ASYNCHRONOUS HMM WITH APPLICATIONS TO SPEECH RECOGNITION

Ashutosh Garg, Sreeram Balakrishnan, Shivakumar Vaithyanathan

IBM Almaden Research Center, San Jose, CA 95120

ABSTRACT

We develop a novel formalism for modeling speech signals which are irregularly or incompletely sampled. This situation can arise in real world applications where the speech signal is being transmitted over an error prone channel where parts of the signal can be dropped. Typical speech systems based on Hidden Markov Models, cannot handle such data since HMMs rely on the assumption that observations are complete and made at regular intervals. In this paper we introduce the asynchronous HMM, a variant of the inhomogenous HMM commonly used in Bioinformatics, and show how it can be used to model irregularly or incompletely sampled data. A nested EM algorithm is presented in brief which can be used to learn the parameters of this asynchronous HMM. Evaluation on real world speech data that has been modified to simulate channel errors, shows that this model and its variants significantly outperforms the standard HMM and methods based on data interpolation.

1. INTRODUCTION

Hidden Markov Models (HMMs) [1] are a popular tool to model time series data and are widely used in fields such as Speech Recognition, computer vision, text analysis, and BioInformatics. While modeling the data using HMMs, it is assumed that there is an underlying Markov process that generates the hidden state sequence and observations are made at *regular* intervals conditioned on these states. As we argue in this paper, for a number of reasons the latter assumption may not always hold. For example, when speech data is transmitted over a noisy channel before recognition then some of the frames might be lost. The approaches mentioned in literature to tackle these problems can be broadly divided into two categories -

1 If the actual time-stamps of the missing frames are available, then interpolated values can be used to fill the missing observations. Once predicted, data is decoded using standard HMM.

2 Modify the structure of the underlying HMM by adding skip-arcs (allow certain states to be skipped). The weights of the skip-arcs are either learnt [2] or chosen in some ad-hoc way.

In this paper we introduce the asynchronous HMM that directly models the uncertainty associated with missing observations without assuming the availability of time-stamps (required for interpolation). This is achieved by actually modeling the time stamps associated with each observation as a hidden variable. As a consequence of this the time interval between each pair of observations may not be the same, which results in the transition matrix becoming time dependent. We show that if the time gap is k, then the effective transition matrix is A^k. This makes the asynchronous HMM a special case of inhomogeneous HMM. However, unlike the more general case of inhomogeneous HMM where the transition matrix varies with time, here the underlying transition matrix is fixed and the variability arises only due to the irregularity in the sampling process (leading to missing observations). Our results on real speech data demonstrate that the asynchronous HMM is an effective method for handling data that has been irregularly or incompletely sampled, and even outperforms alternatives based on interpolation of the missing data.

The remainder of the paper is structured as follows. In Section 2 we introduce the notation and give a brief description of the standard HMM formulation. Section 3 introduces asynchronous HMM and discusses various special cases. In Section 4 we give a brief description of a novel, nested EM algorithm for Maximum Likelihood training of the new model. Section 5 is concerned with how the Asynchronous HMM can be implemented efficiently using some alternative formulations. Finally in Section 6 we present results on a speech recognition task.

2. NOTATIONS AND PRELIMINARIES

Fig. 1 gives the standard HMM rolled out in time. Here $S_i \in \{1, ..., N\}$ and $O_i \in \{1, ..., M\}$ are random variables referred to as hidden states and observations. The corresponding sequences of random variables are denoted as $\mathbf{S} = \{S_1, ..., S_T\}, \mathbf{O} = \{O_1, ..., O_T\}$. The standard HMM is characterized by parameter vector $\lambda = (A, B, \pi)$, where

- A - Transition probability matrix. It is a $N \times N$ matrix where $a_{ij} = A(i, j) = P(S_t = i | S_{t-1} = j)$.
- B - Observation probability matrix. It is a $M \times N$

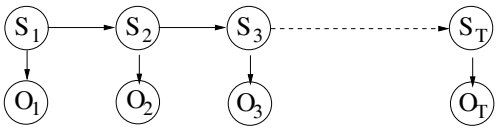

Fig. 1. Standard Hidden Markov Model

matrix where $b_{ij} = B(i,j) = P(O_t = i|S_t = j)$.

- π - Initial state probability matrix. $\pi(i) = P(S_1 = i)$. It is a $N \times 1$ matrix.

3. ASYNCHRONOUS HMM

Consider an observation sequence $\widetilde{\mathbf{O}} = \{\widetilde{O}_1, \widetilde{O}_2, ..., \widetilde{O}_K\}$. Let $\widetilde{O}_k$ was observed at time C_k where $C_K \leq T$ but the actual value of C_k's are unknown. There are $Choose(T,K)$ possible ways in which the time index can be assigned to the observation sequence. However, many of these choices may not be feasible due to the constraints imposed by the observation and transition probability matrices. Moreover, we might want to incorporate prior information in the form of a distribution over the length of the sequence or over the time gaps between individual observations. It is precisely this problem with missing observations that asynchronous HMMs are designed to tackle.

For the observation sequence $\widetilde{\mathbf{O}}$, if $\forall k: \ C_{k+1} = C_k + 1$ (i.e., there are no missing observations) then the problem reduces to that of the standard model. In our model, we allow C_k to take on any values under the constraint $C_1 < C_2, \ldots, C_K \leq T$. It is therefore conceivable that for some values of k, the difference between successive C_k's can be greater than one. This model is further generalized if instead of the actual values of C_k, only a prior distribution over the values that C_k can take is assumed to be known. Note that C_{k+1} is not independent of C_k since due to the temporal constraint $C_{k+1} > C_k$.

Fig. 2 represents an asynchronous HMM. Given that $\widetilde{S}_k, \widetilde{O}_k$ are observed at time C_k, we will use S_{C_k} and O_{C_k} interchangeably with $\widetilde{S}_k$ and $\widetilde{O}_k$. The additional parameters needed to characterize asynchronous HMM are:

- $P(C_{k+1}|C_k)$ - the probability distribution over the values taken by C_{k+1} conditioned on the values of C_k. This is our prior model for the sequence $C_1 \cdots C_K$. For simplicity we make a first order Markov assumption about C_{k+1}. In addition $P(C_{k+1}|C_k)$ can also be used to impose the constraint that $C_{k+1} > C_k$ while at the same time $C_K \leq T_{\max}$ (the maximum length of the ground truth observation sequence.) W.l.o.g we assume $P(C_1 = 1) = 1$.

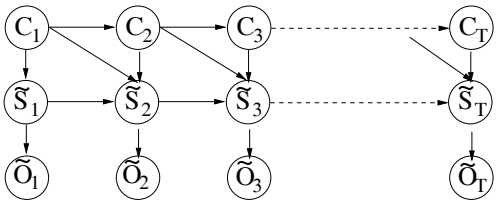

Fig. 2. Asynchronous HMM.

- $P(\widetilde{S}_{k+1} = j|\widetilde{S}_k = i, C_{k+1}, C_k)$ - Since, C_k is the time-stamp associated with $\widetilde{S}_k$, we have

$$P(\widetilde{S}_{k+1} = j|\widetilde{S}_k = i, C_{k+1}, C_k) = [A^{(C_{k+1}-C_k)}]_{ij} \quad (1)$$

A is the transition probability matrix of the asynchronous HMM. The ij^{th} element of $A^{C_{k+1}-C_k}$ is obtained by summing the probabilities of all state sequences of length $C_{k+1} - C_k$ that start at state i and end at state j. It is this particular choice of parameter that distinguishes asynchronous HMM from other extensions of HMM such as factorial [3] or coupled HMM [4].

Let $\widetilde{\lambda}$ denote the parameters of this model. The joint probability of the observation sequence, state sequence and time index sequence of length K, can be written as

$$P(\widetilde{\mathbf{O}}, \widetilde{\mathbf{S}}, \mathbf{C}|\widetilde{\lambda}) = P(\widetilde{O}_1, \widetilde{S}_1) \prod_{k=2}^{K} P(\widetilde{O}_k|\widetilde{S}_k, C_k) P(\widetilde{S}_k, C_k|\widetilde{S}_{k-1}, C_{k-1})$$

$$= P(\widetilde{O}_1, \widetilde{S}_1) \prod_{k=2}^{T} P(O_{C_k}|S_{C_k}) P(C_k|C_{k-1}) [A^{(C_k - C_{k-1})}]_{S_{C_{k-1}} S_{C_k}}$$

Once the parameters of the asynchronous model are estimated, decoding can be accomplished using a two step process - Obtain the most likely time index sequence (C_k) followed by obtaining the hidden state sequence S_k. Estimating the parameters of this model is non-trivial and in the next section we discuss the challenges involved and present a nested EM algorithm as a possible solution.

3.1. Special Case - Skip Arc Model

There are a number of cases (arising either due to the data generation process itself or as a result of various simplifying assumptions made) which allows one to simplify the computational effort associated with the asynchronous HMM. In particular the most interesting case is of skip arc model. In this model, it is assumed that the $P(C_{k+1}|C_k)$ is independent of k. That is the probability of missing an observation at any given time is independent of the actual time at which the observation is made. It can be shown that this scenario can be modeled by the standard HMM by choosing the transition matrix as $\tilde{A} = \sum_{n=1}^{\infty} P(n) A^n$. Note that in this case one doesn't have control over the actual number of missing frames which is important as evident from the results obtained on the speech data.

4. NESTED EM ALGORITHM FOR TRAINING

The asynchronous HMM is a model proposed primarily to tackle missing observations. If there are no missing observations in the training data then data can be modeled using a standard HMM. However, if we then encounter missing observations during decoding the standard HMM has to be enhanced by additional parameters ($P(C_k|C_{k-1})$) to form an asynchronous HMM. Since these were not present during training they have to specified by some prior knowledge. If, on the other hand, the training data does contain missing observations then these extra parameters, of the asynchronous HMM, can be learned. The EM learning algorithm used for standard HMM, unfortunately, cannot be directly used to learn an asynchronous HMM. At a high level, the nested EM algorithm adopted for this purpose is an iterative algorithm that iterates over these three steps -

- E-Step 1 - Estimate the hidden state time index sequence $C_1, \ldots, C_k$.
- E-Step 2 - Based on estimated time indexes, obtain the hidden state sequence $\widetilde{S}_1, \ldots, \widetilde{S}_K$.
- M-Step - Estimate the parameters that maximize the probability.

It can be easily shown that the algorithm is guaranteed to converge and can be seen as a special case of generalized EM algorithms. In the interest of the space, the details and actual update equations are ignored.

5. IMPLEMENTATION ISSUES

The simplest way to implement an asynchronous HMM is to define a new state variable Q that is the product of the state spaces of $\tilde{S}$ and C. If $\tilde{S} \in 1, \ldots, N$ and $C \in 1, \ldots, T$ then we can define

$$Q \in 1, \ldots, NT \text{ where } Q_k = N(C_k - 1) + \tilde{S}_k$$

We can then obtain a new transition matrix for Q from

$$\begin{aligned} P(Q_k|Q_{k-1}) &= P(\tilde{S}_k = j, C_k|\tilde{S}_{k-1} = j, C_{k-1}) \\ &= P(\tilde{S}_k = j|\tilde{S}_{k-1} = i, C_k, C_{k-1})P(C_k|C_{k-1}) \\ &= [A^{C_k - C_{k-1}}]_{ij} P(C_k|C_{k-1}) \end{aligned} \quad (2)$$

Hence using equation (2) the asynchronous HMM can be implemented as normal HMM with an extended state space and a transition matrix computed from (2). Unfortunately it is easy to see that even for small T this leads to a huge increase in the number of states. We propose two ways of handling this issue: (1) dynamically expand the state space and (2) simplify the asycnhronous HMM, by redefining C to reduce its range.

5.1. Dynamic state space expansion

Instead of statically expanding the product of the state spaces of S and C to Q, we can use a dynamic expansion that only creates the states that are going to have high probability. Dynamic HMM implementations are common in Large Vocabulary Speech Recognition systems [5]. For the particular case of the asynchronous HMM we can use beam pruning in the search to limit the underlying set of values for C_k that are kept alive for each state S_k.

5.2. Redefining C to reduce its range

Another way of reducing the amount of computation required is to change the definition of C so that it has a smaller range, but in a way that preserves the underlying ability to model irregularly sampled and incomplete data. One way of doing this is to define $C_k - 1$ to be the difference between the time k the observation is made and the real time t the observation was generated. Hence $C_k - 1 = t - k$ and $\tilde{O}_k = O_t$ where $t = k + C_k - 1$. Also

$$P(S_k = j|S_{k-1} = i, C_k, C_{k-1}) = [A^{C_k - C_{k-1}+1}]_{ij}$$

We can now impose a restriction that the maximum delay should be less than D, i.e. that $C \in 1, \ldots, D$. Since D can be much smaller that maximum value of t the total number Q space states becomes much more manageable.

As a concrete example, if we set $D = 2$ and define Q as in (2) then it can be seen that the effective transition matrix in Q space, A_Q is given by:

$$A_Q = \begin{bmatrix} A_S[A_C]_{11} & A_S^2[A_C]_{12} & A_S^3[A_C]_{13} \\ Z & A_S[A_C]_{22} & A_S^2[A_C]_{23} \\ Z & Z & A_S[A_C]_{33} \end{bmatrix} \quad (3)$$

where A_S is the transition matrix in S space, Z is a matrix of zeros and $[A_C]_{ij} = P(C_k = j|C_{k-1} = i)$.

A further simplification can be made if we assume that there is only a single contiguous block of missing observations. In this case

$$A_C = \begin{bmatrix} a_0 & a_1 & a_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

This transition matrix for C ensures that once a transition from $C = 1$ has been made, then no further transitions in the value of C are possible except self transitions. i.e. it can model a single block of $1 \ldots D$ missing frames. For this particular case C can be further redefined as follows:

$$C = \begin{cases} 1 & \text{if there are no missing frames} \\ 2 & \text{if a block of } \leq D \text{ frames is missing} \end{cases}$$

The effective matrix in Q space assuming now becomes

$$A_Q = \begin{bmatrix} A_S a_1 & \sum_{d=2}^{D} A_S^d a_d \\ Z & A_S \end{bmatrix} \quad (4)$$

6. RESULTS

We tested the asynchronous HMM on a speech recognition task from our test database collected in a car [6]. We report word error rates on a test set comprised of small vocabulary grammar based tasks (addresses, digits, command and control). Data for each task was collected at 3 speeds: idling, 30mph and 60mph. There are 147 combinations of speaker, task and environment in the test set, and for each combination there are 100 test utterances, giving a total of 73743 test words. The observations consisted of 39 dimensional cepstra+delta+accelation at 67 frames per second, and the recognition system had $10K$ Gaussians. The baseline WER for this system was 2.36%.

To simulate conditions for which the data is irregularly or incompletely sampled, we randomly removed a single block of frames of random length from each utterance. The maximum length of the block was varied between 20 and 40 frames. We computed results for two scenarios

Location of missing block *unknown* - in this scenario the following systems were compared

(a) A standard HMM

(b) A Full Asynchronous HMM with no prior knowledge of the location or size of the missing block - the transition matrix defined by (4) was used for all frames (with $D = 15$ and $a_d = 0.8^d$)

Location of missing block *known* - here we compared the following techniques that require knowledge of the location of the missing block

(a) An Interpolation of the data followed by a standard HMM - The interpolation was performed by linearly interpolating the state observation probabilites of the first frame before and after the deleted block

(b) A Restricted Asynchronous HMM used only for frames straddling the missing block of data - The transition matrix and definition of C from equation (4) was used for the transition between the frames staddling deleted block. For all other frames the $\sum_{d=2}^{D} A_S^d a_d$ part of (4) was replaced by a matrix of zeros

It can be seen from the results in Table 1 that with no knowledge of the location of the deleted block, the Asynchronous HMM achieves significant reduction in error rate versus the standard HMM, especially as the size of the block increases. The only drawback is that for no deletions it has a worse performance. When the location of the missing block is known, techniques such as interpolation can be used. However if the Asynchronous HMM is restricted to the transition between the frames straddling the missing block and a standard HMM transition used elsewhere, then this system either matches or significantly outperforms interpolation.

Location of Block	Max Size of Deleted Block	0	20	30	40
Unknown	Standard HMM	2.36	6.15	9.41	12.47
Unknown	Full Async	4.64	5.80	7.21	9.23
Known	Interpolate Data	2.36	4.06	6.35	8.92
Known	Restricted Async	2.36	4.14	5.56	7.44

Table 1. Comparison of Word Error Rates (%) For standard HMM versus full asynchronous HMM with no knowledge of missing block and for Interpolation of Data followed by Standard HMM versus Restricted Asynchronous HMM (used only where data is missing)

6.1. Conclusion

In this paper we presented a new model, the Asynchronous HMM, that extends HMMs to model irregularly sampled data, and we provided a novel nested EM training algorithm. We showed with experiments on speech data with randomly deleted blocks of data, that the asynchronous HMM outperforms the both standard HMM and techniques such as interpolation. In future we plan to explore the potential gains that may arise from retraining asynchronous model on the speech data. Additionally, the flexibility of the asynchronous HMM may prove useful in other applications.

7. ACKNOWLEDGEMENTS

The authors wishes to thank Ramesh Gopinath for his helpful feedback.

8. REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, 1989.

[2] M.D. Monkowski, M.A. Picheny, and P.S. Rao, "Context dependent phonetic duration models for decoding conversational speech," in *Proceedings of International Conference on Accoustics, Speech and Signal processing*, 1995, pp. 528–531.

[3] Z. Ghahramani and M. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–273, 1997.

[4] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proceedings of Computer Vision and Pattern Recognition*, 1997, pp. 994–999.

[5] S. Ortmanns, H. Ney, F. Seide, and I. Lindam, "A Comparison of Time Conditioned and Word Conditioned Search Techniques for Large Vocabulary Speech Recognition," in *Proc. int. Conf. on Spoken Language Processing*, 1996.

[6] P. Olsen and R. A. Gopinath, "Modeling inverse covariance matrices by basis expansion," in *Proceedings of International Conference on Accoustics, Speech and Signal processing*, 2002, vol. 1, pp. 945–948.