Coherent Time Modeling of semi-Markov Models with Application to Real-Time Audio-to-Score Alignment

Philippe Cuvillier, Arshia Cont

2014 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2014, REIMS, FRANCE

COHERENT TIME MODELING OF SEMI-MARKOV MODELS WITH APPLICATION TO REAL-TIME AUDIO-TO-SCORE ALIGNMENT

Philippe Cuvillier, Arshia Cont

MuTant Project-team – UPMC, Inria, Ircam & CNRS
1, place Igor-Stravinsky, Paris 75004, France

ABSTRACT

This paper proposes a novel insight to the problem of duration modeling for recognition setups where events are inferred from time-signals using a probabilistic framework. When a prior knowledge about the duration of events is available, Hidden Markov or Semi-Markov models allow the setting of individual duration distributions but give no clue about their choice. We propose two criteria of temporal coherency for such applications and prove they are fulfilled by statistical properties like infinite divisibility and log-concavity. We conclude by showing practical consequences of these properties in a real-time audio-to-score alignment experiment.

Index Terms— Hidden Markov model, semi-Markov chains, alignment, score following

1. INTRODUCTION

Many natural phenomena exhibit a latent temporal structure and may be observed through a temporal signal, e.g. music, speech or text. They are often structured as time-contiguous *events* which generate specific observations. In music, basic events may be notes - pitched sounds - and silences.

To recognize the sequence of events that generates the observed signal, probabilistic models [1] are relevant when statistical relationships between observation, events and any additional variables 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. HMMs represent possible states as vertices of a graph whose transitions are possible evolutions of the state. Once this state-space is specified, the 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 symbolic score. Since ordering of events is known, recognition boils down to alignment. Among the numerous applications of HMM, music has an outstanding characteristics: 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 the nominal duration. This investigation is built on the framework of hidden semi-Markov models (HSMM) as it provides explicit choice of the duration model. In section 2 we 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 [2, 4, 5, 6]. This paper presents an alternative based on a theoretical study of the inference equations.

In section 3, we state our first coherency criterion with aggregates and show how it leads to the property of infinite divisibility [7]. In section 4, we state our second criterion, leading to specific statistical properties like log-concavity [8]. Afterwards, we compare our prescriptions with the choices of duration models we have found in the MIR literature. In section 5, we make a comparative test to illustrate the practical benefits of taking into account these theoretical properties.

2. BACKGROUND & MOTIVATION

2.1. Background: semi-Markov models for alignment

Hidden semi-Markov models were introduced in [9] as a generalization of HMM that offer explicit duration probability densities. 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 finite state-space $E \stackrel{\text{def}}{=} \{1, 2, \ldots, J\}$. Since its realizations $(s_t)_{t\in\mathbb{N}^*}$ are not observable, they are called hidden states. The observation $(o_t)_{t\in\mathbb{N}^*}$ is considered as a realization of the second process $(O_t)_{t\in\mathbb{N}^*}$. We denote $\mathbb{N} \stackrel{\text{def}}{=} \{0, 1, \ldots\}$ and $\mathbb{N}^* \stackrel{\text{def}}{=} \{1, 2, \ldots\}$.

In probabilistic models the duration spent on a state j is a random variable L_j. Its law is called the *occupancy distribution* $d_j(u) \stackrel{\text{def}}{=} \mathbb{P}(S_{t+u+1} \neq j, S_{t+2}^{t+u} = j \mid S_{t+1} = j, S_t \neq j)$. It is only defined for $u \in \mathbb{N}^*$ so it must check $d_j(0) = 0$. The explicit choice of d_j sets HSMM apart from regular HMM modeling. For a Markovian state with self-transition p, d_j implicitly follows a geometric law $d_j(u) = (1-p)p^{u-1}$.

This work was partially funded by the French National Research Agency ANR-11-JS03-005-01.

978-1-4799-3694-6/14/$31.00

Any semi-Markov chain consists of two additional choices per state j: the initial probability $\pi(j) \stackrel{\text{def}}{=} \mathbb{P}(S_1 = j)$, and the transition probabilities $p_{ij} \stackrel{\text{def}}{=} \mathbb{P}(S_{t+1} = j \mid S_{t+1} \neq i, S_t = i)$ with $p_{ii} = 0$ for semi-Markov states.

Hidden model paradigm describes how states (S_t) influence observations (O_t) using *observation probabilities* $b_j(o_t^{t+u}) \stackrel{\text{def}}{=} \mathbb{P}(O_t^{t+u} = o_t^{t+u} | S_t^{t+u} = j)$. We make the usual Markovian assumption of conditional independence: $\forall t, u, \quad b_j(o_t^{t+u}) = \prod_{v=0}^{u} b_j(o_{t+v})$.

Markov models were originally designed for recognition. In alignment tasks the prior information of ordering is conveniently modeled by left-to-right topologies of transition probabilities p_{ij}. We exclusively deals with the simplest topology, the *linear semi-Markov chains*: $\forall i, j, \quad p_{ij} = \delta_{i,i+1}$.

2.2. Right-censored Forward inference for real-time estimation

Offline inference uses the Viterbi algorithm [2] at final time T to decode the most likely state sequence $\hat{s}_1^T$. But for real-time alignment several estimators compete. Systems like [6, 10, 11] prefer using the Forward algorithm at each time t to estimate the most likely current state $\hat{s}_t$. We follow them and exclusively study this inference mode.

The *right-censored* Forward algorithm for semi-Markov states as proposed in [12] provides suitable inference mechanism for real-time applications. It suggests computing the probabilities of the events $\{S_t = j\}$ instead of the events $\{S_t = j, S_{t-1} \neq j\}$. This detail turns out to be crucial for real-time applications since at the current time t one cannot know if the current event S_t is over yet. The algorithm introduces the quantities $F_j(t) \stackrel{\text{def}}{=} \mathbb{P}(S_t = j \mid O_1^t = o_1^t)$, $F_j^{\text{o}}(t) \stackrel{\text{def}}{=} \mathbb{P}(S_{t+1} \neq j, S_t = j \mid O_1^t = o_1^t)$ and $F_j^{\text{i}}(t) \stackrel{\text{def}}{=} \mathbb{P}(S_{t+1} = j, S_t \neq j \mid O_1^t = o_1^t)$. To compute them, a key feature is the *survivor distribution* $D_j(u) \stackrel{\text{def}}{=} \mathbb{P}(S_{t+2}^{t+u} = j \mid S_{t+1} = j, S_t \neq j)$. It is induced by d_j owing to the relationship $D_j(u) = \sum_{v \geq u} d_j(v)$. The Forward inference consists of the following recursions over time t

$$\begin{aligned} F_j(t) &= \sum_{u=1}^{t} b_j(o_{t-u+1}^t) F_j^{\text{i}}(t-u) D_j(u) \\ F_j^{\text{o}}(t) &= \sum_{u=1}^{t} b_j(o_{t-u+1}^t) F_j^{\text{i}}(t-u) d_j(u) \end{aligned} \tag{1}$$

while $F_j^{\text{i}}(0) = \pi(j)$ and $F_j^{\text{i}}(t) = \sum_{i \neq j} p_{ij} F_i^{\text{o}}(t)$. The estimated current state is implemented as $\hat{s}_t \stackrel{\text{def}}{=} \arg\max_j F_j(t)$.

2.3. Motivation: modeling prior information of duration with HSMMs

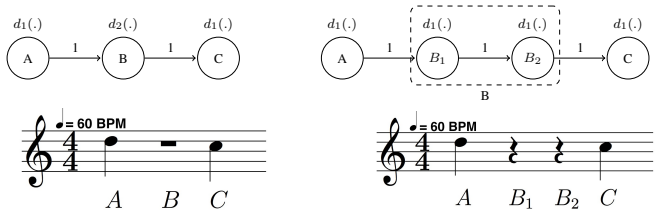
Using semi-Markov models to decode event sequences requires a careful design of the Bayesian prior d_j for each state. Most systems tune these with statistical learning using solutions like the HSMM version of the Baum-Welch algorithm [12]. But estimating a distribution d_j per state j would require huge training datasets. Consequently most systems learn the occupancy distributions with regularization techniques over parametric classes of pdfs [4]. This has led to ad hoc engineering choices to reduce the parameter space. This paper focuses on this tradition, and studies their mathematical behavior, application consequences to get insights on these choices.

Our study starts from an outstanding property of our MIR 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 [13] willingly discard this prior information, this work considers durations as an explicit element of modeling.

Accounting for nominal durations requires the following strong assumption: **two events with identical nominal durations 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 pdfs $(d_l)_{l \in L}$ such that for all state j, $l_j \in L$ and $d_j = d_{l_j}$. This framework sharpens the problematic: are there relevant choices of the mapping from nominal durations l to distributions d_l?

A review of literature shows that many modeling heuristics compete. Some of them are listed in section 4.5. Up to now no investigation has been done to compare such choices. The next two sections develop two criteria so as to justify or disqualify such heuristics.

3. CRITERION 1: COHERENT TIME-AGGREGATES



(a) Three events of duration 1,2,1 (b) Four events of duration 1

Fig. 1: Two equivalent event sequences and their graphical models. The silence B aggregates the short ones (B_1, B_2).

We introduce our first criterion of coherency and show that it leads to the notion of infinite divisibility and convolution semigroups. The motivation comes from aggregates of events. Figure 1 illustrates the concept with two toy sequences. The music score 1b informs of four events with equal nominal duration of 1 sec. In the music score 1a one silence of duration 2 replaces the two ones of duration 1. These two sequences would generate the same observations (O_t) since no physical signal could distinguish consecutive

silences. The "one semi-Markov state per event" strategy would map 1a to the 3 states graphical model (A, B, C) and 1b to the 4 states one (A, B_1, B_2, C), so state B carries the same prior as the aggregate (B_1, B_2). Coherency would ask that these two different graphical models infer the same values for (B_1, B_2) and B respectively.

Coherency criterion 1. An aggregate of N states with nominal duration $l_1, l_2, \ldots$ and identical observation functions $b_1 = b_2 = \ldots = b$ induces the same inference quantities as a single state with duration $l_{1:N} \stackrel{\text{def}}{=} l_1 + \ldots + l_N$ and observation b.

To achieve this criterion, we study duration prior of linear aggregates. The duration spent in state j is a random variable L_j whose law is d_j. Intuitively the duration spent in two consecutive states i.e. the aggregate $(j, j+1)$ is the sum of their individual durations: $L_{(j,j+1)} = L_j + L_{j+1}$. As the Markovian assumption makes these two random variables independent, the law of an aggregate is the convolution of individual laws: $d_{(j,j+1)} = d_j * d_{j+1}$.

Remark: the discrete convolution product $*$ is defined as $[f * g](t) \stackrel{\text{def}}{=} \sum_{u \in \mathbb{Z}} f(u) g(t-u)$.

Actually, whether this intuition is valid depends on the inference mode. Simple algebraic computations prove it does not hold for the Viterbi algorithm whereas it does for the Forward one. Details are given in the following proposition.

Proposition 1 (Consistency with aggregates)**.** For the Forward inference a linear aggregate $(B_1, B_2, \ldots, B_N)$ is equivalent to a single state B if and only if

1. the observation probabilities are identical: $b_B = b_{B_1} = b_{B_2} = \ldots$ and check the Markovian assumption - see section 2.1,
2. the initial probabilities check $\pi(B_i) = \pi(B)\, \delta_{1,i}$,
3. the occupancy distributions check

$$d_B = d_{B_1} * d_{B_2} * \ldots * d_{B_N}.$$

This proposition proves that the family $(d_l)_{l \in L}$ respects criterion 1 if

$$\forall l_1, l_2 \in L, \qquad l_1 + l_2 \in L \tag{2}$$

$$d_{l_1 + l_2} = d_{l_1} * d_{l_2} \tag{3}$$

It turns out that equation (2) is the exact definition of L being an additive subsemigroup of $\mathbb{R}_+$ while equation (3) is the exact definition of $(d_l)_{l \in L}$ being a *convolution semigroup* [7]. Now we study the question: do such families exist? The answer depends on the structure of L.

Case A. One could say that $L = \{l_0, 2l_0, 3l_0, \ldots\}$ for some $l_0 > 0$. This base duration l_0 is called the *tatum*, or time quantum in the MIR literature [14]. To build the associated semigroup, one can choose d_{l_0} as any valid pdf then compute its successive convolution powers.

Case B. One could say that L contains the subdivisions $l_0, l_0/2, l_0/3, \ldots$ of some l_0. Indeed music scores are written with rational subdivisions of a base duration and may virtually contain *all* rational values. This implies that

$$\forall n \in \mathbb{N}^*, \quad d_{l_0} = \underbrace{d_{l_0/n} * d_{l_0/n} * \ldots * d_{l_0/n}}_{n \text{ times}}.$$

This latter property is the exact definition of d_{l_0} being an *infinitely divisible distribution* [7]. As this reference explains, such a distribution induces a convolution semigroup $(d_l)_{l>0}$ on the "full" set $L = \mathbb{R}_+$. The discrete class of such distributions are known as *compound Poisson distributions*. Unfortunately none of them check the requirement $d(0) = 0$ to define a valid occupancy distributions so no valid family respects the exact condition (3). In practice we prescribe choosing a true convolution semigroup $(d_l)_{l>0}$ and truncating the values at 0. This approximation of (3) is all the better as durations l are long since for all convolution semigroups, $d_l(0) = (d_1(0))^l \underset{l \to \infty}{\to} 0$.
Moreover, this problem of existence is due to time discretization. It could be solved by considering continuous-time HSMMs [9] but these models bring about implementation difficulties.

4. CRITERION 2: COHERENCY UNDER NON-DISCRIMINATIVE OBSERVATION

4.1. Non-discriminative observation

In this section we derive a second criterion by extrapolating the situation of "repeated events" that motivates previous section. What would happen if *all* events share the same observation probabilities? We call *non-discriminative observation* a model where $b_1 = b_2 = \ldots$ or equivalently $b_j \equiv 1$. This assumption may hold in other realistic situations of Bayesian inference such as missing observations [15] and outliers. It also approximates observation probabilities b_j being "too close" because of noisy training data for example. Consequently, we argue that the behavior of inference under non-discriminative observation assesses its robustness.

4.2. Inference in non-discriminative observation

We introduce our second criterion and show that it leads to new prescriptions inspired by precise notions of statistics. If the observation probabilities do not discriminate states, then the inference should respect the states ordering and their nominal duration as these are the only available information.

Coherency criterion 2. On a linear graphical model with non-discriminative observation, the inference successively decodes states $1, 2, \ldots$ and assigns to each state j a duration which is equal to its nominal duration l_j.

This criterion is not validated by most existing systems and is worth a theoretical investigation. The Forward estimation respects criterion if and only if

$$\forall j \in E, \qquad \frac{F_{j+1}(t)}{F_j(t)} \begin{cases} \leq 1 & \text{if } t < l_1 + l_2 + \ldots + l_j \\ > 1 & \text{else} \end{cases}.$$

The assumption of non-discriminative observation leads to successive convolutions over Forward equations (1). A simple induction over states j proves that

$$F_j^{\mathrm{o}} = d_1 * d_2 * \ldots * d_j$$
$$F_j = \begin{cases} d_1 * d_2 * \ldots * d_{j-1} * D_j & \text{if } j > 1 \\ D_1 & \text{else} \end{cases}$$

Now we are able to compare the Forward quantities between successive states $F_j(t), F_{j+1}(t)$. Since state 1 is the most likely one at first time step $t=1$, we begin with a comparison between states 1 and 2.

4.3. The case of two-state chains

To achieve criterion 2, we successively present partial results on the evolution of the following quantity: $\frac{F_2(t)}{F_1(t)} = \frac{[d_1 * D_2](t)}{D_1(t)}$. To begin with, the *median* gives a universal lower bound for the duration assigned to state 1 and this bound is tight.

Proposition 2. 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 distributions d_2 such that

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

Proof. The sufficient condition is straightforward. 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 $d_2(t) = \delta_{\mathbf{m}[d_1]}(t)$. □

Proposition 2 provides half of criterion 2: state 1 is more likely than state 2 while $t < \mathbf{m}[d_1]$. But will state 2 be more likely afterwards? In the general case the answer is negative: there are distributions d_2 such that $d_1 * D_2(t) < D_1(t)$ for all time t. Fortunately, we have found out that standard properties of statistics are relevant for this question.

Definition 1. 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.

Next proposition reveals the interest of IHR for our problem. Its proof is straightforward and is also a consequence of [16, Corollary 2.3].

Proposition 3. If d_1 is IHR then $\frac{d_1 * D_2(t)}{D_1(t)}$ is an increasing function of t for any distribution d_2.

When this proposition holds, then by the monotone convergence theorem the limit $f_{1,2} \stackrel{\text{def}}{=} \lim_{t\to\infty} \frac{F_2(t)}{F_1(t)} \in [0, +\infty]$ exists. Two cases appear: If $f_{1,2} \leq 1$ then state 2 is never decoded and the criterion 2 is not fulfilled. Else there exists a unique time $\mathbf{t}_1$ such that state 1 is more likely before and state 2 likewise after $\mathbf{t}_1$.

So now criterion 2 is almost reached. The remaining condition $\mathbf{t}_1 = l_1$ can be numerically checked on the chosen couple (d_1, d_2). We left a deeper investigation and only make two assertions. Firstly, $f_{1,2} = +\infty$ holds if d_1 has equal or faster tail decay than d_2. Secondly, typical values of $\mathbf{t}_1$ are located on the median $\mathbf{m}[d_1]$ or nearby for many couples (d_1, d_2). In conclusion, we prescribe to calibrate the median of all distributions d_l on l.

4.4. Generalization to N-state chains

The previous arguments cannot be generalized to more than two states without further assumptions. We present one sufficient condition which is informative although restrictive.

Definition 2. A discrete distribution d is *log-concave* if for all n in $\mathbb{N}^*$, $\quad d(n)^2 \geq d(n-1)\, d(n+1)$.

It is noteworthy that all log-concave distributions are IHR.

Proposition 4. If all states share the same occupancy distribution, i.e. $d_1 = d_2 = \ldots = d$ and if d is log-concave then all states are decoded in their ordering.

Proof. Again we only sketch the proof. Firstly, for all j the quantities $\frac{F_{j+1}(t)}{F_j(t)}$ are non-decreasing. The proof is similar to proposition 3. Then these quantities all diverge to $+\infty$ so state $j+1$ becomes more likely than state j after some time. Finally, they check the following property: $F_{j+1}(t) < F_j(t)$ implying $F_{j+2}(t) < F_{j+1}(t)$. This ensures that each state j is decoded before state $j+1$. □

Log-concavity plays an important role in many fields of statistics but has been scarcely studied on HSMMs. It has been highlighted in [17] for improving computational efficiency of the Viterbi inference. Proposition 4 also prescribes this property for all d_l as it provides theoretical coherency to the Forward inference.

4.5. Comparaison with pre-existing heuristics

Previous sections lead to prescriptions on the family $(d_l)_{l\in L}$. We confront them with a list of duration models from the MIR literature. Refer to [8] for properties of the usual probability laws.

- [18] chooses geometric laws with mean fitted on l: $d_l \sim \mathcal{G}(1 - 1/l)$. The family $(d_l)_{l>1}$ respect criterion 2 but not criterion 1 as it is not a convolution semigroup;

- [3] chooses exponential laws with mean fitted on l: $d_l \sim \mathcal{E}(1/l)$. The family $(d_l)_{l>0}$ respect no criteria. It is not a convolution semigroup and the median of d_l is not l but $(l \ln 2)$.

- [19] chooses log-normal distributions with constant shape parameter $\sigma > 0$ and log-scale fitted on l: $d_l \sim \ln\mathcal{N}(\ln l, \sigma)$. This family is not a convolution semigroup and no d_l is IHR.

- [13] uses normal distributions with mean equal to l and standard deviation proportional to l: $d_l \sim \mathcal{N}(l, l^2\sigma^2)$ for some $\sigma > 0$. [20] makes the same choice but sets the *variance* proportionally to l: $d_l \sim \mathcal{N}(l, l\sigma^2)$. Only the latter choice gives a convolution semigroup and it respects all prescriptions.

- [11] uses negative binomial distributions with mean fitted on l: $d_l \sim NB(l(1-p)/p, p)$ for some $p \in (0, 1)$. This defines a convolution semigroup that do respect our prescriptions for long enough l. This might explain why the authors does observe their inference working well with repeated events.

Alternative approaches like [21, 22] define the state-space E on the continuous real line and events as contiguous intervals. Inference is implemented by particle filtering methods. These models turn out to respect criterion 1. Indeed their occupancy distributions are implicitly defined as *first-passage times* of a diffusion process and such variables are always infinitely divisible [7, Section 7].

5. RESULTS & EXPERIMENTS

Accounting results of section 3 and 4, we recommend the following choice for duration model: the family of Poisson laws $(d_l)_{l>0} \sim (Po(l))_{l>0}$. This choice is the simplest compound Poisson semigroup and is optimal for all criteria exposed previously. Indeed, [23, Theorem 2] has proven that it is the only semigroup whose distributions are log-concave for all $l > 0$. Moreover the median of d_l is close to l for all l.

To illustrate the benefits of our proposal, we expose comparative results of real-time audio-to-score alignment on an example that challenges state-of-the-art real-time systems. The example is a 1939 piano performance of Chopin's *Mazurka Op. 17 No. 2* by A. Rubinstein from the *Mazurka Project* dataset [24]. For this experiment, we use exactly the same HSMM system proposed in [3] (the *Antescofo* system which constitutes both practical and scientific state-of-the-art in this application domain). The default setting for d_l consists of exponential laws and does not respect our prescriptions as discussed in section 4.5. In our proposal we set d_l to Poisson laws but do not change anything else. Figure 2 compares the result of the alignment of this audio to its music score

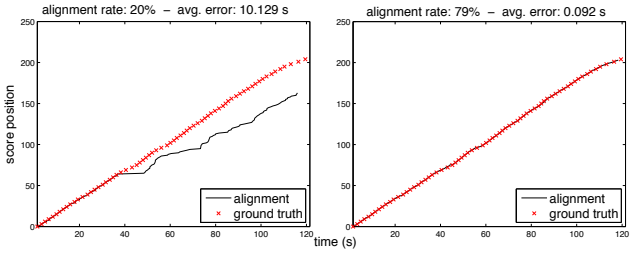


Fig. 2: Comparison of alignments. Left: d_l are exponential laws. Right: d_l are Poisson laws. Reference events are music beats. For each beat the error is the absolute time lapse between the estimate and true time of the beat. Alignment rate is the percentage of beats whose error is below 100 ms.

with these two settings[1]. The default system gets lost after some seconds whereas the proposal manages the entire piece without losing track and with good precision. Similar enhancements have been observed on the whole mazurkas dataset but we leave deeper quantitative evaluation for further publications.

From this experiment, we conclude that the coherency of the duration model compensates the weakness of the observation model. Indeed usual performances of mazurkas make a deep use of the piano pedal that adds some pitches of a chord to the consecutive ones. These unexpected pitches "blur" probabilities between neighboring states. As a result the observation loses its the discriminative power and the inference rely more deeply on its duration model.

6. CONCLUSION & PERSPECTIVES

This papers introduces two criteria of time-coherent modeling in semi-Markov models for alignment. First condition is inference consistency with aggregates and subdvisions; second condition is estimation coherency under non-discriminative observation. We show that coherency is theoretically guaranteed if the chosen probability distributions has specific statistical properties and that respecting these prescriptions experimentally improves real-time alignments.

This short study calls for further theoretical and experimental developments. More necessary and sufficient conditions related to the criteria can be derived. The framework can be extended to other estimators such as Viterbi and 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. Finally, compound Poisson distributions have been extensively studied in statistics and econometrics. The consequences of their numerous properties for Bayesian inference would be worth deeper investigations.

[1] Files, video of the experiments and more details on the results are accessible on `http://repmus.ircam.fr/mutant/mlsp14`.

7. REFERENCES

[1] Kevin P. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. thesis, UC Berkeley, Computer Science Division, July 2002.

[2] Lawrence R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[3] Arshia Cont, "A coupled duration-focused architecture for real-time music-to-score alignment," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 974–987, June 2010.

[4] Carl D. Mitchell and Leah 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.

[5] Christopher Raphael, "Automatic segmentation of acoustic musical signals using hidden Markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 360–37–0, 1999.

[6] Nicola Orio and Francois Déchelle, "Score following using spectral analysis and hidden Markov models," in *Proc. of the International Computer Music Conference (ICMC)*, Habana, Sept. 2001.

[7] Fred W. Steutel and K.V. Harn, *Infinite Divisibility of Probability Distributions on the Real Line*, Chapman & Hall/CRC Pure and Applied Math. Taylor & Francis, 2003.

[8] Norman L. Johnson, Samuel Kotz, and Adrienne W. Kemp, *Univariate Discrete Distributions*, Wiley Series in Probability and Statistics. Wiley-Interscience, 2 edition, Feb. 1993.

[9] Stephen E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Comput. Speech Lang.*, vol. 1, no. 1, pp. 29–45, Mar. 1986.

[10] Diemo Schwarz, Nicola Orio, and Norbert Schnell, "Robust polyphonic midi score following with hidden Markov models," in *Proceedings of the ICMC*, Miami, Florida, Nov. 2004.

[11] Nicola Montecchio and Nicola Orio, "A discrete filter bank approach to audio to score matching for polyphonic music," in *Proceedings of ISMIR 2009*, Oct. 2009, pp. 495–500.

[12] Yann Guédon, "Estimating hidden semi-Markov chains from discrete sequences," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 604–639, 2003.

[13] Cyril Joder, Slim Essid, and Gaël Richard, "An improved hierarchical approach for music-to-symbolic score alignment," in *ISMIR*, J. Stephen Downie and Remco C. Veltkamp, Eds., 2010, pp. 39–45.

[14] Anssi Klapuri, Antti J. Eronen, and Jaakko Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.

[15] Shun-Zheng Yu and Hisashi Kobayashi, "A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking," *Signal Process.*, vol. 83, no. 2, pp. 235–250, Feb. 2003.

[16] James Lynch, Gillian Mimmack, and Frank Proschan, "Uniform stochastic orderings and total positivity," *Canadian Journal of Statistics*, vol. 15, no. 1, pp. 63–69, 1987.

[17] David Tweed, Rober Fisher, José Bins, and Thor 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.

[18] Tomohiko Nakamura, Eita Nakamura, and Shigeki Sagayama, "Acoustic score following to musical performance with errors ad arbitrary repeats and skips for automatic accompaniment," in *Proceedings of Sound and Music Computing*, 2013.

[19] Haruto Takeda, Takuya Nishimoto, and Shigeki Sagayama, "Rhythm and tempo analysis toward automatic music transcription," in *ICASSP (4)*. 2007, pp. 1317–1320, IEEE.

[20] Christopher Raphael, "Aligning music audio with symbolic scores using a hybrid graphical model," *Machine Learning*, vol. 65, no. 2-3, pp. 389–409, Dec. 2006.

[21] Zhiyao Duan and Bryan Pardo, "Soundprism: An online system for score-informed source separation of music audio," *J. Sel. Topics Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[22] Nicola Montecchio and Arshia Cont, "A unified approach to real time audio-to-score and audio-to-audio alignment using sequential Monte Carlo inference techniques," in *ICASSP*, Prague, Czech Republic, May 2011, pp. 193–196, IEEE.

[23] Toshiro Watanabe, "On the strong unimodality of Lévy processes," *Nagoya Mathematical Journal*, vol. 121, pp. 195–199, 1991.

[24] Craig Sapp, "The Mazurka Project," `http://www.mazurka.org.uk`, 2010.