}, s_t) \equiv p(y_t \,|\, y_{1:t-1}, s_t)$ if $t \leq R$ and $p(s_1 \,|\, s_0) \equiv p(s_1)$. The emission probability, corresponding to Eq. 1, is given by

$$p(y_t \,|\, s_t, \tilde{\mathbf{y}}_t) \propto \exp\left\{-\frac{1}{2\sigma_{s_t}^2}\big(y_t - \tilde{\mathbf{y}}_t^{\mathsf{T}}\mathbf{c}_{s_t}\big)^2\right\} \qquad (4)$$

where $\tilde{\mathbf{y}}_t = [y_{t-1} \dots y_{t-R}]^{\mathsf{T}}$ and $c_{s_t} = [c_1(s_t) \dots c_R(s_t)]^{\mathsf{T}}$.

In practice it is not desirable to allow the switch state to change at each time step because we expect the dynamics to last for a minimal amount of time—1.75 ms in our case[1]. In the SAR-HMM, the speech signal is therefore considered as the concatenation of N fixed-length segments over which the state cannot change. This corresponds to the joint distribution

$$p(y_{1:T}, s_{1:N}) = \prod_{n=1}^{N} p(s_n \,|\, s_{n-1}) \prod_{t=t_n}^{t_{n+1}-1} p(y_t \,|\, s_n, \tilde{\mathbf{y}}_t) \qquad (5)$$

where t_n is the time step at which the n-th segment starts.

Gain Adaptation in the SAR-HMM

Given a sequence of samples $y_{1:T}$, GA is performed in the SAR-HMM by replacing the state innovation variance σ_s^2 in Eq. 4 by the *per segment and state* variance σ_{ns}^2 which maximise the likelihood of the observed sequence $y_{1:T}$, i.e.,

$$\sigma_{ns}^2 = \frac{1}{T_n} \sum_{t=t_n}^{t_{n+1}-1} \big(y_t - \tilde{\mathbf{y}}_t^{\mathsf{T}}\mathbf{c}_s\big)^2$$

where $T_n = t_{n+1} - t_n$ is the length of the n-th segment.

3. THE BAYESIAN SAR-HMM

In the SAR-HMM the AR coefficients $\mathbf{c}_s$ and innovation variances σ_s^2 are considered as free parameters that have to be learned from data. In the proposed Bayesian approach we treat them as *random variables* whose probability distributions are controlled by hyper-parameters. Fig. 1 shows the

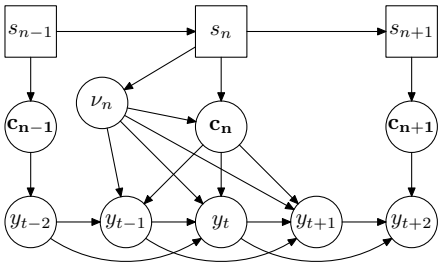

Fig. 1. DBN representation of the Bayesian SAR-HMM. The graph represents a model with segments of 3 samples and an AR process of order 2. The index n represents the segment number. Squares and circles represent discrete and continuous variables respectively.

Dynamical Bayesian Network (DBN) representation of the Bayesian SAR-HMM. A particular segment n is modelled by an R-th order AR process whose coefficients $\mathbf{c}_n$ and inverse innovation variance[2] ν_n are drawn randomly from a prior distribution conditioned on the switch state s_n. Formally the Bayesian SAR-HMM defines the joint distribution

$$p(y_{1:T}, \mathbf{c}_{1:N}, \nu_{1:N}, s_{1:N}) = \qquad (6)$$
$$\prod_{n=1}^{N} p(\mathbf{y}_n \,|\, \mathbf{c}_n, \nu_n, \tilde{\mathbf{y}}_{t_n})\, p(\mathbf{c}_n, \nu_n \,|\, s_n)\, p(s_n \,|\, s_{n-1})$$

which is a temporal extension of [5]. Explicitly,

$$p(\mathbf{y}_n \,|\, \mathbf{c}_n, \nu_n, \tilde{\mathbf{y}}_{t_n}) = \prod_{t=t_n}^{t_{n+1}-1} p(y_t \,|\, \mathbf{c}_n, \nu_n, \tilde{\mathbf{y}}_t). \qquad (7)$$

The new factor

$$p(\mathbf{c}_n, \nu_n \,|\, s_n) = p(\mathbf{c}_n \,|\, \nu_n, s_n)\, p(\nu_n \,|\, s_n)$$

defines priors on the AR coefficients and the inverse innovation variance of the n-th segment. In order to keep the model tractable, we use the conjugate priors[3]

$$\mathbf{c} \,|\, \nu, s \sim \mathcal{N}(\boldsymbol{\mu}_s, \nu^{-1}\boldsymbol{\Sigma}_s) \quad \text{and} \quad \nu \,|\, s \sim \gamma(\alpha_s, \beta_s)$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $\gamma(\alpha, \beta)$ is the gamma distribution defined as

$$\gamma(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \nu^{\alpha-1} e^{-\beta\nu}.$$

4. TRAINING

The free parameters of the Bayesian SAR-HMM are, $\boldsymbol{\mu}_s$, $\boldsymbol{\Sigma}_s$, α_s, β_s, for each state s, and the transition probability $a_{ij} \equiv$

[1] This corresponds to 140 samples at a sampling frequency of 8 kHz.

[2] To ease notation we prefer using the inverse variance $\nu = 1/\sigma^2$.

[3] The segment number has been dropped to simplify the notation.

$p(s_n = j \,|\, s_{n-1} = i)$ for each pair (i, j) of switch states. Training the model consists of maximising the likelihood of the observed training data

$$p(y_{1:T}) = \sum_{\mathbf{c}_{1:N}, \nu_{1:N}, s_{1:N}} p(y_{1:T}, \mathbf{c}_{1:N}, \nu_{1:N}, s_{1:N}). \quad (8)$$

To achieve this, we use the standard *Expectation Maximisation* (EM) algorithm: given the current setting of the model parameters ϕ, an updated setting $\hat{\phi}$ is found by maximising (M-step) the expected complete log-likelihood (E-step)

$$\left\langle \log p\big(y_{1:T}, \mathbf{c}_{1:N}, \nu_{1:N}, s_{1:N} \,|\, \hat{\phi}\big) \right\rangle_q \quad (9)$$

where $\langle \cdot \rangle_q$ is the average with respect to the posterior

$$q \equiv p(\mathbf{c}_{1:N}, \nu_{1:N}, s_{1:N} \,|\, y_{1:T}, \phi). \quad (10)$$

The formulae for the posterior and the updated parameter settings are given in Appendices A and B respectively. A detailled derivation can be found in [6].

5. PERFORMANCE

We compared the Bayesian SAR-HMM to the original SAR-HMM proposed in [2], with and without gain adaptation, and also against a standard feature-based HMM. The task was to recognise isolated digits pronounced by various male speakers from the TI-DIGITS database [7]. The training/test sets were composed of 110/112 utterances for each of the eleven digits (0–9 and 'oh'), spoken by 55/56 different speakers respectively. Each digit class was modelled by a separate SAR-HMM and recognition was performed by associating the utterance to the digit whose model had the highest likelihood. Whilst this speech classification problem is relatively easy, the effective amplitude of each utterance is different so that, for AR-based models, some form of GA is crucial for good performance.

The three types of SAR-HMMs were composed of $S = 10$ states, a left-right transition matrix and 10-th order AR processes. This corresponds to the setting proposed in [2]. The Bayesian SAR-HMM was initialised with a uniform left-right transition matrix, i.e, $p(s_{t+1} \,|\, s_t) = 0.5$ only if $s_{t+1} \in \{s_t, s_t + 1\}$. For each state s, the model parameters where initialised as follows: (i) each speech utterance of the training set was split into S sequences of equal length, (ii) all the s-th sequences were gathered together and used to train an AR process for state s, (iii) the shape of the Gamma prior was arbitrarily set to $\alpha_s = 10$ and β_s was set such that the mean of the Gamma distribution matched the inverse innovation variance $1/\sigma_s^2$ obtained by training the AR process, i.e., $\beta_s = \alpha_s \sigma_s^2$, (iv) the AR coefficients $\mathbf{c}_s$ obtained were used as the mean in the Gaussian prior, i.e., $\boldsymbol{\mu}_s = \mathbf{c}_s$, (v) the covariance of the AR coefficients was set to the identity matrix, i.e, $\sigma_s^2 \boldsymbol{\Sigma}_s = \mathbf{I}$, (vi) a new state segmentation was obtained by doing Viterbi decoding with the so-defined Bayesian SAR-HMM and steps (ii) to (vi) were then repeated three times. The feature-based HMM was composed of 18 states with a left-right transition matrix, a mixture of three Gaussians per state and used 13 MFCC features, including energy. It was implemented using HTK [8].

Model	Word Accuracy
HMM (HTK)	100%
SAR-HMM (no gain)	88.3%
SAR-HHM (gain)	97.2% (98.5%)
Bayesian SAR-HMM	98.7%

Table 1. Word accuracy of three different models on a single digit recognition task on the TI-DIGITS database; *gain* and *no gain* indicates whether or not gain adaptation has been used. The performance of the gain adapted SAR-HMM reported in [2] is indicated between parenthesis.

Table 1 shows the word accuracy of each model. The performance of the gain adapted SAR-HMM is reproduced from [2]. All the other results have been obtained by our own implementation of the respective models. That the accuracy we obtained for the gain adapted SAR-HMM is slightly below that reported in [2]—this is likely to be due to differences in the intialisation or in the stopping criterion used. The Bayesian and gain adapted SAR-HMM have a word accuracy which is 10% higher than that of the non gain adapted SAR-HMM. This demonstrates that dealing with the gain problem is crucial to ensure good performance. The performance of the Bayesian SAR-HMM demonstrates that the Bayesian approach is an alternative principled alternative to ad-hoc maximum likelihood gain adaptation.

6. CONCLUSION

Modelling the raw acoustic signal is an alternative strategy to using feature based HMMs for speech recognition. A motivation for this is that strong signal models may be used to remove noise, and can also form the basis of powerful hierarchical models of the signal. However, signal models based on AR-processes are over-sensitive to signal amplitude, and this problem is typically healed using ad-hoc GA methods. In contrast, our Bayesian approach provides a statistically principled and straightforward exact alternative to standard Maximum Likelihood Gain Adaptation. The result is a simple update formula which correctly deals with the uncertainty in the parameter estimates from the training set, and automatically computes the posterior distribution of parameters in light of test data. This is an encouraging step towards the development of more complex signal and noise models, in which the flexibility of the models is ever increasing.

Code implementing the standard and Bayesian SAR-HMM

is available from `http://www.idiap.ch/~bmesot`.

A. INFERENCE

The posterior distribution is obtained using a forward-backward algorithm. The forward pass calculates the filtered posterior[4] $p(\mathbf{c}_n, \nu_n, s_n \,|\, \mathbf{y}_{1:n})$ and the backward pass finds the posterior $p(\mathbf{c}_n, \nu_n, s_n \,|\, \mathbf{y}_{1:N})$ by correcting the filtered posterior.

Forward Pass

The filtered posterior $p(\mathbf{c}_n, \nu_n, s_n \,|\, \mathbf{y}_{1:n})$ for the n-th segment is proportional to

$$p(\mathbf{c}_n, \nu_n \,|\, s_n, \tilde{\mathbf{y}}_{t_n}, \mathbf{y}_n)\, p(\mathbf{y}_n \,|\, s_n, \tilde{\mathbf{y}}_{t_n}) \times \sum_{s_{n-1}} p(s_n \,|\, s_{n-1})\, p(s_{n-1} \,|\, \mathbf{y}_{1:n-1}). \tag{11}$$

The mean $\boldsymbol{\mu}_n$ and covariance $\nu^{-1}\boldsymbol{\Sigma}_n$ of $\mathbf{c}_n$ are obtained by iterating, for $t_n \le t < t_{n+1}$:

$$\sigma_t^2 = \tilde{\mathbf{y}}_t^\mathsf{T} \boldsymbol{\Sigma}_{t-1} \tilde{\mathbf{y}}_t + 1, \qquad \mathbf{K}_t = \frac{1}{\sigma_t^2} \boldsymbol{\Sigma}_{t-1} \tilde{\mathbf{y}}_t,$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \mathbf{K}_t \big(y_t - \tilde{\mathbf{y}}_t^\mathsf{T} \boldsymbol{\mu}_{t-1}\big), \quad \boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{t-1} - \mathbf{K}_t \tilde{\mathbf{y}}_t^\mathsf{T} \boldsymbol{\Sigma}_{t-1}$$

where the recursion is initiated with $\boldsymbol{\mu}_{s_n}$ and $\boldsymbol{\Sigma}_{s_n}$. Similarly, $p(\nu_n \,|\, s_n, \tilde{\mathbf{y}}_{t_n}, \mathbf{y}_n)$ is a Gamma distribution with parameters

$$\hat{\alpha} = \alpha + \frac{T_n}{2} \quad \text{and} \quad \hat{\beta} = \beta + \sum_t \frac{1}{2\sigma_t^2} \big(y_t - \langle y_t \rangle\big)^2.$$

Integrating $p(\mathbf{y}_n, \mathbf{c}_n, \nu_n \,|\, s_n, \tilde{\mathbf{y}}_{t_n})$ over $\mathbf{c}_n$ and ν_n, we obtain

$$p(\mathbf{y}_n \,|\, s_n, \tilde{\mathbf{y}}_{t_n}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\hat{\alpha})}{\hat{\beta}^{\hat{\alpha}}} \prod_t \frac{1}{(2\pi\sigma_t^2)^{1/2}}.$$

The filtered state posterior $p(s_n \,|\, \mathbf{y}_{1:n})$ is obtained by integrating (11) over $\mathbf{c}_n$ and ν_n.

Backward Pass

The posterior $p(\mathbf{c}_n, \nu_n, s_n \,|\, \mathbf{y}_{1:N})$ is given by

$$p(\mathbf{c}_n, \nu_n \,|\, s_n, \tilde{\mathbf{y}}_{t_n}, \mathbf{y}_n) \sum_{s_{n+1}} p(s_n \,|\, s_{n+1}, \mathbf{y}_{1:n})\, p(s_{n+1} \,|\, \mathbf{y}_{1:N})$$

with $p(s_n \,|\, s_{n+1}, \mathbf{y}_{1:n}) \propto p(s_{n+1} \,|\, s_n)\, p(s_n \,|\, \mathbf{y}_{1:n})$.

B. PARAMETER UPDATING

Differentiating 8) with respect to the updated mean $\hat{\boldsymbol{\mu}}_s$ and covariance $\hat{\boldsymbol{\Sigma}}_s$ and setting the result equal to zero, gives the following update formulae[5]

$$\hat{\boldsymbol{\mu}}_s = \langle \mathbf{c}_n \rangle_{\tilde{q}(s)}, \quad \hat{\boldsymbol{\Sigma}}_s = \big\langle \nu_n \big(\mathbf{c}_n - \boldsymbol{\mu}_s\big)\big(\mathbf{c}_n - \boldsymbol{\mu}_s\big)^\mathsf{T} \big\rangle_{\tilde{q}(s)}.$$

where

$$\langle \cdot \rangle_{\tilde{q}(s)} \equiv \frac{1}{\sum_n q(s_n = s)} \sum_n q(s_n = s)\, \langle \cdot \rangle_{q(\mathbf{c}_n, \nu_n \,|\, s_n = s)}$$

Similarly, optimising over $\hat{\beta}_s$ gives $\hat{\beta}_s = \hat{\alpha}_s / \langle \nu_n \rangle_{\tilde{q}(s)}$. Differentiating with respect to $\hat{\alpha}_s$ gives

$$\log \hat{\alpha}_s - \psi(\hat{\alpha}_s) = \log \langle \nu_n \rangle_{\tilde{q}(s)} - \big\langle \log \langle \nu_n \rangle_{q(\nu_n \,|\, s_n = s)} \big\rangle_{\tilde{q}(s)}$$

where $\psi(\hat{\alpha}_s)$ is the digamma function. Whilst no explicit formula for $\hat{\alpha}_s$ exists, the equation is well-behaved and can be solved using Newton-Raphson's method, for example. The updated transition distribution is given by

$$\hat{a}_{ij} \propto \sum_n q(s_{n-1} = i, s_n = j).$$

[4] Notationally, $p(\cdot \,|\, \mathbf{y}_{1:n}) \equiv p(\cdot \,|\, y_{1:t_{n+1}-1})$.

[5] This is presented for a single training example. The extension to multiple examples is straightforward.

C. REFERENCES

[1] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proceedings of the IEEE International Confererence on Acoustics, Speech, and Signal Processing*, May 1982, vol. 7, pp. 1291–1294.

[2] Y. Ephraim and W. J. J. Roberts, "Revisiting autoregressive hidden Markov modeling of speech signals," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 166–169, February 2005.

[3] Y. Ephraim, "Gain–adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Transaction on Signal Processing*, vol. 40, no. 6, pp. 1303–1316, June 1992.

[4] K. Y. Lee and J. Lee, "Recognition of noisy speech by a nonstationary AR HMM with gain adaptation under unknown noise," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 741–746, October 2001.

[5] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Proceedings of NIPS 2001*, 2001.

[6] B. Mesot and D. Barber, "A Bayesian alternative to gain adaptation in autoregressive hidden Markov models," Tech. Rep. 55, IDIAP, 2006.

[7] R. G. Leonard, "A database for speaker independent digit recognition," in *Proceedings of ICASSP84*, 1984, vol. 3.

[8] "The Hidden Markov Model Toolkit," Available at: `http://htk.eng.cam.ac.uk/`.