# Report

# Time modeling in Hidden Markov Models

Version 1.0 24. 09 2004

Hagen Kaprykowsky\* IRCAM Centre Pompidou

\*<kaprysonne@web.de>

# Contents

1	Introduction	<b>2</b>		
	1.1 Durational behaviour in HMMs	2		
	1.2 Explicite state duration modeling (ESDM)	3		
	1.3 Implicite state duration modeling (ISDM)	3		
<b>2</b>	Relations between ISDM and ESDM			
	2.1 Forward Variable ISDM	4		
	2.2 ISDM as LTV system	7		
3	From ISDM to ESDM	8		
4	Decoding	9		
<b>5</b>	Conclusion	10		

# List of Figures

1	Single state	2
2	Cluster of n states	3
3	Cluster $n = 8$	3
4	Macro state	5
5	Forward Variable $\alpha$ in a cluster with the input $\alpha_t(j-1)$ and the output $\alpha_t(j)$	6
6	Impulse responses for different $n$ and $p$	8

#### Abstract

A classical problem with the HMM approach lies in its temporal modeling. Perhaps the major weakness of conventional HMMs is the modeling of state duration [Rabiner, 1989]. In this report a connection between implicit state duration modeling in HMMs, explicit state duration modeling, and time invariant linear systems will be given. The work takes place in the context of the Ircam score follower while most approaches are given in speech recognition. The maximum state duration measured as the number of observations in speech recognition is typically 32 frames of 10 ms. [Mitchell et al., 1993]. The maximum state duration of a note is typically much higher. This has to be taken in account in the temporal modeling of the HMM of the score follower.

### 1 Introduction

#### Score following

Score following is the synchronisation of a computer with a performer playing a known musical score. The score follower that we developed at Ircam is based on a Hidden Markov Model and on the modeling of the expected signal recieved from the performer. The model works on two levels. The lower level computes the features of the incoming signal with the expected ones. Groups of the lower level are embedded in states at higher level, which are used to model the performance by taking into account the possible errors a performer may make. The score follower works in realtime and due to the nature of a music performance only transitions to previous and next states are allowed and essentially assures a temporal left-right flow on the score model [Orio and Déchelle, 2001, Orio et al., 2003, Cont, 2004].

#### The current state durational modeling in the IRCAM score follower

The temporal modeling in the score follower is based on implicit state duration modeling (ISDM). ISDM is an efficient approach of time modeling developed for speech recognition. The approach has been chosen because of the real time conditions of the system [Mouillet, 2001]. Because of the maximum state duration up to 1000 observations for a long note in the HMM of the score follower, ISDM provides an unsatisfying durational modeling and maximum state constraints because of the real time conditions.

#### Exploration of explicit state duration modeling (ESDM) :

Explicit state duration modeling is the most general approach of modeling time behaviour in HMMs. High calculation and storage costs are the main problem in speech recognition of this approach. Based on the conditions of a musical model, ESDM for score following will be analysed and a connection between ISDM, ESDM and linear time invariant (LTI) systems will be given.

#### 1.1 Durational behaviour in HMMs

Single state :



Figure 1: Single state

The durational probability density function (pdf) of a single state i of an HMM is:

$$p_i(d) = p^{d-1}(1-p) \qquad d \ge 1$$
 (1)

 $p_i(d)$  denotes the probability of staying in state *i* for exactly *d* time steps, and *p* is the selfloop probability of state *i*. This is the geometrical distribution. It gets to its maximal value at the minimal duration d = 1, and decays exponentially as *d* increases [Wang, 1997]. This durational behaviour is regarded as improper modelling the duration of notes.

#### 1.2 Explicite state duration modeling (ESDM)

A common idea in speech recognition concerning the durational modeling in HMM is to replace the pdf of (1) with well-chosen pdf close to the durational distribution of speech segments. HMMs with such an explicity added state duration pdf are called hidden semi-Markov models (HSMM), because the transition properites are no longer governed by a markov process. The selfloops are removed: p=0. The main problem of explicit state duration modeling is, that both the memory space and calculation costs increase with a factor from 2 to a few hundred, as compared to that of HMM [Wang, 1997].

#### 1.3 Implicite state duration modeling (ISDM)

:

 $Cluster \ of \ n \ states$ 



Figure 2: Cluster of n states

A more efficient, but less flexible way to model non-geometric waiting times is to replace each state with n new states, each with the same emission probabilities as the original state. The durational probability density function of a cluster model of an HMM is:

$$p(d) = \begin{pmatrix} d-1\\ n-1 \end{pmatrix} p^{d-n} (1-p)^n$$
(2)

This is the negative binomial distribution [Murphy, 2002].

#### The timemodeling of the IRCAM score follower :

Each note is modeled by a cluster of maximum 8 states. The expacted duration of the note i corresponds to the  $\mu_i$  of the cluster. The relations between  $\mu$ ,  $\sigma^2$ , n and p are:  $\mu = \frac{n}{1-p}$  and  $\sigma^2 = \frac{np}{(1-p)^2}$  [Wang, 1997]. To achieve the expected  $\mu$  (duration) of each note we have to calculate p. Figure 3 shows the relations between  $p(\mu)$ ,  $\mu(p)$  and  $\sigma^2(p)$  for a Cluster of n = 8 states:



Figure 3: Cluster n = 8

A cluster with fixed numbers of states,  $\mu$  and  $\sigma^2$  has dependence of p as follows:

$$\Rightarrow \quad \mu \sim \frac{1}{1-p} \quad and \quad \sigma^2 \sim \frac{p}{(1-p)^2} \tag{3}$$

A long note modeled by a cluster with few number of states results in a high variance. A higher number of states would reduce the variance but increases calculation costs and introduces a minimum of state duration. The connection between ISDM and ESDM and the comparison of the calculation costs for a musical model will be made. As follows we give an example of a score used in the IRCAM score follower.

#### **Example:** 'Enecho' extract (Manoury)

This is a part of a score used for score following at IRCAM: (Notation of the Score: duration + name + octave)

	$(1.5_{do4})$
	$1.2_{do4}$
	$0.2_{si3}$
	$0.2_{mi4}$
	$0.3_{si3}$
	$0.3_{sib3}$
score =	$0.3_{la3}$
	$0.3_{sold3}$
	$0.3_{re4}$
	$0.3_{fa3}$
	$1.6_{mi4}$
	$1.2_{mi4}$

(4)

The nominal relation among performed note length and corresponding note values in a score and its tempo can be related as follows: The duration d [sec] of a note in the performance is related both to the intended note value q [beats] that appear in the score and to the tempo variable  $\tau$  [beats/sec]. This is formulated by:  $d [sec] = \tau [sec/beats] q [beats]$  [Takeda et al., 2004]. Here: absolute duration in sec.: We take the note with the maximum duration of this part:  $1.6_{mi4}$ . The expected tempo of the piece is given, so we can calculate the expected duration  $\mu$  of this note in number of observations:  $\mu = 276$ . Given the maximum number of states in a cluster fixed as n=8 we get: p=0.971,  $\sigma^2 = 9243$  and  $\sigma = 91.14$ .  $\sigma$  is approximately 35 percent of  $\mu$ . sigma could be interpreted as variation of the tempo. The time modeling is unsatisfying .The current choice of n and p is a compromise between the tolerance for the tempo and the accuracy of time modeling. Nevertheless for long notes we obtain large  $\sigma$  and the influence of time modeling is negligable compared to that of the observations and the HMM is driven by the observations only.

### 2 Relations between ISDM and ESDM

#### 2.1 Forward Variable ISDM

While being used in real-time, the sore follower is examining the probability that a sequence of events would occur. In the classical HMM literature an efficient procedure called Forward-Backward procedure computes this variable which takes into account both probability of the state before and after a certain state for computation. Clearly, since we are dealing with real time situations we can have no chance of observing the future states and therefore, we use only the Forward Procedure which takes into account the observation probabilities for all states from the observation block, transition probabilities obtained from the music model and previous sequence probabilities [Cont, 2004]. The forward variable  $\alpha_t(j)$  is defined as:

$$\alpha_t(i) = P(O_1, O_2, ..., O_t, q_t = S_i | \lambda)$$
(5)

i.e. the probability of the partial observation sequence,  $O_1, O_2, ..., O_t$ , (until time t) and state  $S_i$  at time t, given the model  $\lambda$ . We can solve for  $\alpha_t(i)$  inductively, as follows:

$$\alpha_1(i) = \pi_i b_i(O_1) \tag{6}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(j) \ a_{ij}\right] \ b_j(O_{t+1}) \tag{7}$$

[Rabiner, 1989]

The transition probabilities  $a_{ij}$  of a cluster model are:

$$a_{ij} = \begin{cases} p & \text{if } j = i \\ 1 - p & \text{if } j = i + 1 \\ 0 & \text{else} \end{cases}$$

$$\tag{8}$$

The observations for one cluster are:  $b_j(O_{t+1}) = b(O_{t+1})$ . We start in the first state of the Cluster  $\Rightarrow \pi_1 = 1$ . We obtain the formula for the forward variable  $\alpha_t(i)$  in a cluster.

$$\alpha_1(1) = b_1(O_1) \tag{9}$$

$$\alpha_{t+1}(j) = [\alpha_t(j) \ p + \alpha_t(j-1) \ (1-p)] \ b(O_{t+1})$$
(10)

We can express the formular of the  $\alpha_i(t)$  by using a matrix multiplication:

$$\begin{pmatrix} \alpha(1) \\ \vdots \\ \alpha(n) \end{pmatrix}_{t+1} = \underbrace{\begin{pmatrix} p & 0 & \dots & \dots & 0 \\ 1-p & p & \ddots & \ddots & \vdots \\ 0 & 1-p & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1-p & p \end{pmatrix}}_{trans'} \begin{pmatrix} \alpha(1) \\ \vdots \\ \alpha(n) \end{pmatrix}_{t} b(O_{t+1})$$
(11)

In the context of the HMM a cluster can be represented as a macro state shown in Figure 4:



Figure 4: Macro state

The calculation of  $\alpha$  for a macro state can be represented as follows:



Figure 5: Forward Variable  $\alpha$  in a cluster with the input  $\alpha_t(j-1)$  and the output  $\alpha_t(j)$ 

#### 2.2 ISDM as LTV system

In order to express a cluster of states as a discrete state space model we can write as follows:

$$\begin{pmatrix} \alpha(1) \\ \vdots \\ \alpha(n) \end{pmatrix}_{t+1} = \left[ \underbrace{\begin{pmatrix} p & 0 & \cdots & \cdots & 0 \\ 1-p & p & \ddots & & \vdots \\ 0 & 1-p & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1-p & p \end{pmatrix}}_{A} \begin{pmatrix} \alpha(1) \\ \vdots \\ \alpha(n) \end{pmatrix}_{t} + \underbrace{\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{B} \alpha_{t}(j-1) \right] b(O_{t+1}) D_{t}$$

$$y_t = \underbrace{\left(\begin{array}{ccc} 0 & \dots & 0 & 1-p\end{array}\right)}_C \left(\begin{array}{c} \alpha(1) \\ \vdots \\ \alpha(n) \end{array}\right)_t = \alpha_t(j)$$
(13)

x is the state space vector which corresponds to the different  $\alpha s$  of the cluster. u is the input of the state space model and y is the output. The equations (12) and (13) can be expressed as follows:

$$x_{t+1} = A(t) x_t + B(t) u_t$$
(14)

$$y_t = C x_t \tag{15}$$

A series of states in the ISDM can be modeled as a LTV system. The time variance is based on the time dependent observations. We can seperate this system into a part which corresponds to the duration modeling and a multiplication with the observations. In the following we will represent the duration modeling by a LTI system G(z). We get:

$$x_{t+1} = A x_t + B u_t \tag{16}$$

$$y_t = C x_t \tag{17}$$

$$G(z) = C (zI - A)^{-1}B$$
(18)

and obtain:

$$G(z) = \frac{(1-p)^n}{(z-p)^n}$$
(19)

In Figure 6 we show the impulse responses of clusters with different n and p. Obviously the smallest sequence this can generate is of length n. Any path of length d through the model has probability  $p^{d-n}(1-p)^n$ ; the number of possible paths is  $\binom{d-1}{n-1}$ , so the total probability of a path of length d is:

$$p(d) = \begin{pmatrix} d-1\\ n-1 \end{pmatrix} p^{d-n} (1-p)^n$$
(20)

This is the negative binomial distribution. [Murphy, 2002].

We observe the minimal duration fixed by the parameter n. The dependence of the variance by the parameter p is shown in Figure 6.

The calculation of  $\alpha_t(j)$  using G(z) (order n) is the convolution  $\alpha_t(j) = p_j(t) * \alpha_t(j-1)$ .

 $p_j(d)$  is the impulse response of the cluster and respresent the discrete probability density function of duration of state j.



Figure 6: Impulse responses for different n and p

Introducing the observations directly in this formular, we obtain a linear time invariant system. The impulse response is truncated at a maximum duration value N and normalized:  $\sum_{d=1}^{N} p_j(d) = 1$ . We obtain the formula as follows:

$$\alpha_t(j) = \sum_{d=1}^N p_j(d) \; \frac{\alpha_{t-d}(j-1)}{\prod_{d=1}^{t-d} \; b(O_s)} \prod_{s=1}^t \; b(O_s) \tag{21}$$

$$= \sum_{d=1}^{N} p_j(d) \ \alpha_{t-d}(j-1) \prod_{s=t-d+1}^{t} b(O_s)$$
(22)

The impulse response of a cluster (negative binomial distribution) is zero for n < d.

$$\alpha_t(j) = \sum_{d=n}^N p_j(d) \; \alpha_{t-d}(j-1) \prod_{s=t-d+1}^t \; b(O_s) \tag{23}$$

# 3 From ISDM to ESDM

The formular of  $\alpha_t(j)$  of the ESDM is:

$$\alpha_t(j) = \sum_{i=1}^N \sum_{d=1}^D \alpha_{t-d}(i) a_{ij} p_j(d) \prod_{s=t-d+1}^t b_j(O_s)$$
(24)

$$\alpha_t(j) = P(O_1, O_2, ..., O_t, the stay in state j ends at t|\lambda) =$$
(25)

$$\alpha_t^*(j) = \sum_{i=1}^N \alpha_t(i)a_{ij} \tag{26}$$

$$\alpha_t^*(j) = P(O_1, O_2, ..., O_t, stay \text{ in state } j \text{ starts at } t+1|\lambda)$$
(27)

$$\alpha_t(j) = \sum_{d=1}^{D} \alpha_{t-d}^*(j-1) \, p_j(d) \prod_{s=t-d+1}^{t} b_j(O_s)$$
(28)

[Rabiner, 1989]

 $\alpha_t^*(j)$  is the input  $\alpha_t(j-1)$  of the discrete state space model and  $\alpha_t(j)$  correspond to the output of the model. We obtain the result that the formular of ESDM correspond to the equation () of the ISDM.

#### Musical model: case $a_{j-1,j} = 1$ :

In the case, that each state is only connected to the next state (common for musical models) we get:

$$\alpha_t(j) = \sum_{d=1}^{D} \alpha_{t-d}(j-1) p_j(d) \prod_{s=t-d+1}^{t} b_j(O_s)$$
(29)

#### Decoding 4

#### Decoding in the current score follower :

After computing all possible  $\alpha$  variables  $(\alpha_t(i) = P(O_1, O_2, ..., O_t, q_t = S_i | \lambda))$  for the audio frame in consideration, we can solve individually most likely state  $q_t$ , as

(30) $q_t = argmax_{1 \le i \le N}[\alpha_t(i)]$ 

Equation (30) is referred to as decoding and specifies the most likely high-level state depending on previous matches, new observation probabilities as well as the score and in real time [Orio and Déchelle, 2001].

#### New Decoding :

The forward variable for an ESDM  $\alpha_t(j) = P(O_1, O_2, ..., O_t, the stay in state j ends at t|\lambda)$  is not adapted for the decoding. We define a new variable:

 $\widehat{\alpha}_t(j) = P(O_1, O_2, \dots, O_t, q_t = S_i | \lambda).$ 

Г

Based on the durational modeling with a LTI system we calculate  $\hat{\alpha}_t(j)$  as follows;

$$\widehat{\alpha}_t(j) = \sum_t \alpha_t(j-1) - \sum_t \alpha_t(j) = \sum_t (\alpha_t(j-1) - \alpha_t(j))$$
(31)

$$\widehat{\alpha}_{z}(j) = \frac{1}{z-1} \left( \alpha_{z}(j-1) - \alpha_{z}(j) \right)$$
(32)

$$G(z) = \frac{\alpha_z(j)}{\alpha_z(j-1)} = \frac{(1-p)^n}{(z-p)^n} \Rightarrow \hat{\alpha}_z(j) = \frac{1}{z-1} (\alpha_z(j-1) - \frac{(1-p)^n}{(z-p)^n} \alpha_z(j-1))$$
(33)

$$\widehat{\alpha}_z(j) = \frac{1}{(z-1)} \left( 1 - \frac{(1-p)^n}{(z-p)^n} \right) \alpha_{in}(z)$$
(34)

Based on the discrete state space model (12) (13) we express  $\hat{\alpha}_t(j)$  as follows:

$$\begin{pmatrix} \alpha(1) \\ \vdots \\ \alpha(n) \end{pmatrix}_{t+1} = \left[ \underbrace{\begin{pmatrix} p & 0 & \cdots & \cdots & 0 \\ 1-p & p & \ddots & \vdots \\ 0 & 1-p & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1-p & p \end{pmatrix}}_{A} \begin{pmatrix} \alpha(1) \\ \vdots \\ \alpha(n) \end{pmatrix}_{t} + \underbrace{\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{B} \alpha_{t}(j-1) \right] b(O_{t+})$$

$$y_{t} = \underbrace{\begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}}_{C} \begin{pmatrix} \alpha(1) \\ \vdots \\ \alpha(n) \end{pmatrix}_{t} = \widehat{\alpha}_{t}(j)$$
(36)

$$\widehat{\alpha}_t(j) = (\widehat{\alpha}_{t-1}(j) + \alpha_{t-1}(j-1)) b_j(O_t) - \alpha_t(j)$$
(37)

$$\widehat{\alpha}_{t}(j) = \left( \left( \widehat{\alpha}_{t-1}(j) + \alpha_{t-1}(j-1) \right) b_{j}(O_{t}) - \sum_{d=1}^{D} p_{j}(d) \alpha_{t-d}(j-1) \prod_{s=t-d+1}^{t} b_{j}(O_{s}) \right)$$
(38)

After computing all possible  $\hat{\alpha}$  for the audio frame in consideration, we can solve individually most likely state  $q_t$ , as

$$q_t = \arg\max_{1 \le i \le N} [\widehat{\alpha}_t(i)] \tag{39}$$

Equation (39) is referred to as decoding and specifies the most likely state depending on previous matches, new observation probabilities as well as the score and in real time

## 5 Conclusion

The starting point of this was the ISDM of the score follower with a cluster of maximum 8 states. Time modeling for long notes with a cluster of few states has as consequence a large  $\sigma$ , which corresponds to a large variation of the tempo. For long notes the time modeling is unsatisfying. To obtain a more accurate state duration modeling, we would have to raise the number of states in the cluster which raises the calculation costs. Starting by developing a general ISDM, the relations between ISDM and ESDM has been shown and an approach by using a discrete state space model has been given. A decoding technique for ESDM has been developed.

## References

Arshia Cont. Improvement of observations for scorefollowing. Rapport de stage, IRCAM, 2004.

- C. D. Mitchell, R. A. Helzerman, L. H. Jamieson, and M. P. Harper. A parallel implementation of a hidden markov model with duration modeling for speech recognition. In *Proceedings of the Fifth IEEE Symposium* on Parallel and Distributed Processing, Dallas, Texas, pages 298–306, 1993.
- Vincent Mouillet. Prise en compte des informations temporelles dans les modles de markov cachs. Rapport de stage, IRCAM, 2001.
- Kevin Murphy. Dynamic Bayesian Networks: Representation, Inference and Learning. PhD thesis, UC Berkeley, Computer Science Division, 2002.
- Nicola Orio and F. Déchelle. Score Following Using Spectral Analysis and Hidden Markov Models. In *Proceedings of the ICMC*, Havana, Cuba, 2001.
- Nicola Orio, Serge Lemouton, Diemo Schwarz, and Norbert Schnell. Score Following: State of the Art and New Developments. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, Montreal, Canada, May 2003.
- L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–285, 1989.
- Haruto Takeda, Takuya Nishimoto, and Shigeki Sagayama. Rythm and tempo recognition of music performance from a probabilistic approach. In *ISMIR*, 2004.
- Xue Wang. Incorporation knowledge on segmental duration in Hmm-based continuous speech recognition. PhD thesis, University of Amsterdam, 1997.