Submitted to NIP98, Algorithms and Architectures, poster presentation.

Event-coupled hidden Markov models

Trausti T. Kristjansson
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
kristjan@uiuc.edu

Brendan J. Frey (corresponding author)
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
405 North Mathews Avenue, Urbana, Illinois, 61801
bfrey@uiuc.edu

Thomas Huang
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
huang@where.csl.uiuc.edu

Abstract

Inferences from time-series data can be greatly enhanced by taking into account multiple modalities. In some cases, such as audio of speech and the corresponding video of lip gestures, the different time-series are tightly coupled. We are interested in loosely-coupled time series where only the onset of events are coupled in time. We present an extension of the forward-backward algorithm that can be used for inference and learning in event-coupled hidden Markov models and give results on a simplified multi-media indexing task where the objective is to detect an event whose onset is loosely coupled in audio and video.

1 Foreground

The combination of multiple modalities for inference has proven to be a very powerful way to increase detection and recognition performance (Yuhas *et al.* 1988; Becker and Hinton 1992; Bregler *et al.* 1994; de Sa and Ballard 1998). By combining the soft information provided by models of the different modalities, weakly

incorrect evidence in one modality often can be corrected by another modality. This "diversity effect" lies at the heart of many powerful algorithms in a variety of areas, including the best error-correcting decoding algorithm (see Frey 1998 for a review).

In general, the state spaces representing the different modalities may be richly interdependent, making exact inference intractable. Methods for approximate inference include Markov chain Monte Carlo (Neal 1992), recognition models in Helmholtz machines (Dayan *et al.* 1995; Hinton *et al.* 1995), variational techniques (Jordan *et al.* 1998) and iterated inferences based on local independence assumptions (Frey and MacKay 1998).

In time-series data, each modality is often modeled with a hidden Markov model (HMM) and the states of the different HMMs are coupled in time. If the coupling is highly synchronized in time, then the modalities can be modeled using a single HMM with a number of states equal to the product of the numbers of states in the separate HMMs. If exact inference in the total state space is too computationally burdensome, variational methods can be used to couple the modalities using a mean field approximation (Jordan *et al.* 1997). If the coupling between the modalities is widely distributed in time, exact inference is intractable. Again, the inter-modality influences can be approximated by mean fields (Saul and Jordan 1996).

While the variational approximations described above are potentially very powerful, they also suffer from spontaneous symmetry breaking and restrictive assumptions about the form of the posterior distribution being inferred. In this paper, we focus on modeling the *onset* of events in different modalities. We expect this type of model to work well on data with random temporal transients whose onsets are well coupled. An example of data where the difference of onset times in two modalities is important is in distinguishing between /p/ and /b/ speech sounds. These sounds have similar audio and video characteristics, especially if the audio is noisy. However, the onset of voicing, as measured from the opening of the mouth, is later for /p/ than /b/. In this case, it is possible to discern between the sounds by relating the opening of the mouth to the onset of voicing.

Another important application is the detection loosely-coupled events in audio and video for the purpose of multi-media indexing and retrieval. These events include people talking or walking, a hand knocking on a door, a batter hitting a baseball, a tennis racquet hitting a ball, cars driving, an automobile accident, etc. For example, Fig. 1a and b show the posterior probabilities that an explosion occurred in a movie clip at or before time t under an audio and a video HMM that were each trained on examples of explosions. In this case the sound of the explosion begins roughly 0.15 seconds later than the video of the explosion.

In exchange for not modeling coupling that takes place after the onset of each event, we gain the ability to perform exact inference. The hope is that by taking one step back from the brink of intractability, we can take two steps forward in performance.

2 Event-coupled hidden Markov Models (ECHMMs)

A graphical model that describes event-coupled audio and video HMMs is shown in Fig. 1a. $\{x_1^{\mathrm{A}}, x_2^{\mathrm{A}}, \ldots, x_M^{\mathrm{A}}\}$ is the observation sequence for the audio and $\{s_1^{\mathrm{A}}, s_2^{\mathrm{A}}, \ldots, s_M^{\mathrm{A}}\}$ is the state sequence for the audio, whereas $\{x_1^{\mathrm{V}}, x_2^{\mathrm{V}}, \ldots, x_N^{\mathrm{V}}\}$ is the observation sequence for the video and $\{s_1^{\mathrm{V}}, s_2^{\mathrm{V}}, \ldots, s_N^{\mathrm{V}}\}$ is the state sequence for the video. Notice that the the two modalities may be sampled at different rates.

We assume that the state space for each modality can be partitioned into a subspace that models "nonevent" dynamics and a subspace that models the event. Consider

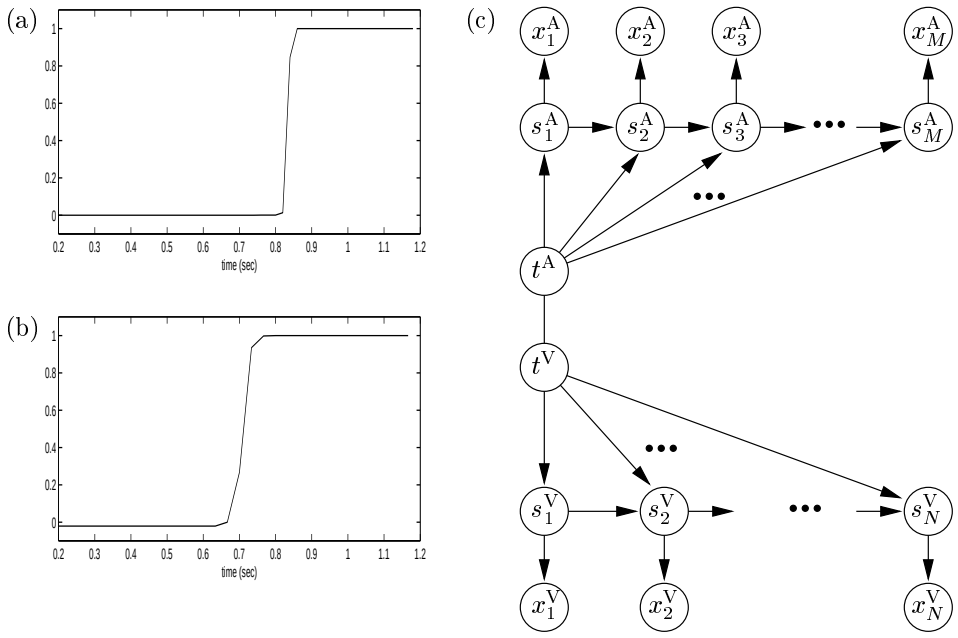

Figure 1: (a) and (b) show the posterior probabilities that an explosion occurred in a movie clip at or before time t under an audio and a video HMM. (c) A graphical model for two event-coupled HMMs.

ECHMMs where an event is assumed to begin at some point in time in both modalities. The time index at which the HMM for the audio makes a transition from the nonevent subspace to the event subspace is t^{A}. t^{V} is the corresponding transition time in the video. (To drop the assumption that an event occurs, a special "no-event" value for t^{A} and t^{V} can be introduced.)

The undirected link between t^{A} and t^{V} in Fig. 1a represents a joint probability distribution over the the times at which the events occur in the two modalities. So, the overall distribution is

$$\begin{aligned} &P(t^{\mathrm{A}}, t^{\mathrm{V}}, s_1^{\mathrm{A}}, \ldots, s_M^{\mathrm{A}}, x_1^{\mathrm{A}}, \ldots, x_M^{\mathrm{A}}, s_1^{\mathrm{V}}, \ldots, s_N^{\mathrm{V}}, x_1^{\mathrm{V}}, \ldots, x_N^{\mathrm{V}}) \\ &\quad = P(t^{\mathrm{A}}, t^{\mathrm{V}}) \prod_{i=1}^{M} P(s_i^{\mathrm{A}} | s_{i-1}^{\mathrm{A}}, t^{\mathrm{A}}) \prod_{i=1}^{M} P(x_i^{\mathrm{A}} | s_i^{\mathrm{A}}) \prod_{i=1}^{N} P(s_i^{\mathrm{V}} | s_{i-1}^{\mathrm{V}}, t^{\mathrm{V}}) \prod_{i=1}^{N} P(x_i^{\mathrm{V}} | s_i^{\mathrm{V}}). \end{aligned} \tag{1}$$

The computational complexity of inference grows with the size of the support (number of nonzero values) of $P(t^{\mathrm{A}}, t^{\mathrm{V}})$, so it is important to keep the coupling local. Writing $P(t^{\mathrm{A}}, t^{\mathrm{V}}) = P(t^{\mathrm{A}})P(t^{\mathrm{V}}|t^{\mathrm{A}})$, we see that the second term should introduce only local coupling to keep the size of the support small.

3 Inference in ECHMMs

Here, we focus on the computation of the marginal probability of the observed sequences, $P(x^{\mathrm{A}}, x^{\mathrm{V}})$. Posterior distributions over states and event times can be computed in a similar fashion. This computation can be used in the E-step in an exact EM algorithm for fitting event-coupled HMMs.

Equation 1 gives the probability of the event times, the state sequences and the observation sequences. To calculate $P(x^{\mathrm{A}}, x^{\mathrm{V}})$ directly, we need to sum over all state sequences and transition times. This can be accomplished more efficiently by defining a forward probability $P(x^{\mathrm{A}}, x^{\mathrm{V}}, s_t^{\mathrm{A}}, s_{t'}^{\mathrm{V}})$, where x^{A} here denotes the audio observation sequence from the beginning to time t, x^{V} denotes the video observation sequence until the time t', s_t^{A} denotes a state in the audio sequence at time t and similarly for video.

The observation probability is obtained by summing the forward probability over the states. A backward probability can be defined for the same purpose. This is equivalent to what is done for conventional HMMs.

Since ECHMMs also include a temporal relation that imposes a transition constraint, we must sum over all possible t^{A} and t^{V}:

$$P(x^{\mathrm{A}}, x^{\mathrm{V}}, s^{\mathrm{A}}, s^{\mathrm{V}}) = \sum_{t^{\mathrm{A}}} \sum_{t^{\mathrm{V}}} P(t^{\mathrm{A}}, t^{\mathrm{V}}, x^{\mathrm{A}}, x^{\mathrm{V}}, s^{\mathrm{A}}, s^{\mathrm{V}}), \tag{2}$$

which can be rewritten as:

$$\sum_{t^{\mathrm{A}}} P(t^{\mathrm{A}}) P(x^{\mathrm{A}}, s^{\mathrm{A}} | t^{\mathrm{A}}) \sum_{t^{\mathrm{V}}} P(t^{\mathrm{V}} | t^{\mathrm{A}}) P(x^{\mathrm{V}}, s^{\mathrm{V}} | t^{\mathrm{V}}) \tag{3}$$

The $P(x^{\mathrm{V}}, s^{\mathrm{V}} | t^{\mathrm{V}})$ can be thought of as constrained forward probabilities. They define the forward probability given that t^{V} was the last nonevent state. In an HMM where state 1 captures the nonevent state and state 2 captures the first state of the event, the constrained forward probability is calculated by constraining the model to stay in state 1 until time t^{V} and then forcing a transition to state 2 and using the regular forward algorithm for the rest of the sequence.

A direct implementation of equation 3 is comprised of two loops. The inner loop multiplies the time dependent values of $P(t^{\mathrm{V}} | t^{\mathrm{A}})$ with $P(x^{\mathrm{V}}, s^{\mathrm{V}} | t^{\mathrm{V}})$. This calculation can be made more efficient by precalculating the $P(x^{\mathrm{V}}, s^{\mathrm{V}} | t^{\mathrm{V}})$ and storing them, since the same values are used for each outer loop for the outer sum. By limiting the support of $P(t^{\mathrm{A}}, t^{\mathrm{V}})$, the efficiency of the computation can be enhanced further.

For the purpose of training, a backward probability can be defined and implemented in similar manner.

4 Results

The task we tried was the detection of explosions in videos. The reason for choosing explosions was that they have clear onsets both in audio and video. Explosions can be adequately characterized in video by global color and therefore do not require region segmentation or more complex video preprocessing. (There is also ample data available at the local video rental store.)

Data for training the explosion model was collected from 2 action films. The non-explosion data that was used as negative examples, consisted of waterfall and river sequences. We chose these as "negative" examples because they shared audio features (thundering of water) and video features (pans from sunsets to waterfalls) with the explosion sequences, but with different timing characteristics. A total of 13 instances were used for training the audio model. For the training the video model, 30 sequences were used. Nine sequences were used for testing

There were fewer audio examples because many clips included music that masked the audio. In an automated system, auditory and visual occlusion (*e.g.*, due to

editing) will have to be taken into account. Because of the relative independence of audio and video in the ECHMM this can be done, say, by using only the video part of the model if music is present.

The audio was preprocessed by using a 20 bin filter bank with evenly spaced bins in the range from 0 to 22kHz. Features were computed at 20 millisecond intervals. The video features were the histograms of the HSV components of each frame. An 8 bin histogram was computed for each component and these features were augmented with the histogram difference between frames, giving feature vectors of length 48. A feature vector was computed for every frame. The frame rate was 30 frames per second. As can be seen from this example, the sampling rate for the audio and video models do not have to be the same, allowing the use of the optimum feature rate for each modality.

The component HMM models were trained separately and just a single state was used to model the nonevent observations. A discretized Gaussian was used to relate the transition times from state 1 to state 2 in the component HMMs. The mean of this Gaussian was set to 0 and the standard deviation was set to 0.1 seconds. Other probability distributions or mixtures of Gaussians can be used.

For testing, clips of 9 explosions were used. The negative examples consisted of 7 clips from the same action movies, as well as 18 clips of river and waterfall sequences. The lengths of the clips ranged from 10 to 50 video frames. Due to shot boundaries, the sequence length had to be limited.

To evaluate the performance of the model, observation log-probabilities for the explosion model were calculated for each of the test sequences. The log-probability under a nonexplosion model (the same as state 1 in the explosion model) was subtracted from each of these to normalize for different sequence lengths. Then, a threshold was applied to label the sequence as one containing an explosion or not.

Fig. 2 shows the false negative rates (curves dropping to the right) and the false positive rates (curves rising to the right) as a function of detection threshold for four different methods: audio detection alone (green), video detection alone (blue), audio plus video without temporal coupling (black), and the ECHMM (red). In the third method, the probability of the audio sequence under the audio HMM was simply multiplied by the probability of the video sequence under the video HMM, before the threshold was applied. (This is equivalent to setting $P(t^{\mathrm{A}}, t^{\mathrm{V}}) = const.$ in the ECHMMs.) Notice that the ECHMM can be viewed as this method but with an additional Gaussian penalty for the time difference between transitions. This causes a relative shift to the left of both the false negative and false positive error rate curves. However, the false negative curve shifts *more* to the left, giving improved discrimination.

5 Future work

We are currently collecting a larger data set in order to more clearly show the improvement ECHMMs make over uncoupled HMMs. We are also exploring recognizing commands from features extracted from both audio and video. An example of this is a "halt" command which is followed, or preceded by a raised and extended palm.

Although we focussed on inference in ECHMMs in this paper, learning in ECHMMs is also tractable. The inference method described can be used as the E-step in an exact EM algorithm for learning the coupled model. Future work will involve implementing a coupled forward-backward training algorithm for ECHMMs.

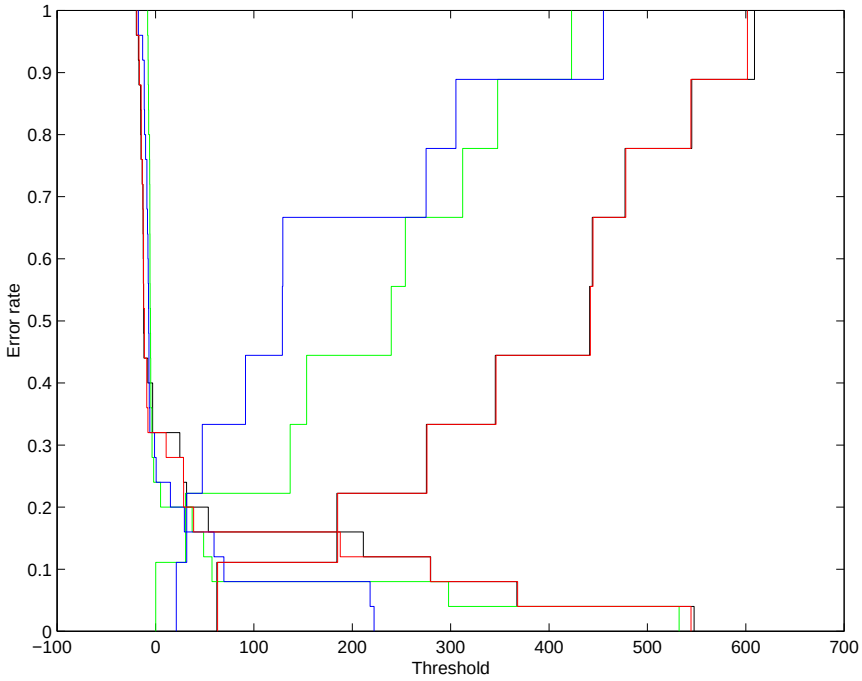

Figure 2: Misclassification rate as a function of log-probability threshold for negative cases (curves dropping to the right) and positive cases (curves rising to the right) for audio alone (green), video alone (blue), audio plus video without temporal coupling (black) and the ECHMM, which includes temporal coupling (red).

Acknowledgements

We thank Milind Naphade for extracting video features and training the video HMMs. This research was supported by a grant from the Arnold and Mabel A. Beckman foundation and the NCSA Isaac project. Trausti Kristjansson is a Fulbright Scholar; Brendan Frey is a Beckman Fellow.

References

S. Becker and G. E. Hinton 1992. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature* **355**, 161 – 163.

C. Bregler, S. Omohundro and Y.Konig 1994. A hybrid approach to bimodal speech recognition. In *Proceedings of the 28th Annual Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA.

P. Dayan, G. E. Hinton, R. M. Neal and R. S. Zemel 1995. The Helmholtz machine. *Neural Computation* **7**, 889–904.

B. J. Frey 1998. *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, MA.

B. J. Frey and D. J. C. MacKay 1998. A revolution: Belief propagation in graphs with cycles. In M. I. Jordan, M. I. Kearns and S. A. Solla (eds) *Advances in Neural Information Processing Systems 10*, MIT Press, Cambridge, MA.

G. E. Hinton, P. Dayan, B. J. Frey and R. M. Neal 1995. The wake-sleep algorithm for unsupervised neural networks. *Science* **268**, 1158–1161.

M. I. Jordan, Z. Ghahramani and L. K. Saul 1997. Hidden Markov decision trees. In M. C. Mozer, M. I. Jordan and T. Petsche (eds), *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, MA.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola and L. K. Saul 1998. An introduction to variational methods for graphical models. In M. I. Jordan (ed) *Learning and Inference in Graphical Models*, Kluwer Academic Publishers, Norwell MA.

R. M. Neal 1992. Connectionist learning of belief networks. *Artificial Intelligence* **56**, 71–113.

V. R. de Sa and D. Ballard 1998. Category Learning through Multi-Modality Sensing. To appear in *Neural Computation* **10:5**.

L. K. Saul and M. I. Jordan 1996. Exploiting tractable substructures in intractable networks. In D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (eds), *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge, MA.

B. Yuhas, T. Goldstein, T. Sejnowski and R. Jenkins 1988. Neural network models of sensory integration for improved vowel recognition. *Proceedings IEEE* **78**, 1655–1668.