PARCOURS MASTER 2
ATiAM

TELECOM ParisTech

Rapport de Stage

Master 2 ATIAM 2011-2012

Détection automatique de caractéristiques rythmiques dans un flux audio

---

Laure Cornu

Encadré par **Thomas Fillon**
Bertrand David

Remerciements

Je tiens tout d'abord à remercier mes deux encadrants Thomas Fillon et Bertrand David qui m'ont permis de réaliser ce stage dans l'enceinte de Télécom Paristech.

Merci à Thomas qui est à l'initiative du sujet et qui a su tout au long de ce stage m'encourager dans les différentes pistes qui se sont proposées à moi.

Merci à Bertrand pour sa participation enthousiaste, ses conseils, et sa rigueur scientifique qui m'ont été souvent bien utiles.
Enfin merci à tous les deux pour les relectures attentives et conseils pour l'écriture de ce rapport.

Merci à Jérôme et Floriane, pour leur soutien quotidien et échanges enrichissants. Merci à mon père pour la relecture de se rapport.

Enfin je me permets de saluer la joyeuse compagnie de l'équipe AAO, Aymeric, Anne-Claire, François, Angélique, Antoine L., Antoine F., Manu, Benoit, Rémi, Sébastien, Nicolas, Mounira, Honoré et Slim sans qui ce stage n'aurait pas été aussi plaisant.

Table des matières

Chapitre 1

Contexte

L'extraction automatique de la description du contenu d'un morceau de musique (*content-based* description en anglais) a fait l'objet d'un nombre important de recherches ces dernières années. Parmi les informations extraites, la structure rythmique est particulièrement informative et peut servir de pré-traitement à de nombreux autres algorithmes d'extraction automatique de contenu. Dans ce travail, on s'intéressera plus particulièrement à l'estimation automatique de la position des temps, des premiers temps et du tatum (*beat/downbeat*) d'un morceau de musique ainsi qu'à l'estimation sous-jacente du tempo. Pour ce faire, des modèles probabilistes à états cachés permettant de relier les observations rythmiques et tonales extraites du signal à l'estimation de la position des temps et des premiers temps seront abordés. Quelques unes des théories perceptives qui m'ont permis d'accéder à une meilleure compréhension des éléments à estimer seront présentés sec. [1.1].
Puis un état de l'art qui permettra de situer cette étude parmi les différentes approches, singularisées par les applications visées, les *a priori* musicaux et les natures des données analysées (symbolique/Midi/flux audio). [1.2]

1.1 Le rythme musical

D'après Parncutt, "*le rythme musical est une séquence acoustique évoquant une sensation de pulsation*". En effet, lorsque ces durées sont récurrentes et tendent vers une reproduction cyclique, on perçoit alors une pulsation qui est a l'origine des instants qui vont être marqués en accompagnant le morceau de musique en frappant des mains, tapant du pied, claquant des doigts...
Mais l'apparente facilité à extraire des informations musicales telles que la pulsation ou la métrique d'un morceau de musique cache une grande complexité du système perceptif mis en jeu. Cette complexité explique en partie les difficultés rencontrées lorsqu'on cherche à construire un système informatique capable de comprendre et de représenter le rythme.

1.1.1 Groupement et métrique

L'organisation temporelle d'un rythme possède deux attributs différents, distingués dans la littérature par les termes de *groupement* et de *métrique*.

Le *groupement* se réfère à la ségrégation d'un morceau en motifs musicaux courts composés de groupes de notes et de leurs assemblages à différents niveaux temporels permettant de constituer des phrases musicales puis des mouvements et enfin d'extraire une structure analytique du morceau complet.

La *métrique* est la dénomination d'une hiérarchie des éléments détectés selon l'analyse des régularités temporelles des événements sonores présents dans le morceau.
Ce deuxième attribut (bien que souvent les deux hiérarchisations sont considérées comme complémentaires) et les définitions des différents niveaux hiérarchiques nous permettront de renseigner notre démarche computationnelle. Néanmoins la définition des mécanismes de groupement permet d'introduire la définition d'une hiérarchie métrique.

1.1.1.1 Groupement

La théorie de la Gestalt est considérée aujourd'hui comme une contribution majeure de la psychologie moderne, offrant une compréhension de la perception visuelle et auditive. Cette théorie se fonde sur l'hypothèse que la perception ne se limite pas à une détection de stimuli élémentaires isolés, mais consiste au contraire en une construction d'une représentation, d'une forme, à partir du phénomène observé. L'approche gestaltiste a influencé certaines théories musicales du 20ème siècle dont l'une des plus populaires est la Théorie Générative de Musique Tonale (TGMT). Proposée par ses auteurs Lerdhal et Jackendoff (1983), cette théorie développe un modèle perceptif de la musique qui repose sur les concepts d'acculturation et d'attente. Selon cette théorie, un événement musical n'a de signification que parce qu'il est orienté vers un autre événement musical attendu. Ainsi le retour au ton principal procure un sentiment de résolution tonale ou de détente. A contrario, la tension naît d'une attente contrariée. L'appréciation de ces attentes dépend en partie de l'acquisition des connaissances musicales de l'individu, développées par acculturation dans un milieu donné. A ce sujet voir Frances [8]. De la succession de ces attentes contrariées et résolues naît la perception d'une structure musicale. Selon la TGMT, l'expérience perceptive des formes dépend en premier lieu du traitement des informations à un niveau local et en second lieu de la qualité des relations entre les unités locales pour que se forment des connexions à un niveau supérieur.

C'est donc l'ordonnancement temporel (présence/absence) des *événements* musicaux qui va nous permettre de définir une hiérarchie métrique. Bien que les indices temporels, durées et moments d'apparition des événements musicaux soient essentiels à l'établissement d'une hiérarchie métrique ceux-ci sont a interpréter de concert avec des événements des indices de hauteur, par exemple la stabilité tonale des événements pour la détermination de la structure sonore du morceau (TGMT).

1.1.1.2 Métrique

L'invariant du système métrique Dans le but de caractériser la structure métrique, il est intéressant d'en caractériser l'invariant, l'unité de construction. La structure métrique permet de caractériser l'interprétation d'un rythme. Il est important de différencier ce qui est attenant au rythme (les différentes valeurs rythmiques), et à la métrique. Olivier Lartillot dans sa thèse [20] nous permet de définir le rythme :

> Là où la hauteur peut être envisagée de manière absolue, la valeur rythmique, au contraire, ne peut être estimée en terme de durée absolue. Car contrairement à la hauteur, une telle valeur rythmique absolue n'offre pas de composante symbolique, et ne peut être évaluée aisément par l'auditeur.

Cette citation fait écho à ce qui a été énoncé plus haut : la rythmique d'un morceau ne peut être extraite à l'échelle de la note, il faut prendre en compte un groupement de notes. En général la valeur rythmique est considérée comme un rapport à une *pulsation de base*. L'unité de base du système métrique est donc la pulsation principale du morceau.
Il est possible de reproduire un même rythme à un tempo différent. En définir les durées exactes n'a pas de sens, pour les raisons rapportées plus haut.

La vitesse du tempo est mesurée par le nombre de battements par minutes (Bpm). On le mesure parfois en secondes, on désignera alors le temps qui s'écoule entre deux pulsations. Un tempo à 120 Bpm est un tempo à 500ms.

L'inférence d'une pulsation Lorsque la musique est jouée, l'auditeur est capable de battre du pied "en rythme" à différentes vitesses et en suivant la métrique du morceau. Il est capable de faire la différence entre une valse à trois temps et un rock binaire... Il est capable d'inférer une pulsation sur le morceau qu'il écoute, ce qui nécessite de sa part une capacité d'analyse du contexte permettant l'adaptation d'une pulsation convenable à ce dernier. A ce sujet Drake & Parncutt (2001) [5] définissent deux aptitudes nécessaires au processus cognitif de l'analyse d'une métrique. La première est la mémoire musicale permettant de situer chaque élément sonore par rapport à ceux qui se sont déjà produits, la deuxième est l'attente musicale qui permet de le comparer aux instants de pulsations futures attendues.

Cette approche permet d'expliquer la robustesse de la perception d'un tempo aux interruptions. Par exemple lors d'une brève coupure dans un morceau de musique, l'auditeur anticipe une continuité dans la sensation de pulsation, et son attention est dirigée sur les instants naturellement soulignés par la pulsation attendue.

1.1.1.3 La structure métrique

Comme il a été dit plus haut, l'auditeur est capable de battre du pied en rythme et ce à différentes vitesses. Ceci est dû au fait que la structure métrique d'un morceau peut se décomposer en différentes "couches rythmiques". Définir la structure métrique et les différentes couches rythmiques se fait en considérant les deux assertions suivantes rapportées de la thèse de M. Alonso [6] :

1. Chaque impulsion dans une couche métrique donnée, coïncide avec une impulsion de la couche métrique inférieure.
2. Le rapport entre deux couches métriques différentes est d'un facteur de deux (on parle alors d'une musique "binaire") ou trois ("ternaire").

L'existence de plusieurs couches métriques introduit dans l'analyse de l'auditeur plusieurs candidats au tempo. Plusieurs études se sont penchées sur le choix et donc sur la prédominance de la pulsation du tempo. Il apparait que l'auditeur tend à préférer un niveau de tempo modéré :

- autour de 500ms (120bpm) d'après Lerdahl et jackendoff [9]

- qui affecterai l'interprétation d'une pièce : d'après Handel [7] les distributions des cadences de battements en musique ne sont pas liées aux tempi annotés dans les partitions mais plutôt à la cadence du tempo perceptif. En d'autres termes, l'auditeur a tendance à nucléariser la mesure lorsque le tempo est lent. Au contraire, il les concatène lorsque le tempo est rapide.
- d'après la TGMT : Les battements espacés de plus de 1,5 secondes ne sont pas perceptibles. La battue (i.e détection. de régularité en frappant à la vitesse de la représentation symbolique annotée) est plus facile à la noire qu'à la blanche, qu'à la ronde. Elle est impossible pour des unités supérieures.

La prédominance du tempo influencera ensuite l'interprétation des couches métriques inférieures et supérieures en forçant la détection de périodicités dont le rapport au tempo est un multiple de deux ou trois.
La définition d'une hiérarchie sonore nécéssite l'énumération du vocabulaire afférent. On désignera par trois termes différents les trois principaux niveaux hiérarchiques du rythme.
La pulsation de référence est, comme nous l'avons dit plus haut, le *tactus*, *beat*,*tempo* qui sont autant de synonymes pour définir la couche intermédiaire. Les autres nivaux sont :

- Le *tatum*, qui est le niveau de périodicité de plus rapide détecté dans un flux sonore. La dénomination provient de la concaténation des termes latins *temporal atum*, qui est aussi un clin d'œil à la virtuosité du célèbre pianiste de jazz Art Tatum.
- La mesure qui est le niveau de périodicité le plus lent.

Les trois couches métriques sont représentées sur la Figure [1.1].

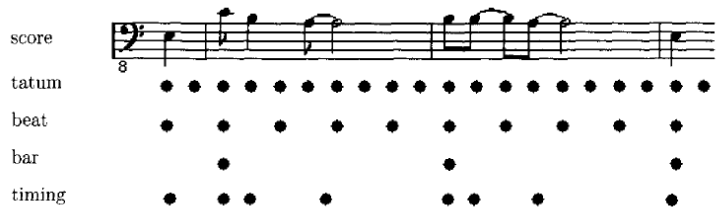


Figure 1.1 – Relations entre le Tatum, Tactus/Beat et la Mesure, figure extraite de Hainsworth dans [2]

Dans le cas de la valse, l'auditeur repèrera de manière plus aisée la noire sur les trois temps de la valse (*tactus*), puis il interprétera un des battements sur trois comme étant le plus important (*mesure*) : les premiers temps de la valse sont alors désignés par le terme de "temps fort" et les deux autres temps comme des "temps faibles". Cette alternance de temps forts et temps faibles établit une hiérarchie métrique qui permettra à l'auditeur de guider son écoute.

Remarque : Ce schéma et les définitions précédentes semblent indiquer que la détermination de la couche métrique supérieure obéit aux mêmes lois perceptives que le *tatum* ou bien le *beat*. En fait c'est le niveau qui est le plus difficile à déterminer perceptivement. Parfois cette tâche est impossible. En effet selon les genres musicaux, ce ne sont pas les mêmes temps qui

sont accentués musicalement. Et cela peut induire un biais dans l'*a priori* musical encodé, nécessaire à la détermination automatique de tempi Laroche [10]. Les compositeurs classiques avant le XXème siècle, et la majorité de la musique rock et pop actuelle,cultivent (pour une métrique 4/4) l'accentuation des premiers et troisièmes temps, que l'on appellent pour cette raison des temps forts. A contrario des musiques plus récentes accentuent les *temps faibles* ou *contretemps*, recherchant un effet de déséquilibre rythmique ou effet "swing" tel que le jazz, le funk ou le reggae.

1.1.2 Accent dynamique

En effet il a été montré, dans des expériences perceptives ultérieures, que l'accent rythmique est plus important que l'accent mélodique dans la reconnaissance de mélodies. Enfin, si l'accent mélodique est moins important, il est quand même à prendre en compte comme l'a fait M. R. Jones en 1987 [12] pour l'élaboration d'un modèle représentant les attentes musicales de l'auditeur. Il définit dans la structure unitaire d'accents. Elle est élaborée à l'aide de la superposition d'accents mélodiques et temporels. Les instants du morceaux qui sont marqués par un accent mélodique et temporel seraient pressentis comme plus importants que les instants marqués par un seul des deux accents. Par conséquent ils sont de bons candidats pour définir les barres de mesures.

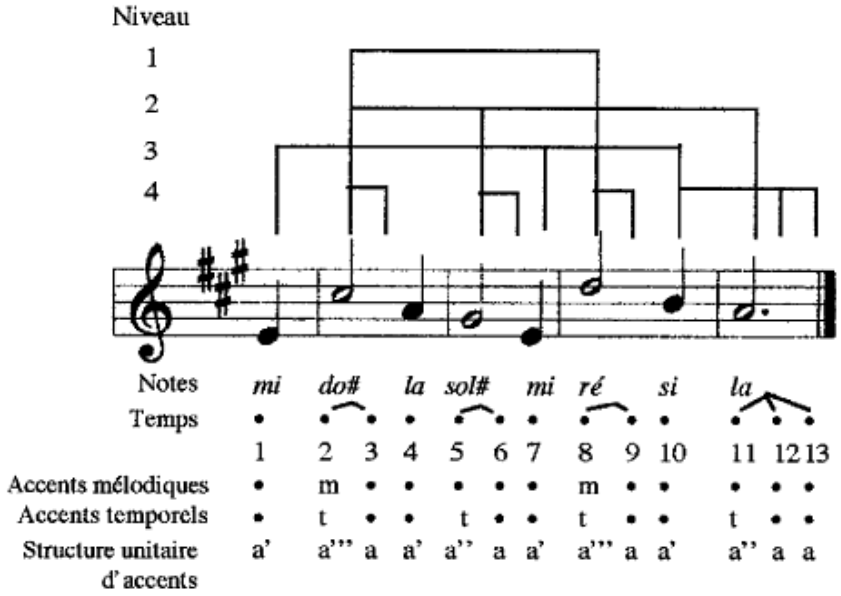


FIGURE 1.2 – Exemple d'une structure unitaire d'accent sur une courte mélodie, extrait de [13]

1.1.3 Résumé

En résumé les informations qui nous permettront de déterminer une métrique sur flux audio sont :

– Les accents mélodiques qui sont caractérisés dans la section [**??**]

– Les accents rythmiques qui sont caractérisé dans la subsection [**??**]

Ces accents permettent de caractériser la métrique en hiérarchisant les périodicités concernant :

– *le tactus* : fait un compromis entre un tempo de référence aux alentours de 500ms et la périodicité correspondant aux éléments désignés par le premier chiffrage sur la partition
– *le tatum* : Périodicité perceptive la plus petite de rapport entier avec le *tactus*
– *La mesure* : L'inférence perceptive en est difficile, elle correspond à la périodicité des éléments correspondant au deuxième chiffrage sur l'annotation d'une partition.

Ces hypothèses — assez réductrices, il est vrai — ont, dans le cadre de cette étude, une vertu essentiellement pragmatique, car elles assurent un fonctionnement élémentaire du système computationnel proposé dans les chapitres suivants.

1.2 État de l'art

Afin de présenter les différentes méthodes qui ont été développées dans le contexte de l'extraction de tempo, il semble pertinent de les ségréger en deux approches : l'approche symbolique et l'approche dite de traitement du signal. Les deux approches ne traitent pas des données de même nature. L'approche symbolique traite des informations de haut niveau de représentation du signal, telle que les hauteurs de notes, leurs durées, leurs temps de début et de fin, décrites dans un format compact tel que le MIDI. Dans la littérature actuelle, les deux approches ont des objectifs différents. Beaucoup de modèles symboliques sont développés dans des domaines de recherches tels que l'étude de la perception musicale ou des mécanismes cognitifs. La plupart des modèles acoustiques, sont, au contraire, développés dans le but de résoudre des taches d'ingénierie telles que la transcription de la musique ou le suivi de partition. Dans la majorité des cas la détection de rythme applique dans l'ordre les taches

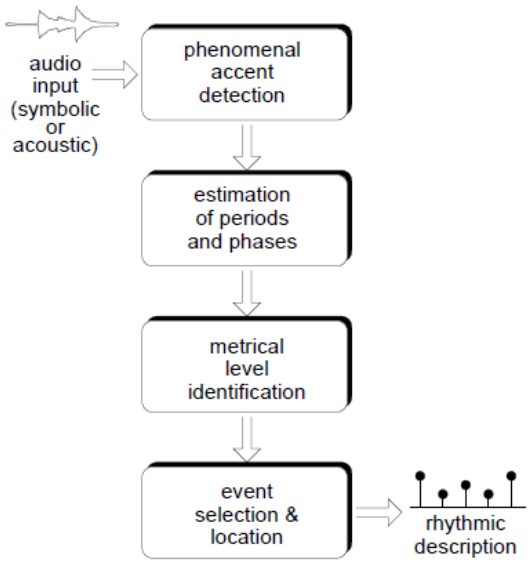


FIGURE 1.3 – Schéma descriptif des traitements en série pour l'extraction du rythme (fig. extraite de [6])

montrées dans le schéma.

1. Un premier traitement qui permet de détecter les éléments sonores importants. On extrait une fonction qui exprime le degré d'accentuation musicale en fonction du temps.
2. Ensuite il faut extraire les périodicités ressenties.Plusieurs méthodes le feront de concert avec l'estimation du temps exact auxquelles elles apparaissent. Les méthodes qui permettent de le faire sont nombreuse : Banc de filtres oscillants, autocorrélation, somme spectrale, produit spectral, densité de puissance du spectre...
3. Définition de la métrique du morceau. Souvent un modèle d'*a priori* permet de détecter les différents niveaux métriques.
4. Sélection du meilleur chemin. En reliant les estimations de métriques aux données de représentation du signal.

1.2.1 Approche symbolique

Dans sa thèse, B. Meudic [14], utilise des données MIDI (les onsets sont connus). Ses méthodes sont basées sur l'estimation de la prédominance d'accents. Il propose une fonction de détection musicale qui comprend les trois items suivants : la détection de régularité (avec I_n l'intervalle entre deux onsets, $I_n = I_{n-k} \pm 0.2$), la longueur de la note et le nombre de notes au moment de l'instant musical. Il déduit ensuite plusieurs candidats de tempi, en retirant et ajoutant des items à la liste formée. Puis il extrait le tempo en faisant un compromis entre le beat qui a le poids le plus important et celui qui le plus de sous-périodicités.

Méthodes probabilistes Cemgil dans [16] propose une approche Bayésienne pour l'écriture de partition automatique, d'après des données MIDI. C'est un modèle probabiliste qui permet de traiter les déviations rythmiques dans la musique live. Il formule les approximations sous la forme d'une liste d'*a priori* permettant de rester robuste au bruit de quantification. Il propose dans son article un modèle qui permet d'estimer la probabilité que le rythme soit binaire ou ternaire. Pour cela il estime l'expansion binaire et ternaire des estimations, et pénalise la métrique qui est la moins adéquate. Puis il résout son système en utilisant la méthode des Markov Chain Monte Carlo (MCMC).

1.2.2 Approche acoustique

fonction de détection d'accentuation musicale La détermination de l'accent musical est différente dans l'approche acoustique, en effet contrairement au format MIDI ou on a accès au notes, la fonction de détection d'accent musical va être mise en œuvre pour détecter les onsets. Ce qui est appelé un "onset" est le temps de début d'un événement musical présent dans le flux sonore.

L'article de Scheirer [15], qui s'impose comme référence dans le domaine, propose de décomposer le signal en différentes bandes de fréquences, de déterminer les sauts d'énergies dans chacune d'elles et d'en faire la somme pondérée par les différents poids perceptifs appliqués aux différentes bandes de fréquences (notre oreille favorisant les fréquences voisines du pitch de la voix). C'est cette somme qui sera la fonction d'onset résiduelle. Le paradoxe est que si la découpe par bande est fine elle permet de détecter les changements de notes *legato* , mais un onset percussif, pourtant présent sur toutes les bandes aura une énergie moins présente sur la somme résiduelle. L'article de Klapuri et Eronen [1] propose une alternative : une découpe très fine (suivant une distribution perceptive dite ERB), suivie d'une somme parcellaire (afin d'obtenir et d'étudier les périodicités de plusieurs flux).

Flux spectral Dans sa thèse M. Alonso [6] propose de faire une détection d'onsets à l'aide d'un flux spectral d'énergie. L'approche est semblable que celle de Klapuri et Eronen [1] mais le traitement s'applique en parallèle sur un signal qui a été séparé sous la forme harmonique plus bruit. Ensuite les spectres sont réassignés (STFT renseignée par les valeurs des phases calculées sur la fft du signal).

1.2.2.1 Détection des périodicités

Comme on l'a vu plus haut, la détection de périodicités peut se faire de différentes manières. Dans son article Laroche [10] utilise une méthode de corrélation croisée entre la fonction de détection normalisé et plusieurs peignes $E_{R,t}$ de tempos R et de phases t, variables. Il estime donc les périodicités et phases conjointement.

1.2.2.2 Induction du tempo

L'induction du tempo se fait souvent de concert avec un *a priori* permettant de sélectionner plusieurs tempi candidats. Dans son article [16] G. Peeters propose d'évaluer les différents tempi en fonction de plusieurs textita priori dont celui de la balance spectrale qui est l'estimation, pour un tempo candidat, de la répartition alternée des basses fréquences et hautes fréquences sur une représentation temps-fréquence du signal.

1.2.3 Contexte applicatif

1.2.3.1 Approches globales : système d'interaction musicale

Analyse des positions des temps forts/ temps faibles L'analyse de la synchronisation des temps faibles et temps fort permet de développer différentes applications, plus particulièrement des systèmes d'interaction musicale. On peut citer à titre d'exemple :

- L'alignements de partition. Par exemple le projet Antescofo propose d'évaluer en temps réel l'avancée de l'interprétation d'un musicien par rapport à la lecture de la partition de ce morceau. On peut comparer son principe avec le principe de fonctionnement de la "sympathie des horloges", un phénomène de physique découvert au $17e$ siècle par Christian Huygens. Ce dernier avait montré que deux pendules oscillants finissaient par se balancer au même rythme s'ils étaient reliés avec des lames de profil correct. [17]
- Les systèmes interactifs d'improvisation musicales.

Outils de traitement des temps forts/temps faibles L'utilisation des outils de détection de métrique peut être utilisée à des fins de synthèse sonore (content based audio effect). Les applications permettent alors :

- l'étirement la métrique,
- son changement,
- ou sa ségrégation.

1.2.3.2 Approche locale : MIR

La recherche d'information musicale (Music Information Retrieval) est la détermination des différents attributs symboliques du signal. Ceci permet l'indexation de grande bases de données ou peut servir à des tâches plus compliquées, parmi lesquelles :

- la reconnaissance de mélodie
- la détermination d'accords
- les arrangements automatiques
- la reconnaissance de musique
- la segmentation de morceaux

Chapitre 2

Méthodes développées

Au cours de ce travail nous avons choisi comme méthode de référence la méthode proposé dans [3]. La méthode développée se présente sous la forme d'une succession de traitements décrits dans l'ordre chronologique ici. La fonction de détection puis l'estimation des périodicités et enfin l'estimation des phases. On présentera en dernier l'état de recherche de l'induction de métrique déterminé grâce à l'implémentation de la reconnaissance d'accents mélodiques.

2.1 Fonction de Détection

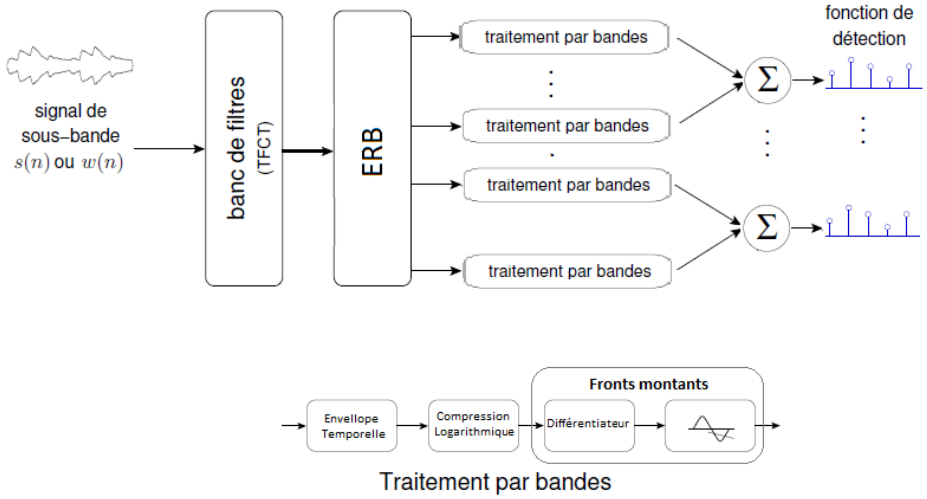


FIGURE 2.1 – Schéma général du fonctionnement de la fonction de détection du système de référence.

La fonction de détection d'*onset* s'attache à détecter dans un flux audio les accents perceptifs appelés dans la littérature anglaise *phenomenal accent*. Les accents perceptifs sont les événements sonores qui permettent de marquer certains instants dans le flux audio. On peut par exemple lister les débuts de tous les évènements sonores discrets dans un morceau de

musique, plus particulièrement les débuts de longues notes, les changements harmoniques ou les changement d'énergie ou de timbre.

La fonction de détection est une partie qui est très sensible dans la détection de tempo, afin de pouvoir tester la fiabilité du système de référence on a souhaité comparer les résultats d'estimation avec ceux fournis avec l'implémentation de la fonction de détection M. Alonso [6].

2.1.1 Système de référence

2.1.1.1 TFCT

La Transformée de Fourier à Court Terme permet d'obtenir une représentation temporelle du spectre du signal audio. Le principe est décrit en Annexe1. Pour obtenir une représentation du rythme il est important d'avoir une bonne résolution temporelle. Il est donc choisi d'avoir des trames de signal d'une durée de 23ms avec un overlapp de 75 %. Cherchant à analyser les saut d'énergie présents dans le signal on utilisera le spectrogramme (module de la TFCT).

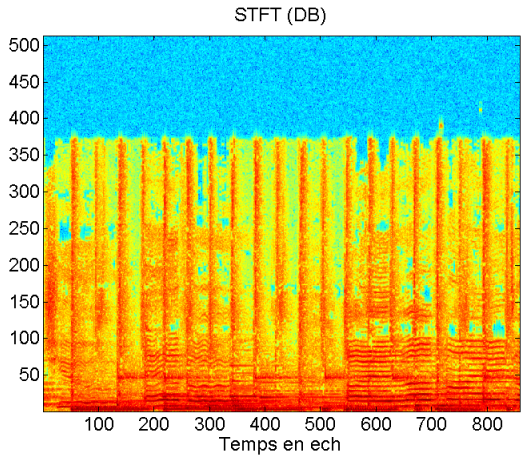


Figure 2.2 – Représentation de la TFCT en décibels des 5 premières secondes de la chanson *'Amado mio'* du groupe Pink Martini. Pour des fenêtres de 23ms, et une fréquence d'échantillonage de 172Hz.

2.1.1.2 Découpe en ERB (Equivalent Rectangular Bandwith)

L'utilisation de TFCT nous permet d'opérer une première analyse du signal en bandes fréquentielles du nombre de la longueur (en échantillons) de la fenêtre d'analyse ($0.023 * Fe$ dans notre exemple). Ceci impliquant que le signal est analysé à partir d'une échelle linéaire et uniforme en fréquence. Cependant, le système auditif humain ne perçoit pas les signaux selon une échelle uniforme. Notamment, notre système auditif effectue une analyse plus précise dans les fréquences graves que dans les bandes hautes du spectre.

Réalisant l'analyse spectrale selon une échelle en fréquence conforme à notre perception auditive, l'échelle ERB (Equivalent Rectangular Bandwith) est la largeur des bandes critiques, elle modélise la résolution fréquentielle des filtres auditifs. La courbe de des fréquences centrales associées à leurs équivalent ERB est représentée dans la figure extraite de [21].

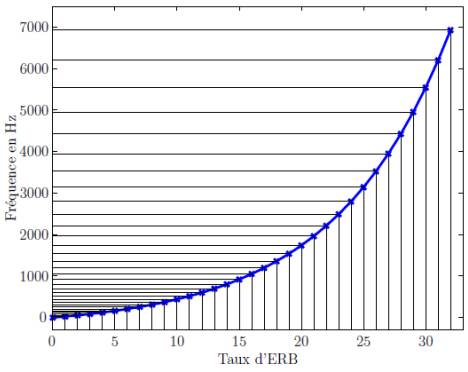


FIGURE 2.3 – Relation de correspondance entre la fréquence centrale et son équivalent ERB

L'échelle des fréquences (Hz) et l'échelle ERB sont liées par la relation suivante : $ERB = 21.4 \log\left(\frac{4.37.f}{1000} + 1\right)$ qui nous permet de passer d'une échelle à l'autre et de déterminer le contour des filtres passes bandes qui nous permettent de passer d'une échelle linéaire à une échelle ERB. On choisi de découper les fréquences en 36 sous bandes de 20Hz à 20kHz, à l'aide d'une banque de filtre ERB triangulaires.

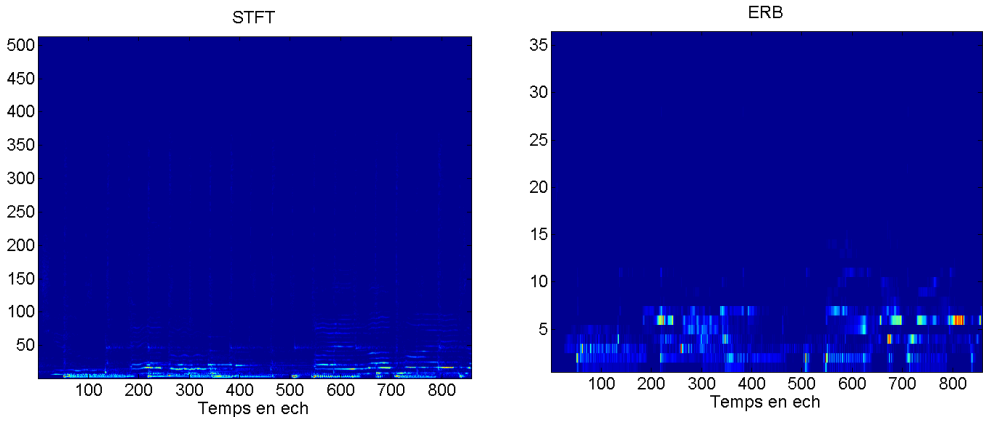


FIGURE 2.4 – Représentation de Tfct et de la sortie du banc de filtre passes bandes ERB, des 5 premières secondes de la chanson 'Amado mio' du groupe Pink martini, pour des fenêtre de 23ms et une fréquence d'échantillonage de 172Hz.

Je vais changer les représentations et les mettre en log.

2.1.1.3 Compression logarithmique

La détection d'un changement d'intensité ΔI est inversement proportionnelle à l'intensité I du signal. A Δ_I constant, la perception de cette variation sera plus aisée sur un signal de faible énergie. Afin de rétablir une attribution plus proche de la perception humaine des événements marquants dans un flux sonore, il est important de rétablir les poids perceptifs des variations d'intensité. La fraction de weber $\frac{\Delta I}{I}$ est approximativement stable dans la bande de

fréquence (20Hz - 20kHz) concernée. Elle peut s'écrire autrement $\frac{\partial x(f,t)}{\partial_t} = \frac{\partial ln(x(f,t))}{\partial_t}$, ce qui justifie l'emploi d'une fonction logarithmique. Cependant une telle fonction n'est pas définie au voisinage de zéro. Une possibilité est d'utiliser une compression μ-law qui est linéaire au voisinage de zéro.

$$y = \frac{ln(1+\mu x)}{ln(1+\mu)}, \tag{2.1}$$

Le paramètre de compression μ permet de choisir une répartition de poids qui va de logarithmique (μ grand) à linéaire. Dans l'article d'Erronen and al [3] il et fixé à 100, dans l'article de scheirer [1] et de goto, il est de 0.1. D'autres types de compression peuvent être choisis dans l'article de Laroche [10] la répartition de poids est déterminée par $x^{1/2}$.

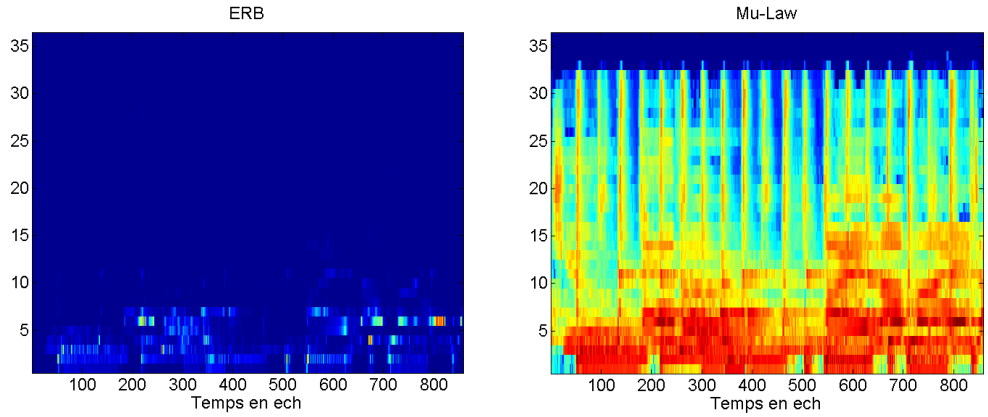


Figure 2.5 – Comparaison de banc de filtres avant et après la compression logarithmique.

Remarque : La compression logarithmique est similaire à l'échelle des décibels, les deux étant de forme logarithmique, ceci implique que la figure de droite de [2.7] est similaire à la figure [2.4]

2.1.1.4 Enveloppe temporelle

Système de référence Afin de ne prendre en compte que la tendance des variations d'énergie présentes dans le signal, un filtre passe bas est appliqué sur toutes les bandes temporelles, cf Fig. [2.7]. Dans cette implémentation le filtre choisi est un filtre de Butterworth d'ordre 6 de fréquence de coupure 10Hz.

Apport Scheirer dans [1] propose d'extraire une enveloppe temporelle en filtrant le signal temporel par une demie fenêtre de Hann, la taille de la fenêtre est calculée de manière à ce que la fréquence de coupure du filtre passe-bas soit de 10Hz. Le résultat obtenu paraît plus lisse.

2.1.2 Fronts montants

Pour chaque bande temporelle on détermine les sauts d'énergie en dérivant la courbe des variations temporelle d'énergie. On récupère ensuite les valeurs positives qui marquent les

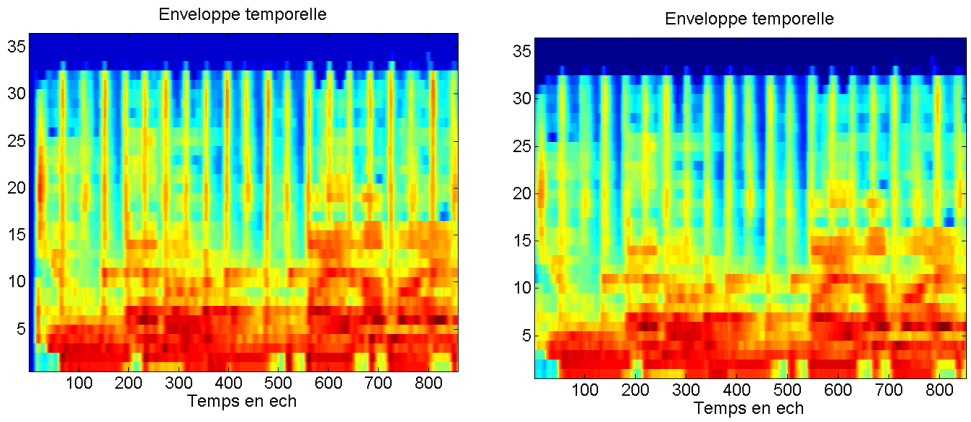


FIGURE 2.6 – Comparaison des deux méthodes d'extraction d'enveloppe, 1. à droite avec le filtre de Butterworth (Référence), 2. à gauche avec la demie fenêtre de Hann. (Scheirer)

débuts des événements marquants (*onset*) présents dans le signal, en appliquant un seuillage à zéro, cf Fig. [2.7]. A contrario les valeurs négatives indique la fin d'un événement marquant (*offset*) sont, comme on l'a vu plus haut, perceptivement moins marquantes dans l'analyse rythmique d'un morceau.

Système de référence Dans l'implémentation proposée par l'article la dérivée est obtenue en faisant la différence des valeurs successives du signal.

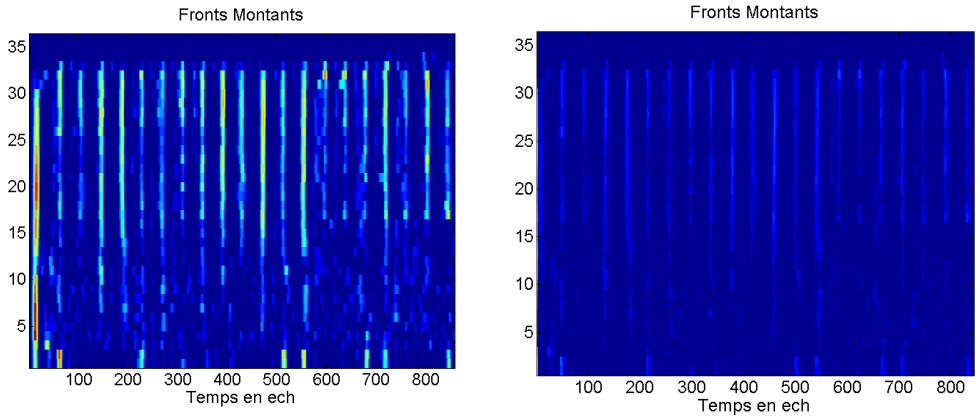


FIGURE 2.7 – Comparaison des deux profils de fronts montants extraits, 1. Méthode de référence, 2. Demie fenêtre de Hann et filtre dérivateur.

Apport personnel Pour résoudre le problème de la dérivée, ayant des données numériques discrètes, on peut utiliser un filtre dérivateur. On dit d'un filtre qu'il est dérivateur si :

$$z'_b(t) = \frac{dz_b(t)}{dt} \tag{2.2}$$

où, z'_b est le signal de sortie et z_b est le signal d'entrée. Ce qui signifie que la fonction de transfert du filtre est de la forme (Dérivateur idéal) :

$$H(j\omega) = j\omega \tag{2.3}$$

Ceci implique que différencier une fonction dans l'espace temporel revient à multiplier son spectre par une fonction linéaire de pente 2π , ou à la filtrer par un filtre dont la réponse en fréquence est $H(j\omega) = j\omega$. Ce qui est fait dans le code à l'aide de la fonction matlab *firpm* et qui permet d'obtenir une approximation de $j2\pi\nu$.

Les auteurs [3] proposent de compenser les effets de bords d'une estimation de la dérivée en ajoutant un petite partie du signal originel, on note :

$$u_b(n) = (1-\lambda)z_b + \lambda\frac{f_r}{f_{LP}}z'_b(n) \tag{2.4}$$

où, λ est fixé à 0.8, f_r est la fréquence d'échantillonnage de z_b et f_{LP} la fréquence de coupure du filtre passe bas, le rapport $\frac{f_r}{f_{LP}}$ permet de compenser la perte d'énergie du signal filtré. On note u_b, le signal sortant de ce traitement.

2.1.3 Somme des sous bandes

La fonction de détection est obtenue en sommant l'énergie des sous bandes du signal. Lorsqu'il y a peu de sous bandes la détection s'applique sur des événements sonores dont le spectre est à large bande par exemple les *onsets* percussifs. Au contraire une discrétisation fine en fréquence favorise les *onsets* marquant des changements plus fin, par exemple harmoniques, et est plus adaptée au traitement de la musique classique. La particularité de la méthode de référence [1] est de faire une détection fine en proposant une discrétisation fréquentielle de 36 bandes (scheirer [1], dans son article de référence en propose seulement 6 !) de faire une somme parcellaire qui permet d'obtenir quatre fonctions de détections.

$$v_c(t) = \sum_{b=(c-1)m_0+1}^{cm_0} u_b(t), \tag{2.5}$$

$$c = 1, 2, 3, 4 \tag{2.6}$$

Où $m_0 = 36/4$ est le nombre de sous bandes a sommer pour chacune des quatre fonctions de détection obtenues.

L'observation des deux résultats de fonctions de détection montrent que la deuxième méthode à attribué une énergie très forte sur le premier onset. La surestimation de l'énergie d'un onset ne détériore pas les résultats des détections de périodicités.

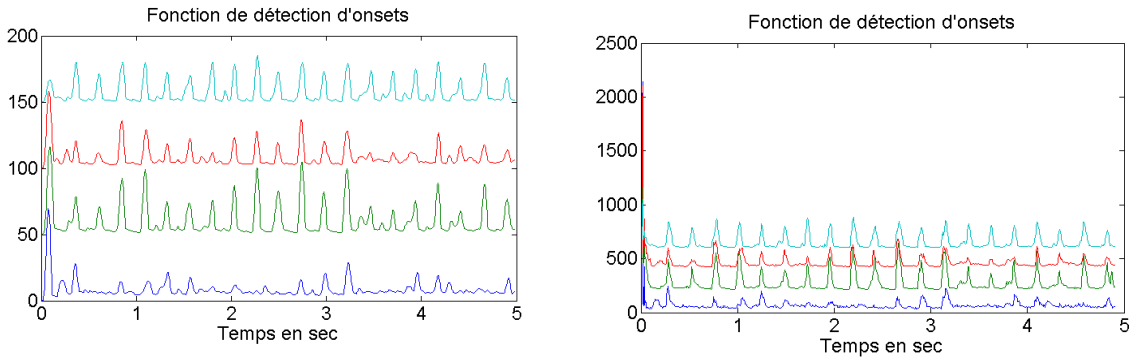


FIGURE 2.8 – Comparaison des résultats des deux fonctions de détection d'onsets, à gauches celle obtenue avec le système de référence à droite celle obtenue avec le filtre dérivateur et le filtre passe bas de profil demi-Hann. Les résultats sont ordonné des basses fréquences (en bas) aux hautes fréquences.

2.2 Observation de périodicités

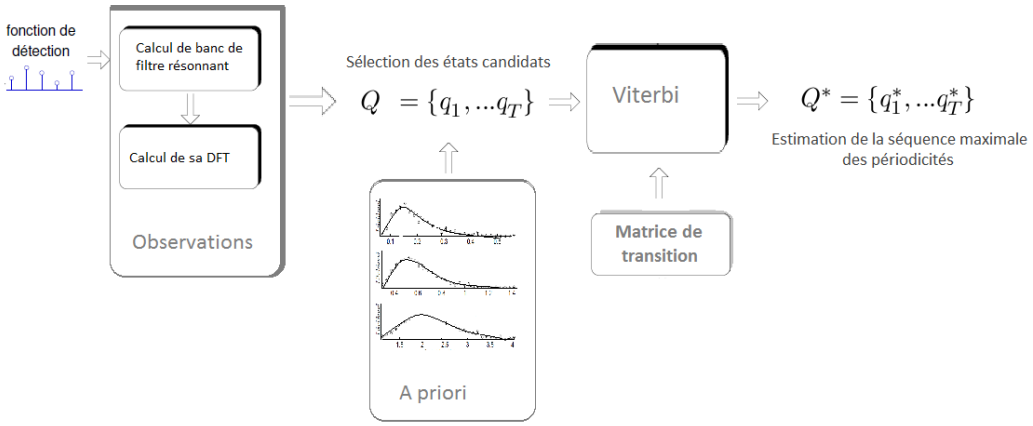


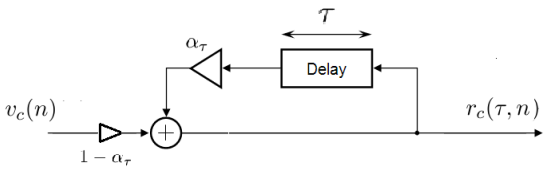
FIGURE 2.9 – Schéma de fonctionnement de la détections de périodicités

Dans cette partie on s'intéresse à l'estimation de la séquence la plus probable des périodicités présentes dans le morceau. On cherche l'information du *tempo* du morceaux en fonction du temps. Pour cela on analyse dans un premier temps la périodicité de la fonction de détection v_c. En fonction de cette information a priori on effectue une analyse de la séquence la plus probable des tempi des trois périodicités recherchées le *tatum*, *tactus*, *mesure*. Cette analyse se fait au regard de règles musicales sur la régularité de ces deniers, sur leurs probabilités conjointes et sur leurs durées moyennes. Le modèle probabiliste qui permet de modéliser ces dépendance est un modèle de Chaines de Markov Cachées (Hidden Markov Model en anglais ou HMM).

2.2.1 Observation des périodicités de la fonction v_c

Afin de déterminer les périodicités du tactus et de la mesure présentes dans le morceau on effectue une première analyse de la fonction v_c en analysant les sorties du filtrage de cette dernière par un banc de filtres résonnants.

2.2.1.1 Fonctionnement du banc de filtres résonnants



Le schéma de fonctionnement de ce filtre est montré sur la figure [2.2.1.1], son équation est la suivante :

$$r_c(\tau, n) = (1 - \alpha_\tau)v_c(n) + \alpha_\tau r_c(\tau, n - \tau) \tag{2.7}$$

Banc de filtre résonnant et AR1 L'équation précédente [2.7] peut s'écrire sous la forme :

$$r_c(\tau, n) - \alpha_\tau r_c(\tau, n - \tau) = (1 - \alpha_\tau)v_c(n). \tag{2.8}$$

La transformée en z de l'équation s'écrit alors :

$$(1 - \alpha_\tau z^{-\tau})R_c(z) = (1 - \alpha_\tau)V_c(z). \tag{2.9}$$

La fonction de transfert de ce filtre s'écrit :

$$H(z) = \frac{R_c(z)}{V_c(z)} = \frac{1 - \alpha_\tau}{1 - \alpha_\tau z^{-\tau}}. \tag{2.10}$$

$$H(z^\tau) = (1 - \alpha_\tau)\frac{1}{1 - \alpha_\tau z^{-1}}. \tag{2.11}$$

où on reconnait la fonction de transfert d'un filtre AR1, multiplié par un facteur $1 - \alpha_\tau$.

Fonctionnement du filtre sur une peigne de périodicité τ Le paramètres d'un filtre résonnant sont donc α et le retard τ. La réponse impulsionnelle du filtre pour un retard de 86 échantillons est montrée sur le figure [2.10]. On peut la comparer à la réponse impulsionnelle d'un AR1 pour mettre en évidence le fait que c'est AR1 décimé de τ échantillons.

La figure [2.11] montre les réponses de trois filtres résonnants de paramètres respectifs $\alpha = 0.5$ et $\tau = [43, 86, 172]$ d'un peigne de fréquence 2Hz avec une fréquence d'échantillonnage de 172.

Le principe des filtres résonnants est d'opérer la somme du signal avec lui même transposé d'un nombre τ d'échantillons et d'appliquer à chaque opération d'addition un poids multiplicatif α correspondant à la réponse impulsionnelle d'un filtre AR1 de coefficient α.

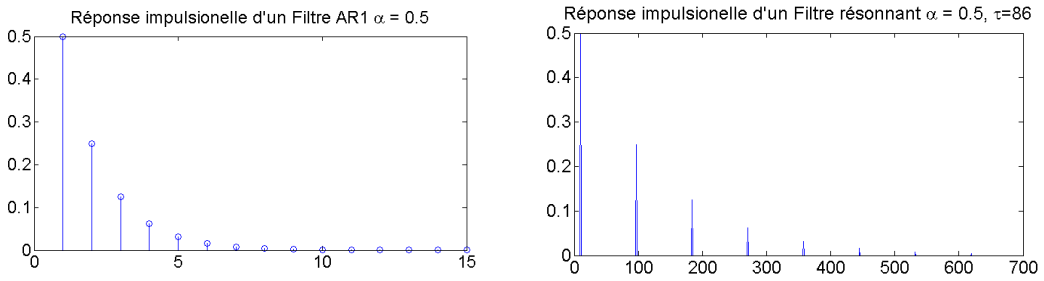


FIGURE 2.10 – Comparaison des réponses impulsionnelles d'un AR1 (fig. gauche) , et d'un filtre résonnant de retard $\tau = 86$ et de paramètre α

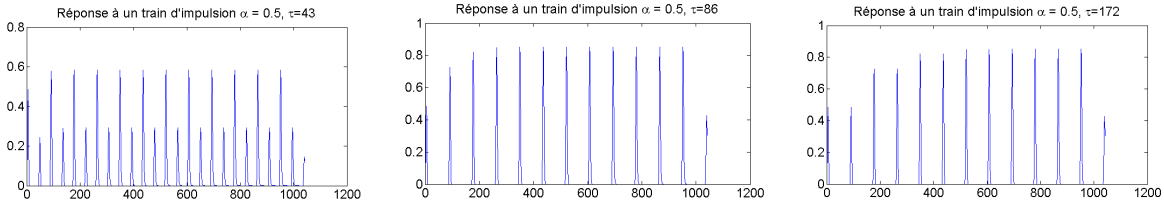


FIGURE 2.11 – Comparaison de la réponse de trois filtres résonnants à un train d'impulsion de 2Hz échantillonnée à 172Hz (périodicité de 86 ech.), en fonction de leurs temps de retard, de gauche à droite $\tau = 43$,$\tau = 86$, et $\tau = 172$.

Lorsque le retard du filtre est accordé à périodicité du signal, on obtient un signal avec la même périodicité que le signal d'entrée mais l'amplitude des pics augmente jusqu'à atteindre un maximum au bout d'un certain temps, appelé le temps de montée du filtre, cf fig.[2.11, fig. centrale]. Lorsque le retard est deux fois plus petit que la période du signal, la figure, fig.[2.11, fig. gauche], permet d' illustrer le fonctionnement du filtre. La transposition étant deux fois plus petite que la période, l'opération de somme va former des interpics, toutes les demi périodes. Lorsque le retard du filtre est k fois plus grand que la période considérée le temps de montée est k fois plus long, cf. fig.[2.11, fig. droite].

2.2.1.2 Implémentation

On souhaite avoir un analyse de toutes les périodicités présentes dans le morceaux, pour cela on enregistre dans une matrice la sortie des filtres résonnants de la fonction de détection v_c, pour tous les retards τ compris entre 1 et 400 échantillons.
La fonction de détection est échantillonnée à 172 Hz, elle permet de tester les tempi de 25.8 Bpm ($172/400 * 60$) jusqu'à 10320 Bpm. Ce qui n'a pas de sens pour le calcul des périodicités du *tactus* et de la *mesure*, mais ces évaluation seront utiles pour le calcul des périodicités du *tatum*).

temps de monté Afin d'avoir des évaluations aux amplitudes comparables, on ajuste le coefficient α en fonction du temps de monté, considérant qu'un temps de monté de $3s$ est suffisamment court pour réagir a des variation de tempi, et suffisamment long pour estimer

correctement des tempi réguliers de plus $4s$. Ici le temps de monté est le temps nécessaire (en échantillons) pour que l'amplitude de la réponse du filtre soit diminué de moitié. On note α_τ le coefficient du filtre pour un retard de τ échantillons :

$$T_0 = 3 * Fe \tag{2.12}$$
$$\alpha_\tau = 0.5^{\tau/T_0} \tag{2.13}$$

Normalisation Les filtres n'ayant pas le même coefficient α ils n'ont pas le même gain. Pour rétablir une détection correcte il faudra diviser chaque sortie de filtre par son gain. Pour cela on intègre la réponse impulsionnelle de chaque filtre. Ayant remarqué que la réponse impulsionnelle du filtre est une succession de Dirac d'amplitude décroissante d'un facteur α l'impulsion précédente.

$$\gamma(\alpha) = (1-\alpha)\sum_{n=0}^{N}\alpha^n \tag{2.14}$$
$$\gamma(\alpha) = \frac{1-\alpha}{1+\alpha} \tag{2.15}$$

La matrice $R_c(n,\tau)$ qui est la sortie des filtres résonnants, présente un profil $Rn(n,\tau)$ à τ fixé qui est d'amplitude moyenne supérieure aux autres périodes, si la période τ est présente dans le signal. Les sommets permettront de localiser les instants précis où sont placés les accents dans le morceau. Nous nous intéressons ici à l'amplitude moyenne qui nous permettrait de dire qu'à une date n la période τ est ressentie dans le morceau.
Pour cela il faut extraire l'amplitude moyenne tout au long du morceau pour chaque période τ. Le procédé se déroule en deux étapes :

- Filtrer le signal avec un filtre moyenneur de la taille de la période considérée sur toutes les bandes de périodes de la matrice :

$$\hat{r}_c(\tau,n) = \frac{1}{\tau}\sum_{i=n-\tau+1}^{n} = r_c(\tau,i).^2 \tag{2.16}$$

  Cette opération est équivalente à convoluer le signal avec une fenêtre rectangulaire de la taille de τ pondérée par $\frac{1}{\tau}$
- Normaliser les sorties en fonction des différences de gain de chaque filtre résonnant :

$$s_c(\tau,n) = \frac{1}{1-\gamma(\alpha_\tau)}\left(\frac{\hat{r}_c(\tau,n)}{\hat{v}_c(n)} - \gamma(\alpha_\tau)\right) \tag{2.17}$$

Les figures [2.2.1.2,2.13] montrent la sortie des filtres résonnants ayant en entrée une fonction de détection d'onsets synthétiques (dans ce cas un train d'impulsion d'une fréquence de 2Hz convolué avec une fenêtre de Hann), pour différentes étapes de normalisation (décrites plus bas).

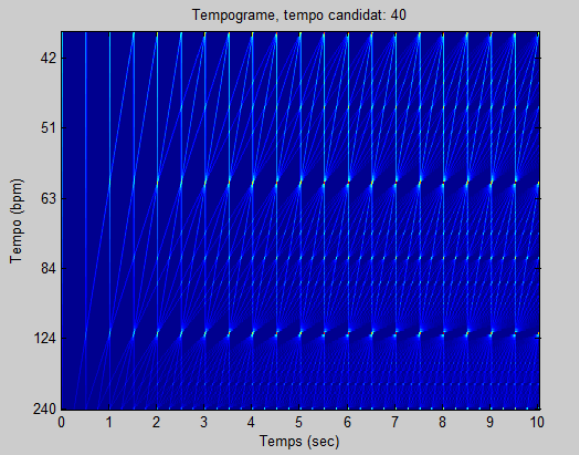


FIGURE 2.12 – Sortie des filtres résonnants non normalisée $R_c(n, \tau)$

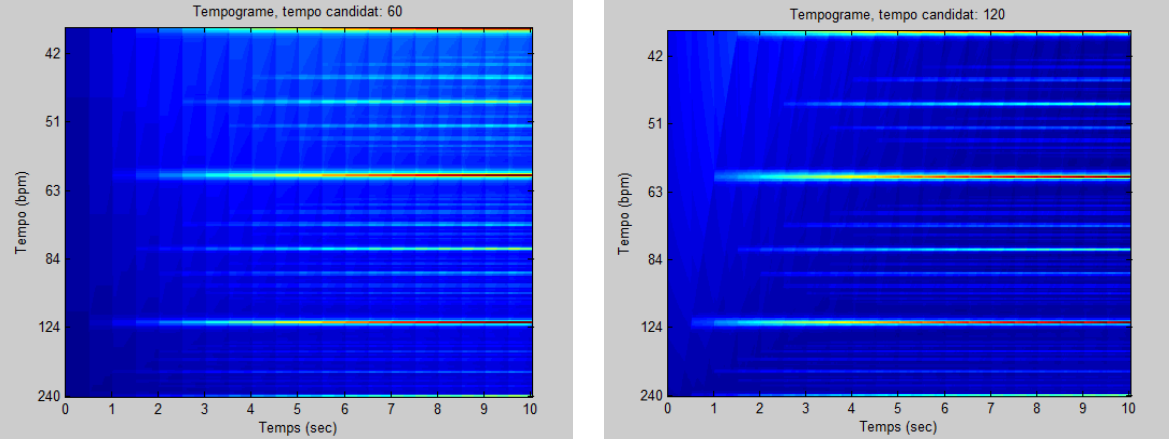


FIGURE 2.13 – La figure de gauche représente la sortie des filtres résonnants dont l'énergie est extraite avec un filtre moyenneur $\hat{R}_c(n, \tau)$, la figure de droite représente la sortie des filtres résonnants normalisée $S_c(n, \tau)$

2.2.1.3 Périodicité du Tatum

La périodicité du *tatum* est évaluée à l'aide de la *Densité Spectrale de Puissance* des sorties normalisées des bancs de filtres. Cette grandeur permet de représenter les différentes composantes spectrales d'un signal et d'en effectuer l'analyse harmonique. Seules les fréquences inférieures à 20Hz sont utilisées comme observations les *Tatum* ayant une périodicité supérieure sont très rare (20Hz correspond à 1200 bpm). Il est déterminé avec l'équation suivante :

$$S(f, n) = f\left|\frac{1}{\tau_{max}} \sum_{\tau=1}^{\tau_{max}} (s(\tau, n))\zeta(\tau) e^{-i2\pi f(\tau-1)/\tau_{max}}\right|^2 \tag{2.18}$$

où, $\zeta(\tau)$ est une demi fenêtre de Hann.

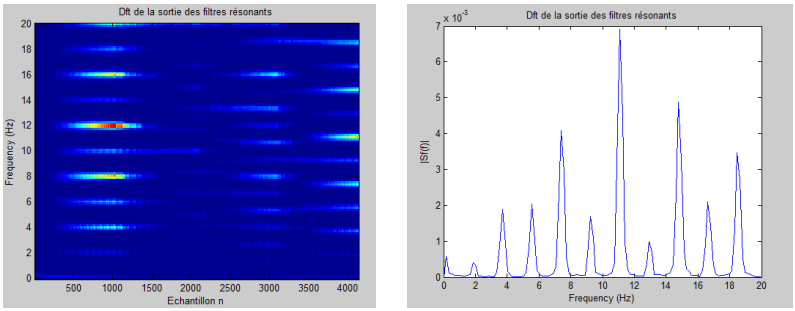


Figure 2.14 – Dft de la sortie des filtres normalisé à droite matrice des estimation pour tous les temps, à gauche profil d'estimation pour un temps fixé

2.2.2 Induction du tempo

Il est important de souligner le fait que ce que l'on cherche à estimer sont les périodicités les plus probables du *Tactus*, *Tatum* et de la *Mesure* en fonction du temps du morceau (ici toutes les secondes). On se servira des profils $s_c(\tau, n)$ et $S(f, n)$ calculés précédemment qui seront les observations d'un système qui résout l'estimation la plus probable d'une séquence de périodicités au vu de règles musicales modélisant la dépendance de la probabilité à un instant t connaissant la probabilité à l'instant $t-1$, et les dépendances entre les périodicités du *tactus*, *tatum* et *mesure*.

2.2.2.1 Chaine de Markov d'ordre 1

Une chaîne de Markov d'ordre 1 est un processus aléatoire qui est caractérisé par la propriété de Markov faible : la distribution conditionnelle de probabilité de distribution de l'état présent ne dépend que de l'état précédant. Dans le cas de notre implémentation les données sont discrètes, (i.e les variables aléatoires qui représentent les états de la chaîne sont à valeur dans un espace d'états discret).

Propriété de Markov faible :

$$P(q_t = Q_j | q_{n-1} = Q_i, q_{n-2} = Q_k, ...) = P(q_n = Q_j | s_{n-1} = Q_i). \tag{2.19}$$

où, q_n est l'état dans lequel se trouve le système à l'instant n qui est à valeur dans Q qui est l'espace des états.

Propriété d'homogénéité On suppose en plus que la chaîne est homogène ce qui implique que les valeurs de transitions a_{ij} de l'état i vers l'état j sont indépendantes de l'état n.

$$\forall t \geq 1, \forall (i,j) \in Q^2, \tag{2.20}$$

$$P(q_n = Q_j | q_{n-1} = S_i) \quad = \quad P(q_1 = Q_j | q_1 = Q_i). \tag{2.21}$$

Caractérisation d'une chaine de Markov Une chaine de Markov est donc caractérisée par :

- Sa matrice de transition : $A = a_{ij}$ qui est une matrice stochastique ($\sum_j a_{i,j} = 1$).
- Sa distribution initiale $\pi_i(q_1 = Q_i)$, $\forall i \in Q$.

2.2.2.2 Chaines de Markov Cachées

Les Chaines de Markov Cachées (en anglais Hidden Markov Model, HMM) est un double processus aléatoire. Composé d'un premier modèle aléatoire sous-jacent non observable à estimer (caché, ce sont les états), et qui ne peut être observé qu'à travers un autre processus stochastique qui produit la séquence d'observations [19]
Cela induit l'existence d'une matrice d'émission qui permet de faire le lien entre les deux processus, en répertoriant la probabilité des états cachés connaissant les observations.

Paramètres caractéristiques d'un HMM Une chaine de Markov cachée est définie à l'aide du quintuplet $\lambda(A, B, \pi)$. De la même manière que pour une chaine de Markov cachée, la matrice A désigne la matrice de transition des états *cachés*, π la distribution de probabilité initiale. Les autres paramètres sont :

- N le nombre d'états (cachés) dans le modèle, les états sont notées $Q = \{q_1, ...q_N\}$
- T, le nombre d'observations, notées $O = \{o_1, ...o_{,T}\}$
- $B : b_i(o_t) = P(o_t|q_n = Q_i)$, $\forall i \in \{1, ..N\}$, $\forall t \in \{1, ..T\}$

Résolution d'un HMM Pour implémenter notre algorithme on aura besoin de répondre à deux problèmes parmi les trois listés par Rabiner :

1. Étant donné un HMM $\lambda = (A, B, \pi)$ et une séquence d'observations $O = \{o_1, o_2, .., o_T\}$, comment calculer la vraisemblance de $P(Q|\lambda)$ de la séquence d'observations connaissant le modèle ?
2. Étant donné un HMM, $\lambda = (A, B, \pi)$ et une séquence d'observations $O = \{o_1, o_2, .., o_T\}$ comment trouver la séquence d'états $Q = \{q_1, q_2, .., q_T\}$ qui décrit au mieux les observations ?

La résolution de ces deux problèmes et la mise en contexte avec notre implémentation fait l'objet des deux prochaines sections.

Mise en contexte On recherche la probabilité des observations O en fonction des états Q et du modèle.

Hypothèse : les données sont indépendantes entre elles conditionnellement aux états.
On peut donc écrire :

$$P(O|Q, \lambda) = \prod_{n=1}^{N} P(o_t|q_t, \lambda), \tag{2.22}$$

$$P(O|Q, \lambda) = \prod_{n=1}^{N} b_n(o_t). \tag{2.23}$$

La probabilité d'apparition des états Q est :

$$P(Q|\lambda) = \pi_i a_{12} a_{23} ... a_{T_1 T} \tag{2.24}$$

La probabilité jointe de Q et$O$ s'écrit :

$$P(Q, O|\lambda) = P(Q|O, \lambda)P(O, \lambda), \tag{2.25}$$

La probabilité d'obtenir la séquence d'observations O est la somme de tous les chemins d'états possibles :

$$P(O|\lambda) = \sum_Q P(Q|O,\lambda)P(O,\lambda) \tag{2.26}$$

$$P(O|\lambda) = \sum_Q \pi_1 b_1(q_1) a_{12} b_2(q_2) ... a_{T-1T} b_T(q_T). \tag{2.27}$$

2.2.2.3 Problème posé

On cherche à estimer les trois états conjoints, les périodicités du *tatum*, du *tactus* et de la *mesure*. On note respectivement τ_n^A, τ_n^B,τ_n^C, les trois paramètres à estimer au temps n, et donc q_n l'état au temps n. Les observations sont la matrice des sorties des filtres résonnants s_c (pour le tactus et mesure), et leurs spectres de puissance S_f (tatum). On note s_n les ob
servations au temps n
On écrit alors l'équation :

$$P(Q,O) = P(q_1)p(s_1|q-1)\prod_{n=2}^{N} P(q_n|q_{n-1})p(s_n|q_n) \tag{2.28}$$

On modélise aussi une loi de dépendance *intra* états entre les variables τ^A, τ^B et τ^C. Afin de simplifier les équations, on ne conserve que les dépendances avec le *tactus*. ainsi :

$$P(q_n|q_{n-1}) = P(\tau^B|q_{n-1})P(\tau_n^A|\tau_n^B, q_{n-1})P(\tau_n^C|\tau_n^B, \tau^A, q_{n-1}), \tag{2.29}$$

devient :

$$P(q_n|q_{n-1}) = P(\tau^B|\tau_{n-1}^B)P(\tau_n^A|\tau_n^B, \tau_{n-1}^A)P(\tau_n^C|\tau_n^B, \tau_{n-1}^C). \tag{2.30}$$

Matrice d'émission Dans l'article la vraisemblance des estimations connaissant les états est simplifiée en considérant que les états sont indépendants entre eux, les relations de dépendances inter-états apparaitront dans le calcul de $P(\tau_n^A|\tau_n^B, \tau_{n-1}^A)$ et de son homologue $P(\tau_n^A|\tau_n^B, \tau_{n-1}^A)$. La vraisemblance des états connaissant les estimations s'exprime comme le produit des probabilités des états connaissant les observations, c'est à dire des matrices d'émissions qui sont ici s_c et S_f :

$$p(s|q_n = [j,k,l]) \propto s_c(k,n)s_c(l,n)S_f(\frac{1}{j}, n) \tag{2.31}$$

Il est important de remarquer que l'égalité est ici une égalité en loi car $s_c(k,n)s_c(l,n)S_f(\frac{1}{j}, n)$ n'est pas une probabilité, ses valeurs ne sont pas comprise entre 0 1.

Matrice de transition Les dépendances inter états sont donc modélisées dans la matrice de transition par la probabilité $P(\tau_n^A|\tau_n^B, \tau_{n-1}^A)$, pour simplifier l'écriture des calculs on écrit $P(A|B,C)$. L'estimation de trois états dépendants peux se décomposer de la manière suivante :

$$P(A,B,C) = P(A,B|C)P(C), \tag{2.32}$$

$$P(A,B,C) = P(A|B,C)P(B|C)P(C), \tag{2.33}$$

$$\Rightarrow P(A|B,C) = \frac{P(A,B|C)}{P(B|C)}. \tag{2.34}$$

En remplaçant A, B, C par $\tau_n^A, \tau_n^B, \tau_{n-1}^A$ on obtient l'écriture de la probabilité inscrite dans [3] :

$$P(\tau_n^A|\tau_n^B, \tau_{n-1}^A) = \frac{P(\tau_n^A, \tau_n^B|\tau_{n-1}^A)}{P(\tau_n^B|\tau_{n-1}^A)}, \tag{2.35}$$

$$P(\tau_n^A|\tau_n^B, \tau_{n-1}^A) = P(\tau_n^A|\tau_{n-1}^A)\frac{P(\tau_n^A, \tau_n^B|\tau_{n-1}^A)}{P(\tau_n^A|\tau_{n-1}^A)P(\tau_n^B|\tau_{n-1}^A)}, \tag{2.36}$$

$$\tag{2.37}$$

qui permet de décomposer en deux parties la probabilité recherchée :

- La probabilité de transition, modélisée par $P(\tau_n^i|\tau_{n-1}^i)$, est une distribution normale centrée dont le paramètre est le logarithme du rapport entre les estimations des deux périodes. La variance est choisie de manière à ce qu'un rapport double ou moitié entre $P\tau_n^i\ \tau_{n-1}^i$ aie une probabilité nulle. Si elle sont identiques le paramètre sera nul et la distribution sera maximale . Une représentation de cette loi est donnée fig. [2.15].

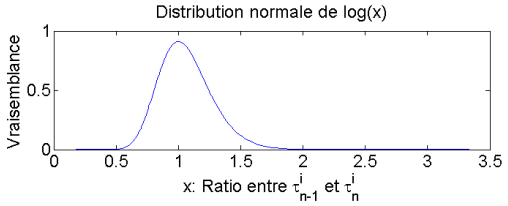


FIGURE 2.15 – Ratio entre τ_n^i et τ_{n-1}^i

- La relation de dépendance inter états, est donnée par : $\frac{P(\tau_n^A,\tau_n^B|\tau_{n-1}^A)}{P(\tau_n^A|\tau_{n-1}^A)P(\tau_n^B|\tau_{n-1}^A)}$ en posant une égalité en loi avec une distribution $g(\frac{\tau_n^B}{\tau_n^A})$ modélisant le rapport entre les deux période. Elle est modélisée par un mélange de gaussienne dont les poids sont attribués en fonction d'*a priori musicaux* sur ratios des périodicités les plus courantes en musique. Elle est représentée en fig. [2.16]

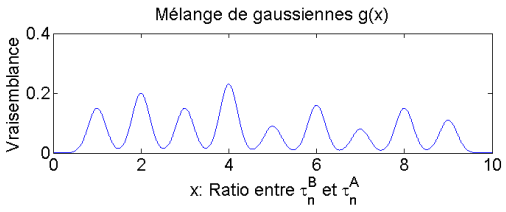


Figure 2.16 – Ratio entre τ_n^i et τ_{n-1}^i

Sélection des états Les états candidats sont obtenus en maximisant la vraisemblance de la multiplication d'une loi *a priori* $P(\tau^i)$ et de la probabilité $P(\tau_n^i|s) = s(\tau_n^i, n)$, $i \in \{A, B, C\}$ et $s = [s_c, S_f]$. $P(\tau^i)$ modélise la distribution *attendue* en fonction de mesures réalisées sur une base de données. Chaque $P(\tau^i)$ est modélisé par une loi log-normale dont les paramètres d'écart type et de variance sont estimés sur un base de données et illustré sur la figure Fig.[2.17].

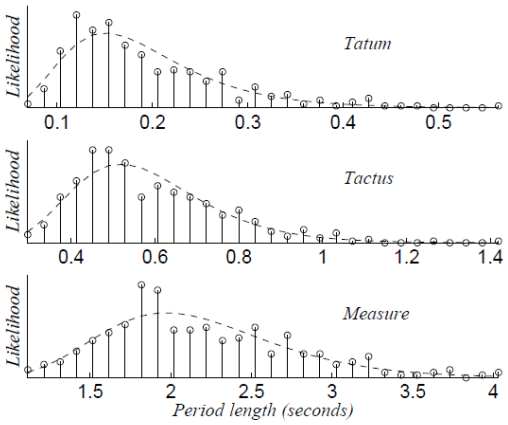


Figure 2.17 – Estimation par inférence de la probabilité $P(\tau^i)$, figure extraite de [3]

On sélectionne, à chaque instant n (toutes les secondes), les 5 maximums *Greedy* de $P(\tau^i)P(\tau_n^i|s)$. Pour ne pas obtenir les valeurs voisines correspondant au même pic, on sélectionne le maximum et on force les k valeurs voisines à zéros, où, k est égal respectivement pour le *tatum*,*tactus*, *mesure*, à $\{1, 3, 5\}$. La maximisation est représentée sur Fig.[2.18].

On fait une estimation des périodicités toutes les secondes en utilisant les matrices de transition et d'émission décrites plus haut. Pour cela on utilisera l'algorithme de Viterbi.

2.2.2.4 Résolution du second problème : l'algorithme de Viterbi

On cherche la séquence maximale qui maximise la probabilité $P(O|Q, \lambda)$ ce qui est équivalent à maximiser $P(O, Q|\lambda)$. La solution adoptée est l'algorithme de Viterbi. C'est un algorithme qui permet d'optimiser le temps de calcul du meilleur chemin en proposant un cout de calcul de $O(T \times N^2)$. Une approche gourmande en temps serai de calculer l'ensemble des chemins

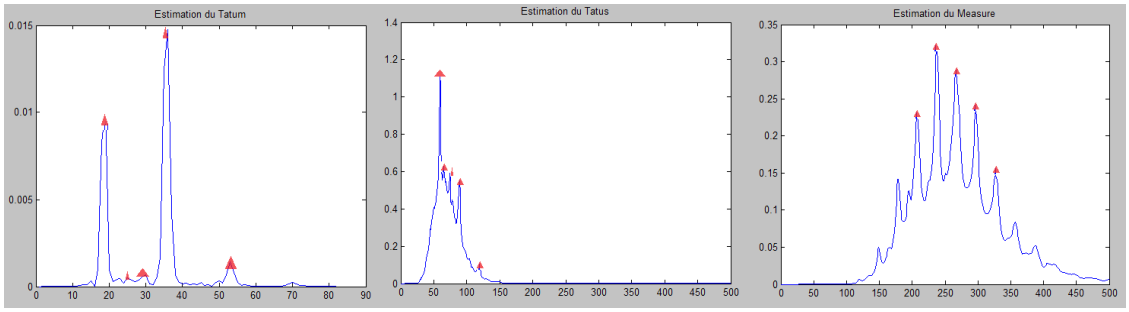


FIGURE 2.18 – Estimation *Greedy* des états sur $P(\tau^i)P(\tau_n^i|s)$ sur le morceau 'amado mio' du groupe Pink Martini

ayant un cout de calcul $O(N^T)$. Le principe est de ne garder pour chaque estimation d'état q_n, que l'antécédent q_{n-1}, qui maximise la probabilité du chemin $\{q_1...q_n\}$. On pose :

$$\delta_t(i) = \max_{q_i,..q_T-1} P[o_1,..o_t = i, q_1, ..., q_t|\lambda]. \tag{2.38}$$

Probabilité maximale, à l'instant n, à travers *un seul chemin*. Elle représente les n observations et finit à l'état q_i. Par itération :

$$\delta_{t+1} = \max_i[\delta_t(i)a_{ij}b_j(q_{t+1})] \tag{2.39}$$

Probabilité du meilleur parent q_n. Pour pouvoir récupérer la séquence d'états, on conserve les *arguments* qui maximisent (2.39). Ce sont des pointeurs sur les états majorants (2.38). Il sont gardés dans la matrice $\psi_t(j)$. L'implémentation de l'algorithme est décrite par le pseudo code extrait de [19] :

Algorithme de Viterbi

Entrées : $\lambda = \{A, B, \pi\}$, $O = \{o_1, ...o_T\}$

pour $i = 1 \rightarrow N$
$\delta_1(i) = \pi_i b_i(q_1)$
$\psi_1(i) = 0$
fin pour

pour $t = 2 \rightarrow T$
pour $i = 1 \rightarrow N$
$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] b_j(q_t)$
$\psi_t(j) = \arg\max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}]$
fin pour
fin pour

$P^* = \max_{1 \leq i \leq N} [\delta_T(i)]$
$q_T^* = \arg\max_{1 \leq i \leq N} [\delta_T(i)]$

pour $t = T - 1 \rightarrow 1$
$q_t^* = \psi_{t+1}(s_{t+1}^*)$
fin pour

Sorties : $Q^* = \{q_1^*, ...q_T^*\}$

2.2.3 Estimation des phases

2.2.3.1 *Tatum*, *Tactus*

Le terme de *phase* dans le langage de la détection de tempo désigne les instants de la pulsation. Pour estimer les phases on se servira des matrices Rn des sorties des bancs de filtres résonnants non normalisées.

On estimera les phases sur la somme pondérée des 4 matrices Rn : attribuant un poids plus important pour la sortie des filtres en basses fréquences.

Si la sortie du filtre $Rn(\tau = j, n)$ est de forte amplitude, alors $[\tau = j]$ est la période d'une pulsation présente dans le morceau à cet instant (sur une durée d'au moins $4s$ ref.[2.2.1.2]). La phase de la pulsation du morceau coïncide alors avec les pics d'amplitudes de $Rn(\tau = j, n - vois)$ ou $n - vois$ représente un segment d'échantillons voisins au temps n. La phase ϕ_n est l'échantillon correspondant au pic d'amplitude maximum de $Rn(\tau = j, n - \tau_n)$.

Considérant alors que ces instants ont une période $\hat{\tau}_t^i$, déterminer la phase du *tatum* et du *tactus* revient à chercher le pic d'amplitude maximum sur un horizon de la taille de la période. Ce procédé est illustré sur la figure suivante, fig. [2.19].

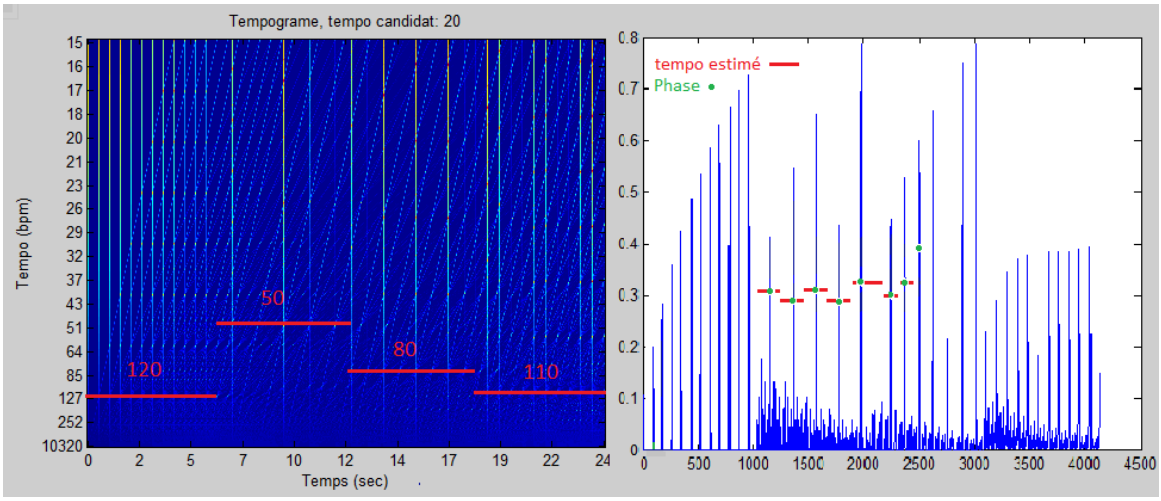


FIGURE 2.19 – Procédé d'évaluation des phases sur un signal synthétique en peigne avec fréquences variables : 1.estimation des périodicités, 2.estimation des phases.

Viterbi Afin de d'imposer une régularité dans l'estimation de la phase, on utilise un algorithme de Viterbi. La matrice de transitions impose une régularité en pénalisant les estimations de phases adjacentes trop éloignées. Ce qui permet d'estimer une séquence de phases présentant un compris entre l'énergie du banc de philtre à l'instant ϕ_n et la régularité de l'attribution des phases (conforme à la période estimée).

Matrice de transition La loi de transition est une loi exponentielle qui permet de pénaliser un terme d'erreur e qui est corrélé à la distance entre deux phases adjacentes $|\phi_n^i - \phi_{n-1}^i|$. La représentation de la matrice de transition est représentée sur la figure [2.20]. Le calcul de son expression est :

$$P(\phi_n^i|\phi_n^i) = \frac{1}{\sigma\sqrt{2\pi}} exp\Big(-\frac{e^2}{2\sigma^2}\Big) \tag{2.40}$$

$$e = \frac{1}{\hat{\tau}_n^i}\Big\{\Big[\Big(\Big|\phi_n^i - \phi_{n-1}^i\Big| \mod \tau_n^i\Big)\Big] - \frac{\hat{\tau}_n^i}{2}\Big\}. \tag{2.41}$$

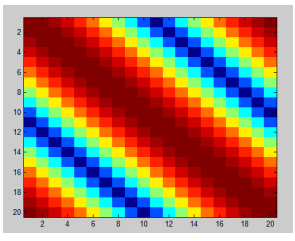


FIGURE 2.20 – Matrice de transition pour l'estimation de la séquence des phases ϕ_i

2.2.3.2 Problèmes de phases

Problème La matrice de transition permet d'estimer une phase qui est dans le voisinage de la phase attendue, en respect de la phase précédente, et en fonction du *modulo* de la périodicité estimée.
Il est nécessaire d'avoir une estimation *modulo*afin d'être plus robuste aux problèmes d'estimation des périodicités. Mais cela encourage la double annotation d'une phase (périodicités trop courtes), ou bien le saut de phase (périodicités trop longues).

Solution Une solution simple est de faire un post-traitement des phases afin de détecter les erreurs de doubles phases et manquement de phases, puis de rétablir une séquence plus régulière.

Une solution plus sophistiqué est d'estimer les phases sous la forme d'un cycle Estimation-Maximisation :

1. E : Estimation des phases
2. M : post-traitement des phases et obtention d'une nouvelle séquence $\hat{\tau_n}$ de périodicité.
3. retour en 1.

Cette approche a été implémentée et à donné dès la deuxième itération une distribution des phases très irrégulières ne respectant plus aucune périodicité. Cela permet de rendre compte de la grande dépendance entre l'évaluation des périodicités et l'évaluation des phases.

Poids de la matrice de transition Lorsque le poids de la matrice de transition est plus important, la dépendance aux périodicités estimées décrois. Les premiers résultats obtenus sont encourageants. Ci-dessous une figure qui montre l'estimation de la phase en fonction des observations, et du paramètre α qui est le poids d'attribution de la mesure.

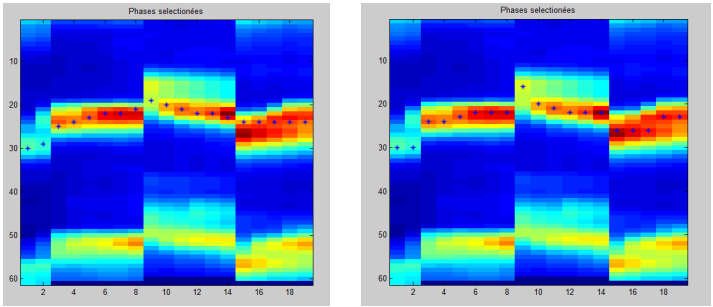


Figure 2.21 – Estimation des phases en fonction du paramètre α

2.2.3.3 Estimation des phases de la mesure

L'estimation des phases de la mesure se fait grâce à un *a priori* musical sur les pulsations en fonction des différentes bandes de fréquences.
L'exemple le plus efficace est le coup caisse claire qui marque le début des mesures dans la musique rock (hautes fréquences). La périodicité du *beat* est ressentie sur toutes les bandes

de fréquences avec un poids décroissant en fonction de la fréquence.

l'*a priori* est modélisé en sous la forme des deux matrices suivantes :

$$\eta^1_{ck} = \begin{bmatrix} 12 & 1 & 0 & 5.7 \\ 0 & 2 & 0 & 2 \\ 0 & 3 & 0 & 3 \\ 0 & 4 & 0 & 4 \end{bmatrix} \eta^2_{ck} = \begin{bmatrix} 10 & 0 & 1.4 & 1.3 \\ 0 & 0 & 2.8 & 0.8 \\ 0 & 0 & 4.3 & 1.2 \\ 0 & 0 & 5.8 & 1.5 \end{bmatrix} \tag{2.42}$$

où c est le numéro de la fonction de détection allant de $\{1, ..4\}$, respectivement des basses fréquences aux fréquences plus élevées, k est le numéro de *beat*.

2.2.4 Apport Mélodique

Déterminer les accents mélodiques devrait permettre de pouvoir définir un a priori musical permettant de représenter les mécanismes perceptifs à l'œuvre lors de la définition des mesures. celles ci se trouvant aux endroits conjoints des accents métriques et mélodiques.

2.2.4.1 Traitement sur flux audio

Voici une proposition de méthode permettant de déterminer le contour de la mélodie sur un signal monophonique. Les traitements doivent être traités en parallèle pour le tactus et le tatum.

Tfct La Tfct est réalisée aux phases du tactus ou du tatum évaluées. La longueur de la fenêtre est celle de la différence maximum entre deux instants consécutifs de la phase de l'item évalué.

Filtre médian Un filtre médian est appliqué sur le spectre du signal permettant de déterminer la courbe d'un seuil qui permettra de discriminer la partie bruit de la partie harmonique. La médiane est une mesure statistique qui représente une alternative robuste à la moyenne, c'est la valeur "du milieu" de l'échantillon ordonné. L'utilisation d'un filtre médian pourra remplacer celle d'un filtre moyenneur, il sera moins sensible aux valeurs extrêmes du signal ne les prenant pas toujours en compte.

Représentation des hauteurs de pitch midi On définit la banque de filtres à bandes passantes, qui permettent d'obtenir des bandes de fréquences alignées sur la gamme tempérée.

"Débruitage" Pour éviter que la détection soit perturbée par la présence de pics de bruit pendant une période de silence entre deux notes, on considère que la note continue pendant cette période. Pour cela on s'aide du formalisme de Goto dans [22]

Sélection de la mélodie La sélection de la mélodie se fait à l'aide d'un algorithme de Viterbi qui retiendra un chemin qui fait un compromis entre faibles variations et énergie importante du pitch.

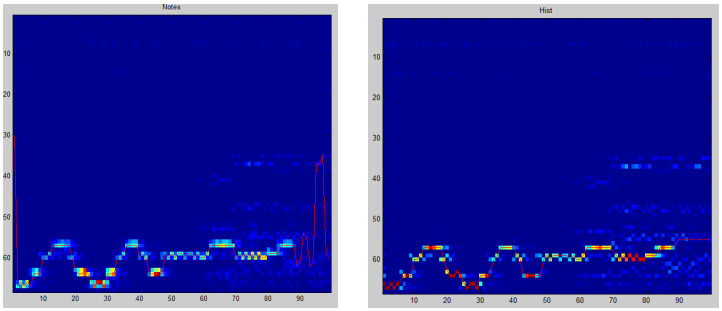


Figure 2.22 – Dessin d'une mélodie monophonique sur l'estimation de notes et sur l'histogramme

2.2.4.2 Détection de mélodie existante

Le traitement sur flux audio proposé n'est pas résistant aux signaux polyphoniques, il faut faire pour cela en amont et/ou pendant la détection de mélodie une séparation de source. C'est une tâche de traitement du signal délicate et il a été préférable d'utiliser un algorithme existant [14]. L'algorithme que l'on utilise est un algorithme de séparation de source appliqué à une représentation du signal CQT (qui est une représentation temps fréquence du signal avec une échelle de répartition des fréquences qui est celle de la gamme tempérée à facteur Q-constant), toujours dans l'optique d'extraire une mélodie qui s'approche de celle que l'on perçoit. La séparation permet de scinder la mélodie de son accompagnement, en définissant des paramètres différents pour les deux entités.

- La partie d'accompagnement est décomposé avec une NMF (Non négative Matrix Factorisation) sans information sur le dictionnaire de décomposition sinon le nombre de canaux R souhaités.
- La partie mélodique à un a priori sur la forme spectrale du contenu, pour chaque trame de temps il est exigé d'avoir une estimation du pitch et de l'*enveloppe spectrale* de la trame. L'enveloppe spectrale est la somme pondérée des harmoniques du pitch

L'algorithme EM est ensuite exécuté en deux temps : une première fois de manière aveugle, puis, afin de s'assurer que le résultat n'offre qu'une seule estimation du pitch par bande de fréquence un algorithme de Viterbi permettant de choisir une chemin en faisant un compris entre une énergie importante des pitchs sélectionnés et une trajectoire avec peu de variations. Les hauteurs éloignées de plus d'un demi ton sont remises à zéro et une deuxième estimation est lancée.
Ceci nous permet d'obtenir la trajectoire du pitch en fonction du temps avec une précision temporelle de 10ms.

2.2.4.3 Déterminer l'accent mélodique

Les relations locales de hauteur, entre notes voisines, peuvent être envisagées de manière encore plus simple, par la simple prise en compte du contour, c'est- à-dire du sens de variation de l'intervalle mélodique : ascendant, descendant ou stagnant. Comme rapporté dans [23] la seule prise en compte de l'intervalle mélodique rapporte les meilleurs résultats parmi plusieurs modèles perceptifs testés. Le modèle que l'on utilise ici est le modèle de Thomassen [fig.2.23] qui attribue un poids en fonction du sens de variation de l'intervalle mélodique : l'accent

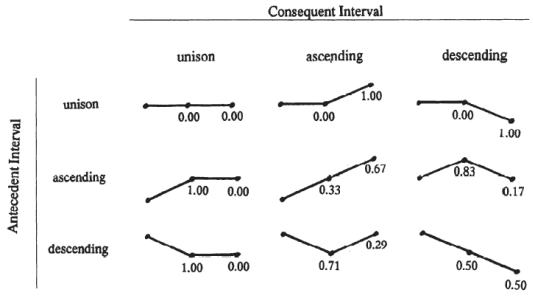


FIGURE 2.23 – Model théorique des accents mélodiques Thomassen

obtenu d'après la mélodie est montré sur la figure [2.23].

Des accents mélodiques aux phases de la mesure La détermination de l'accent mélodique se fait en utilisant l'article de [22] qui permet d'évaluer la redondance sur le temps d'une mesure ou du beat de l'accentuation apporté par les changements harmoniques.
L'accent mélodique peut être inclus dans le modèle présenté en attribuant un poids relatif à son contour à la sortie des filtre résonnants $R_c(\tau, n)$ pour l'estimation de la mesure.

L'insertion de l'apriori dans le modèle n'a pas encore été réalisé. Mais cela ouvre des perspectives sur de nouveaux travaux. Il serai intéressant de comparer les scores avec ceux de la méthode [22] permettant de quantifier les changements harmoniques.

Chapitre 3

Évaluation

La base de données qui a servi à l'évaluation se se limite au rock, pop rap, RnB. Elle ne contient pas de musique classique, ni de jazz pour lesquels il est difficile d'évaluer un tempo : dans le premier cas, à cause de l'absence d'événements percussifs et la relative absence d'accent sur les premiers temps de mesures ; dans le second cas à cause des rythmiques souvent compliquées car variables et au chiffrage parfois improbable, souvent ternaires, et de l'interprétation swing qui peut perturber la détection du beat.

3.1 Présentation des méthodes

Il y a plusieurs difficultés ou impasses auxquelles les évaluations doivent être robustes.

- La comparaison de la sortie d'algorithmes et des annotations doit prendre en compte une fenêtre de tolérance. (Il est difficile d'avoir une précision plus fine que 50ms), elle peut être définie en temps absolu ou en temps relatif.
- La méthode d'obtention des vérités terrain est caractérisée par l'application que l'on cherche à tester. Elle peut être orientée sur la détection d'un tempo perceptif (équivalent à celui qu'on obtient naturellement en tapant dans nos mains), ou bien sur un tempo quantitatif qui est celui que l'on trouvera sur l'annotation d'une partition.
- Lorsque l'on se réfère à une base d'annotations perceptives obtenues en enregistrant la battue des annotateurs, l'une des plus grandes diversité présente est le décalage d'un demi- beat, certains auront tendance à battre le tempo *off-beat* et d'autres *on-beat*.
- La deuxième variation importante d'un annotateur à l'autre est la période à laquelle les beats sont détectés, parfois ils seront battus au double tempo, ou au demi tempo. Selon la métrique utilisée cette erreur peut être négligée (binaire) ou dénoter d'un véritable problème d'interprétation (ternaire).
  Au vu des difficultés énoncées on va comparer plusieurs méthodes qui permettent de dresser un profil complet de l'évaluation du beat. Les notations sont celles de [18]. on note γ la séquence des B beats évalués avec l'algorithme de détection du beat, avec γ_b le b^{ieme}. De manière respectives les valeurs des annotations sont la séquence a de J échantillons a_j.

3.1.1 Différentes mesures

3.1.1.1 F-mesure

La F-mesure est l'évaluation générique souvent utilisée en recherche d'informations musicales. Pour l'évaluation de beat la F-mesure est calculée à l'aide de trois paramètres : c qui est le nombre de beats correctement alloués, f^- qui est le nombre de beats qui sont placés avant le beat annoté, et f^+ le nombre de beats qui sont placés après. La mesure se calcule en respectant l'équation :

$$F = \frac{2c}{2c + f^- + f^+} \times 100 \tag{3.1}$$

La F-mesure est donc une mesure exprimée en pourcentage, qui pénalise avec un score de zero pour une estimation à contre temps, et pénalisera en proportion du nombre de beats additionnés ou manquant les battues au multiple ou diviseur de la période.

3.1.1.2 Cemgil et al.

La mesure proposée par Cemgil est proportionnelle à la distance du beat estimé au beat annoté le plus proche.

$$W(x) = exp(-x^2/2\sigma_e^2), \tag{3.2}$$

où, $x = \gamma_b - a_j$, la distance ici n'est pas une valeur absolue, ceci impliquant que les erreurs de phases positives ne sont pas traitées de la même manière que les erreurs de phases négatives. Les erreurs de phases négatives telles que $x > \sigma_e$ impliquent que la distance W tend vers 0. Les erreurs de phases positives $x < 0$: la mesure sera dévaluée par décimation d'un facteur$B$ (nombre de beats détectés) plus grand dans la formule ci-dessous :

$$Cem_acc = \frac{\sum_j max_b W(\gamma_b - a_j)}{(B + J)/2} \times 100 \tag{3.3}$$

Par ailleurs, l'écart type σ_e est suffisamment petit (40ms) pour que l'évaluation du beat à contre temps soit pénalisée par une mesure à zero et ce pour des tempi allant jusqu'à 240 bpm.

3.1.1.3 Pscore

La méthode de Pscore attribue la valeur de la corrélation croisée entre un peigne déterminé par le temps des beats annotés et un peigne déterminé par les beats estimés. Un certain nombre d'ajustements sont réalisés afin de faciliter le calcul et d'autoriser un certaine variabilité. La mesure de Pscore a été développée pendant la campagne MIREX 2006 par McKinney sur une vérité terrain déterminée avec le concours des enregistrements des battues de tempo par 40 annotateurs. Les cinq premières secondes des deux trains d'impulsions sont retirés (pour éviter les erreurs d'annotation sur les premier beats). Les peignes sont alignés avec une valeur toutes les 10ms. Une variabilité de 20% sur les corrélations des annotations et estimations, est permise en prenant le résultat de la corrélation sur une fenêtre w de taille fixée par rapport à la médiane des inter-annotations Δ_j : $w = 0.2\text{median}(\delta_j)$, l'opération de corrélation croisée avec variabilité w est noté $\star_{(w)}$. La somme est normalisée par le plus grand nombre d'annotations J ou estimations B.

$$Pscore = \frac{\sum_w T_a \star_{(w)} t_\gamma}{max(J, B)} \times 100 \tag{3.4}$$

3.1.2 Histogramme

Les mesures Pscore et F-mesure attribuent à chaque beat évalué une valeur binaire selon qu'il est ou non présent dans la fenêtre de tolérance centrée sur le beat annoté. Ceci implique que cette mesure ne permet pas de faire la différence entre une méthode oracle et une méthode pour laquelle les phases des beats seraient légerement décalées. La méthode de Cemgil et al. présente une évaluation plus fine qui prend en compte la distance au beat annoté, cependant l'ensemble des trois méthodes donne un score nul pour un rythme qui est évalué à contre temps et donne des faibles score pour une évaluation aléatoire. La méthode proposée permet a contrario :

- De prendre en compte la distance du beat aux annotations,
- d'avoir un meilleur score pour une évaluation à contre temps que pour une évaluation sur des beats distribués aléatoirement,
- de plus la représentation sous forme d'histogramme permet d'avoir une meilleure idée des défauts éventuels de la méthode à estimer (ce qui est intéressant pour l'évaluation de méthodes à performer.)

La méthode estimée propose d'évaluer les distances temporelles entre les beats et les annotations. Ces valeurs sont assignées dans un histogramme, et on les compare à la distance entre l'histogramme obtenu et un histogramme dont les phases du beats sont distribuées aléatoirement. Cette méthode consiste à déterminer la séquence des temps γ_q des beats estimés compris dans la fenêtre $[\Delta^*_{j-1}\Delta^*_j]$ centrée en a_j. On écrit : $\gamma_q = \gamma_b \quad : \quad a_j - \Delta^*_{j-1} \leq \gamma_b \leq a_j - \Delta^*_j$.

Construction de l'histogramme La distance de chaque beat à son annotation la plus proche est normalisée de façon à obtenir des valeurs comprises entre -0.5 et 0.5 On obtient ainsi une séquence correspondant au nombre de beats estimés que l'on note $\zeta_{\gamma|a}$.
Cependant, si le tempo estimé est deux fois trop lent mais que les phases des beats estimés sont en correspondance avec celles des annotations, on ne détectera pas d'erreur. Il est donc proposé de faire les mêmes calculs pour les annotations en fonction des estimations.
On notera la nouvelle séquence obtenue $\zeta_{a|\gamma}$. Les valeurs contenues dans les deux séquences sont assignées dans un histogramme. Les distances sont regroupées en $K = 41$ valeurs allant de -0.5 à 0.5. La probabilité estimée $p_\zeta(z_k)$ de la distance centrale z_k avec $z_k \in [-0.5\ 0.5]$ et $k \in 1, ..., K$ est conditionnée par le fait que $\sum_{k=1}^{K} = p_\zeta(z_k) = 1$.

Définition de la mesure inter histogramme La divergence de Kulback-Leibler entre la distribution $p_\zeta(z_k)$ et la distribution uniforme d'un histogramme avec K valeurs de poids $1/K$, s'écrit :

$$D = \sum_{k=1}^{K} p_\zeta(z_k) log_2\Big(\frac{p_\zeta(z_k)}{1/K}\Big) \tag{3.5}$$

$$D = \sum_{k=1}^{K} p_\zeta(z_k) log_2(p_\zeta(z_k)) + log_2(K) \tag{3.6}$$

elle est comprise entre les valeurs $[0\ log_2(K)]$.

3.1.3 Évaluation de nos méthodes

Ce rapport rend compte de résultats qui sont intermédiaires, ils ne sont pas comparables avec l'état de l'art. Cependant ils peuvent permettre de donner des pistes d'améliorations du système. On fait une étude comparée sur 5 morceaux.

3.1.3.1 Évaluation qualitative : les histogrammes

Le système de référence montre des résultats très variables selon les différents morceaux. Le premier morceau *with or without you* du groupe U2 est bien estimé par notre méthode de référence, ce qui permettra d'étalonner les différentes modifications.

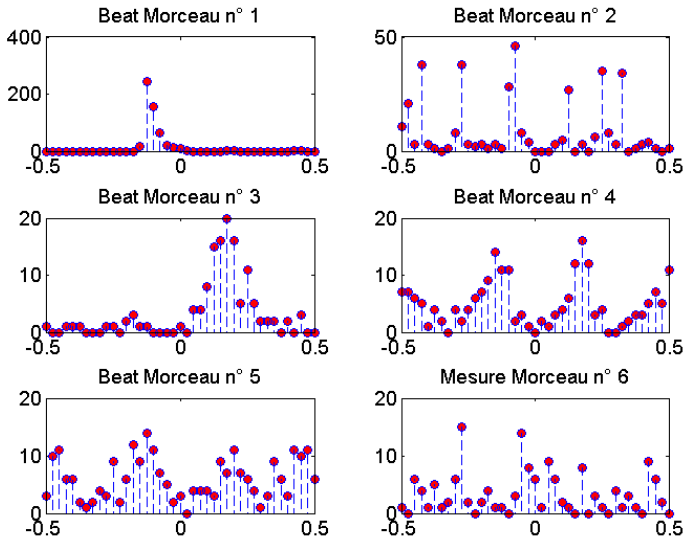


Figure 3.1 – Résultats obtenus avec la méthode de référence, la figure 6 représente la phase de la mesure du premier morceau

Système de référence L'estimation de la phase de la mesure ne donne pas de bons résultats malgré le fait que sur ce morceau le beat soit bien estimé.

Amélioration sur la fonction de détection L'amélioration donnant les meilleurs résultats est celle faite sur la fonction de détection à l'aide du traitement en série du filtre passe-bas avec une fenêtre de Hann et du filtre dérivateur.

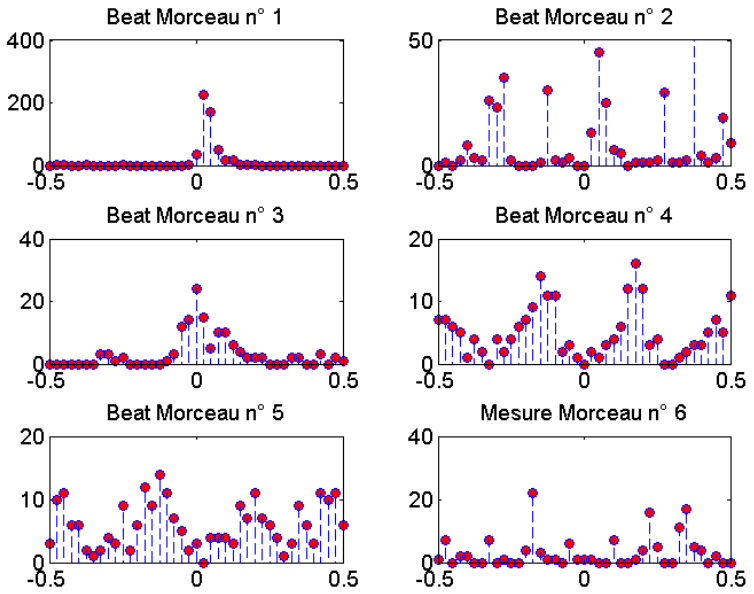


FIGURE 3.2 – Résultats obtenus avec le traitement Hann+filtre Diff, la figure 6 représente la phase de la mesure du premier morceau

Tests On ne présente les résultats que sur l'estimation de la chanson passant le test avec le plus de succès. Les autres résultats n'apportent pas d'informations supplémentaires.
Les résultats des deux premiers tests effectués ne permettent pas de tirer des concludions.

- La première figure montre les performances du système en utilisant la fonction de détection de M. Alonso. Les mauvais résultats de la fonction sont étonnants et inattendus. Cela vient du fait que la fonction de détection développée dans ce rapport est spécifique à cette application et qu'il aurait fallut changer les paramètres de la fonction de détection de M. Alonso afin de la rendre adaptée au problème.
- Le deuxième résultat est obtenu en forçant la dépendance inter-états pour la recherche de périodicités. Les états sont constitués des 5 meilleurs estimations du tatum et de leurs multiples. Cet a priori est trop fort et l'estimation en est fortement dégradée.
- La troisième méthode fournit des résultats qui améliorent le système de base. Le poids de la matrice de transition d'estimation des phases est augmenté d'un facteur 10. On ne peut rien conclure au vu de la quantité d'estimation mais cela pourrait être un champ d'investigations pour des travaux futurs.

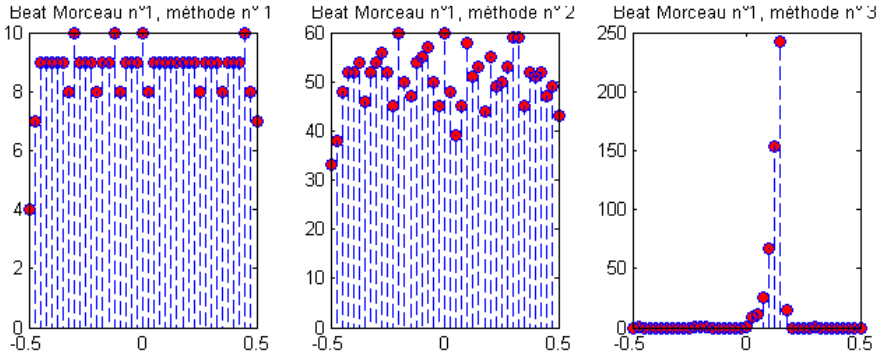


Figure 3.3 – Comparaison des trois méthodes 1. Fonction de détection de M. Alonso, 2. Changement des états. 3. Changement du poids de la matrice de détection.

3.1.3.2 Comparaison quantitative : Fmesure, Pscore et Cemgil

Conclusions pour le premier morceau

- La méthode de référence propose une périodicité qui est proche de la périodicité des estimations : la mesure $Pscore$ est très élevée 99.2. Les beats estimés ne sont pas proches des phases annotées avec une mesure $Cemgill$ basse 29.7. Les distances aux phases annotées sont régulières (concentrées sur peu de bins de l'histogramme $D = 3.29$). Le résultat d'estimation final est moyen $Fmesure = 54.6$.
- En comparant avec la méthode de référence, la méthode $Hann + Diff$ obtient de bien meilleurs résultats $Fmesure = 98.2$. Cela est dû au fait que les beats estimés sont plus proches des annotations $Cemgill = 87.1$. Bien que les périodicités correspondent moins aux annotations $Pscore = 98.8$ (les phases étant mieux estimées le $Pscore$ devrait être majoré).
- Augmenter le poids de la matrice de transition pour l'estimation des phases permet d'en améliorer le score 30.2 contre 29.7 et améliore légèrement l'estimation globale.

Remarque : Ces conclusions sont établies sur la mesure d'un seul morceau et sont uniquement prospectives, elle seraient à vérifier avec des estimations réalisées sur une base de données plus importante.

Conclusions sur les phases des mesures

- La mesure de référence semble faire une estimation hasardeuse $D = 0.75$, avec des périodicités qui sont peu corrélées avec les annotations $Pscore = 45.3$ mais certaines phases sont en accord avec les mesures. Cela pourrait être dû au fait que les phases sont choisies parmi les quatre beats de la mesures.
- La mesure $Hann + Diff$ semble faire une estimation plus régulière $D = 1.35$ mais elle est presque constamment à côté des phases estimées $Cemgil = 0.5$ ce qui donne un résultat très mauvais $Fmesure = 0.7$.

Remarque : Pour le troisième morceau la méthode $Hann + Diff$ donne un bon score $Fmesure = 80.4$ ce qui semble être dû, à nouveau à une meilleure estimation de la phase $Cemgill = 66.5$, certainement centrée puisque l'histogramme indique que les distances à la

	Morceaux	Fmesure	Cemgil	Pscore	D
Syst. de Référence	n° 1	54.8	29.7	99.2	3.29
	n° 2	28.0	12.3	50.1	1.2
	n° 3	36.6	21.0	68.9	1.39
	n° 4	17.1	11.0	47.1	0.48
	n° 5	22.3	13.2	41.1	0.27
Hann+Diff	n° 1	98.2	87.1	98.8	3.19
	n° 2	28.8	19.5	32.3	1.30
	n° 3	80.4	66.5	82.9	1.44
	n° 4	17.1	11.2	47.1	0.48
	n° 5	22.3	13.2	41.1	0.27
Transition : $\alpha = 10$	n° 1	56.7	30.2	99.2	3.26
Miguel détection	n° 1	1.0	0.1	0.8	0.01
États	n° 1	17.2	11.5	18.6	0.01
Mesures Référence	n° 1	6.6	5.8	45.3	0.78
Mesures Hann+Diff	n° 1	0.7	0.5	40.0	1.35

FIGURE 3.4 – Tableau de report des résultats d'estimations des mesures *Fmesure*, *Cemgil* et *Pscore* sur les différentes méthodes implémentées.

phase estimée sont étalées sur plusieurs bins 1.44. Ce qui est confirmé par l'histogramme.

Conclusion

Ce document décrit l'implémentation d'un système permettant d'extraire les différentes couches métriques *tatum*, *tactus*, et *mesure*. Pour cela il s'inspire de l'article [3] décrivant le procédé d'extraction des éléments rythmiques de la manière suivante :

- la détection des débuts de notes dans le flux audio : en sommant les enveloppes énergétiques sur différentes bandes de fréquences.
- la recherche des périodicités sur : -un banc de filtre résonnant et -l'ajout d'un a priori musical modélisé sous la forme d'une Chaine de Markov Cachée, et résolution avec l'algorithme de Viterbi.
- l'estimation des phases se fait de manière indépendante à l'aide d'une seconde modélisation sous forme de Chaine de Markov, sur les périodes estimées.

Il présente une phase de recherche sur l'extraction de l'accent mélodique, qui pourrait être formulée comme un *a priori musical* permettant d'améliorer l'estimation des phases.

Enfin l'observation des différents résultats permet de conclure sur quelques pistes d'améliorations :

- L'estimation des onsets. Il apparaît que la fonction de détection est la pièce maîtresse de la détection de rythme.
- Déterminer un poids α de la matrice de transition, maximisant l'estimation des phases.
- Enfin, la variabilité des résultats obtenus en fonction des différentes mesures d'évaluations a de quoi laisser sceptique sur la comparaison des résultats obtenus avec une seule méthode d'évaluation. La combinaison des trois méthodes et la définition de la mesure D est très informative sur les problèmes d'estimations d'un système donné. Il n'en reste pas moins que l'idée d'une évaluation perceptive permettrait (selon les applications) de véritablement donner un score qui permettrait de comparer les méthodes.

Bibliographie

[1] Scheirer E. D., Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, 1998.

[2] Hainsworth S., Beat Tracking and Musical Metre Analysis. *Signal Processing Methods for Music Transcription* 2006, Part II, $101 - 129$.

[3] Klapuri A. P., Eronen A.J., Analysis of the meter of acoustic musical signals 2006, Vol.14., p$342 - 355$.

[4] Fuentes B., Liutkus A., Badeau R, Richard G.,Probabilistic model for main melody extraction using constant-Q transform, *37th International Conference on Acoustics, Speech, and Signal Processing* ICASSP'12, Kyoto, Japon, March 25 - 30, 2012, pp. 5357-5360.

[5] Parncutt, R., & Drake, C. (2001) Psychology : Rhythm. In S. Sadie (Ed.), *New Grove Dictionary of Music and Musicians, 20. London*, $535 - 538, 542 - 555$.

[6] M.A., Alonso, *Extraction d'information rythmique à partir d'enregistrement musicaux*, mémoire de thèse, Telecom-Paristech, 2006.

[7] Handel, S. The effetct of the tempo and tone duration on rythmique discrimination, *Perception & Psychophysics*, $54(3) : 370 - 382$.

[8] Francès R., *Psychologie de l'esthétique*, Presse universitaires de France, 1968.

[9] Lerdahl & jackendoff, *A génerative theory of tonal music*, MIT Press, Cambridge, MA, USA.

[10] Laroche, J. Efficient Tempo and beat tracking in audio recording, *Journal of the audio Engineering society* $51(4) : 226 - 233$

[11] W.J. Dowling et D.L. Harwood. *Music Cognition.* Series in Cognition and Perception. Academic Press, 1986.

[12] Mari Riess Jones Dynamic pattern structure in music : Recent theory and research, *Attention, Perception,& Psychophysics*,Volume 41, Number $6(1987), 621 - 634$

[13] M. Pineau & B. Tiellmann,*Percevoir la musique une activité cognitive*, l'Harmattan, 2001.

[14] Meudic, B., *Détermination automatique de la pulsation, de la métrique, dans des interprétations à tempo variables d'œuvres polyphoniques*, Mémoire de thèse, Université Pierre et marie Curie, Paris 6.

[15] Cemgil, A. T. & Kappen, B. Monte Carlo methods for tempo tracking and rythm quantization, *Journal of New Music reserche0* $28(4) : 259 - 273$.

[16] Peeter G., & Papadopoulos H., Simultaneous Beat and Downbeat-Tracking Using a Probabilistic Framework : Theory and Large-Scale Evaluation, *Proc. IEEE Speech and langage*, VOL. 19, NO. 6, Aout 2011.

[17] Cont A. *L'ordinateur qui joue comme un musicien*, La recherche, Juin, 2012.

[18] Davies E.P.M., Degara N., Plumbey M., Evaluation for musical Audio Beat Tracking Algorithms, technical Report, Queen Mary University, Centre for Digital Music, 2009.

[19] Rabiner. R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *In Proceedings of the IEEE*, 1989.

[20] Maazaoui M., *Classification audio par approche hybride SVM-HMM* , mémoire de stage, Telecom Paristech, 2008.

[21] Fillon T., Traitement numérique du signal pour une aide aux malentendants, *mémoire de thèse*, Telecom-Paristech, 2004.

[22] Goto M., & Muraoka, Y., Real-time Rhythm Tracking for Drumless Audio Signals Chord Change Détection for Musical Décisions, *Proc. IJCAI-97 Worshop on Computational Auditory Scène Analysis*

[23] Huron D., Royal M., What is melodic accent ? Converging Evidence from Musical practice, *Music Perception*, $Vol.13, No.4, 489 - 516$, 1996.