Suivi de partition : étude du cadre multi-objets pour l'inférence de position

par

Philippe Cuvillier

Directeur de stage :	Arshia Cont
Période :	07 mai – 22 octobre 2012
Organisme :	Unité Mixte de Recherche Ircam-CNRS-UPMC
Lieu :	Équipe Représentations Musicales, Ircam — Centre Pompidou 1, place Igor-Stravinsky, 75004 Paris, France.

Table des matières

Introduction

Le suivi de partition est un problème d'informatique musicale qui consiste à synchroniser une exécution musicale avec la partition du morceau joué. Parmi les nombreuses approches proposées depuis 1984, les plus performantes sont les méthodes probabilistes, et plus particulièrement celles d'inférence bayésienne sur modèles génératifs. Elles consistent à proposer deux modèles, l'un expliquant les évolutions possibles de l'état caché, l'autre expliquant la génération des observations par ces états, afin d'inférer la probabilité qu'à la cible d'occuper chaque état. À cette fin, de nombreuses approches emploie l'hypothèse d'une évolution markovienne ainsi qu'un espace d'état où la position est discrétisée, tel qu'un un modèle graphique gauche-droite [17], une chaîne de Markov cachée [14] ou une chaîne semi-Markov [3]. Toutefois, tous rencontrent des difficultés en situation polyphonique, car ils supposent une synchronie rigoureuse entre voix musicales. En pratique, cette synchronie est soit intermittente, soit illusoire comme dans le cas de polyrythmies complexes.

De ce constat est né l'intuition qu'un cadre multi-objets pourrait étendre le suivi aux signaux musicaux asynchrones. Le *multi-target tracking* [2] est un sujet de recherche actif dans la communauté de la surveillance visuelle ou radar, contexte où non pas une mais plusieurs cibles génèrent simultanément des images dans le champ de vision d'un capteur. Inférer les cibles à partir des images par un filtrage bayésien requiert une stratégie de *data association* qu'un cadre mathématique développe avec succès, les *Random Finite Sets* (RFS) [10]. Aussi, ce stage est parti à la découverte de la littérature du suivi RFS afin d'étudier la faisabilité d'une formulation multi-objet du suivi de partition. De là, deux conclusions ont structuré notre approche. En ce qui concerne le modèle d'évolution, ce cadre incite à l'inférence approchée par simulations stochastiques ; nous nous sommes donc aguerris à ces méthodes en reprenant les travaux de N. Montecchio [12, 13], les premiers à proposer le filtrage particulaire pour le suivi de position. En ce qui concerne le modèle d'observation, le cadre RFS n'a de véritable apport que si l'observation du signal est multi-objet ; nous nous sommes donc intéressés au suivi du signal MIDI [18] et à la possibilité de son extension sur des signaux audio par l'emploi d'un estimateur de fréquences fondamentales multiples [4].

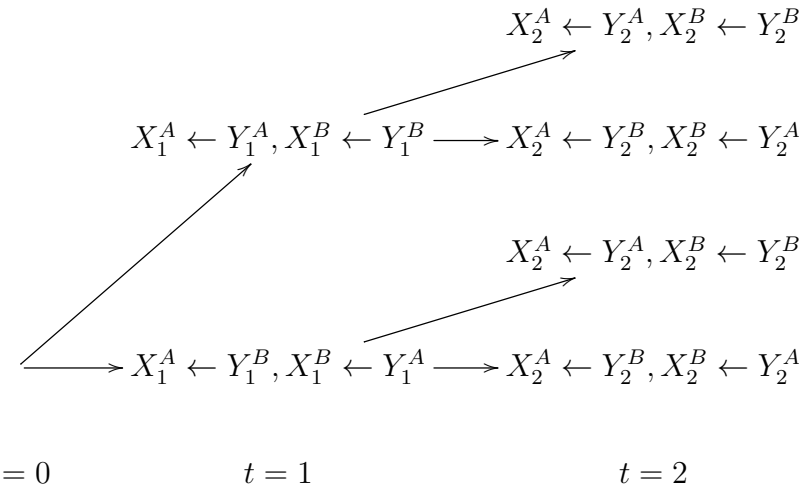
Ainsi, dans notre travail sur le filtrage particulaire, nous prouvons que si leur inférence est efficace, la méthode d'estimation employée souffre de biais inévitables à cause de l'indifférenciation des hypothèses qu'elle induit, puis nous montrons comment au contraire une gestion plus fine hypothèses étend l'utilisation du suivi sur des performances dont le point de départ est inconnu. Après avoir défini l'asynchronie, nous élaborons un modèle RFS d'observation MIDI adapté à la gestion de celle-ci. Puis continuant notre recherche de modélisation RFS, nous montrons que contrairement à l'idée de départ, au suivi multi-voix musicales et qu'ils ne peuvent pas étendre le suivi de position classique. Nous montrons que leurs hypothèses les rend adéquat à un suivi multi-notes pour une classe particulière d'instruments, ce qui nous permet d'esquisser un modèle de suivi simple dans le cas du MIDI mais adaptable à une future utilisation d'estimation multi-F0.

1 Présentation du suivi multi-objet

1.1 Cadre mathématique pour le multi-objet

1.1.1 Présentation des ensemble finis aléatoires RFS

Introduction au paradigme RFS Le problème de l'estimation multi-objets est avant tout le phénomène d'explosion combinatoire. Dans un modèle de Markov caché, l'observation Y_t est à chaque instant associée à l'état X_t. Si maintenant l'on reçoit N observations $Y_t^1, \dots, Y_t^N$ pour N états $X_t^1, \dots, X_t^N$, utiliser le modèle markovien requiert au préalable une association bipartite entre X^i et Y^i , soit $N!$ alternatives. De plus, pour chacune d'elles, $N!$ nouvelles alternatives se représentent alors à l'instant suivant, et ainsi de suite. La figure suivante illustre cette croissance exponentielle de cas. En outre, une difficulté conceptuelle supplémentaire apparaît dès que le nombre d'observations diffère de celui d'états suivis.



C'est pourquoi le cadre *Random Finite Set* (RFS) propose de suivre non plus un point de l'espace d'état E, mais un **sous-ensemble fini** de points de E, et en lui associant l'unique RFS d'observation Y au RFS d'état X. Leur efficacité calculatoire repose sur une nouveauté conceptuelle simple, qui est de ne pas réaliser explicitement d'association entre cibles et observations, mais plutôt de considèrer les contributions potentielles de chaque cible à chaque image, et de sommer ces contributions. Ainsi, à chaque instant, l'association est remise à plat.

Définition des RFS On définit ainsi un RFS X sur $E \subset \mathbb{R}^d$ comme une variable aléatoire à valeurs dans $\mathcal{F}(E)$, qui est l'espace des sous-ensembles finis de E. Cet espace, et en particulier sa probabilisation, fait l'objet d'un traitement mathématique particulier. Nous nous contentons dans la section en cours de les synthétiser, et renvoyons à la littérature [11, 21, 15] pour de plus amples détails.

Manier $\mathcal{F}(E)$, espace des sous-ensembles finis, est particulièrement compliqué. Il est de très grande taille, et n'est pas un espace vectoriel. De plus, ses éléments sont de dimension hétérogène, ce qui complique la définition d'une mesure répondant à notre intuition physique. C'est pourquoi cet espace a fait l'objet d'une nouvelle théorie, la *Finite Set Statistics* (FISST). Elle munit $\mathcal{F}(E)$ d'une topologie particulière, appelée topologie de Mathéron, et d'une mesure d'un genre nouveau, désignée par μ.

Mesure sur l'espace des ensembles finis La théorie FISST munit l'espace topologique $\mathcal{F}(E)$ d'une mesure positive particulière, appelée "mesure non-normalisée d'un processus ponctuel de

Poisson", définie sur tout borélien $\mathcal{T}$ de $\mathcal{F}(E)$ par

$$\mu(\mathcal{T}) = \sum_{i=0}^{\infty} \frac{1}{i!} \lambda^i \left(\{ (x_1, \dots, x_i) \in E^i \mid \{x_1, \dots, x_i\} \in \mathcal{T} \} \right)$$

Cette formule utilise les mesures de Lebesgue adimensionnées λ^i, et les E^i, i$^{\text{e}}$ produit cartésien de E. Le facteur $i!$ joue le rôle de "normalisation de la dimension", permettant d'ajouter les mesures d'objets de différentes dimensions. Toutefois, la mesure μ est sans dimension, et ses valeurs dépendent de l'unité-étalon de base choisie sur E, que l'on note K. Il s'agit d'une différence importante avec les mesures euclidiennes classiques.

Distributions RFS Soit X un Random Finite Set. Sa loi est alors caractérisée par sa distribution de probabilité P_X, définie sur la tribu borélienne de $\mathcal{F}(E)$. (ici, un borélien est donc un ensemble ouvert d'ensemble finis)

$$P_X(\mathcal{T}) := \mathbb{P}(X \in \mathcal{T})$$

Par suite, si la distribution de probabilité P_X est absolument continues par rapport à cette mesure μ, alors il existe une densité de probabilité RFS $p_X : \mathcal{F}(E) \mapsto \mathbb{R}_+$ qui vérifie pour tout borélien $\mathcal{T} \subset \mathcal{F}(E)$:

$$P_x(\mathcal{T}) = \int_{\mathcal{T}} p_X(\Xi) \mu(\mathrm{d}\Xi) \tag{1}$$

$$= \sum_{i=0}^{\infty} \frac{1}{i!} \int_{E^i} \mathbf{1}_{\mathcal{T}}(\{x_1, \dots, x_i\}) f(\{x_1, \dots, x_i\}) \lambda^i(\mathrm{d}x_1, \dots, \mathrm{d}x_i) \tag{2}$$

Annexe : définition des densités par dérivation fonctionnelle L'espace $\mathcal{F}(E)$ n'est pas un espace de Banach ni même un espace vectoriel, et sa tribu borélienne de $\mathcal{F}(E)$ est complexe à manipuler. Nous ne pouvons donc définir les densités de probabilités comme différentielle de fonctionnelles à valeur scalaire. Toutefois, une théorie nommée FISST, a été développé pour donner un sens à la "dérivée par rapport à un ensemble fini" des fonctions scalaires $F : \mathcal{C}(E) \to \mathbb{R}_+$ définies sur $\mathcal{C}(E)$, ensemble des fermés de E. Elle donne ainsi deux définitions alternatives aux opérations classiques.

– La *set-dérivation*, ou dérivation par rapport à un ensemble fini, est définie par récurrence sur le cardinal de celui-ci

$$\text{Initialisation :} \qquad (\mathrm{d}F)_x(S) = \lim_{\lambda(\Delta_x) \to 0} \frac{F(S \cup \Delta_x) - F(S)}{K \lambda(\Delta_x)} \tag{3}$$

où $\Delta_x \subset E$ désigne un voisinage de $x \in E$.

$$\text{Hérédité :} \qquad (\mathrm{d}F)_{\{x_1, \dots, x_n\}}(S) = (\mathrm{d}(\mathrm{d}F)_{\{x_1, \dots, x_{n-1}\}})_{x_n}(S) \tag{4}$$

L'unité de $(\mathrm{d}F)_X(S)$ est $K^{|X|}$, et dépend donc du cardinal du RFS X.

– La *set-intégration* est définie par :

$$\int_S f(X) \delta X = \sum_{i=0}^{\infty} \frac{1}{i!} \int_{S^i} f(\{x_1, \dots, x_i\}) \, \lambda_K^i(\mathrm{d}x_1, \dots, \mathrm{d}x_i)$$

(où K est l'étalon choisi sur E pour la mesure de Lebesgue)

Un résultat important justifie cette construction. Il affirme que les set-derivation et l'absolue continuité d'une fonction par rapport à la mesure μ introduite plus haut coïncident, à l'unité près. Soit X un RFS de distribution P_X. Cette distribution induit une fonction $\beta_X(\cdot)$, appelée *belief-mass function*, définie ainsi :

$$\forall S \in \mathcal{C}(E), \quad \beta_X(S) = \mathbb{P}(X \subset S)$$

Le résultat s'exprime de la manière suivante. Si P_X admet une densité p_X par rapport à μ, alors

$$\frac{\mathrm{d}P_X}{\mathrm{d}\mu}(Y) = K^{|Y|}(\mathrm{d}\beta_X)_Y(\varnothing)$$

Autrement dit, sur tout ensemble fermé S,

$$\mathbb{P}(X \subset S) = \int_S p_X(\Xi)\delta\Xi$$

Ces outils fournissent une représentation bien plus commode des RFS, car elles ne manipulent que des sous-ensembles fermés de E et non pas des sous-ensembles de $\mathcal{F}(E)$. Elles permettent ensuite d'établir une formulation différente des RFS, qui s'avère très commode pour définir des propriétés statistiques.

RFS et Point process theory Les RFS possèdent une formulation équivalente, bien plus proche de l'intuition. Étant données les zones de l'espace ambiant E, combien de points vont se réaliser dans chacune d'elles ? On imagine facilement qu'en recensant ces réalisations dans des zones de plus en plus en petites, on obtient une caractérisation complète du RFS.

Ainsi, à un RFS X correspond une et une seule **mesure de comptage aléatoire** $\mathcal{X}$ telle que pour tout fermé $S \subset E$,

$$\mathcal{X}(S) = |X \cap S|$$

L'étude de ces processus particuliers, aussi appelés *simple-finite point process*, fait l'objet de cette *Point process theory*. Son intérêt majeur est d'apporter des outils statistiques classiques qui manquaient aux RFS. Tel quel, on ne pouvait parler d'espérance d'un RFS, car l'addition n'est pas définie sur ces ensembles.

1.1.2 Moments statistiques d'un RFS

Aussi, le point de vue *Point process theory* donne un sens aux notions statistiques de moment, et fonction génératrice des moments, pour un RFS. Le premier moment statistique est particulièrement important pour l'algorithmique de filtrage RFS, aussi nous le présentons-nous ici.

Le moment d'ordre 1 d'un RFS est défini par une mesure positive sur l'espace ambiant E, appelée *mesure d'intensité* :

$$\forall S \subset E, \quad M(S) = \mathbb{E}(|X \cap S|)$$

Si la mesure est absolument continue par rapport à la mesure de Lebesgue sur E, alors elle admet une densité $\gamma : E \to \mathbb{R}_+$, appelée *fonction d'intensité*, définie ainsi :

$$\forall S \subset E, \quad M(S) = \int_S \gamma(x)\mathrm{d}x$$

À cette densité correspond une densité de probabilité :

$$\forall x \in E, \quad \eta(x) = \gamma(x)/M(E)$$

Intuitivement, la valeur $\gamma(x)\mathrm{d}x$ correspond au nombre de valeurs espéré dans un voisinage infinitésimal de x. La fonction d'intensité permet donc une localisation d'un ensemble fini aléatoire.

1.1.3 Classes remarquables de processus RFS

La distribution d'un RFS pouvant être très compliquée, il est capital de considérer des familles semi-paramétriques de RFS.

- Un RFS de Bernouilli est un RFS vide avec une probabilité $1-q$, ou contenant un unique élément avec une probabilité q. Cet élément est alors tiré selon une densité de probabilité $p(\cdot)$. Un tel RFS est donc paramétré par $(q, p(\cdot))$. La densité d'un RFS de Bernouilli est

$$p(X) = \begin{cases} 1-q, & X = \{\varnothing\} \\ q \times p(x), & X = \{x\} \\ 0, & |X| > 1 \end{cases}$$

  Un processus de Bernouilli décrit bien une particule qui soit disparaît, soit évolue selon une certaine équation de transition.

- Un RFS multi-Bernoulli est l'union d'un nombre fixé de RFS de Bernouilli indépendants.

$$X = \bigcup_{i=1}^{N} X^{(i)}$$

  Ces RFS sont donc paramétrés par $\{(q^{(1)}, p^{(1)}), \ldots, (q^{(N)}, p^{(N)})\}$ où un couple $(q^{(i)}, p^{(i)}(\cdot))$ est le paramètre de $X^{(i)}$. La densité d'un RFS multi-Bernouilli est

$$p(X) = \begin{cases} (1-q)^n q^{N-n} \times \sum_{1 \leq i_1 < \ldots < i_n \leq N} p^{(1)}(x_{i_1}) \ldots p^{(n)}(x_{i_n}), & X = \{x_1, \ldots, x_n\} \\ 0, & |X| > N \end{cases}$$

  On remarque que cette expression est fondée sur les polynômes symétriques élémentaires à N variables :

$$\sigma_n(X_1, \ldots, X_N) = \sum_{1 \leq i_1 < \ldots < i_n \leq N} X_{i_1} \ldots X_{i_n}$$

- Un RFS de Poisson est paramétré par sa **fonction d'intensité** $\gamma(.)$, qui, comme vu précédemment, se décompose en un nombre $M(E)$ et une densité de probabilité sur E, $\eta(.)$. Une réalisation d'un RFS de Poisson se déroule en deux étapes :

  1. tirage du nombre entier N de points, suivant la loi de Poisson de paramètre $M(E)$;
  2. tirage de N points i.i.d. dans E selon la densité de probabilité $\eta(.)$.

  La densité d'un RFS de Poisson vaut :

$$p(X = \{x_1, \ldots, x_n\}) = e^{-M(E)} \gamma(x_1) \ldots \gamma(x_n)$$

  Les RFS de Poisson ont donc deux des comportements des variables de Poisson usuelles. Ils donc entièrement caractérisés par leur moment statistique, i.e. leur mesure d'intensité. Ils bénéficient du théorème de superposition/décomposition ; ainsi, la somme de processus de Poisson indépendants forme un nouveau processus de Poisson, d'intensité égale à la somme de celles précédentes.

- Un RFS *cluster*, ou RFS de Poisson généralisé est paramétré par sa **distribution de cardinalité** $\rho(\cdot)$ et sa densité de probabilité d'intensité $\eta(\cdot)$. Une de leur réalisation se déroule en deux étapes :

  1. tirage du nombre entier d'objets N dans $\mathbb{N}$, suivant la distribution discrète ρ ;

2. tirage de N points i.i.d. dans E selon la densité spatiale η.

La distribution de cardinalité détermine la masse globale : $\mathbb{E}(|X|) = \mathbb{E}(N)$, et la densité η correspond bien à celle induite par la fonction d'intensité γ, i.e.

$$M(E) = \int_E \gamma(x)\mathrm{d}x = \sum_{n\in\mathbb{N}} n\rho(n)$$

1.2 Formulation multi-objet du suivi bayésien

1.2.1 Phase 0 : modèle multi-cibles / multi-observations

Le filtre recherche la solution d'un **modèle espace-état classique** (équation de transition et équation d'observation), mais qui serait **parcouru par plusieurs cibles** et détecterait plusieurs observations. Le modèle suppose qu'en plus de cette évolution, les cibles apparaissent et disparaissent selon un **cycle de vie indépendant**.
Les processus stochastiques d'état $(X_t)_t$ et d'observation $(Y_t)_t$ repèrent à chaque instant t un nombre fini N_t de cibles et un autre nombre fini M_t d'observations.

$$X_t = \{x_{t,1}, x_{t,2}, \ldots, x_{t,N_t}\} \in \mathcal{F}(E_s)$$

$$Y_t = \{y_{t,1}, y_{t,2}, \ldots, y_{t,M_t}\} \in \mathcal{F}(E_o)$$

Hypothèses minimales de suivi Tous les modèles et approximations présentés reposent sur les hypothèses *a minima* suivantes :

- les cibles et observations sont ponctuelles, et ne se superposent pas ;
- une cible évolue et génère des observations indépendamment des autres cibles ;
- l'évolution d'une cible est markovienne, et consiste soit à disparaitre, soit à transiter vers un nouvel état ;
- les processus de naissance, de prolifération et d'évolution sont indépendants entre eux ;
- les fausses observations (*clutter*) apparaissent indépendamment de celles générées par les cibles.

Nous détaillons maintenant ces hypothèses, selon qu'elles concernent les cibles ou les observations.

Hypothèses sur les cibles Le cadre du suivi RFS propose d'étendre la dynamique markovienne mono-objet, en modélisant la transition entre deux instants t et $t+1$ comme la superposition de **trois phénomènes**.

- l'**évolution** individuelle de chaque cible. Une cible meurt avec une probabilité $1 - p_s^t(x_t)$, ou bien transite vers un nouvel état x_{t+1} tiré selon la densité $f_{t+1}(\cdot \mid x_t)$.
- La **naissance** spontanée de nouvelles cibles, regroupées dans l'ensemble fini Γ_{t+1}, qui définissent ainsi le RFS des naissances ex nihilo.
- La **prolifération** (*spawning*) de cibles. En plus des disparitions et apparitions *ex nihilo*, chaque cible est suspectible d'engendrer une nouvelle souche de cibles.

Cette dynamique particulière se résume par la formule suivante.

$$\begin{aligned} X_{t+1} &= T_{t+1|t}(X_t) \cup S_{t+1|t}(X_t) \cup \Gamma_{t+1} \\ &= \left(\cup_{x\in X_t} T_{t+1|t}(x)\right) \cup \left(\cup_{x\in X_{t+1}} S_{t+1|t}(x)\right) \cup \Gamma_{t+1} \end{aligned}$$

Hypothèses sur les observations Comme pour un modèle espace-état classique, on suppose que la connaissance de l'état courant donne toute l'information disponible sur son image. Le modèle RFS suppose de plus que certaines mesures sont des leurres à rejeter. Ainsi, les mesures valides sont les images des cibles, tirées selon une densité $g_t(y_t \mid x_t)$, tandis que les leurres, appelés *clutter*, sont générés par un processus RFS intempestif K_t.

$$\begin{aligned} Y_{t+1} &= \Theta(X_{t+1}) \cup K_t \\ &= \left(\cup_{x \in X_{t+1}} \Theta(x)\right) \cup K_t \end{aligned}$$

Le suivi temps réel consiste à estimer à chaque instant l'état X_t le plus probable en fonction de l'échantillon $Y_{1:t} = (Y_1, \ldots, Y_t)$. Dans notre modèle génératif, cette opération revêt la dénomination de filtrage bayésien, et prend la forme de deux équations récursives que nous décrivons maintenant.

1.2.2 Phase 1 : récursion bayésienne

Récursion bayésienne RFS directe L'inférence bayésienne classique consiste à déterminer la densité de probabilité des états possibles connaissant la réalisation de notre échantillon d'observations. Pour cela, elle utilise les deux fonctions du modèle génératif :

- la densité de transition RFS : $X_t \mapsto f_t(X_t \mid X_{t-1})$;
- la vraisemblance RFS : $Y_t \mapsto g_t(Y_t \mid X_t)$.

Le filtre bayésien optimal est donnée par la récursion suivante. Hormis le fait que l'intégration porte sur l'espace mesuré $(\mathcal{F}(E_s), \mu_s)$, la théorie FISST aboutit à une formulation est identique à celle classique.

$$\text{densité prédite :} \qquad p_{t|t-1}(X_t \mid Y_{1:t-1}) = \int_{\mathcal{F}(E_s)} f_t(X_t \mid X) p_t(X \mid Y_{1:t-1}) \mu_s(\mathrm{X})$$

$$\text{densité filtrante :} \qquad p_t(X_t \mid Y_{1:t}) = \frac{g_t(Y_t \mid X_t) p_{t|t-1}(X_t \mid Y_{1:t-1})}{\int_{\mathcal{F}(E_s)} g_t(Y_t \mid X) p_{t|t-1}(X \mid Y_{1:t-1}) \mu_s(\mathrm{X})}$$

Implémentations possibles de la récursion Quel que soit le modèle de la densité filtrante, les équations de récursion multi-objet ne sont pas directement calculables. C'est pourquoi pour chaque modèle de filtre proposé, deux implémentations sont envisagables, soit deux approximations différentes.

Les **simulations stochastiques**, de type filtrage particulaire. Très génériques, elles sont utilisables sur tout modèle de fonctions f, g. Elles reviennent à approximer les densités de probabilité par une somme pondérée de fonctions de Dirac.

Les **approximations paramétriques**. La plus courante est de supposer le modèle linéaire-gaussien, qui rend la récursion calculable directement.

1.2.3 Phase 2 : Extraction des cibles

Après la phase de propagation de la densité $p_t(X_t \mid Y_{1:t})$ vient celle d'estimation des positions optimales. La procédure pour aller de la densité filtrante p_t à l'estimation $\hat{X}_t$ ne va pas de soi, et nous allons expliquer pourquoi.

Faillite de l'estimation MAP L'estimateur MAP recherche le maximum de la densité filtrante :

$$\hat{X}_t = \underset{X \in \mathcal{F}(E_s)}{\operatorname{argmax}} \, \pi_{t|t-1}(X \mid Y_{1:t})$$

Il s'agit de la plus simple des méthodes et son emploi est très immédiat en mono-objet. Toutefois, l'estimateur MAP en multi-objets est inconsistant, du fait de la dépendance des densités RFS à l'étalon choisi. Cette inconsistance est illustrée à l'aide de l'exemple suivant, repris de [21].

Prenons comme étalon le mètre. Soit X un RFS de Bernouilli de densité

$$\pi(X) = \begin{cases} 0.5 & X = \{\varnothing\} \\ 0.25 & X = \{x\}, 0 \leq x \leq 2 \\ 0 & |X| > 1 \end{cases}$$

L'estimateur MAP donne alors $\hat{X} = \{\varnothing\}$

Prenons maintenant comme étalon le kilomètre.La densité de X devient alors

$$\pi(X) = \begin{cases} 0.5 & X = \{\varnothing\} \\ 250 & X = \{x\}, 0 \leq x \leq 0.002 \\ 0 & |X| > 1 \end{cases}$$

Et l'estimateur MAP donne alors un résultat différent $\hat{X} = \{x\}$ pour n'importe quel $x \in [0, 0.002]$. Ainsi, l'étalon change du tout au tout l'estimation MAP, qui est donc inconsistante.

L'*intensity-measure-based estimation* De nombreuses estimations se fondent sur l'étude du premier moment statistique γ, qui associe à chaque zone de l'espace le nombre espéré de cibles évoluant dans cette zone. L'énorme avantage de ce procédé est de travailler avec une champ scalaire sur E, et non pas sur $\mathcal{F}(E)$. Une telle estimation se déroule alors généralement en deux temps :

1. estimer un nombre de cibles présente $\hat{N}_t = |X_t|$.
2. fractionner la densité η autour de $\hat{N}_t$ centres $\{\hat{x}_{i,t}\}_{i=1}^{\hat{N}_t}$, qui seront les cibles estimées.

Décrivons maintenant le déroulement habituel de ces deux étapes.

Estimation du nombre de cibles Là encore, les deux choix les plus naturels sont l'estimateur MAP $\hat{N} = \arg\sup_{\mathbb{N}} \rho(\cdot)$ ou bien l'estimateur EAP $\hat{N}_t = \lceil M(E) \rfloor$. La littérature recommande le premier [21].

Extraction de la position des cibles La méthode la plus directe, appelée *marginal multi-object estimator* (MaM), revient à choisir l'estimation MAP.

$$\hat{X}^{MaM} = \underset{X:|X|=\hat{N}}{\arg\max}\ \pi(X \mid Y_{1:t})$$

Ce choix implique de connaître et analyser toute la distribution a posteriori. C'est pourquoi de nombreux algorithmes préfèrent la méthode dite de *first moment visualization*, qui consiste à choisir les $\hat{N}$ premiers maximums de la fonction d'intensité γ, par une méthode de type *peak extraction* ou *clustering*.

La méthode de segmentation la plus populaire est celle des K-moyennes, qui minimise l'erreur quadratique moyenne. Ce choix générale celui, dans le cas d'une seule cible présente, du barycentre spatial de l'intensité : $\hat{x}_t = \int_E x\, \eta_t(x) \mathrm{d}x$.

Enjeu : augmenter l'espace pour discriminer Computationnellement parlant, l'algorithme des K-moyennes est coûteux, car NP-complet, tandis que calculer une moyenne spatiale est explicite. L'idée est de diviser ou d'augmenter l'espace d'une ou plusieurs dimensions discrètes, numérotées en strates, de manière à concentrer les zones d'intensité saillante. Même en l'absence d'informaton supplémentaire, un choix se présente naturellement, celui de la "date de naissance" de la particule. Cette idée est utilisée dans [9], sous le nom de *track association information*, à des fins de suivi de voix multi-locuteurs. Nous décrivons l'algorithme :

1. Considérer à l'instant t l'espace $E_t = E \times [\![0;t]\!] \subset E \times \mathbb{N}$, puis les mesures "partielles" $\{M^{(t)}, M^{(t-1)}, \ldots\}$ définies par
$$\forall S \subset E, \quad M^{(t')}(S) = M(S \times \{t'\})$$
.
2. Supposer que si $M^{(t')}(E) > 0.5$ alors une mesure est apparue à l'instant t'. L'article fait l'hypothèse qu'à chaque instant, pas plus d'une cible peut naître. Mais le nombre de particules nées à l'instant t' est estimé par $\hat{N}^{(t')} = \lceil M^{(t')}(E) \rfloor$; après quoi une méthode K-moyenne ou autre est appliquée sur chaque $M^{(t')}$.
3. Si une cible est née à l'instant t', l'estimer par le barycentre de $M^{(t')}(\cdot)$.

Ainsi, nous entrevoyons les enjeux RFS de modélisation de l'espace d'état caché, afin de permettre une estimation efficace. À cette fin, ce dernier procédé amène une question mathématique intéressante : quel est l'effet d'augmenter l'espace sur la variance de l'estimateur choisi du nombre de cibles présentes d'un tel ajout de dimension ? A-t-on $\mathrm{Var}(\hat{N}) < \mathrm{Var}\left(\sum_{t' \leq t} \hat{N}^{(t')}\right)$? Nous laissons cette question théorique pour des travaux futurs et nous tournons maintenant du côté des principaux filtres d'inférence RFS employés dans la littérature.

1.3 Présentation des filtres multi-observation classiques

1.3.1 Hypothèses générales des filtres

Trois modèles de filtres particuliers ont été développés : PHD, CPHD, CB-MeMBer. Ces filtres sont essentiellement adaptés au cas multi-cibles et multi-observations, et reposent tous sur la simplification des calculs entraînées par les hypothèses suivantes.

- les cibles sont ponctuelles et ne se superposent pas ;
- les apparitions, disparitions surviennent pour chaque cible indépendamment des observations ;
- les observations sont ponctuelles et ne se superposent pas ;
- une observation est soit l'image d'une seule cible, soit une fausse observation ;
- une cible est soit détectée, et produit alors une image indépendamment des autres cibles, soit non-détectée ;
- les fausses observations (*clutter*) apparaissent indépendamment de celles des cibles ;
- les probabilités de disparition et de non-détection ne dépendent que de l'état de la cible.

La présente section décrit successivement ces trois filtres. Leurs caractéristiques sont résumées dans le tableau suivant. Pour chaque filtre, il est possible d'implémenter une inférence soit approximative par un filtrage particulaire, soit exacte sous l'hypothèse d'un modèle linéaire-gaussien. La complexité indiquée ici est celle de ce modèle d'inférence paramétrique.

Filtre	Hypothèses	Récursion	Complexité
PHD	naissance et clutter : Poisson	γ	linéaire en $\lvert Y_t \rvert$
CPHD	naissance et clutter : Poisson généralisé	(γ, ρ)	cubique en $\lvert Y_t \rvert$
CB-MeMBer	naissance : multi-Bernoulli clutter : Poisson	$\{(q^{(i)}, p^{(i)}(\cdot))\}_{i=1}^{M_t}$	linéaire en $\lvert Y_t \rvert$ et M_t

1.3.2 Présentation du filtre PHD

Le filtre *Probability Hypothesis Density* (PHD) est un filtre bayésien sous-optimal, conçu pour les situations **à grand nombre de cibles simultanées**. Son calcul repose sur une approximation des densités filtrantes par des processus qui modélisent les situations où la position et la dynamique d'un objet sont indépendants des positions des autres objets. Comme cette forme de processus est stable par l'équation de mise à jour, et est caractérisée par leur intensité $\gamma(\dot{})$, l'inférence peut être réduite à cette seule fonction.

Hypothèses Les hypothèses du filtre PHD se résument mathématiquement de la manière suivante :

- Le RFS X_t des cibles à l'instant t se présente comme la superposition de trois RFS indépendants conditionnellement à X_{t-1}.
$$X_t = T_{t|t-1}(X_{t-1}) \cup S_{t|t-1}(X_{t-1}) \cup \Gamma_t$$
- Conditionnellement à X_{t-1}, $T_{t|t-1}(X_{t-1})$ est un processus RFS multi-Bernouilli de paramètre
$$\{\big(p_t^s(x_{t-1}), f_{t|t-1}(\cdot \mid x_{t-1})\big) \mid x_{t-1} \in X_{t-1}\}$$
- Conditionnellement à X_{t-1}, $S_{t|t-1}(X_{t-1}) = \bigcup_{x_{t-1}\in X_{t-1}} S_{t|t-1}(x_{t-1})$, où chaque $S_{t|t-1}(x)$ est un processus de Poisson, indépendant des autres, de fonction d'intensité $b_{t|t-1}(\cdot \mid x)$.
- Γ_t est un processus de Poisson de fonction d'intensité $\mu_t(\cdot)$, indépendant de X_{t-1}.
- Le RFS Y_t des observations à l'instant t se présente comme la superposition de deux RFS indépendants, conditionnement à X_t.
$$Y_t = \Theta_t(X_t) \cup K_t$$
- Conditionnellement à X_t, Θ_t est un processus RFS multi-Bernouilli de paramètre
$$\{\big(p_t^d(x_t), g_t(\cdot \mid x_t)\big) \mid x_t \in X_t\}$$
- K_t est un processus de Poisson de fonction d'intensité $h_t(\cdot)$, indépendant de X_t.
- Les fonctions $p_t^d, p_t^d : E_s \to [0,1]$, les densités positives $b_{t|t-1}, \mu_t$ et les densités de probabilités $f_{t|t-1}(\cdot \mid x), g_t(\cdot \mid x)$ sont choisies *a priori*, et peuvent dépendre du temps.

Équations de la récursion bayésienne L'inférence bayésienne se réduit ainsi à une récursion sur les intensités prédites et a posteriori, donnée par les équations suivantes qui forment le filtre PHD :

$$\text{prédiction PHD :} \quad \gamma_{t|t-1}(x) = \int_{E_s} \left[p_t^s(u) f_{t|t-1}(x \mid u) + b_{t|t-1}(x \mid u)\right] \gamma_{t|t-1}(u)\mathrm{d}u + \mu_t(x)$$

$$\text{filtrage PHD :} \quad \gamma_{t|t}(x) = (1 - p_t^d(x))\gamma_{t|t-1}(x) + p_t^d(x) \sum_{y\in Y_t} \frac{g_t(y \mid x)\gamma_{t|t-1}(x)}{h_t(y) + \int_{E_s} p_t^d(u) g_t(y \mid u)\gamma_{t|t-1}(u)\mathrm{d}u}$$

Inteprétation et utilisation pour le suivi Le filtre PHD effectue la même simplification que le filtre de Kalman, en réduisant le problème à la propagation du moment d'ordre 1 de la distribution du processus $(X_t)_t$, ce qui revient à négliger les moments statistiques d'ordre supérieur. Il calcule en fait la meilleure approximation de $(X_t)_t$ par un processus de Poisson RFS, conditionnellement à l'échantillon d'observations.

Nous rappelons que tel quel, le filtre PHD n'effectue donc pas de suivi de position ; il actualise à chaque instant une estimation de la fonction d'intensité γ, qui permet d'employer des méthodes d'estimations telles que décrites dans la section précédente.

Implémentation 1 : Sequential Monte Carlo (SMC-PHD) La stratégie du filtrage particulaire est d'approcher la fonction d'intensité γ par une somme pondérée de diracs :

$$\gamma_t = \sum_{j=1}^{L_t} w_t^{(j)} \delta_{x_t^{(j)}} \quad , x_t^{(j)} \in E_s$$

Les équations de récursion sont donc aisément calculables.

Implémentation 2 : Gaussian mixtures (GM-PHD) La récursion bayésienne est calculable explicitement pour un modèle linéaire-gaussien, tel qu'introduit par B.-T. Vo [19, 20] :

- La densité de transition d'une cible est linéaire-gaussienne : $f_{t|t-1}(x_t \mid x_{t-1}) = \mathcal{N}(x_t; F_{t-1}x_{t-1}, Q_{t-1})$
- La densité de vraisemblance de la mesure l'est aussi : $g_t(y_t \mid x_t) = \mathcal{N}(y_t; H_t x_t, R_t)$
- les probabilités de survie et de détection sont uniformes sur l'espace d'état : $p_t^d(x) \equiv p_t^d$, $p_t^s(x) \equiv p_t^s$
- les intensités de naissance et de prolifération RFS sont des mélanges de gaussiennes :

$$\mu_t(x_t) = \sum_{i=1}^{J_{\mu,t}} w_{\mu,t}^{(i)} \mathcal{N}(x_t; m_{\mu,t}^{(i)}, P_{\mu,t}^{(i)})$$

$$b_{t|t-1}(x_t \mid \xi) = \sum_{j=1}^{J_{B,t}} w_{B,t}^{(j)} \mathcal{N}(x_t; F_{\mu,t-1}^{(j)}\xi + d_{B,t-1}^{(j)}, Q_{\mu,t-1}^{(j)})$$

- Remarque : les différentes matrices et probabilités sont fixées a priori et peuvent dépendre du temps.

Sous ces hypothèses, la fonction d'intensité du RFS cible est un mélange de gaussiennes. $\gamma_{t|t-1}$ et $\gamma_{t|t}$ sont récursivement calculés par des équations similaires au filtre de Kalman.

Remarquons que le nombre de gaussiennes croît naturellement à une allure exponentielle. En effet, **chaque gaussienne correspond à une hypothèse**, et à chaque instant le processus de naissance ajoute des hypothèses là où il est non-nul. L'emploi de ce filtre requeirt donc une stratégie d'élagage/fusion de gaussiennes, qui en pratique n'est fondée sur des heuristiques. Le livre [16] traite cette opération de façon exhaustive.

L'hypothèse importante est le fait que le nombre de cibles présente suit une loi de Poisson, aussi parle-t-on souvent de filtre *Poisson PHD*. Comme cette hypothèse estsouvent peu fidèle à la réalité, le filtre CPHD ci-après a été développé de manière à la relâcher.

Extension : filtre CPHD Le filtre CPHD étend le filtre PHD au cas où le nombre d'apparitions de cibles, ainsi que celui de fausses observations, suivent des lois quelconques et non des lois de Poisson. Pour cela, le filtre CPHD repose sur une généralisation des processus de Poisson, les RFS *clusters*. Notons de plus qu'il ne traite pas le phénomène de prolifération. Il émet donc exactement les mêmes hypothèses que le filtre PHD, hormis les deux suivantes :

- Les RFS de naissance Γ_t et de *clutter* K_t sont des processus de Poisson généralisés, de lois de cardinalité respectives $\rho^{(B)}(\cdot)$ et $\rho^{(C)}(\cdot)$.
- Il n'y a pas de prolifération : les RFS $b_{t|t-1}$ sont tous vides.

Alors que l'inférence PHD ne porte que sur le moment ρ, celle du CPHD porte alors sur les deux fonctions (γ, ρ). L'extraction des cibles à partir de $\gamma_{t|t}$ doit être précédée d'une estimation du

nombre de cibles présentes $\hat{N}$ à partir de $\rho_{t|t}$. Il est pour cela recommandé ([21]) d'utiliser un MAP plutôt qu'un EAP.

$$\hat{N}_t = \arg\max_{\mathbb{N}} \rho_{t|t}(\cdot)$$

1.3.3 Présentation du filtre CB-MeMBer

Contrairement aux filtres PHD et CPHD, dont les hypothèses impliquent une conservation de la forme de la densité filtrante, le filter CB-MeMBer réalise à chaque pas de temps une approximation de celle-ci par une forme simple, celle d'une densité multi-Bernoulli. Il est adapté au cas où le processus de naissance est un RFS multi-Bernoulli, plutôt que Poisson généralisé.

Hypothèses Les hypothèses du filtre CB-MeMBer se résument mathématiquement de la manière suivante :

naissance : Γ_t est un RFS multi-Bernoulli indépendant de X_{t-1}, de paramètres $\{(q^{(i)}_{\Gamma,t}, p^{(i)}_{\Gamma,t})\}_{i=1}^{M^\Gamma_t}$.

prolifération : aucune.

***clutter* :** K_t est un processus de Poisson indépendant de X_{t-1}, de fonction d'intensité $h_t(\cdot)$, supposée de faible amplitude.

détection : les probabilités de détection p^d_t sont supposées proches de 1.

image : conditionnellement à X_t, Θ_t est un RFS multi-Bernouilli de paramètre

$$\{\big(p^d_t(x_t), g_t(\cdot \mid x_t)\big) \mid x_t \in X_t\}$$

transition : conditionnellement à X_{t-1}, $T_{t|t-1}(X_{t-1})$ est un processus RFS multi-Bernouilli de paramètre

$$\{\big(p^s_t(x_{t-1}), f_{t|t-1}(\cdot \mid x_{t-1})\big) \mid x_{t-1} \in X_{t-1}\}$$

Récursion de l'approximation Le filtre CB-MeMBer réalise l'approximation suivante. Si à un instant $t-1$, la densité filtrante est multi-Bernoulli $\pi_{t-1} \sim \{(q^{(i)}_{t-1}, p^{(i)}_{t-1})\}_{i=1}^{M_{t-1}}$, alors à la prochaine observation Y_t,

1. la prédiction ajoute les modes du RFS de naissance :

$$\pi_{t|t-1} \sim \{(q^{(i)}_{t|t-1}, p^{(i)}_{t|t-1})\}_{i=1}^{M_{t|t-1}} = \{(q^{(i)}_{t-1}, p^{(i)}_{t-1})\}_{i=1}^{M_{t-1}} \cup \{(q^{(i)}_{\Gamma,t}, p^{(i)}_{\Gamma,t})\}_{i=1}^{M^\Gamma_t}$$

2. la mise à jour modifie les modes prédits et ajoute autant de modes que d'observations $|Y_t|$:

$$\pi_t \sim \{(q^{(i)}_t, p^{(i)}_t)\}_{i=1}^{M_t} = \{(q^{(i)}_{L,t}, p^{(i)}_{L,t})\}_{i=1}^{M_{t|t-1}} \cup \{(q^{(i)}_{D,t}, p^{(i)}_{D,t})\}_{i=1}^{|Y_t|}$$

L'évolution du nombre de modes est donc résumée par

$$M_{t|t-1} = M_{t-1} + M^\Gamma_{t-1} \tag{5}$$
$$M_t = M_{t|t-1} + |Y_t| \qquad = M_{t-1} + (M^\Gamma_{t-1} + |Y_t|) \tag{6}$$
$$\tag{7}$$

Ainsi, à chaque pas de temps la prédiction les modes du RFS de naissance. Ensuite, pour chaque observation, un *measurement-updated mode* (D) est créé et regroupe la somme des contributions de chaque modes prédits, cette contribution s'apparentant à la vraisemblance.

L'extraction des cibles est aisée, car le calcul de la fonction d'intensité et de l'espérance du nombre de cibles est immédiat :

$$\gamma_t = \sum_{i=1}^{M_t} q_t^{(i)} p_t^{(i)}$$

$$\hat{N}_t = \sum_{i=1}^{M_t} q_t^{(i)}$$

L'interprétation des modes en termes d'hypothèses est donc naturelle. Un mode i augmente l'espérance du nombre de cible d'une fraction $q_t^{(i)}$. Dans le cas d'une probabilité de détection suffisamment bonne et une intensité de fausses mesures suffisamment faibles, les *legacy modes* tendent à perdre considérablement en probabilité. Là encore pour ce filtre, une stratégie de gestion du nombre croissant de modes doit être gérés, par exemple en élaguant ceux dont la probabilité $q_{L,t}^{(i)}$ devient trop faible.

1.3.4 Analyse des filtres RFS classiques

Quel que soit leur point de vue les filtres RFS infèrent tous une densité de probabilité **multi-modale**. Que ces modes soient des particules ou des gaussiennes, ils ne font que représenter des profils élémentaires d'**hypothèses** sur la localisation d'une cible.

Le suivi multi-objet nous semble d'implémentation difficile à cause de la dérive naturelle du nombre des hypothèses qu'il doit considérer. L'étude du fonctionnement des filtres montre qu'à chaque étape du temps, l'étape de prédiction ajoute un mode pour chaque zone où une cible est susceptible de naître ; et qu'ensuite chaque localisation possible de cible rajoute un mode. Or comme les possibilités de naissance et de non-détection sont combinales, même en absence de signal, à chaque instant se créent des hypothèses de nouvelles cibles qui ont une vraisemblance non-nulle et doivent être considérées. Prenons par exemple le filtre PHD. Nous faisons le calcul suivant ; en supposant une absence totale d'observations, des probabilité de détection et de survie uniformes, le nombre espéré de particules $M(E) = \int_E \gamma(x)\mathrm{d}x$ suit l'équation de récurrence suivante.

$$M_t(E) = (1 - p_t^D)(p_t^S M_{t-1}(E) + M_t^\Gamma(E))$$

où M_t^Γ est l'intensité du RFS de naissances. Ainsi, même en l'absence d'observations, ce filtre considère des hypothèses de naissance à chaque isntant, ce qui amène son nombre espéré de particules non-nulles.

Aussi, tous ces filtres requiert un post-traitement pour élaguer les hypothèses de particules qui d'une part naissent à chaque instant à l'étape de Quel que soit son nom, *track management*, *label association*,*gaussian mixture fusion*, il traduit la **dynamique des hypothèses** que l'on souhaite gérer. Ce modèle dynamique constitue le point sensible de la modélisation multi-objet, qu'il faut penser en fonction des particularités du champ d'application.

Dans cette optique, nous trouvons intéressant le modèle multi-Bernoulli d'approximation de la densité inférée, car un mode de Bernoulli $(r, p(\cdot))$ représente un **groupe d'hypothèses**. En effet, du point de vue de l'implémentation, la seule différence entre une densité formée par N particules est équivalente à un densité et une autre formée par 2 modes de Bernoulli à $N/2$ particules chacun réside dans le groupage des hypothèses induit par cette dernière, opération intéressante car permettant de localiser l'extraction des cibles.

2 Le suivi mono-cible de position

2.1 L'inférence position-tempo par filtrage particulaire

Le suivi de position consiste à estimer l'évolution temporelle d'une variable aléatoire cachée, appelée **état** x, en fonction d'une variable aléatoire mesurée, appelée **observation** y. Dans un suivi de position, la finalité est l'estimation de la position sur la partition s_t, exprimée en nombre de pulsations (*beats*).

Présentation du filtrage particulaire Afin de nous préparer à leur utilisation dans un modle multi-objets, nous nous sommes intéressé à l'inférence approchée par simulations stochastiques méthode qui excelle dans de nombreux domaines mais quasiment pas investie pour notre problème. Aussi, nous sommes partis des deux articles [12, 13] de N. Montecchio. Ils proposent une implémentation directe de l'algorithme dit de **condensation** [1]6, aussi appelé SIR, qui est le plus simple des méthodes de filtrage particulaire. Son principe est d'approximer à chaque instant la densité filtrante par une mesure discrète, représentée par des couples $\{x_k^i, w_k^i\}_{i=1}^{N_s}$, où les états x_i sont appelés **particules**, et les scalaires w^i sont leurs **poids**. À chaque nouvelle observation y_k, la nouvelle densité filtrante se calcule en mettant les particules à jour.

1. étape de prédiction : les particules avancent, tirée aléatoirement suivant la densité de transition $f(x_k \mid x_{k-1})$.
$$x_{k-1}^i \longrightarrow x_k^i \simeq f(x_k \mid x_{k-1}^i)$$
2. étape de mise à jour : les poids sont corrigés selon la fonction de vraisemblance $g(y_k \mid x_k)$, puis normalisés.
$$w_{k-1}^i \longrightarrow w_k^i = \frac{\tilde{w}_k^i}{\sum_j \tilde{w}_k^j}, \qquad \text{avec } \tilde{w}_k^i = w_{k-1}^i g(y_k \mid x_k^i)$$

Chaque particule représente une hypothèse sur l'état courant, et son poids sa vraisemblance par rapport aux autres hypothèses courantes. Un filtrage de qualité cherche donc à éviter les échantillonnages dégénérés, où peu d'hypothèses priment sur toutes les autres. Cette dégénérescence s'estime par le **nombre efficace** de particules $N_{eff} = \left(\sum_j (w_k^j)^2\right)^{-1}$. Le rééchantillonnage est déclenché que cette valeur devient plus basse qu'un seuil choisi a priori.
Montecchio utilise la méthode la plus simple, connue sous le nom de **rééchantillonnage par importance** (SIR). Elle consiste à remplacer l'ensemble courant de particules par N tirages indépendants dans ce même ensemble, considéré comme un espace discret probabilisé, puis à uniformiser leur poids en posant $w^i = 1/N_s$.

Modèle d'état L'enjeu d'un modèle d'état est d'inclure suffisamment d'informations pour rendre réaliste l'hypothèse d'évolution markovienne. Aussi, les actuels suiveurs de de partition suivent simultanément la position courante s et le tempo t, avec un vecteur d'état à deux dimensions $x = (s, t)$. Le tempo a en effet une grande valeur *explicative* : il permet de décomposer la position comme somme de la *position sur la grille de pulsation* et de l'*écart à la pulsation*. C'est cet écart à la pulsation que l'on peut assimiler à une perturbation à faible variabilité, et approcher par un bruit blanc. De plus, le tempo a une grande valeur *prédictive*, puisqu'il donne la vitesse d'avancée de la position.

La représentation de la partition se fait en deux étapes. D'abord, un modèle graphique gauche-droite est construit ; les notes sont agrégées en chaînes d'états polyphoniques, de plus courte durée,

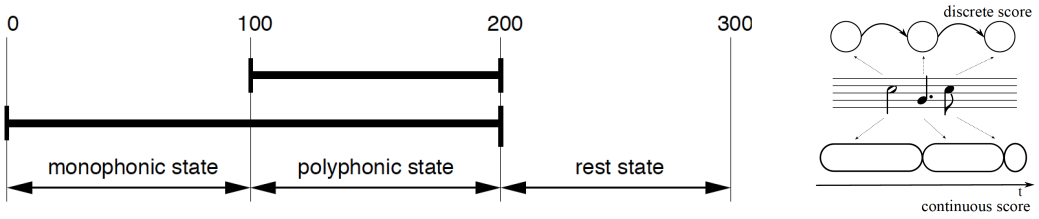


FIGURE 1 – Représentation de la partition dans un modèle de suivi de position. À gauche, agrégation et fragmentation de la partition en chaîne d'états polyphoniques. À droite, segmentation de l'espace continu des positions selon la chaîne d'états.

délimité par chaque événement de la partition (début ou fin de note). Ensuite, cette chaîne d'états est transposé en modèle continu. La partie entière de la variable s de position donne le numéro de la pulsation, et la partie fractionnaire correpond à l'avancement au sein d'une pulsation. Ces deux étapes sont illustrées par la figure 1.

L'autre originalité du travail de Montecchio est de choisir un modèle continu. Il présente un intérêt calculatoire immédiat, car il permet raisonnablement d'utiliser un modèle linéaire-gaussien (aussi introduit dans [17]) pour l'évolution de l'état x.

$$\begin{cases} s_k = s_{k-1} + \Delta T t_{k-1} & +\sigma_s\sqrt{\Delta T}\mathcal{N}(0,1) \\ t_k = t_{k-1} & +\sigma_t\sqrt{\Delta T}\mathcal{N}(0,1) \end{cases}$$

Ce choix revient à modéliser le tempo par un mouvement brownien. Remarquons que le choix de variance proportionnelle au pas de temps ΔT est le seul qui garantit qu'indépendamment du pas de temps, la variance du tempo est proportionnelle au temps écoulé : $t_\tau \sim \mathcal{N}(t_0, \sigma_t^2\tau)$.

Modèle d'observation L'implémentation de Montecchio reprend le modèle d'observation d'Antescofo, décrit dans [3] et dérivé de [17]. La vraisemblance est une mesure de similarité entre le spectrogramme à court terme observé, et un patron (*template*) de spectrogramme. Ces patrons sont pré-définis pour chaque fréquence fondamentale, et son superposés en cas d'état polyphonique. La mesure de similarité choisie est la divergence de Kullback-Leibler entre les spectrogrammes préalablement normalisés.

Nous observons que ce choix fait perdre son avantage à un modèle continu, car la vraisemblance est fonction de l'état discret. Son calcul implique un passage de la position continue s (en nombre de pulsations) à la position discrète i (en nombres de notes écrites). Un espace continu de position implique une fonction de vraisemblance en escalier, donc fortement non-linéaire par rapport à la position s, comme le montre la figure 2.

Implémentation et paramètres Notre travail a d'abord consisté à ré-implémenter cet algorithme SIR en MATLAB, afin de l'expérimenter sur les extraits audios de notre choix. Nous récapitulons ici la liste des paramètres numériques à définir :

- N_s : le nombre de particules, i.e. d'hypothèses d'état testées à chaque instant ;
- σ_s^2, σ_t^2 : les variances des positions et tempos, normalisées par seconde ;
- $N_e^{th}ff$: le seuil inférieur du nombre efficace de particules déclenchant le rééchantillonnage.

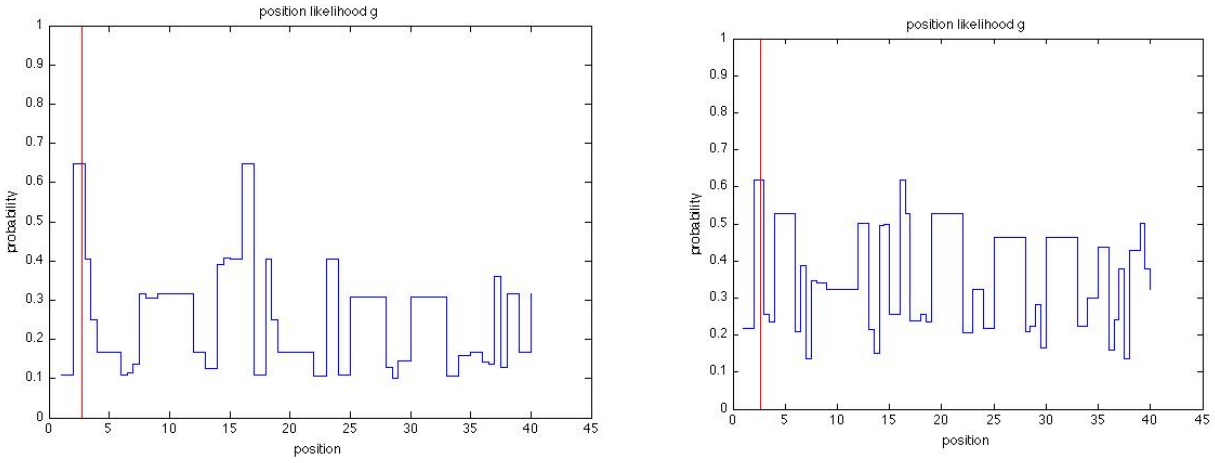


FIGURE 2 – Exemple de vraisemblance $g(s \mid y)$ du choral *Christ, der du bist Tag und Licht* de J.S. Bach, durant la note dont la position est marquée d'un trait. À gauche, les quatre voix ensembles. À droite, les voix ténor et basse ensembles.

2.2 Proposition d'extension du filtrage particulaire SIR

2.2.1 Enjeu de la validation du modèle de transition

Vraisemblance des paramètres Nous avons remarqué la grande sensibilité de la qualité du suivi au choix des paramètres de variance. Aussi, nous voyons comme enjeu, dans la conception d'un système de suivi, de parvenir à prévoir son efficacité sur une exécution donnée. C'est pourquoi nous proposons de jauger la qualité prédictive d'un modèle en calculant la vraisemblance de la suite des instants d'attaque sur les morceaux de musiques étiquetés. Nous voyons là une manière de profiter pleinement du modèle linéaire-gaussien, qui permet l'emploi du protocole suivant :

1. considérer les vrais instants d'*onsets* τ_k d'une suite de position y_k étiquetées ;
2. considérer ces instants comme des observations de la position s_k ;
3. inférer par filtre de Kalman la suite de tempo correspondant t_k ;
4. vérifier la vraisemblance des vrais *onsets* étant donné ce tempo.

Le décodage du tempo à partir de la donnée des *onsets* est calculable explicitement par application du filtre de Kalman ([7]) dans le modèle espace-état suivant.

$$\begin{aligned} \text{transition :} \quad & X_{k+1} = A_k X_k + Q_k \mathcal{N}(0,1) \\ \text{observation :} \quad & Y_{k+1} = H X_{k+1} \end{aligned}$$

où

$$X_k = \begin{pmatrix} s_{\tau_k} \\ t_{\tau_k} \end{pmatrix}, \quad Y_k = s_{\tau_k}, \quad \Delta T_k = \tau_{k+1} - \tau_k$$

$$A_k = \begin{pmatrix} 1 & \Delta T_k \\ 0 & 1 \end{pmatrix}, \quad Q_k = \begin{pmatrix} \Delta T_k \sigma_s^2 & 0 \\ 0 & \Delta T_k \sigma_s^2 \end{pmatrix}, \quad H = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

L'algorithme de filtrage consiste à estimer simultanément le tempo t_k et la matrice de covariance de l'erreur commise P_k. Dans notre cas, l'erreur commise ne porte sur le tempo, donc la matrice P_k n'a qu'un terme non nul. on suppose de plus la position initiale et le tempo connu. Les calculs

se simplifient alors, pour donner :

$$\begin{aligned}
\text{début de l'étape } k+1 : \quad & P_k = \begin{pmatrix} 0 & 0 \\ 0 & p_k \end{pmatrix} \\
\text{prédiction de la position : } \quad & X_{k+1|k} = A_k X_{k|k} = \begin{pmatrix} s_{\tau_k} + \Delta T_k \hat{t}_k \\ \hat{t}_k \end{pmatrix} \\
\text{covariance de l'erreur de prédiction : } \quad & P_{k+1|k} = A_k P_k A_k^T + Q_k = \begin{pmatrix} p_k \Delta T_k^2 + \Delta T_k \sigma_s^2 & p_k \Delta T_k \\ p_k \Delta T_k & \Delta T_k \sigma_t^2 + p_k \end{pmatrix} \\
\text{calcul du gain de Kalman : } \quad & K_{k+1} = P_k H^T \left(H P_{k+1|k} H^T \right)^{-1} = \begin{pmatrix} 1 \\ \frac{p_k}{p_k \Delta T_k + \sigma_s^2} \end{pmatrix} \\
\text{décodage du tempo : } \quad & \hat{t}_{k+1} = \hat{t}_k + (s_{\tau_{k+1}} - (s_{\tau_k} + \Delta T_k \hat{t}_k)) \frac{p_k}{p_k + \sigma_s^2 / \Delta T_k} \\
\text{décodage de l'erreur commise : } \quad & p_{k+1} = \Delta T_k \sigma_t^2 + p_k (1 - \frac{p_k}{p_k + \sigma_s^2 / \Delta T_k})
\end{aligned}$$

La vraisemblance de l'échantillon vaut celle des erreurs de prédiction en tant que tirages indépendants de lois normales dont la matrice de covariance est celle prédite.

$$\epsilon_{k+1|k} := X_{k+1} - X_{k+1|k} = \begin{pmatrix} s_{\tau_{k+1}} - (s_{\tau_k} + \Delta T_k \hat{t}_k) \\ (s_{\tau_k} + \Delta T_k \hat{t}_k) \frac{p_k}{p_k + \sigma_s^2 / \Delta T_k} \end{pmatrix}, \quad p^{11}_{k+1|k} = p_k \Delta T_k^2 + \Delta T_k \sigma_s^2$$

La log-vraisemblance des *onsets* étiquetés conditionnellement aux paramètres (σ_s^2, σ_t^2) s'exprime par

$$l(Y_{0:T}; \sigma_s, \sigma^t) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{k=1}^{T} \log \left(p^{11}_{k+1|k} \right) - \frac{1}{2} \sum_{k=1}^{T} \frac{\epsilon^2_{k+1|k}}{p^{11}_{k+1|k}}$$

Résultats Si ce type d'étude statistique est à mener sur un grand nombre d'interprétations annotées du même morceau de musique, nous l'avons testée dans deux situations d'interprétations du même choral à quatre voix de J.S. Bach, *Ach Lieben Christen*, extrait d'une base de données étiquetées de dix chorals, disponible en ligne [1]. La figure 2.2.1 montre l'évolution de la vraisemblance en fonction de l'écart-type de fluctuation du tempo, σ_t, où l'écart-type de position σ_s a été fixée à une valeur faible (0.05). La courbe de gauche montre l'interprétation conforme à la partition. L'écart-type optimal est $\sigma_t^* = 0.30$ (pour un tempo en BPS et des durées en secondes), ce qui d'ailleurs correspond à la valeur précédemment choisie à la main (0.35). La courbe de droite montre la situation où l'interprétation comporte un point d'orgue à la fin de chaque phrase. Cette simple variation suffit à faire grimper l'optimum à $\sigma_t^* = 1.32$.

En conclusion, ce test nous a révélé tout sensibilité du modèle linéaire-gaussien à ses paramètres de variance. Son hypothèse de volatilité constante du tempo fait chuter la vraisemblance de performances musicales marquant ponctuellement des tenues de note, phénomène qui est pourtant assez commun. De façon plus générale, les tests statistiques sur les performances annotées constituent une pratique intéressante pour l'élaboration de modèles génératifs. Nous regrettons son absence dans la plupart des articles à ce sujet. En particulier, la vérification de l'adéquation entre performances musicales correctement suivies par un algorithme donné, et bonne vraisemblance de ces performances conditionnellement à son modèle de transition, constituerait une extension intéressante aux tests comparatifs menés sur des bases de données communes.

1. http://music.cs.northwestern.edu/index.php

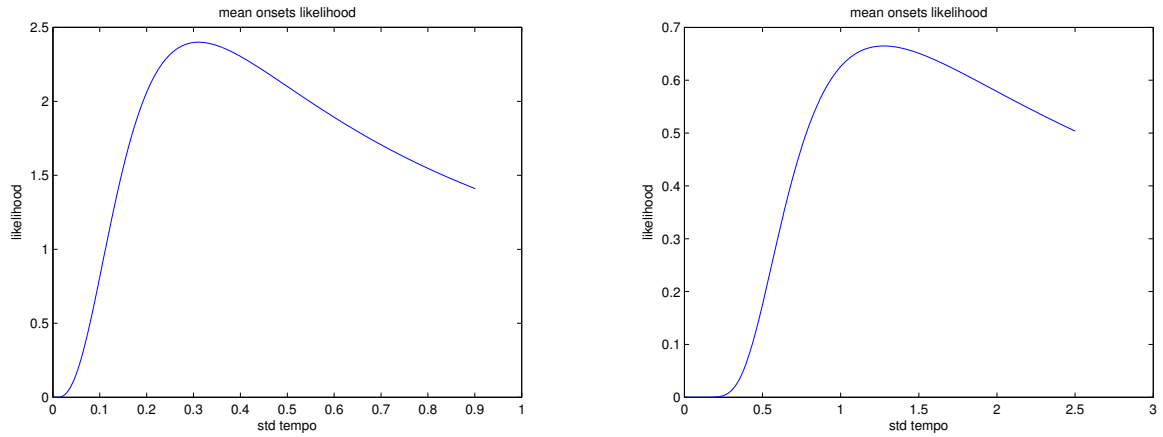


FIGURE 3 – Estimation du paramètre σ_s par maximum de vraisemblance sur le choral *Ach Lieben Christen* de J.S. Bach. À gauche, interprétation conforme à la partition. À droite, interprétation avec points d'orgue à la fin des quatre phrases.

2.2.2 Enjeu du décodage de la position courante

Une fois la distribution discrète propagée, reste l'étape cruciale d'estimation de la position. Peu de choses sont dites au sujet de cette étape qui pourtant ne va pas de soi. Celle retenue par Montecchio est immédiate l'estimation EAP, aussi appelé MMSE. Nous allons rapidement décrire ses atouts avant d'en exposer les faiblesses.

L'estimation EAP L'estimateur EAP choisit l'espérance de la distribution filtrante, approché par le barycentre des particules $\{x_i, w_i\}$.

$$x_t^{EAP} = \mathbb{E}[p(x_t \mid y_{1:t})] \approx \sum_i w_i x_i$$

Elle donne une trajectoire d'estimation lissée, robuste aux aléas de performances. En contrepartie, sa précision instantanée est mauvaise. Ce fait la rend peut apte à détecter les instants précis d'attaque de note, et induit une inertie dans sa réponse aux variations de tempo, qu'elle a tendance à lisser.

Défauts pratiques de l'estimation EAP La figure 4, illustre un défaut du comportement de l'estimation EAP, qui se produit systématiquement sur les événements suffisamment longs, comme la note tenue au début du choral, et les situations où la vraisemblance est très discriminative. Pendant la durée d'une note, les particules ont toutes même vraisemblance, donc le paquet de tempo s'étale jusqu'au passage des particules de tempo rapide sur la note suivante ; à ce moment-là, la vraisemblance de ces dernières s'effondre, ce qui tire le tempo moyen vers le bas. Le phénomène a ensuite lieu en sens inverse lors de l'observation de la seconde note. Le résultat est une estimation systématiquement en dent de scie autour de la vraie position.

Un deuxième problème survient lorsque la densité filtrante $p(x_t \mid y_{1:t})$ est **multimodale**, c'est-à-dire qu'elle comporte plusieurs lobes autour de positions éloignées. Cela arrive dès que des particules atteignent des motifs mélodiques identiques mais éloignés dans la partition.
Le deuxième article [13] de Montecchio évoque à raison l'avantage de l'inférence approchée sur celle exacte. Il est ici facile de supposer la **position de départ inconnue**, en disséminant les particules

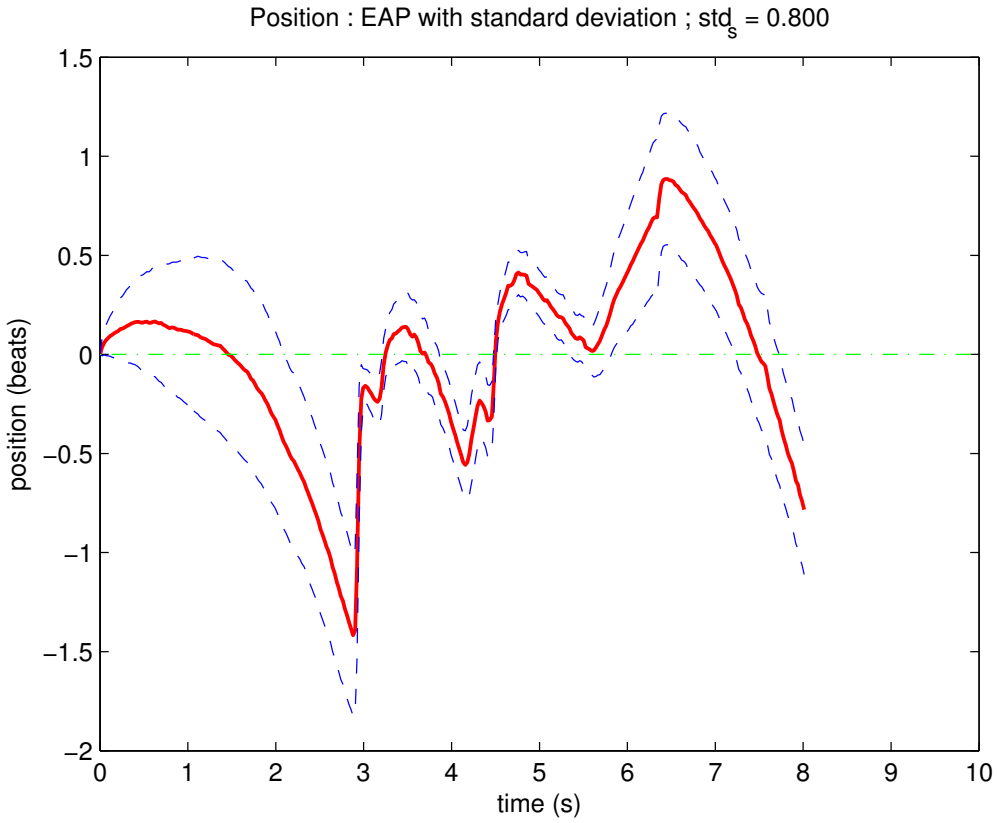


FIGURE 4 – Suivi de position du début du choral *Christ, der du bist Tag und Licht* (voix d'alto seule) de J.S. Bach. Erreurs de position de l'estimation EAP, et écart-type de l'ensemble des particules.

initiales sur la partition. Mais le tri ne s'opèrera pas tant que la vraisemblance de certains points de départ reste forte. Dans la figure 5, nous avons suivi un choral de J.S. Bach en répartissant les particules initiales sur les débuts des quatre phrases musicales qui le composent. Comme la deuxième répète la première, l'estimation EAP est catastrophique durant toute la durée de celle-ci.

Défauts méthodologiques de l'estimation EAP Cette dernière situation met bien au jour le **problème conceptuel** que pose le choix du barycentre des particules, qui "agrège" des chemins logiquement incompatibles. Ainsi, deux points de départs distincts ce sont des éventualités qui s'excluent mutuellement et ne participent à la même estimation. Notre solution consiste à emprunter au suivi RFS l'idée de *track management*, et de donner aux particules le numéro de leur hypothèse de départ. L'estimation se déroule en deux temps : d'abord, sommer les poids w^i hypothèse par hypothèse pour déterminer celle la plus probable ; ensuite, réaliser l'estimation EAP sur le paquet de particules de l'hypothèse la plus vraisemblable. La figure 5 montre le succès de cette estimation en deux temps sur le choral choisi.
Si cette solution est efficace, elle peut instaurer une alternance rapide entre deux hypothèses initiales équiprobables. Une application plus générale demanderait davantage d'heuristiques de *track management*, notamment un modèle de la **dynamique de changement hypothèses**.

Par ailleurs, un problème provient du fait que l'estimation EAP sur des espaces continus ne correspond pas aux méthodes employés sur les modèles discrets. Ceux-ci utilisent principalement l'algorithme de Viterbi pour trouver le chemin $x_{0:t}$ le plus vraisemblable, ou l'algorithme *forward* pour trouver l'état courant le plus vraisemblable. L'estimation EAP pose donc un problème de **protocole de comparaison** des performances entre espaces discret et continu.

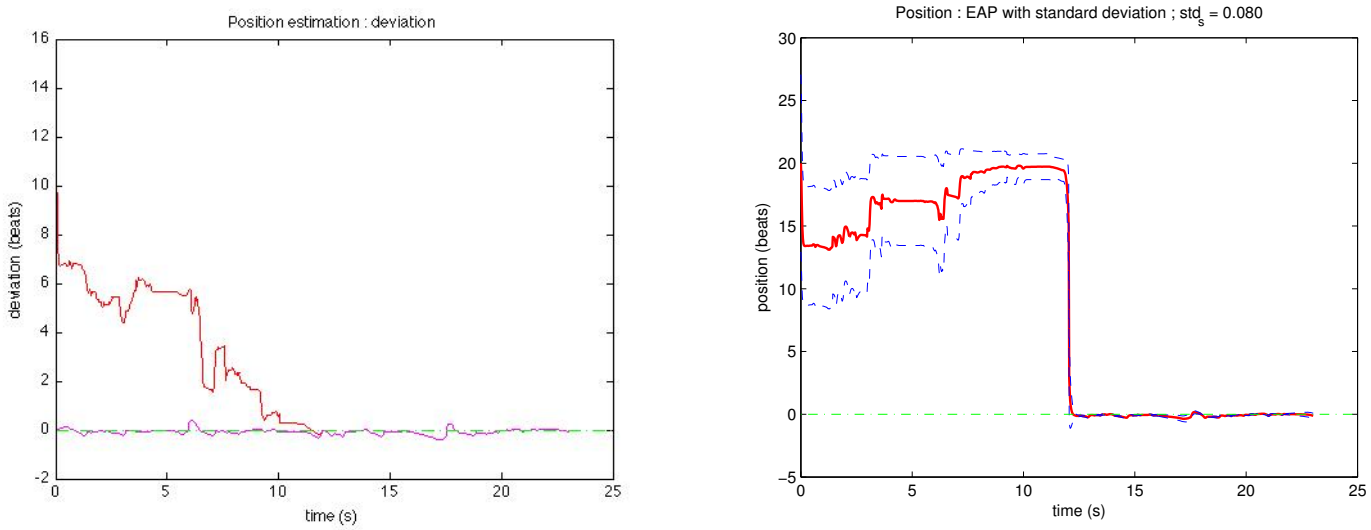


FIGURE 5 – Suivi de position du début du choral à quatre voix *Ach Lieben Christen* de J.S. Bach, avec quatre hypothèses équiprobables de point de départ. À droite, erreur de l'estimation EAP brute et écart-type de l'étalement des particules. À gauche en rose, erreur de l'estimation EAP sur l'hypothèse la plus forte.

2.2.3 Enjeu de l'exploration efficace de la combinatoire

Vers un modèle hybride discret-continu ? Dans ces conditions, le choix d'un modèle continu est-il pertinent ? Le problème de l'estimation le met en doute, et nous allons aborder ici une deuxième raison qui la prolonge. Les particules d'un tel filtre représentent chacune une hypothèse de chemin d'exploration de l'espace des chemins possibles. Pour garder une estimation précise, il est essentiel d'explorer plusieurs directions à partir des positions très probables.
Qu'est-ce qui garantit une exploration efficace de l'espace d'état ? Tout d'abord, l'élagage des chemins sous-optimaux ; lorsque deux chemins se rejoignent, le moins probable peut être définitivement oublié. Ce principe de programmation dynamique, qui donne à l'inférence sur des modèles markovien discrets toute son efficacité, n'a pas lieu sur un espace continu où presque sûrement les chemins ne se croisent jamais. Ces deux raisons plaident en faveur d'un modèle semi-Markovien, tels qu'utilisé dans Antescofo [3], ou la notion d'oubli de chemins sous-optimaux est présente.

À titre de prospective, nous nous sommes intéressé au concept de **stratification**, méthode de réduction de variance bien connue des statisticiens. En effet, la géométrie particulière de notre espace, dont les positions sont découpées en segments de notes, induit naturellement un pavage en strates. C'est pourquoi nous pensons qu'une étude plus approfondie sur le rapport entre modèle markovien discret et stratification d'un modèle markovien continu consistitue une piste de travail intéressante pour l'avenir.

Gestion des hypothèses par la rééchantillonnage D'or et déjà, nous avons utilisé cette idée de stratification en adaptant le *rééchantillonnage résiduel*. L'idée de cette méthode de réduction de variance est de garantir l'équirépartition des N_s tirages uniformes effectués pour remanier l'ensemble des particules. La meilleure variance empirique est atteinte en conservant $\lfloor w^i N_s \rfloor$ copies de chaque particules, puis en tirant aléatoirement celles restantes.
Notre analyse est que le principe de rééchantillonnage résiduel est à appliquer de manière hiérarchique, sur les groupes de particules correspondant à une même hypothèse. Il concourt à la gestion efficace des hypothèses car il conserve l'importance relative de chaque hypothèse et garantit un nombre min-

imal de particules pour chacune d'elle ; de cette manière, les hypothèses les plus fortes seront donc explorées finement, et les hypothèses faibles auront plus de chances d'être conservées. Dans notre algorithme, l'échantillonnage résiduel par hypothèses est utilisé pour deux niveaux hiérarchiques de groupage.

1. groupage des particules par leur hypothèses de position initiale (voire figure 5) ;
2. groupage des particules par micro-état polyphonique, puis rééchantillonnage résiduel.

2.2.4 Enjeu de la robustesse à l'asynchronie

Deux natures d'asynchronie Un système de suivi de position rencontre des difficultés pour suivre une musique polyphonique, car leur modèle suppose les voix parfaitement synchrones. Pourtant, de manière voulue ou non, les notes écrites simultanées peuvent être jouées en décalées. Mais **cette asynchronie existe aussi bien entre voix qu'à l'intérieur d'une voix**. En effet, une note écrite diffère d'une note jouée, qui elle-même diffère d'une note entendue. Du fait du mode de jeu, du temps d'extinction des sons, de la réverbération de la salle ou de la résonance de de l'instrument, **les durées jouées diffèrent des durées écrites**. Une note peut délibérément déborder sur la suivante ou s'arrêter avant leur enchaînement.

Cette incapacité des suiveurs de position provient de l'agrégation verticale des polyphonies. Par conséquent la vraisemblance chute durant toute la durée d'asynchronie, et la robustesse du système n'est garantit que si celle-ci est minime devant celle de synchronie. Les extraits musicaux à débit rapide de notes, ainsi que les polyrythmies sur différentes métriques, sont donc particulièrement problématiques pour les systèmes de suivi de position. (cf. l'exemple infra de la figure 2.3.2)

Possibilité d'extension multi-position de suivi Face à l'asynchronie entre voix, une extension immédiate de notre modèle de suivi de position consiste à **démultiplier par produit cartésien** l'espace d'état, en conservant un tempo global et en démultipliant les positions pour chacune des N voix.

$$E = \{\text{position globale}, \text{tempo}\} \subset \mathbb{R}^2 \longrightarrow \{\text{position 1}, \text{position 2}, \ldots, \text{position N}, \text{tempo}\} \subset \mathbb{R}^{N+1}$$

Une telle extension se heurte à la malédiction de la dimension qui rend prohibitif le calcul de l'inférence exacte sur les modèles discrets évoqués, car le modèle graphique perd son caractère gauche-droite. Des heuristiques d'élagage sont à trouver. C'est ce qu'effectue implicitement le filtrage particulaire, en discrétisant l'espace du même nombre de ses particules quel que soit sa complexité alors qu'en tout logique, ce nombre devrait croître avec la dimension. Nous avons implémenté un suivi multi-positions, mais n'avons pas obtenu des résultats significativement meilleurs sur nos chorals de Bach. Toutefois, d'autres essais restent à faire sur des enregistrements dont l'asynchronie entre voix est bien identifiée.

2.3 Proposition d'observation multi-objets MIDI

Intérêt de l'observation MIDI Afin de disposer d'une observation fondamentalement multi-objets, nous nous sommes intéressés au suivi MIDI. En effet, **le signal MIDI produit un RFS** ; il fournit à chaque instant t un ensemble $Y_t = \{y_1, y_2, \ldots\}$ de *pitchs*, en nombre variable.

La conception d'un modèle génératif pour le signal MIDI n'est en réalité pas simple. En effet, comme cette observation non-bruitée décrit exactement l'état courant, la vraisemblance $g(\text{note entendue} \mid \text{note écrite})$ devrait être à valeur binaire. Ce choix garantit le suivi très précis d'une performance conforme à la partition, mais à la moindre déviation d'interprétation la vraisemblance du vrai chemin risque de chuter à zéro. Ainsi, écouter le signal MIDI déporte tout l'enjeu de

conception sur la manière d'affecter des probabilités aux déviations d'interprétation, afin de rester robuste et pertinent du point de vue musicologique.

Un exemple d'étude : la *Fantaisie-Impromptu* Le cas d'étude qui a motivé notre travail est la *Fantaisie-Impromptu op.66* de Frédéric Chopin. La partition condense les risques d'asynchronie : des doubles croches à la main droite se superposent à des arpèges en sextolets de la main gauche, des accents syncopés incitent au décalage des mains, et le tout se joue à un tempo très rapide. De plus, l'asynchronie au sein d'une même voix est souvent prononcée sur les nombreux arpèges.

Du point de vue musical, l'accompagnement joué par la main gauche est une ligne monophonique au rythme et au phrasé très réguliers, et jouée avec moins de *rubato* que la main droite. Nous pouvons donc dire que son chemin dicte celui global et proposons par suite de décomposer les notes d'une partition de la manière suivante.

- une **voix principale**, composée de la séquence de notes musicalement déterminante pour la pulsation, qui sa causalité dicte celle de la performance globale de l'interprétation. C'est par rapport à elle que s'estiment la position et le tempo courant ;
- une **voix secondaire**, regroupant toutes les autres notes. Nous les supposons jouées avec une causalité plus floue, mais avec une asynchronie bornée par rapport à la voix principale.

Ainsi, notre recherche de modèle est motivée par l'intuition que le suivi rigoureux de position n'est utile que sur la séquence des notes significatives, et qu'une **relaxation des notes non-significatives** rend le suivi plus aisé et plus robuste.

2.3.1 Proposition de modèle d'observation MIDI multi-objets

RFS multi-Bernoulli pour la vraisemblance des notes attendues Nous nous sommes intéressés à l'utilisation d'un RFS multi-Bernoulli pour relier l'observation à l'attente par un modèle purement génératif. Les notes attendues composent les modes d'un $\{(r_t^{(i)}, g^{(i)}(\cdot)\}_{i=1}^N$, dont on considère l'observation $Y_t = \{y_1, y_2, \ldots\}$ comme un tirage aléatoire. Ce modèle permet de représenter chaque note attendue par deux paramètres.

- $r_t^{(i)}$ est la probabilité de détection, i.e. la probabilité que le pitch *attendu* soit dans le signal ;
- $g_t^{(i)}(\cdot)$ est la fonction de vraisemblance de l'image conditionnellement à l'état de la note.

La densité de probabilité RFS d'un multi-Bernoulli à N notes attendues s'écrit de la manière suivante :

$$g_{\text{Attente}}(Y) = \begin{cases} \prod_{i=1}^{N}(1 - r^{(i)}) \times \sum_{1 \leq i_1 < \ldots < i_n \leq N} \prod_{j=1}^{n} \frac{r^{(i_j)}}{1-r^{(i_j)}} g^{(i_j)}(y_j), & Y = \{y_1, \ldots, y_n\} \\ 0, & |Y| > N \end{cases}$$

Dans le cas de l'observation MIDI idéale, l'absence d'ambiguïté sur la hauteur incite à utiliser une vraisemblance $g^{(i)}$ binaire, valant 1 si et seulement si la hauteur observée est celle attendue ; de même, on est incité à définir $r^{(i)}$ sont comme fonction de la position de la voix principale $r^{(i)} := r^{(i)}(x)$, valant identiquement 1 sur l'intervalle de position de la note. Mais cette configuration écarte toute possibilité de déviation d'exécution, car la vraisemblance d'une observation où le *pitch* attendu est absent chute alors à zéro.

Le multi-Bernoulli pour la relaxation en position L'avantage de ce modèle d'observation est de permettre une relaxation fine des la position d'attente des notes, car il suffit de jouer valeur $r^{(i)}$. Plus précisément, la grandeur significative est le quotient $\frac{r^{(i)}}{1-r^{(i)}}$, car il représente l'importance de la situation où le *pitch* attendu est joué par rapport à celle où il est absent.

Notre première idée de relaxation de la voix secondaire est de définir une **fenêtre de position** autour des débuts et fin de notes, pendant laquelle la certitude est fixée est à $r^{(i)} = 1/2$. La figure 2.3.1 l'illustre par un exemple très simple. En effet, cette valeur de r crée une situation où la vraisemblance d'observer le *pitch* de la note est équiprobable à son absence, et provient des chutes drastiques de vraisemblance. Ainsi, nous diminuons les qualités discriminatives du modèle au profit de sa robustesse à l'asynchronie, dans un compromis décidé par La longueur de la fenêtre de position utilisée.

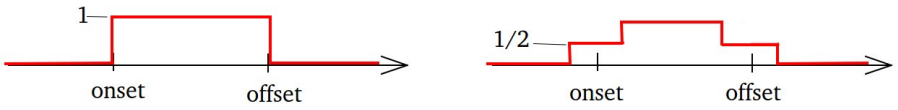


FIGURE 6 – Exemple-jouet de probabilité de détection $r^{(i)}$ en fonction de la position. À gauche, note de la voix principale. À droite, note relaxée de la voix secondaire.

RFS de Poisson pour la vraisemblance du clutter L'emploi d'un modèle multi-objet permet de dissocier le traitement des notes surnuméraires par rapport à celles attendues, en scindant l'ensemble des *pitchs* observés $Y_t = \Theta(X_t) \cup K_t$ en deux RFS indépendants. Sous cette hypothèse la vraisemblance s'écrit comme un produit de convolution sur les parties de l'observation.

$$g(Y_t) = \sum_{\tilde{Y} \subset Y_t} g_{\text{Attente}}(\tilde{Y}) g_{\text{Clutter}}(Y_t - \tilde{Y})$$

Dans notre cas d'observation non-ambiguë sur la hauteur, la combinatoire se réduit sur l'ensemble des *pitchs* observés inclus dans les hauteurs attendues.

Un RFS de Poisson est un choix intéressant pour modéliser le clutter. En effet, s'il est d'intensité λ et de densité spatiale $\kappa(\cdot)$, sa densité de probabilité RFS s'écrit.

$$g_{\text{Clutter}}(Y) = e^{-\lambda} \prod_{y \in Y} \lambda \kappa(y)$$

L'écriture de la densité d'observation se simplifie.

$$g(Y_t) = e^{-\lambda} \sum_{\tilde{Y} \subset Y_t} \frac{g_{\text{Attente}}(\tilde{Y})}{\prod_{y \in \tilde{Y}} \lambda \kappa(y)}$$

Cette expression permet une **interprétation du choix** de la loi de Poisson. Chaque observation non-attendue z_k multiplie a probabilité du chemin d'un facteur $\lambda/\kappa(z_k)$. Par conséquent chaque note non-attendue pénalise la vraisemblance de la même façon, quel que soit le nombre de notes correctement jouées, et ce à chaque pas de temps.
La pénalisation est donc clairement quantifiée. La calibration du paramètre d'intensité λ est donc possible par une étude statistique, comme il correspond au nombre moyen d'erreur par pas de temps ΔT.

2.3.2 Résultats obtenus

Suivi par observation multi-objets Pour valider notre modèle d'observation, nous avons réutilisé l'algorithme d'inférence par filtrage particulaire sur un espace continu position-tempo

$x = (s, t)$, en changeant la fonction de vraisemblance pour celle décrite plus haut $g(Y_t)$, où l'attente est l'ensemble des notes actives à la position s considérée. Nous avons utilisé une transcription extrêmement fidèle de l'exécution par un pianiste concertiste de *Fantaisie-Impromptu*[2] de F. Chopin, et l'avons associé à la partition et supprimé les effets de pédale. La figure 2.3.2 présente le résultat du suivi des trente premières secondes de la performance, avec un nombre de particules $N_s = 2000$, un pas de temps $\Delta T = 20ms$, des écarts-type $\sigma_s = 0.05$ et $\sigma_t = 0.8$ et un seuil $N_{eff} = N_s/10$.

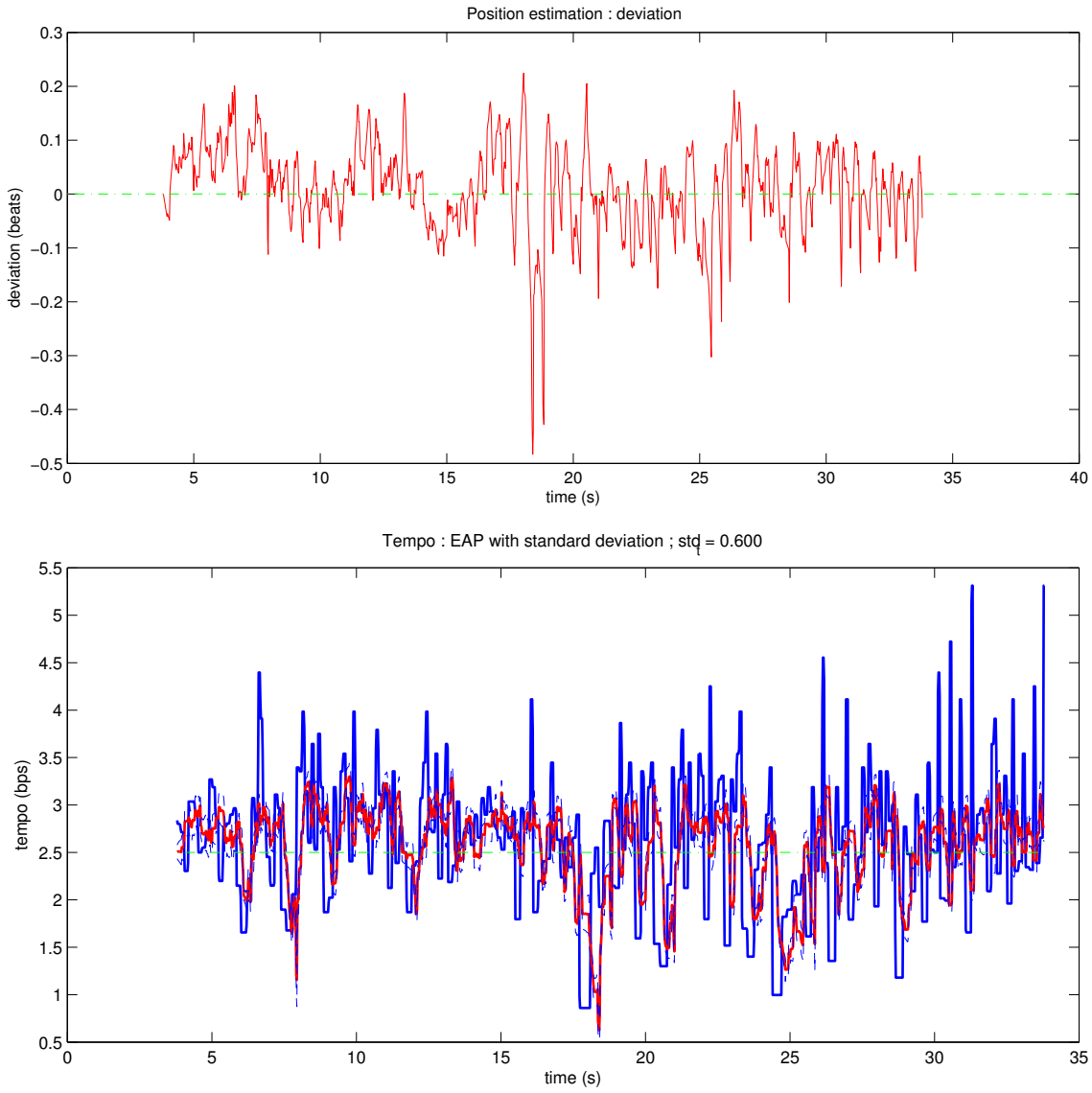


FIGURE 7 – Erreur d'estimation de la position par filtrage particulaire sur un modèle d'observation MIDI multi-objets. Trente premières secondes de la *Fantaisie-Impromptu*. En bas, courbe du vrai tempo instantané (en bleu) de la performance et du tempo décodé (en rouge).

Influence de l'asynchronie Pour jauger de l'effet de l'asynchronie, nous avons analysé l'évolution de la vraisemblance de la *ground truth*, c'est-à-dire la vraisemblance d'observation conditionnelle-

2. disponible sur http://www.classicalmidiconnection.com/chopin.html

ment au vrai chemin de position $t \mapsto g(Y_t \mid x_t^{\text{true}})$. Cette grandeur est très adaptée à l'observation des seuls effets de l'asynchronie car elle neutralise l'influence du modèle de transition.

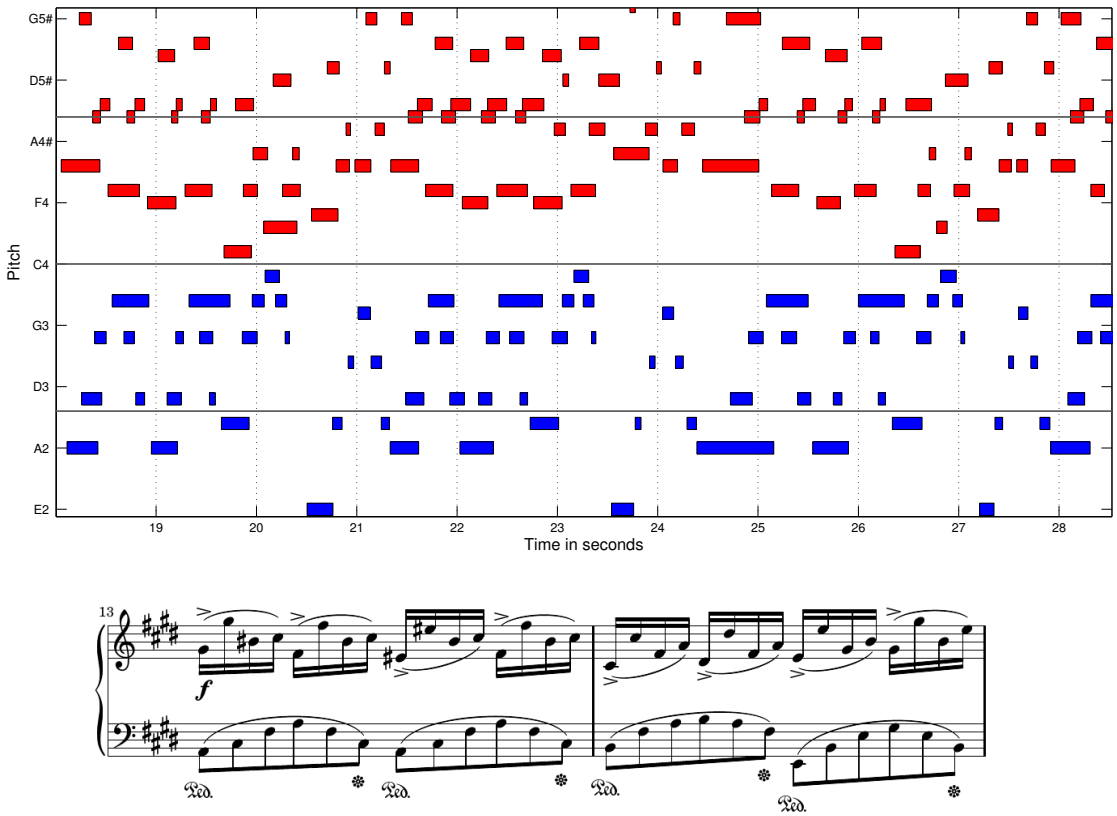


FIGURE 8 – Déviation entre l'interprétation et la partition sur un extrait de la *Fantaisie-Impromptu*, par comparaison entre *Piano roll* de la performance MIDI réelle et partition correspondante.

L'extrait du *piano roll* du fichier MIDI, présenté figure 2.3.2, illustre tous les cas d'asynchronie évoqués : des notes de triolets se chevauchent, d'autres sont raccourcies ; idem pour les double-croches de la main droite, qui sont déphasées par rapport à la main gauche.

La figure 2.3.2 montre la vraisemblance du vrai chemin dans deux situations. D'abord en n'observant que la voix principale (ici la main gauche) pour juger de sa vraisemblance seule ; chaque chute de vraisemblance est le reflet de l'asynchronie interne à la main gauche. Ensuite, les voix ensembles pour juger de la chute produite par l'asynchronie entre main droite et main gauche. On constate que la vraisemblance moyenne du chemin (en moyenne géométrique) diminue de plus de moitié.

Dans les deux simulations, la probabilité de détection r était à 0.95 pour toutes les notes pendant leur durée écrite, et à 0 ailleurs ; l'intensité du clutter était de $\lambda = 0.4$ et sa densité était uniforme à $\kappa \equiv 1$.

Enfin, La figure 2.3.2 montre l'effet de l'intensité du clutter sur la vraisemblance moyenne du vrai chemin. Main gauche seule, l'optimum est à 0.2, ce qui le nombre moyen de notes surnuméraires à chaque instant. Main droite seule, l'optimum monte à 1.2. On en conclut qu'à chaque instant en moyenne, une note de la main droite est asynchrone.

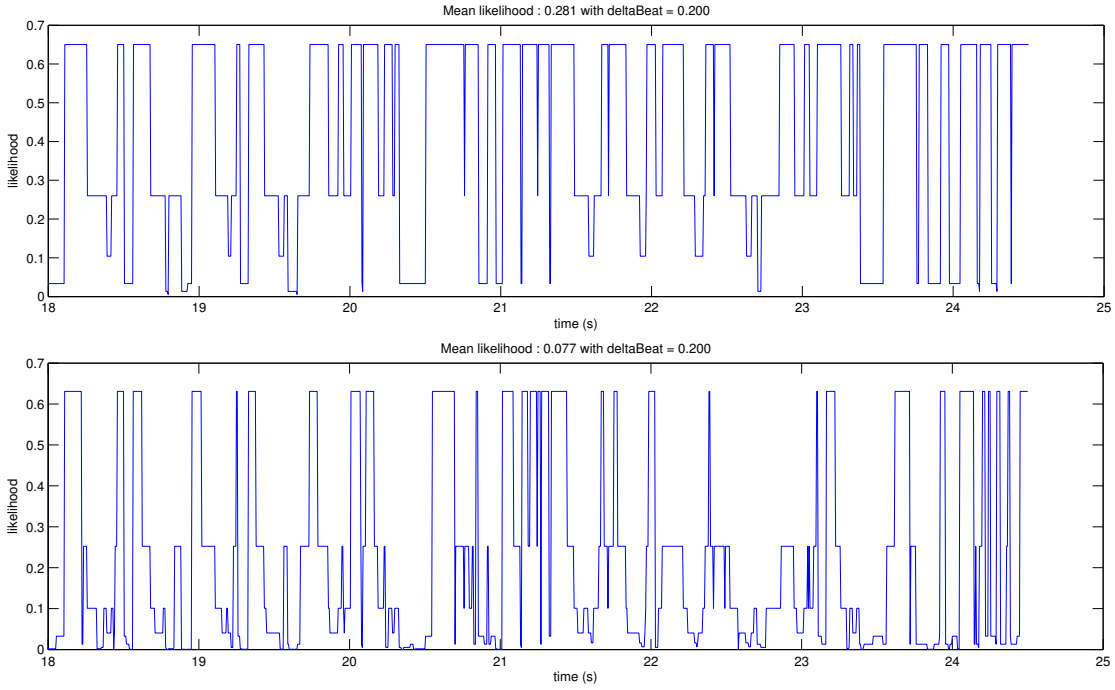


FIGURE 9 – Vraisemblance instantanée de la vraie position d'un extrait de la *Fantaisie-Impromptu* (cf. figure 2.3.2).

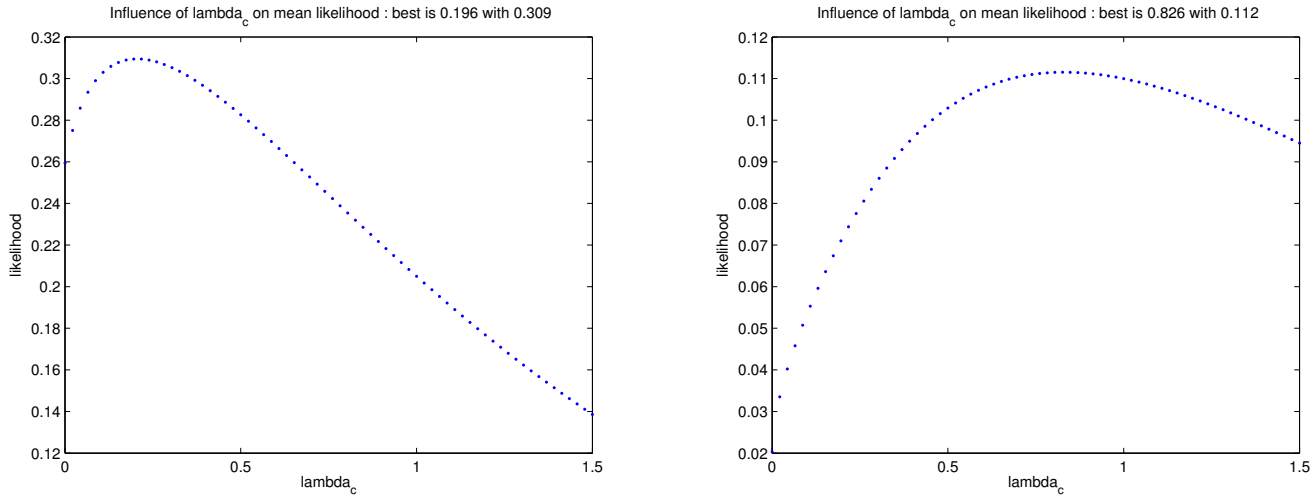


FIGURE 10 – Influence du paramètr eλ d'intensité du clutter sur la vraisemblance moyenne du vrai chemin, prise sur les trente premières secondes de la *Fantaisie-Impromptu*. À gauche, main gauche seule. À droite, mains ensembles.

3 Vers un modèle multi-cibles pour le suivi de partition

3.1 Quels objets musicaux modéliser ?

3.1.1 Modèles de cible : faiblesse du suivi de voix

Maintenant que nous avons manipulé l'observation modélisée par Random Finite Set et l'inférence par filtrage particulaire, il s'agit de concevoir un modèle d'état multi-cibles. Quel objet musical doit-on prendre pour cible, et quelle observation du signal musical est-il judicieux d'utiliser ? Telle est la question ouverte, auquel ce stage a essayé d'apporter une réponse.

Inadaptation des hypothèses RFS aux voix musicales La démultiplication qui nous vient naturellement à l'esprit est celle des voix musicales. Nous pouvons, pour représenter les voix d'une partition par différentes cibles, ajouter à l'espace d'état par une étiquette discrète.

$$E_s = \{(\text{position}, \text{tempo}, \text{voix})\} \subset \mathbb{R} \times \mathbb{R} \times \mathbb{N}$$

Toutefois, cette première idée, pourtant très naturelle, est à écarter car elle ne donne pas d'avantages par rapport à un mono-suivi sur un espace-produit. En effet, on peut résumer le suivi RFS comme un prolongement de celui mono-objet de manière à modéliser les phénomènes suivants :

- un nombre fluctuant de cibles parcourt simultanément le modèle markovien de transition ;
- une cible peut apparaître et disparaître ;
- l'observation est constitué d'un ensemble d'images, en nombre fluctuant ;
- l'image d'une cible peut être ponctuellement manquante ;
- des fausses images peuvent ponctuellement apparaître.

Ce comportement des cibles décrit mal celui des voix. D'abord, elles sont en nombre limité, déterministe, et fixé. Ensuite, une voix musicale ne vie et ne meurt pas ; **un silence ne traduit pas une disparition**, car pendant ce silence la position courante continue d'avancer. Aussi, un suivi de voix s'effectuera avec un nombre de cibles fixes.

Perte de l'intérêt des RFS par rapport au mono-suivi Or, si l'on considère un suivi à cardinalité constante, alors les autres phénomènes de la liste ci-dessus sont directement intégrables dans un modèle markovien mono-objet.

D'abord, la possibilité de disparaître et d'apparaître s'émule ajoutant à l'espace d'état un état-cimetière c (*cemetary state*) $E \longrightarrow E \cup \{c\}$, muni de transitions markoviennes de et vers de tout état x de E. Les probabilités associées à ces transitions donnent celles d'extinction $1 - p^S(x)$ et de naissance $p^\Gamma(x)$ introduites en suivi RFS.

Ensuite la probabilité de non-détection $p^D(x)$ s'émule en étendant l'état par un espace à deux élément $E \longrightarrow E \times \{0, 1\}$, et étendant le modèle de transition par une matrice stochastique de loi invariante $\begin{pmatrix} p^D(x) & 1 - p^D(x) \end{pmatrix}$.

Enfin, nous avons vu que l'extractions des cibles RFS repose la fonction d'intensité γ, très commode car définie sur E et non pas sur $\mathcal{F}(E)$. Son utilisation n'est efficace que si les cibles sont indifférenciées. Or ici, les fonction de transition et de vraisemblance seront différenciées voix par voix, et la situation de présence simultanée de cibles dans la même voix n'a guère de sens. En conclusion, toutes ces raisons nous ont conduit à faire le deuil de la modélisation des voix musicales par cibles RFS.

3.1.2 Modèle d'observation : faiblesse de l'observation directe

De plus, nous allons voir ici que l'utilisation du suivi RFS nécessite de changer de signal observé. L'observation la plus directe d'un signal sonore, qui consiste en l'une de ses représentations temps-fréquence comme le spectrogramme à court terme, nous éloigne du cadre d'utilisations des filtres RFS classiques. Ceux-ci sont en effet conçus pour un capteur multi-observations délivrant des ensembles d'images ponctuelles plutôt qu'une image étendue et globale.

De plus, les représentations temps-fréquence habituelles dérogent doublement aux hypothèses du suivi RFS sur l'observation. D'abord, les images des sons ne sont pas ponctuelles dans l'espace temps-fréquence, se superposent et interfèrent.

Ensuite, elles ne respectent pas l'hypothèse cruciale qu'une cible émette son image indépendamment des autres cibles. En effet, si un modèle génératif utilise représentation temps-fréquence, alors la vraisemblance d'un état x se calcule comme décrite section 2.1, à partir de template T généré comme la somme de ceux de chaque hauteur h_i présente dans l'état.

$$g(y \mid x = \{h_1, h_2, \ldots\}) := g(y \mid T(h_1) + T(h_2) + \ldots)$$

Remarquons toutefois, pour la prospective, qu'un autre schéma de fonction de vraisemblance rendrait possible l'utilisation d'une représentation temps-fréquence, par un filtre mono-observation est possible (cf. l'exemple de [**?**]) Il s'agirait de concevoir une probabilité qu'une représentation du signal "contienne" un template isolé, avec une relaiton fonctionnelle du type suivant.

$$g(y \mid \{h_1, h_2, \ldots\}) = g(T(h_1) \subset y) \cdot g(T(h_2) \subset y) \cdot \ldots)$$

Un tel modèle serait apte à être ainsi "séparable", en particulier lorsque la représentation choisie est supposée linéaire. Remarquons qu'un tel modèle rendrait très pertinent l'intégration d'un suivi de volume sonore de jeu, grandeur sans laquelle le concept de "contenir" un objet sonore n'a pas grand sens. D'ailleurs, un suivi de volume est intéressant pour le modèle actuel. En effet, celui-ci normalise le spectre du signal global avant et superpose les templates de chaque hauteurs en les considérant jouées au même niveau sonore, ce qui peut être grossièrement faux en contexte polyphonique. Nous espérons évaluer l'impact de cette approximation sur la qualité de discrimination de la fonction de vraisemblance lors d'un travail ultérieur.

3.2 Les RFS pour le suivi de notes

3.2.1 Motivations pour une observation indirecte de notes

La partie précédente a dressé le constat que les voix musicales ne sont pas les cibles à modéliser par un RFS. À défaut donc de pouvoir étendre le suivi de position étudié section 2, nous avons dû repartir sur une piste inédite et réfléchir à d'autres candidats de modèles. Cette réflexion nous a amenée à rapprocher les hypothèses du suivi RFS de la physique des sons d'une classe particulière d'instruments de musique, que nous décrivons maintenant.

Instruments "à sons indépendants" Considérons les instruments de musique "à jeu de notes" (instruments à claviers ou jeu de cordes, tels que le piano, le clavecin, la harpe, les claviers à percussions). En première approximation, le musicien ne contrôle l'évolution du son d'une note qu'à deux instants précis, sa production puis son étouffement. En première approximation également, la présence ou non d'autre note n'affectera pas sa résonance décroissante. En conclusion, les notes des instruments "à sons indépendants" évoluent conformément aux hypothèses du suivi RFS : **conditionnellement à son apparition et sa disparition, une note jouée au piano évolue**

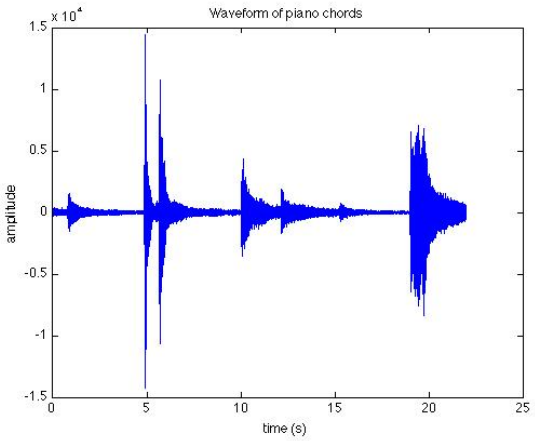


FIGURE 11 – Forme d'onde d'une suite d'accords joués au piano. Extrait de *Pluton* par P. Manoury.

indépendamment de celle des autres. De plus, étant donné ses paramètres d'attaque hauteur et vélocité, le son de chaque note suit le même modèle physique, raisonnablement approchable par un modèle markovien unique. La forme d'onde des notes de piano, retranscrite figure 11, laisse bien à croire à une évolution temporelle déterminée par les seules conditions d'attaque.

Nécessité de l'observation indirecte pour le suivi de notes En conclusion, cette situation permet d'envisager un **suivi de notes**, où le RFS des ibles x_i modélise les notes elles-mêmes.
Le principal défi consiste à trouver une observation multi-objets des notes individuelles et non pas du son global. Il apparaît donc que les RFS sont de bien meilleurs candidats pour l'emploi d'une *observation indirecte*. Nous entendons par là une estimation préalablement menée sur un descripteur du signal plutôt qu'une fonction directement appliquée sur la forme d'onde. Certaines observations indirectes, telles que le chromagramme, ont déjà été employées pour le suivi de position ([5]).

3.2.2 Les RFS pour l'estimation multi-F0

Un langage pour la dynamique des hypothèses La restriction à des instruments à jeu de notes tel que le piano laisse donc bien présager l'emploi d'un estimateur de fréquences fondamentales multiples. D'une manière générale, nous voyons dans les RFS un cadre mathématique prometteur pour l'exploitation des descripteurs MIR (*Music Information Retrieval*). En effet, l'écueils des systèmes utlisant un descripteur est de supposer qu'à chaque instant, son estimation exite et est unique. Or plusieurs valeurs, ou parfois aucune, peuvent être de bons candidats. Ainsi, nous pensons que **l'objet mathématique le plus à même de représenter la sortie d'un estimateur est un RFS**, c'est-à-dire un ensemble fini d'hypothèses co-existant à chaque instant.

Aussi, pour disposer d'une observation indirecte, nous sommes intéressés à l'estimateur multi-F0 développé par l'équipe [4]. Par son emploi d'une décomposition NMF (*Non-negative Matrix Factorization*), il produit en réalité une **décomposition sur un dictionnaire**, dont chaque mot représente le signal d'une note. Il fonctionne ainsi spécifiquement sur les instruments à jeu de notes fixes que nous avons décrit plus haut, Plus précisément, la décomposition NMF fournit des cœfficients d'activation, qui s'apparentent à des niveaux sonores composantes du signal réel.
L'enjeu de la recherche en NMF est de trouver les bonnes normes et fonctions de régularisation qui induisent une estimation sparse, soit dans la dimension des fréquences (nombre de notes apparaissant à chaque instant), soit dans celle temps (fréquence d'apparition des notes). le passage des cœfficients à des notes n'est jamais direct et nécessite un **post-traitement**. La figure 3.2.2 montre à quoi peut ressembler l'estimation sur une suite de croches simples.

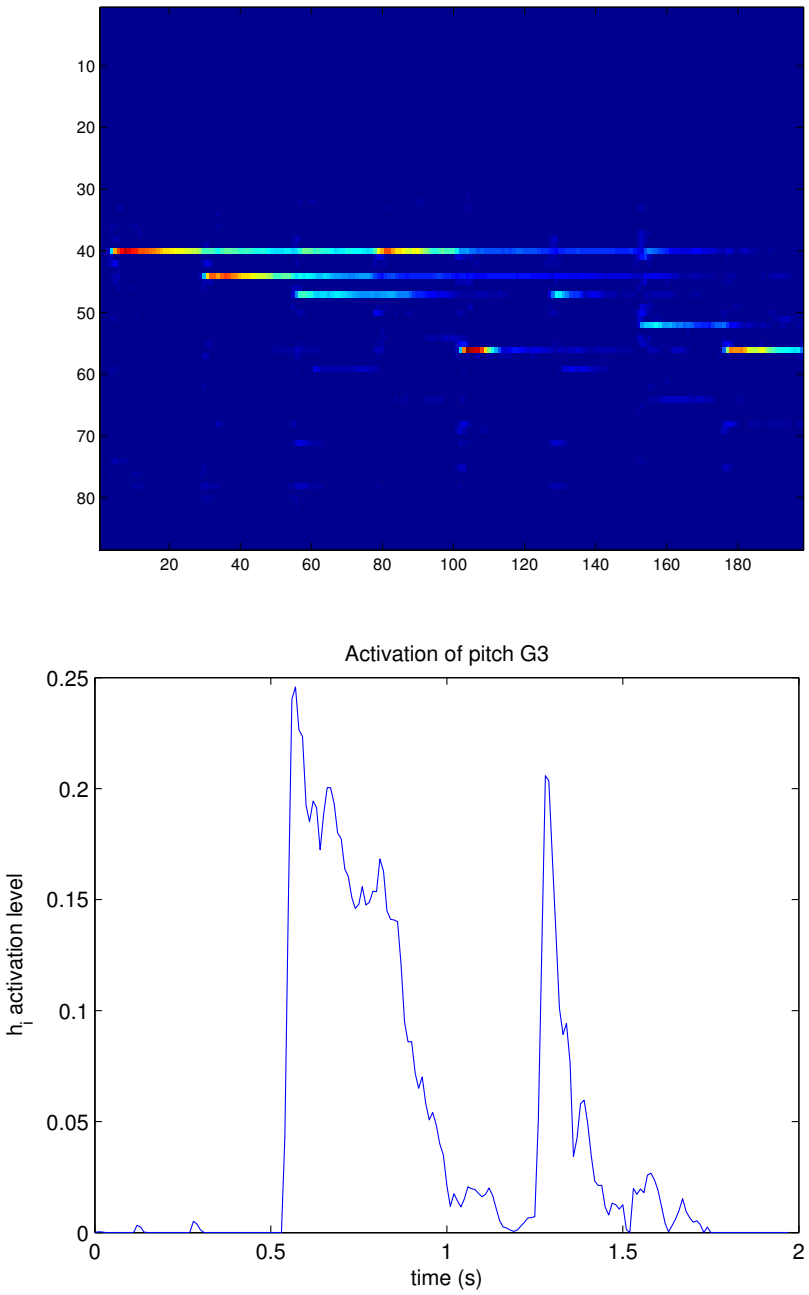


FIGURE 12 – Estimation des fréquences fondamentales sur les deux premières seconde du premier prélude du *Clavier bien tempéré* de J.S. Bach. En haut, ensemble cœfficients d'action. En bas, courbe du cœfficient de la note sol^3.

Le post-traitement d'un estimateur MIR est rarement bien détaillé, à cause de la dépendance au contexte d'utilisation, et du manque de formalisme commun pour exprimer les heuristiques possibles. Notre opinion est que les RFS pourraient apporter un langage commun pour traduire le post-traitement en terme de dynamique des hypothèses induites par l'estimation. En effet, une estimation évolue suivant **deux échelles de temps** distinctes, et le suivi RFS apporte les outils pour modéliser cette double dynamique.

– à l'échelle locale, les hypothèses évoluent individuellement ; cette dynamique est représentée par la fonction de transition $f(x_{t+1} \mid x_t)$;

– à l'échelle globale, de nouvelles hypothèse apparaissent et des anciennes sont abandonnées. Cette enchaînement est modélisé par le processus de naissance Γ_t et la probabilité de disparition p_t^D.

Exemple : heuristiques d'un estimateur multi-F0 Ainsi, les heuristiques simples choisies par les auteurs de [4] peuvent s'exprimer en langage RFS. Nous les citons ici.
– une note de *pitch* i se crée si son niveau h_i dépasse un seuil H_{on} durant une durée de N_{on} échantillons consécutifs ;
– cette disparaît si son niveau repasse en dessous d'un autre seuil H_{off} durant une durée de N_{off} échantillons consécutifs.

Ces hypothèses peuvent aisément s'exprimer par du RFS, en utilisant des observations binaires $(h_i > H_{on}, h_i < H_{off})$. Elles invitent à utiliser comme espace d'état E_s une chaîne de Markov gauche-droite composées N_{on} micros-états pour l'*onset*, de N_{off} micro-états pour l'*offset*, et entre les deux un état semi-markovien dont le temps d'occupation suit une loi uniforme afin de refléter l'absence d'*a priori*.

Utilisation probabiliste L'intérêt de ce modèle d'estimation par NMF réside dans l'évolution des cœfficients d'activation h_i. L'observation de la figure 3.2.2 permet de distinguer une esquisse des trois phases constitutives de la vie d'une note.
– phase d'*onset* : augmentation rapide de h_i ;
– phase de sustain : décroissance lente de h_i ;
– phase de release : chute rapide de h_i.

Bien évidemment, ces tendances sont très bruitées et parfois inexactes. C'est pourquoi l'adjonction d'un algorithme de suivi produirait le filtrage nécessaire. Nous laissons pour la prospective l'utilisation de cet estimateur comme signal observé et nous retournons à l'utilisation du signal MIDI.
En effet, le MIDI constitue l'observation idéale, exacte et non-ambiguë, des phase de la vie d'une note. En effet, l'information qu'il contient ne se réduit pas à la seule donnée instantanée des *pitchs* présents. En effet, tout *pitch* MIDI reçu peut être sans ambiguïté identifié comme un *onset* ou un *offset*, et sinon dans quel intervalle *onset*-*offset* il s'inscrit. Ainsi, le signal MIDI renseigne aussi sur la nature de chaque *pitch* (*onset*, sustain ou *offset*), induit ainsi une information, la délimitation temporelle des notes jouées, qui exprime une échelle de temps supérieure. **Capter cette échelle de temps constitue l'ambition de notre modèle de suivi**.

3.3 Proposition de suivi de notes par observation indirecte

Dans cette section, nous présentons un modèle qui est le fruit de notre réflexion sur l'utilisation des RFS, mais qui n'a pas été encore implémenté. Sa motivation majeure est de gérer l'asynchronie entre voix. Il considère pour cela une voix principale et une voix secondaire afin de mener une inférence de position et de tempo qu'à partir de cette première.

3.3.1 Enjeux de la causalité

Causalité d'une note La faiblesse de notre modèle à espace continu de position est d'être insensible à l'enchaînement causal nécessaire que vit chaque note, résumé par la succession des trois phases *attack* → *sustain* → *release* évoquées précédemment. Ainsi, pour lui, l'observation d'une

hauteur apparaissant et disparaissant à chaque pas de temps 1010101 . . . est aussi probable qu'une phase de présence suivie d'une phase d'absence de cette hauteur 111 . . . 000 . . .. Or ce deuxième exemple est le seul qui forme une évolution plausible d'une note.
L'intérêt de capter cet enchaînement de phases est de **réduire la combinatoire** des chemins possible Pour cela, l'idée est de modéliser l'évolution d'une note individuelle par une chaîne de Markov gauche-droite, telle que montrée figure 3.3.1, et de concevoir des fonctions de vraisemblance distinctes pour chacun de ses états. Ce type de modèle simple est repris par plusieurs algorithmes de la littérature [18, 6], dont celui anciennement développé à l'Ircam [18] auquel nous allons nous référer par la suite.

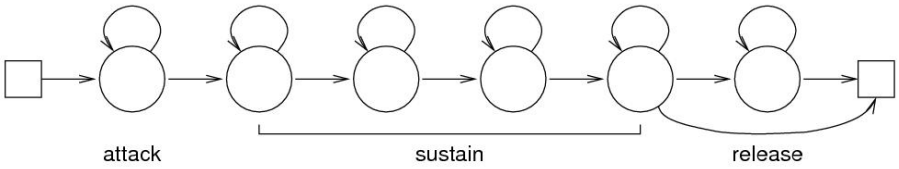


FIGURE 13 – Subdivision d'un note en une chaîne de Markov gauche-droite. Figure tirée du suivi MIDI de l'Ircam.

Causalité(s) d'une partition L'intégration du modèle graphique d'une note dans un modèle graphique de la partition constitue le principe de base d'un système hiérarchique. Dans un modèle qui agrègent les partitions polyphoniques (cf. figure 1), cette intégration se fait par concaténation. Ainsi, la chaîne de Markov élémentaire 3.3.1 ne représente plus une note mais un micro-état polyphonique.

Nous critiquons cette concaténation des modèle graphiques élémentaires en une longue chaîne gauche-droite, car celle-ci impose une **causalité séquentielle** entre les notes. Ainsi, l'*onset* d'une note est nécessairement attendu après l'*offset* de la précédente ; les *onsets* d'un accord sont attendus simultanément ; **la durée physique d'une note jouée est confondue avec la durée musicale de la partition**. Cette causalité n'est pas tenable. Même en contexte monophonique, le phénomène de résonance vient la mettre à mal. Aussi, nous désirons garder la causalité de chaque note, sans la fragmenter. Nous considérons donc un modèle à chaînes simultanées, synchronisées par l'enchaînement de notes écrit sur la partition. La différence de conception entre le modèle classique et notre proposition est montré figure 3.3.1.

Dans un tel modèle, l'inférence globale et unique de la position est accompagnée d'inférences locales et multiples sur chaque note de la partition. L'outil du suivi RFS permet cette inférence multiple et simultanée. En effet, la co-existence de deux échelles de temps distinctes, celle de la trame et celle de l'événement, se retrouve dans deux de ses notions mathématiques :

- à l'échelle locale, la fonction de transition $f(x_{t+1} \mid x_t)$. Elle prédit la dynamique individuelle de chaque événement attendu, ici le devenir d'une note attendue, qui est d'être jouée, de résonner puis de s'éteindre.
- à l'échelle globale, le processus de naissance Γ_t et la probabilité de disparition p_t^D. Elles prédisent l'enchaînement macroscopique des événements. Nous les regroupons dans le concept le *processus d'attente*, qui génère et détruit les cibles que sont les *notes attendues*, représentées par leurs caractéristiques écrites sur la partition.

$$n_i \in \{\text{hauteur} \times \text{position attendue} \times \text{durée attendue}\}$$

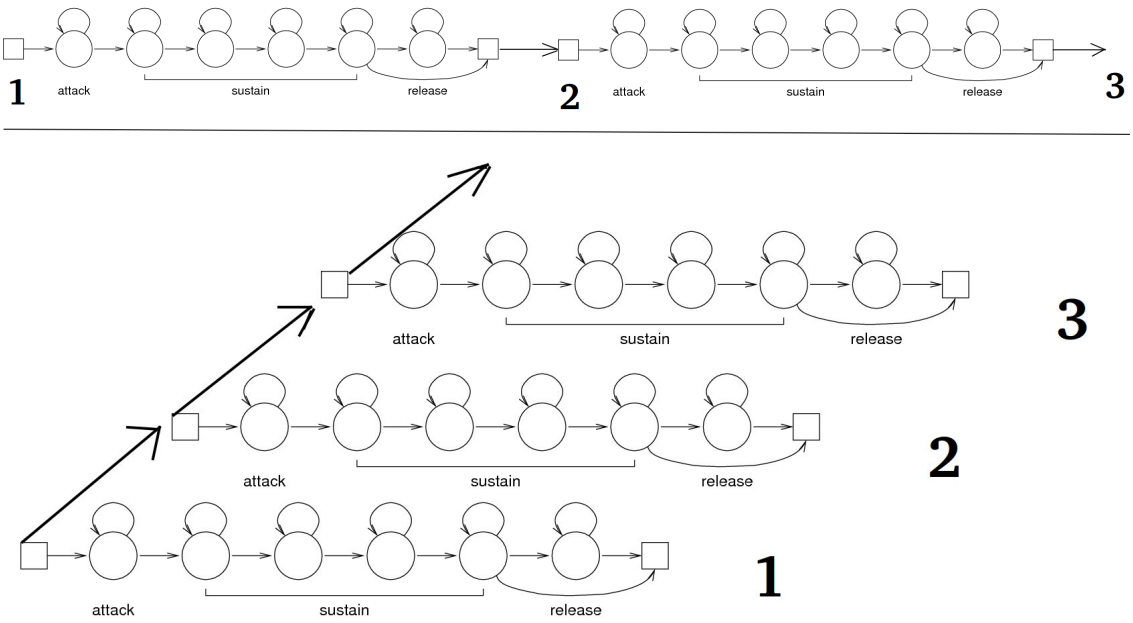


FIGURE 14 – En haut, modèle graphique classique de la partition par concaténation en une chaîne gauche-droite, avec inférence unique. En bas, proposition de modèle graphique par chaînes parallèles, avec inférences multiples.

3.3.2 Modèle d'évolutions des notes

Évolution à l'échelle globale La couche globale consiste à gérer l'apparition et la disparition des notes attendues. L'étude des filtres RFS classiques nous a montré l'écueil majeur de l'inférence multi-objets, à savoir la démultiplication des hypothèses considérées. Celle-ci a notamment lieu à chaque instant où la probabilité de naissance $p_t^{\Gamma}(\cdot)$ n'est pas identiquement nulle.

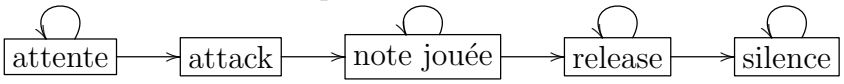
C'est pourquoi notre but est de localiser dans le temps le processus de naissance Γ_t et celui de disparition. Pour cela nous astreindrons notre suivi aux performances respectant l'**hypothèse d'exécution linéaire** (la position inférée x_t est continue et croissante dans le temps). Cette hypothèse garantit une **causalité** simple : si une note attendue car écrite sur la partition, alors elle sera soit manquée, soit jouée puis relâchée dans un horizon restreint autour de la position prévue. L'attente peut donc être localisée sur une fenêtre de position autour de celle écrite sur la partition. Ainsi, pour chaque note i, de position p_i et de fenêtre $[p_i^{min}; p_i^{max}]$, nous pouvons représenter le processus d'attente comme suit.

$$\text{abandon d'attentes} : \Gamma_t(x_t) = \Gamma_{t-1}(x_{t-1}) - \{n_i \in \Gamma_{t-1}(x_{t-1}) \mid n_i \text{éteinte ou } x_t > p_i^{max}\}$$
$$\text{démarrage d'attentes} : \Gamma_t(x_t) = \Gamma_{t-1}(x_{t-1}) \cup \{n_i \in \text{partition} \mid x_t \in [p_i^{min}; p_i^{max}], x_{t-1} < p_i^{min}\}$$

Ainsi, nous supposons que l'inférence de position puisse être menée efficacement même sur un modèle graphique **construit de manière itérative**, au fur et à mesure du décodage de la position. Cette supposition, indispensable pour garder une faible complexité calculatoire, traduit simplement le fait qu'une exécution musicale normale suit linéairement la partition.

Évolution à l'échelle locale Considérons une note attendue repérée par j, de hauteur h_j, de position attendue p_j et de durée attendue d_j. Elle suit nécessairement l'enchaînement causal décrit, que l'on cherche à modélise par la chaîne de Markov suivante. Par rapport à celle de la figure 3.3.1, nous rajoutons un état initial d'attente qui permet la relaxation de la position p_j sur un intervalle

de position $[p_j^{min}; p_j^{max}]$ autour de p_j. Le fait d'ajouter cet état d'attente est capital en terme de réduction de combinatoire, car son absence nécessiterait que le processus de naissance émette l'hypothèse d'apparition de la note à chaque instant.



Ainsi, à chaque instant, le système infère la probabilité de présence sur les états discrets markoviens :

$$x_t^j \in \{\text{attente}, \text{attack}, \text{sustain}, \text{release}, \text{silence}\}$$

Pour cela, deux types de probabilités sont à définir conditionnellement à ces états, celles de transition et d'observation. Celles-ci prennent une forme très simple dans le cas de l'observation MIDI. En effet, comme le signal MIDI est non-bruité, les micro-états sont incompatibles conditionnellement à l'observation. Les probabilités d'observation sont donc élémentaires.

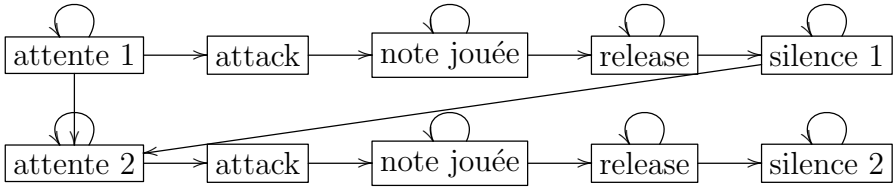
- la vraisemblance d'un *onset* est nulle sauf sur l'état attack. Idem pour *offset* et release ;
- la vraisemblance d'observation du *pitch* est nulle sur l'état silence ;
- la vraisemblance de non-observation du *pitch* sur nulle l'état sustain.

De plus, comme le signal MIDI est parfaitement localisé dans le temps, la chaîne de Markov employée peut être simple.

- les phases d'attack et de release sont réduites à un seul micro-état sans auto-transition ;
- la phase d'attente est équiprobable sur la fenêtre de position $[p_j^{min}; p_j^{max}]$;
- la phase de sustain est équiprobable sur la fenêtre de durée $[d_j^{min}; d_j^{max}]$;

Les distributions équiprobables correspondent à une absence d'a priori. Il est possible de les implémenter à l'aide d'états semi-markoviens, tels qu'employés dans le Antescofo ou bien directement à l'aide d'une variable de mémoire supplémentaire. D'autres distributions peuvent être choisies, comme les lois géométrique ou binomiale qui s'implémentent directement par des micro-états à auto-transition.

Agrégation de causalités Revenons sur notre modèle à l'échelle globale. Le cas de la présence de deux notes proches de même hauteur mérite un traitement particulier. Dans un instrument à jeu de notes comme le piano, l'attaque d'une touche étouffe la note précédemment jouée sur cette touche. Par conséquent, il existe une **causalité séquentielle** entre les notes attendues de même hauteur, que le modèle graphique modelé par le processus d'attente ne doit pas briser. La solution simple consiste à concaténant la chaîne de Markov de la nouvelle note à celle de l'ancienne, en ajouter une transition permettant de court-circuiter la première note.



3.3.3 Inférence à agents multiples

Le système que nous proposons est donc constitué de deux agents en interaction mutuelle :

- l'agent de suivi de position effectue une inférence globale de la position et du tempo ;
- l'agent de suivi des notes effectue des inférences locales sur l'avancement de chaque note attendue.

Dans un sens, les instants d'attaques constatés des notes de la partition permettent d'inférer le couple position-tempo. Dans l'autre, la connaissance de la position permet au processus d'attente de faire évoluer les notes attendues, en initialisant leur cycle de vie individuelle qui permet

l'inférence de leur instants d'attaques. Cette démarche comporte l'originalité d'utiliser un espace d'état hétérogène, composé d'un espace continu donnant la position et le tempo et d'un ensemble fini d'espaces discrets, dont le nombre évolue.

$$E_s = \{\text{position} \times \text{tempo}\} \times \mathcal{F}(\{\text{notes attendues}\})$$

Étant donné un nouveau pas de temps :

- pour chaque note attendue, l'inférence exacte est menée par l'algorithme de Viterbi. La connaissance du chemin optimal donne l'instant de jeu de la note et si celui-ci a eu lieu.
- l'inférence globale de la position et du tempo s'effectue à partir des couples (instant, position) décodés au fur et à mesure. Le filtre de Kalman peut être employé dans le cas où le modèle de transition est linéaire-gaussien comme celui de [12]. Nous pouvons aussi reprendre le modèle non-linéaire utilisé par Antescofo [3].

3.3.4 Critique et extension

Un premier point concerne la fiabilité du système. En somme, le principe de notre système est d'utiliser l'évolution de la position estimée $\hat{x}_t$ pour modifier dynamiquement ses attentes d'observations, de manière à capter la causalité d'une exécution linéaire. Dans son principe, rien ne garantit la robustesse du système de suivi au cas où l'instrumentiste s'arrête ou bien décélère déraisonnablement. Toutefois, le système dispose de mesures de sa fiabilité. L'inférence sur le suivi de notes revient à associer des instants observé aux instants d'*onset* et d'*offset* des notes attendues. Cela donne quatre critères pour mesurer la ressemblance entre la performance et la partition attendue, et relâcher le modèle d'écoute si la ressemblance commence à s'effondrerr :

- l'écart entre *onsets* prévus et *onsets* observés ;
- l'écart entre durée de note prévue et durée de note observée ;
- le nombre de notes attendues mais non-jouées ;
- le nombre de hauteurs surnuméraires (clutter), au fil des trames.

Un second point concerne l'extension aux signaux réels. Ce modèle a la vertu d'être simple dans le cas d'une observation MIDI, et potentiellement adaptable à une observation indirecte plus incertaine, telle que estimateur de fréquence fondamentale multiples décrit précédemment. En effet, adopter une observation plus bruitée revient à avoir à la fois une incertitude et une moins bonne résolution temporelle sur les instants d'*onsets* et d'*offsets*. Un apprentissage statistique pourra être alors utilisé pour optimiser les probabilités d'observation et de transition sur modèle markovien. Un autre angle intéressant sera de suivre le volume sonore. En effet, sur les instruments comme le piano, la causalité se retrouve dans la décroissance du volume sonore durant la période de résonance. Un apprentissage hors-ligne sur les profils de décroissance peut alors être envisagé, afin de l'intégrer dans le modèle markovien. [3]

3. À noter que cette hypothèse de décroissance ressemble aux heuristiques de décroissance exponentielle utilisé dans le suivi MIDI développé à L'Ircam [18].

Conclusion

Ce stage a répondu par la négative au pari de départ, qui portait sur la capacité des RFS à étendre le suivi de position aux situations de signaux musicaux asynchrones. Leur champ d'application spécifique nous a poussé à chercher un modèle qui suivrait les notes jouées simultanément avec la position et le tempo. Cette idée de suivi à double échelle de temps, celui de la trame et celui de la note, est conceptuellement intéressant mais difficile à rapprocher des filtres RFS rencontrés dans la littérature, et ce pour deux raisons. D'une part, nous voulons pour le suivi de notes un modèle de Markov discret voire semi-markovien, et non continu ; d'autre part, nous souhaitons un système ou le suivi multi-objets "orchestré" par suivi mono-objet ; davantage de lectures (telles que [8]) seront à chercher sur ce sujet.

Pour la conception d'un modèle génératif, nous sommes convaincus que l'utilisation du signal MIDI est une première étape très bénéfique. Avec un tel signal, toutes les vraisemblances du modèle d'observations sont prévisibles, ce qui ouvre de grandes possibilités pour la calibration du modèle de suivi. Ici, la calibration des paramètres nous a à deux reprises pousser à considérer le maximum de vraisemblance, procédure qui s'apparente aux algorithmes d'apprentissage tel celui Baum-Welch pour les chaînes de Markov. Pourtant nous sommes aujourd'hui d'avis que la procédure de calibration d'un modèle devrait être tout autre. Dans un modèle génératif, les probabilités ne doivent pas être pensées comme des fréquences d'apparition, mais comme des pondérations relatives entre éventualités ; il s'agit plutôt de savoir quelles erreurs d'interprétation musicale rendront ponctuellement le vrai chemin sous-optimal, et si oui ou non il le restera après retour à la normale. Aussi, un signal parfaitement connu comme le MIDI rend envisageable une analyse casuistique exacte de la robustesse. Le modèle d'observation RFS que nous avons proposé section 2.3.1 produit d'ailleurs une fonction de vraisemblance très adaptée à cet effet, car pour chaque chemin le nombre total d'erreurs commises peut s'y lire.

En outre, beaucoup reste à faire et à expérimenter autour du filtrage particulaire pour le suivi de partition. Nous avons pu constaté qu'il partage les mêmes questions méthodologiques que les RFS, notamment sur la de gestion des hypothèses et sur l'extraction des cibles à partir de la densité filtrante, ce qui justifiait son étude simultanée. Avoir repris les si les simulations stochastiques nous ont semblé très intéressantes, nous doutons maintenant de la pertinence du choix d'un espace continu de position. Il masque la géométrie véritablement discrète d'une partition de musique. Cela provoque un accroissement de la complexité combinatoire, car les chemins sous-optimaux ne sont plus identifiables, et le tirage aléatoire du tempo doit être effectué à chaque pas de temps plutôt qu'à chaque nouvelle note. Aussi, une chaîne d'états semi-markoviens nous semblent bien plus représentatifs d'une partition et d'une performance musicale que le modèle continu repris de [12].

D'une manière générale, toutes nos expérimentations ont été menées sur un nombre restreint de morceau de musique, et nous reconnaissons la nécessité de lancer des tests à plus grande échelle. Nous aurons notamment besoin de constituer une base d'exécutions musicales structurées par leur déviation d'interprétation, notamment leur degré d'asynchronie entre voix, pour lequel des mesures restent à définir.

En conclusion, l'étude des RFS nous a convaincu davantage convaincu de leur potentiel en tant que post-traitement d'estimateurs MIR tels que ceux de fréquence fondamentale. Modélisant les hypothèses, ils permettraient d'ajouter des croyances sur leur dyamiques temporelles, donc d'ajouter la dimension temporelle qui manque pour filtrer une suite brute d'estimations instantanées. Un tel filtrage permettrait notamment l'inférence jointe des caractéristiques structurantes de la musique telles que tempo, la pulsation et la métrique, et constituerait donc le lien manquant entre estimation de *pitchs* et transcription de partition véritable.

Références

[1] M. Sanjeev Arulampalam, Simon Maskell, and Neil Gordon. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50 :174–188, 2002.

[2] Yaakov Bar-Shalom and Thomas E. Fortmann. *Tracking and data association*, volume 179 of *Mathematics in Science and Engineering*. Academic Press Professional, Inc., San Diego, CA, USA, 1987.

[3] Arshia Cont. A Coupled Duration-Focused Architecture for Real-Time Music-to-Score Alignment. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 32 :974–987, June 2010.

[4] Arnaud Dessein, Arshia Cont, and Guillaume Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *International Society for Music Information Retrieval (ISMIR)*, pages 489–494, Utrecht, Netherlands, Août 2010.

[5] Zhiyao Duan and Bryan Pardo. Soundprism : An online system for score-informed source separation of music audio. *J. Sel. Topics Signal Processing*, 5(6) :1205–1215, 2011.

[6] Cyril Joder, Slim Essid, and Gaël Richard. An improved hierarchical approach for music-to-symbolic score alignment. In *ISMIR*, pages 39–45, 2010.

[7] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME – Journal of Basic Engineering*, pages 35–45, 1960.

[8] Feng Lian, Chongzhao Han, Weifeng Liu, Jing Liu, and Jian Sun. Unified cardinalized probability hypothesis density filters for extended targets and unresolved targets. *Signal Process.*, 92(7) :1729–1744, July 2012.

[9] Wing-Kin Ma, Ba-Ngu Vo, Sumeetpal S. Singh, and Adrian J. Baddeley. Tracking an unknown time-varying number of speakers using tdoa measurements : a random finite set approach. *IEEE Transactions on Signal Processing*, 54(9) :3291–3304, 2006.

[10] Ronald P. S. Mahler. The Random Set Approach to Data Fusion. In *Proceedings of the SPIE*, volume 2234, pages 287–295. Sadjadi, F.A., 1994.

[11] Ronald P. S. Mahler. *Statistical Multisource-Multitarget Information Fusion*. Artech House, Inc., Norwood, MA, USA, 2007.

[12] Nicola Montecchio and Arshia Cont. A Unified Approach to Real Time Audio-to-Score and Audio-to-Audio Alignment Using Sequential Montecarlo Inference Techniques. In *ICASSP 2011 : Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 193–196, Prague, Tchèque, République, May 2011. IEEE.

[13] Nicola Montecchio and Arshia Cont. Accelerating the Mixing Phase in Studio Recording Productions by Automatic Audio Alignemtnt. In *International Symposium on Music Information Retrieval (ISMIR)*, Miami, Florida, États-Unis, October 2011.

[14] Nicola Orio and Francois Déchelle. Score following using spectral analysis and hidden markov models. In *ICMC : International Computer Music Conference*, La Havane, Cuba, 2001.

[15] Michele Pace. *Stochastic models and methods for multi-object tracking*. These, Université Sciences et Technologies - Bordeaux I, July 2011.

[16] K. Panta, D. E. Clark, and Ba-Ngu Vo. Data Association and Track Management for the Gaussian Mixture Probability Hypothesis Density Filter. *Aerospace and Electronic Systems, IEEE Transactions on*, 45(3) :1003–1016, July 2009.

[17] Christopher Raphael. Aligning music audio with symbolic scores using a hybrid graphical model. *Machine Learning*, 65(2-3) :389–409, 2006.

[18] Diemo Schwarz, Nicola Orio, and Norbert Schnell. Robust polyphonic midi score following with hidden markov models. In *International Computer Music Conference (ICMC)*, Miami, USA, Novembre 2004.

[19] Ba-Ngu Vo and Wing-Kin Ma. A closed form solution for the probability hypothesis density filter. In *Proceedings of the 8th International Conference on Information Fusion*, 2005.

[20] Ba-Ngu Vo and Wing-Kin Ma. The gaussian mixture probability hypothesis density filter. *IEEE Trans. SP*, pages 4091–4104, 2006.

[21] Ba-Tuong Vo. Random finite sets in multi-object filtering. *Computer Engineering*, October 2008.