Time-coherency of Bayesian priors on transient semi-Markov chains for audio-to-score alignment

Philippe Cuvillier

Time-coherency of Bayesian priors on transient semi-Markov chains for audio-to-score alignment

Philippe Cuvillier

MuTant project-team – Ircam, Inria, UPMC, CNRS
Ircam – 1, place Igor-Stravinsky – Paris, France

Abstract. This paper proposes a novel insight to the problem of real-time alignment with Bayesian inference. When a prior knowledge about the duration of events is available, Semi-Markov models allow the setting of individual duration distributions but give no clue about their choice. We propose a criterion of temporal coherency for such applications and show it might be obtained with the right choice of estimation method. Theoretical insights are obtained through the study of the prior state probability of transient semi-Markov chains.

Keywords: Bayesian inference, sequential estimation, hidden semi-Markov models, semi-Markov chains

INTRODUCTION

Many signals are structured as time-contiguous *events* which generate specific observations, e.g. music, speech or text. In music, basic events may be notes (pitched sounds) and silences. To recognize the sequence of events that generates an observed signal, probabilistic models [1] are relevant when statistical relationships between observation and events are known. In particular, the Hidden Markov Models (HMM) [2] assume that the signal is stationary on time-intervals and identify them with the occupancy of a hidden state. Once the state-space and the statistical priors are specified, Bayesian inference can be readily computed to recognize the state-sequence.

Score alignment [3] is a Music Information Retrieval (MIR) task consisting of synchronizing a musical performance with its score, i.e. the sequence of notes. Since ordering of events is known, recognition boils down to alignment. Among the numerous applications of HMM, music has an outstanding property: a music score assigns to each event its *nominal duration*, i.e. a prior information on their likely duration.

A crucial and undermined question is about the modeling of nominal duration. This investigation is built on the framework of hidden semi-Markov models (HSMM) as it provides explicit choice of the prior duration model. In section 2 we detail this motivation and briefly introduce HSMM. This generalization of HMM involves many Bayesian priors whose tuning is a major issue. To this aim, most probabilistic models rely on learning with training datasets. This paper presents an alternative based on a theoretical study of prior probability distributions of semi-Markov processes.

In section 3, we state our condition of time-coherency, and explain how the Viterbi estimation does not fulfill it. In section 4, we investigate how the Forward estimation may fulfill it or not depending on several distribution properties of the Bayesian priors.

BACKGROUND & MOTIVATION

Semi-Markov models for alignment

Hidden semi-Markov models were introduced in [4] as a generalization of HMM. Both are defined with two stochastic processes [2]. The process $(S_t)_{t\in\mathbb{N}^*}$ is a discrete-time homogeneous Markov chain on a discrete state-space $E \overset{\text{def}}{=} \{1,2,\ldots,J\}$, finite or not ($J=\infty$). Since its realizations $(s_t)_{t\in\mathbb{N}^*}$ are not known, they are called hidden states. The observation $(o_t)_{t\in\mathbb{N}^*}$, e.g. the audio signal, is considered as a realization of the second process $(O_t)_{t\in\mathbb{N}^*}$. We denote $\mathbb{N} \overset{\text{def}}{=} \{0,1,\ldots\}$, $\mathbb{N}^* \overset{\text{def}}{=} \{1,2,\ldots\}$ and $S_t^{t+u} \overset{\text{def}}{=} (S_t, S_{t+1}, \ldots, S_{t+u})$.

In such probabilistic models, the duration spent on a state j is a time-homogeneous random variable. Its law is called the *occupancy distribution* $d_j(u) \overset{\text{def}}{=} \mathbb{P}(S_{t+u+1} \neq j, S_{t+2}^{t+u} = j \mid S_{t+1} = j, S_t \neq j)$ for $u \in \mathbb{N}^*$. For a Markov state with self-transition p, d_j would implicitly be a geometric law $d_j(u) = (1-p)p^{u-1}$. Assuming that $(S_t)_t$ is a semi-Markov chain allows choosing each Bayesian prior d_j as any probabilistic mass function (pmf) on $\mathbb{N}^*$.

A semi-Markov chain consists of two additional choices per state j: the initial probability $\pi(j) \overset{\text{def}}{=} \mathbb{P}(S_1 = j)$, and the transition probabilities $p_{ij} \overset{\text{def}}{=} \mathbb{P}(S_{t+1} = j \mid S_{t+1} \neq i, S_t = i)$ with $p_{ii} = 0$. In alignment tasks, left-to-right topologies of transition probabilities conveniently model the prior information of ordering. This study exclusively deals with the simplest topology, the *linear semi-Markov chains*: $\forall i,j, \quad p_{ij} = \delta_{i,i+1}$ and $\pi_j = \delta_{1,j}$ (see figure 3 for an example).

Moreover, the hidden model paradigm describes how states (S_t) influence observations (O_t) using *observation probabilities* $b_j(o_t^{t+u}) \overset{\text{def}}{=} \mathbb{P}(O_t^{t+u} = o_t^{t+u} | S_t^{t+u} = j)$.

Modeling prior information of duration with HSMMs

Inference with semi-Markov models requires a careful design of the prior distributions d_j for each state. Usual approaches rely on statistical learning. The Baum-Welch algorithm, i.e. the HMM version of Expectation-Maximization (EM), has been adapted to semi-Markov models [5]. But this non-parametric algorithm requires huge training datasets. Consequently, most implementations prefer a parametric EM [6] to learn the occupancy distributions over a parametric family of probabilities, e.g. Gamma, Poisson, log-normal, Negative Binomial laws.

This study aims at elaborating a criterion so as to justify or disqualify such choices. It is built on an interesting property of our application: **musical events are associated with a reference duration**. Indeed a music score provides the prior tempo and prior durations for all notes.

We denote this quantity the *nominal duration* l_j. Although a few music alignment systems like [7] willingly discard this prior information, this work considers duration as an explicit element of modeling and makes the following assumption: two events

with identical nominal duration should get identical occupancy distributions. So the duration model consists of a *set of durations* $L \subset \mathbb{R}_+$ and a *duration-indexed family* of pmfs $(d_l)_{l \in L}$ such that for all state j, $l_j \in L$ and $d_j = d_{l_j}$. This framework sharpens the problematic: are there coherent mappings from nominal durations l to distributions d_l?

CRITERION OF COHERENCY FOR PRIORS OF DURATION

Hypothesis of non-discriminative observation

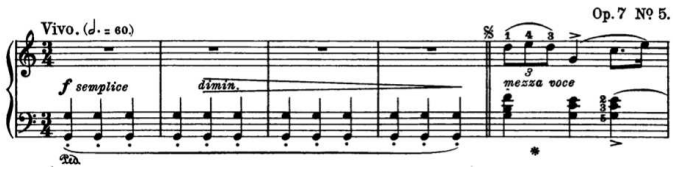

FIGURE 1. Music score of the *Mazurka Op. 7 No. 5* by F. Chopin. It begins with a long sequence of repeated events, i.e. states with identical observation probabilities.

Our definition of time-coherency emerges from the following fact: music scores might be composed of very long sequences of "repeated events" such that the one in figure 1. What would happen if *all* states $j \in E$ share the same observation probabilities? We call *non-discriminative observation* such a model where $b_1 = b_2 = \ldots \overset{\text{def}}{=} b$. Note that this assumption may model other realistic situations of Bayesian inference such as missing observations [8].

Ideal behavior with non-discriminative observation

We state our criterion of time-coherency. Its rationale is simple: if the observation probabilities do not discriminate states, then the inference should respect the states ordering and their nominal durations as these are the only available information.

Time-coherency criterion 1. On a linear chain with non-discriminative observation, the inference successively decodes states $1, 2, 3, \ldots$ at time steps $1, 1+l_1, 1+l_1+l_2, \ldots$ and assigns to each state j a duration which is equal to its nominal duration l_j.

The hypothesis of non-discriminative observations makes the posterior probabilities equal to the prior probabilities:

$$\forall t \in \mathbb{N}^*, \quad \mathbb{P}(S_1, \ldots, S_t \mid O_1, \ldots, O_t) = \mathbb{P}(S_1, \ldots, S_t).$$

Indeed, the Markovian assumption implies that $\mathbb{P}(O_1^t \mid S_1^t) = \prod_{u=1}^t \mathbb{P}(O_u \mid S_u) = \prod_{u=1}^t b_{S_u}(O_u)$. Assuming that $b_{S_u} = b$ gives $\mathbb{P}(O_1^t \mid S_1^t) = \prod_{u=1}^t b(O_u) = \mathbb{P}(O_1^t)$, so (S_t) and (O_t) are independent. Thus, the inferred quantities become independent of the observations. Whether the criterion is fulfilled only depends on the underlying semi-Markov chain $(S_t)_t$ and the estimation method.

Offline alignments estimate the most likely sequence s_1^T at final time T, using the so-called Viterbi algorithm. But online alignments make sequential estimations. At each time $t = 1 \dots T$, they could either estimate the partial sequence s_1^t or the most likely current state $\hat{s}_t$. The *Viterbi alignment* is defined as $\hat{s}_t$ such that $\hat{s}_1^t \overset{\text{def}}{=} \arg\max_{s_1,\dots,s_t \in E^t} \mathbb{P}(S_1^t = s_1^t \mid O_1^t = o_1^t)$. The *Forward alignment* is defined as $\hat{s}_t = \arg\max_{s_t \in E} \mathbb{P}(S_t = s_t \mid O_1^t = o_1^t)$. These quantities are obtained using the recursive equations detailed in [5].

Failure of the Viterbi estimation

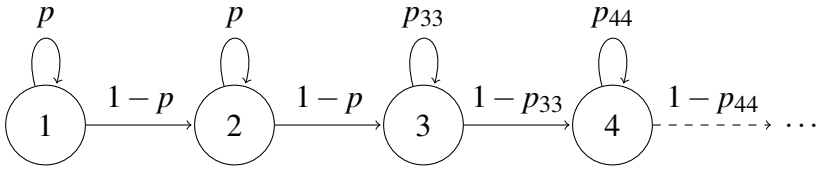

FIGURE 2. Example of a linear Markov chain with identical first two states: $p_{11} = p_{22} = p$.

Our first claim is that the Viterbi alignment fails to be time-coherent. Let us illustrate it with the example in figure 2: a linear Markov chain with identical first two states.

Let $\mathbf{S} = (S_1, \dots, S_{t+1})$ be an admissible (i.e. non-decreasing) path. If $\mathbf{S}$ ends at $S_{t+1} = 1$, then $\mathbb{P}(\mathbf{S}) = p^t$. If $S_{t+1} = 2$, then $\mathbb{P}(\mathbf{S}) = (1-p)p^{t-1}$. So, if $p > 1/2$ then state 1 is more likely than state 2 at all times for the Viterbi estimation, whereas if $p < 1/2$ state 2 is more likely than state 1 at all times $t > 1$.

This simplistic example could be extended to semi-Markov chains for a wide class of occupancy distributions, but it is enough to reveal the lack of coherency of Viterbi alignments: the parameters cannot be tuned to take into account the nominal duration l_1.

COHERENCY OF THE FORWARD ESTIMATION

Our second claim is that the Forward alignment may be time-coherent. This section introduces sufficient conditions on occupancy distributions that imply criterion 1.

Recall that under non-discriminative observation, $F_j(t) = \mathbb{P}(S_t = j)$. Let $\mathbf{F}(t) \overset{\text{def}}{=} (F_1(t), F_2(t), \dots)$ denote the state probability distribution on E, and $\mathbf{M}[\mathbf{F}_t] \overset{\text{def}}{=} \arg\max_{j \in E} F_j(t)$ denote its *mode*.

Criterion 1 has the following translation:

$$\forall t \in \mathbb{N}^*, \qquad \mathbf{M}[\mathbf{F}_t] = j \Leftrightarrow 1 \leq t - (l_1 + \dots + l_j) \leq l_{j+1}$$

On linear chains, the prior state probabilities are given by successive convolutions. Using the recursive equations detailed in [5, Section 3.2], a simple induction over states j proves that

$$F_j(t) = \frac{1}{K(t)} \cdot \begin{cases} d_1 * d_2 * \dots * d_{j-1} * D_j(t) & \text{if } j > 1 \\ D_1(t) & \text{else} \end{cases}$$

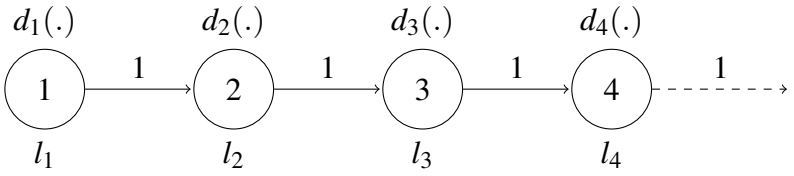

FIGURE 3. Example of a linear semi-Markov chain.

where $D_j(t) = \sum_{u \geq t} d_j(u)$ is the *survivor distribution* associated to d_j, and $K(t) = D_1(t) + d_1 * D_2(t) + d_1 * d_2 * D_3(t) + \ldots$ is an unimportant normalization constant.

The case of first two states

Since state 1 is the most likely one at first time step $t = 1$, we begin with a comparison between states 1 and 2 by studying the evolution of the probability ratio $\frac{F_1(t)}{F_2(t)}$. The Forward estimation respects the criterion 1 for these two states if and only if

$$\forall j \in E, \qquad \frac{F_2(t)}{F_1(t)} \begin{cases} \leq 1 & \text{if } t \leq l_1 \\ > 1 & \text{if } t > l_1 \end{cases}.$$

Proposition 1. Let us denote $\mathbf{m}[d_1] \stackrel{\text{def}}{=} \max\{t \mid D_1(t) \geq 1/2\}$ the median of d_1. Then for any distribution d_2,

$$t \leq \mathbf{m}[d_1] \quad \Rightarrow \quad F_1(t) \geq F_2(t).$$

Reciprocally there exists a distribution d_2 such that $t > \mathbf{m}[d_1] \Rightarrow F_1(t) < F_2(t)$.

Proof. Since $D_2(t) \leq 1$ for all t, $\sum_{u=1}^{t-1} d_1(u) D_2(t-u) \leq \sum_{u=1}^{t-1} d_1(u)$ so $F_2(t) \leq 1 - D_1(t)$ and $F_2(t) - F_1(t) \leq 1 - 2D_1(t)$. Since D_1 is non-increasing, if $t \leq \mathbf{m}[d_1]$ then $D_1(t) \geq 1/2$ and $1 - 2D_1(t) \leq 0$.
For the necessary condition one may consider the trivial distribution $d_2(t) = \delta_{\mathbf{m}[d_1]}(t)$. □

Last proposition tells that the *median* of d_1 is a lower bound for the duration assigned to state 1. Thus it prescribes choosing every distribution d_l such that its median is $\lfloor l \rfloor$. But even if this result provides the first half of the criterion 1, the other half may not be fulfilled in the general case.

Uncoherency of heavy-tailed distributions

First, we give a negative example of distributions that never fulfill the criterion. An important feature of probability distributions is their asymptotic speed of decay. This

feature is related to the *radius of convergence* $R_d \in [1, +\infty]$ of the *probability generating function* of d, denoted $Z[d](z) \stackrel{\text{def}}{=} \sum_{n\in\mathbb{N}} d(n)\, z^n$.

Definition 1. A discrete distribution d is said to be *heavy-tailed* if $R_d = 1$, and *light-tailed* if not.

Convolutions of heavy-tailed distributions have been thoroughly studied. We borrow the following non-trivial result from [9].

Proposition 2. If d_1 is an *heavy-tailed* pmf, then

$$\liminf_{t\to\infty} \frac{\sum_{u=0}^{t-1} d_1(u) D_1(t-u)}{D_1(t)} = 1$$

If the two states have the same heavy-tailed occupancy distribution, then state 1 is decoded an infinite number of time. So, the criterion is never fulfilled. This fact discards using such pmfs as Bayesian priors.

Coherency of IHR distributions

Nevertheless, using light-tailed distributions does not guarantee neither the criterion. But another notion of tail analysis helps checking whether the criterion hold or not.

Definition 2. A distribution d is *Increasing Hazard Rate* (IHR) if its *hazard rate* $h(n) \stackrel{\text{def}}{=} \frac{d(n)}{D(n)}$ is non-decreasing.

Proposition 3. Let d_1 be an IHR pmf. For **all** distribution d_2, $\frac{d_1 * D_2}{D_1}(t)$ is a non-decreasing function of t.

Proof. The proof uses simple algebraic computations. Let us define the functions

$$f_u(t) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } t \geq u \\ \frac{d_1(t-u)}{D_1(t)} D_2(u) & \text{if } 0 < t < u \end{cases}.$$

With this definition we have $\frac{d_1 * D_2}{D_1}(t) = \sum_{u=1}^{t-1} \frac{d_1(t-u)}{D_1(t)} = \sum_{u\in\mathbb{N}^*} f_u(t)$.

Let h be the hazard rate of d_1, and u be in $\mathbb{N}^*$. By definition of h, $\frac{d_1(t-u)}{D_1(t)} = \frac{d_1(t-u)}{D_1(t-u)} \frac{D_1(t-u)}{D_1(t)} = h(t-u) \frac{1}{\prod_{v=1}^{u}(1-h(t-v))}$. Since the function $x \mapsto \frac{1}{1-x}$ is positive and increasing on $[0,1[$, if h is non-decreasing, then $t \mapsto f_u(t)$ is non-decreasing as a product of positive and non-decreasing functions. Consequently, $t \mapsto \frac{d_1 * D_2}{D_1}(t)$ is non-decreasing. □

Monotony is a stronger but very interesting requirement: if the ratio is non-decreasing then the estimation never come back to state 1 after having decode state 2.

Extension to N states

The previous arguments cannot be directly generalized to more than two states without further assumptions. Our next argument extends the idea of monotony that proposition 3 highlights. This approach turns out to be related to the notions of stochastic orderings introduced by [10].

Definition 3. Let p_1, p_2 be two distributions. p_1 is said to be *locally smaller* that p_2, denoted $p_1 \underset{\text{lr}}{\leq} p_2$, if $n \mapsto p_1(n)/p_2(n)$ is non-decreasing on $\text{supp}[p_2]$.

A family $(p_t)_{t\in I}$ of pmfs indexed by $I \subset \mathbb{R}$ is said to be *locally increasing* if $\forall t_1, t_2 \in I, \quad t_1 \leq t_2 \Rightarrow p_{t_1} \underset{\text{lr}}{\leq} p_{t_2}$.

Lemma 1. If $(p_t)_{t\in I}$ is an increasing family, then the mode $\mathbf{M}[p_t]$ of p_t is a non-decreasing function of t.

This straightforward lemma is interesting: if the state probabilities of the semi-Markov chain $(\mathbf{F}(t))_{t\in\mathbb{N}^*}$ constitute an increasing family, then states are decoded with respect to their ordering – although some states might be skipped. Moreover, checking numerically the criterion on a given chain becomes very easy: it holds if and only if $\frac{F_{j+1}}{F_j}(l_1 + \ldots + l_j) \leq 1$ and $\frac{F_{j+1}}{F_j}(1 + l_1 + \ldots + l_j) > 1$ for all j. So, to finish with, next proposition gives a sufficient condition to obtain an increasing process.

Definition 4. A discrete distribution d is *log-concave* if for all n in $\mathbb{N}$, $d(n)^2 \geq d(n-1)d(n+1)$. This is equivalent to $d(.) \underset{\text{lr}}{\leq} d(.+u)$ for all $u \in \mathbb{N}$.

It is noteworthy that all log-concave distributions are IHR. The main point is that log-concavity "preserve" stochastic ordering, as the following lemma explains.

Lemma 2 ([10, Theorem 2.1])**.** Let f, g, h be three distributions. If f is log-concave, then $g \underset{\text{lr}}{\leqslant} h \quad \Rightarrow \quad f * g \underset{\text{lr}}{\leqslant} f * h$.

See the reference for its proof, that is an application of the Binet-Cauchy formula.

Proposition 4. If the semi-Markov chain is linear and all occupancy distributions d_j are log-concave, then the process $(\mathbf{F}(t))_{t\in\mathbb{N}^*}$ is locally increasing.

Proof. Let j be in $\mathbb{N}^*$. Since d_{j-1} is log-concave, it is also IHR and $D_{j-1} \underset{\text{lr}}{\leqslant} d_{j-1} * D_j$.

If $j > 1$, let us consider $\frac{F_{j+1}}{F_j} = \frac{d_1 * \ldots * d_{j-1} * d_j * D_{j+1}}{d_1 * \ldots * d_{j-1} * D_j}$ The class of log-concave distributions is stable by convolution [10], so $d_1 * \ldots * d_{j-1}$ is log-concave. Then, lemma 2 implies that $t \mapsto \frac{F_{j+1}(t)}{F_j(t)}$ is increasing. □

The monotony of Markov processes has been largely studied – see [11] for a survey. Proposition 4 is a first step towards its extension to semi-Markov processes. Moreover, in locally increasing Markov chains, first-time passages $T_j \overset{\text{def}}{=} \inf\{t \mid X_{t+1} \geq j, X_1 = 1\}$ have log-concave pmfs [12]. Proposition 4 looks like a "reverse" counterpart of this result for linear semi-Markov chains, since the pmf of T_{j+1} is $d_1 * \ldots * d_j$ for such chains.

As a conclusion, log-concavity seems to be the most desirable property for prior distributions of duration. While log-concavity plays an important role in many fields of statistics, it has been scarcely studied on HSMMs. It is highlighted by [13] for improving computational efficiency of the Viterbi algorithm. Proposition 4 shows it also provides theoretical coherency to the Forward estimation. Furthermore, experiments show that taking into account these prescriptions do improve the performances of real-time alignment. A comparative test with results and video files can be found on `http://repmus.ircam.fr/mutant/mlsp14`.

CONCLUSION & PERSPECTIVES

This paper introduces a criterion of time-coherent modeling in semi-Markov models for alignment. This criterion is about estimation coherency under non-discriminative observation. We show that coherency cannot be obtained with Viterbi estimation but can be obtained with is the Forward estimation if the chosen probability distributions have some precise properties. This short study calls for further theoretical and experimental developments. More necessary and sufficient conditions related to the criterion can be derived. The framework can be extended to other estimators such as the Forward-backward algorithms. Moreover the proposed prescriptions lead to constraints on the learning parameter space; adding these constraints in HSMM training algorithms would be an interesting issue.

REFERENCES

1. K. P. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. thesis, UC Berkeley, Computer Science Division (2002).
2. L. R. Rabiner, *Proc. of the IEEE* **77**, 257–286 (1989).
3. A. Cont, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **32**, 974–987 (2010).
4. S. E. Levinson, *Comput. Speech Lang.* **1**, 29–45 (1986), ISSN 0885-2308.
5. Y. Guédon, *Journal of Computational and Graphical Statistics* **12**, 604–639 (2003), URL `http://hal.inria.fr/hal-00826992`.
6. C. D. Mitchell, and L. H. Jamieson, "Modeling duration in a hidden Markov model with the exponential family," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, 1993, vol. 2, pp. 331–334 vol.2, ISSN 1520-6149.
7. C. Joder, S. Essid, and G. Richard, "An Improved Hierarchical Approach for Music-to-symbolic Score Alignment," in *ISMIR*, edited by J. S. Downie, and R. C. Veltkamp, International Society for Music Information Retrieval, 2010, pp. 39–45, ISBN 978-90-393-53813.
8. S.-Z. Yu, and H. Kobayashi, *Signal Process.* **83**, 235–250 (2003), ISSN 0165-1684, URL `http://dx.doi.org/10.1016/S0165-1684(02)00378-X`.
9. S. Foss, and D. Korshunov, *The Annals of Probability* **35**, 366–383 (2007), URL `http://dx.doi.org/10.1214/009117906000000647`.
10. J. Keilson, and U. Sumita, *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **10**, pp. 181–198 (1982), ISSN 03195724.
11. M. Kijima, *Journal of Applied Probability* **35**, pp. 545–556 (1998), ISSN 00219002, URL `http://www.jstor.org/stable/3215630`.
12. S. Karlin, Total positivity (1968), URL `http://opac.inria.fr/record=b1089884`.
13. D. Tweed, R. Fisher, J. Bins, and T. List, "Efficient Hidden Semi-Markov Model Inference for Structured Video Sequences," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, 2005, pp. 247–254.