Master 2 Sciences de l'ingénieur
Université Pierre et Marie Curie - Paris 6

parcours ATIAM

Rapport de stage de Master II

Estimation de la position des temps et des premiers temps d'un morceau de musique

---

Auteur :

Simon Durand

Responsables de stage :

Bertrand David - Télécom ParisTech
Gaël Richard - Télécom ParisTech

2 août 2013

Remerciements

Je tiens à remercier tout particulièrement mes deux maitres de stage, Bertrand David et Gaël Richard, enseignants chercheurs à Telecom ParisTech pour m'avoir fait confiance sur ce projet. Je les remercie pour leur écoute, leur bienveillance et leur conseils indispensables à la réussite de stage. Merci également pour la liberté qu'ils m'ont accordée qui est toujours précieuse dans le milieu de la recherche.

Merci à Guillaume Mahenc, étudiant du Master ATIAM qui a partagé mon bureau pendant ces cinq derniers mois. Nos échanges scientifiques et musicaux m'ont beaucoup apporté.

Je remercie également Hélène Papadopoulos, Fabien Gouyon et Thomas Fillon qui m'ont apporté une aide précieuse en me donnant accès à des bases de données et à des algorithmes robustes afin que je puisse évaluer et comparer mes résultats.

Je suis très reconnaissant envers les travaux de la communauté, notamment ceux d'Anssi Klapuri, Geoffroy Peeters, Masataka Goto, Miguel Alonso et Matthew Davies qui m'ont passionné et inspiré.

Je remercie enfin toute l'équipe du groupe Audio Acoustique et Ondes, François, Aymeric, Hequn, Xabier, Angélique, Mounira, Davide, Slim, Yves, Cécilia pour leur accueil, leurs réponses à mes questions techniques ou pratiques et leur bonne humeur qui a offert à ce stage un cadre idéal.

Résumé

Version française :

L'estimation de la position des temps et des premiers temps d'un morceau de musique a de nombreuses applications comme l'aide à la synchronisation en studio ou la génération de listes de lecture adaptées à l'utilisateur et au contexte. C'est également un bon outil pour essayer de comprendre comment les auditeurs captent des informations rythmiques dans un contenu musical. Enfin, cette estimation est une étape importante de nombreux outils de recherche d'informations musicales comme le peut être la transcription automatique de musique sur partition. Après une revue de la littérature sur l'estimation de la position des temps et des premiers temps d'un morceau de musique, il s'agit de développer une méthode de référence sur le sujet. Une méthode d'estimation conjointe de différents niveaux métriques et de la position des temps est présentée par la suite. Enfin, il s'agit de s'intéresser aux premiers temps en identifiant des descripteurs adéquats et en définissant des méthodes permettant de les extraire pour estimer le rythme au niveau de la mesure. Les différentes méthodes mises en place sont évaluées et comparées à différentes méthodes de l'état de l'art sur une base de donnée de dizaines d'extraits musicaux sur plusieurs genres différents.

Mots clés : Rythme musical, premiers temps, analyse de signal acoustique, modèle de Markov caché, descripteurs audio

Version anglaise :

Beat and downbeat estimation of a piece of music is useful in many ways. It can enhance studio track synchonization or generate context or user friendly playlists for example. It is also useful to understand our ability to extract metrical information from a musical content. Several fields of Music Information Retrival such as Automatic Music Transcription or Audio Summary from Signal Analysis are using beat or downbeat estimation. An overview of state-of-the-art algorithms and concepts is firstly given. We then show an implementation of an already existing robust and efficient method and propose another one that jointly estimate the tempo and the beat phase. The downbeat location estimation is also investigated and an algorithm that extract several musically interesting features to choose which beat is a downbeat is described. We then perform an evaluation of beat and downbeat-tracking using a forty songs test-set. In this, we compare our results to five state-of-the-art algorithms.

Keywords : Musical meter, beat, downbeat, Acoustic signal analysis, hidden Markov model, Rhythm description, audio features

Table des matières

Introduction

Cadre du stage

Ce stage a été effectué dans le cadre du Master Acoustique, Traitement du signal et Informatique Appliqué à la Musique de l'Université Pierre et Marie Curie, en partenariat avec l'IRCAM. Il s'est déroulé dans le groupe Audio Acoustique et Ondes (AAO) du département Traitement du Signal et des Images, l'un des quatre départements d'enseignement et de recherche de l'école Télécom ParisTech qui fait partie de l'UMR CNRS 5141 LTCI. Ce groupe vise à développer des méthodes de traitement numérique du signal audio, afin de proposer des solutions aux principales problématiques centrées sur le son (parole, musique,...) dans les applications multimédia.

Le groupe AAO poursuit ses recherches sur l'analyse et la représentation des signaux musicaux notamment à des fins de transcription ou de séparation de sources. Ce stage permettra entre autres d'obtenir des informations sur la structure rythmique de morceaux de musique avec par exemple la position des barres de mesure ou du tempo qui seront utiles à un algorithme de transcription automatique de musique sur partition qui fera l'objet d'une thèse à venir.

Contexte

L'estimation de la position des temps est notre capacité à identifier le tempo ou la battue d'un morceau de musique ainsi que la position de chacun des battements correspondants. L'estimation de la position des premiers temps est notre capacité à identifier les temps placés au niveau des barres de mesure d'une partition. Si la localisation de la position des temps est naturelle chez la plupart des gens, la localisation des premiers temps en revanche l'est beaucoup moins, à part chez les musiciens. Depuis trente ans, de nombreux travaux ont été effectués pour obtenir la position des temps dans un morceau de musique de façon automatique. Si les algorithmes actuels sont efficaces sur de nombreuses musiques largement diffusées, un tempo lent, des rythmes syncopés ou l'absence d'instruments percussifs sont actuellement des causes de nombreuses erreurs d'estimation [HDZ+12].

Rosenthal analyse la difficulté de ces tâches en se référant à nos acquis culturels. Il indique que ce qui nous parait naturel est en fait le fruit d'une évolution longue et sophistiquée de nos capacités et que les processus que nous mettons en jeu pour cette tâche sont loin d'être simples [Ros92]. On peut ajouter que nous sommes performants pour des pièces ou des genres musicaux qui nous sont familiers. La métrique d'un Chaâbi marocain ou du Khush Rank persan [1] déconcertera sûrement un occidental non initié. C'est un véritable

1. Dont le chiffrage est $\frac{17}{8}$. On peut trouver le rythme et des extraits musicaux à cette adresse : `http://www.maqamworld.com/rhythms/muwashahat4.html#khush_rank`

apprentissage que nous mettons en place.

Objectifs du stage

Dans le but d'effectuer une meilleure description du contenu rythmique d'un signal musical, il s'agira :

- D'effectuer un état de l'art des méthodes actuelles d'estimation automatique de la position des temps et des premiers temps,
- D'implémenter certaines de ces méthodes,
- De mettre en place un algorithme d'estimation du tempo et de la position des temps à partir d'un signal audio,
- De mettre en place un algorithme d'estimation de la position des premiers temps à partir de la position des temps,
- D'évaluer ces méthodes sur une base de données de dizaines de morceaux de différents genres à annoter,
- De comparer les résultats avec ceux obtenus par plusieurs algorithmes performants de l'état de l'art à l'aide de mesures reconnues par la communauté scientifique.

Applications

L'estimation automatique du tempo, de la position des temps et des premiers temps d'un contenu musical a de nombreuses applications.

Estimer le tempo d'un contenu musical permet entre autres de créer des listes de lectures adaptées au contexte. Les morceaux rapides par exemple, ne sont-ils pas en général moins adaptés lorsqu'il s'agit de mettre en musique une séance de relaxation ? La recherche du tempo adéquat et également importante lors de la pratique de la danse et de nombreux professeurs et DJs possèdent ou recherchent un accès au contenu par le tempo.

L'estimation des temps et des premiers temps permet une modification fine du signal pour l'adapter à l'utilisateur. Un premier exemple est un karaoke qui s'adapte à la vitesse du chanteur et non l'inverse [Got01]. On peut également mentionner les chaussures de footing qui adaptent la musique écoutée par le joggueur au rythme de ses pas [HWF09].

Le musicien est également concerné. Souligner les temps et les premiers temps d'un extrait musical peut en permettre une meilleure compréhension ou un suivi plus aisé. Le suivi du tempo en fonction du temps permet d'étudier l'interprétation d'une œuvre. La modification du signal sonore au tempo souhaité peut permettre au musicien d'apprendre à jouer le morceau plus facilement. L'estimation du premier temps est une étape importante dans la transcription de partition à partir du signal sonore, afin de connaitre l'emplacement des barres de mesures.

Des applications sont également possibles au niveau de la production et l'interaction avec le matériau sonore. La synchronisation des différentes pistes en studio sera simplifiée en estimant la position des temps et des premiers temps. Le suivi de partition qui consiste à synchroniser une exécution musicale avec la partition du morceau joué est aidé par le suivi du tempo de l'interprétation.

Chapitre 1

Description rythmique computationnelle

Ce chapitre présente l'état de l'art sur la description rythmique computationnelle. Pour cela, nous replacerons les recherches dans leur contexte avec un bref historique, nous définirons plusieurs concepts rythmiques importants et nous effectuerons un état de l'art suivant les grandes étapes des méthodes d'estimation du tempo, de la position des temps et de la position des premiers temps.

1.1 Historique de l'estimation du tempo, des temps et des premiers temps

L'estimation du tempo, de la position des temps et des premiers temps à connu beaucoup de progrès ces trente dernières années. Ces travaux se sont d'abord attachés à reproduire la battue humaine puis des systèmes d'estimation automatique de plus en plus performants ont été mis en place sur des signaux MIDI, puis des signaux audio de plus en plus complexes.

Des modèles auditifs, perceptifs et musicologiques pour commencer

Un des premiers essais pour estimer le tempo remonte au début des années 70 avec l'étude du rythme des fugues de Bach par Longuet-Higgins et Steedman en 1971 [LHS71]. Quinze ans plus tard, Povel et Essen ont cherché le niveau métrique le plus adéquat à un rythme complexe. Ils ont utilisé un ensemble de règles et ont montré l'importance de la durée et de la position des notes [PE85]. Parncutt a également utilisé des modèles musicologiques pour introduire des a priori sur les accents musicaux [Par94]. Large a essayé de se rapprocher des tempo perceptifs et à mis en place des oscillateurs qui rentrent en résonance lorsqu'une périodicité se répète [LK94].

Du MIDI à l'Audio

L'identification du tempo, des accents rythmiques et de la position des temps d'un morceau de musique à d'abord été effectuée à partir d'une représentation symbolique comme la partition ou la transcription MIDI. Cela permet de d'obtenir une représentation claire du signal avec des attaques et des silences bien délimités, plusieurs voix ou instruments qui n'interagissent pas entre eux et un volume de données relativement faible. On trouve plu-

sieurs méthodes d'analyse du contenu rythmique à partir de fichiers MIDI dans les années 90 et le début des années 2000 [AD90] [Rap97] [Rap01] [Dix01].

Les premiers systèmes d'estimation de la pulsation à partir d'un extrait audio complexe arrivent avec Goto qui utilise un système par agents multiples pour la prise de décision et Scheirer qui lui décompose le signal en sous bandes et se sert d'un réseau d'oscillateurs [GM94] [GM99] [Sch98].

Les méthodes probabilistes et dynamiques

De plus en plus de méthodes probabilistes, qui permettent de mieux prendre en compte les variations de tempo sont apparues au milieu des années 2000 avec notamment l'utilisation des modèles de Markov cachés par Klapuri et de la programmation dynamique par Alonso ou Davies [KEA06], [AA06], [DP07]. Les modèles de Markov cachés sont présentés à l'annexe A.

Et maintenant ?

Des modèles probabilistes plus sophistiqués comme les champs conditionnels aléatoires qui permettent de prendre en compte toutes les dépendances du morceau [1] ont été étudiés par Joder et Fillon [2]. On voit également apparaitre des méthodes d'apprentissage avec des bases de données de plus en plus importantes qui permettent d'emmagasiner une grande quantité d'information [HDZ+12]. Cependant, les modèles musicaux sont encore pris en compte. Hamanaka à essayé d'implémenter une méthode générative sur la musique tonal proposée par Lerdahl vingt ans plus tôt qui permet d'accentuer certains temps et d'utiliser un modèle de métrique pour trouver les premiers temps [HHT06] [LJS85].

1.2 Concepts et définitions utiles pour ce stage

Dans le travail présenté ici, le terme temps correspond aux différents temps que comporte une mesure. Il correspond en général à la battue, à la pulsation, ou au tapement du pied que l'on peut faire en écoutant un morceau, mais pas toujours. Par exemple, en Jazz, la battue est souvent faite sur les deuxièmes et quatrièmes temps. Le premier temps est équivalent au premier temps de la mesure. Il correspond à l'endroit où l'on voit une barre de mesure sur une partition. La position temporelle d'un temps est sa phase.

1.2.1 Rythme musical

Il n'existe pas de consensus pour définir de manière univoque le rythme musical. On peut cependant le lier à une mise en relation d'un événement sonore avec ce qui le précède et ce qui le suivra probablement. C'est donc la mémoire et l'attente qui permettent d'organiser les événements dans le temps et d'extraire le rythme.

Un élément important lié au rythme musical est la notion d'accent. Un accent permet de faire ressortir la structure rythmique en insistant sur un événement précis. Un accent peut être l'intensité relative, mais aussi l'articulation (passage de legato à staccato), le changement de timbre (en changeant d'instrument par exemple) ou la durée entre deux

1. Voir l'article de Lafferty sur le sujet [LMP01].
2. L'article correspondant n'est pas encore sorti.

attaques consécutives. Un événement fortement accentué aura plus de chance d'être situé au niveau du temps ou du premier temps.

Évaluer la position des temps et des premiers temps d'un morceau de musique suppose qu'il existe une structure temporelle sous-jacente au delà d'un enchainement note à note. Les travaux de Lerdahl et Jackendoff montrent une organisation hiérarchique du rythme musical en deux catégories que sont la métrique et le groupement [LJS85].

- Le groupement est la combinaison de motifs rythmiques de durée équivalente pour former des unités musicales cohérentes de plus en plus longues jusqu'à atteindre la taille totale du morceau. On peut ainsi passer de motifs courts à des phrases, pour ensuite obtenir des passages plus longs, des mouvements et des œuvres entières. Appliqué à la musique populaire, cette organisation permet une segmentation du morceau en diverses parties telles que l'introduction, le couplet, le refrain, ou le pont.
- La métrique est une notion plus perceptive qui sous entend un découpage temporel régulier de la musique. Un niveau métrique est un découpage régulier de la musique. On perçoit par exemple une pulsation, ou une battue, lorsque des événements ou des motifs se répètent régulièrement. Cette notion est continue car si les événements s'arrêtent momentanément, la même pulsation est toujours perçue.

Ces deux notions peuvent être observées séparément mais sont complémentaires. Elles sont nécessaires à une analyse rythmique complète. Notre étude se concentrera principalement sur la métrique. L'objectif est en effet de représenter l'extrait audio de façon régulière au niveau du temps et du premier temps de la mesure. Le groupement peut par contre être utile pour identifier des motifs musicaux de durée irrégulière et non identique à celle de la mesure. En effet, le premier temps aura plus de chance d'être situé au début d'un motif musical, qu'au milieu de celui-ci.

1.2.2 Structure métrique

Le temps perceptif est lié au moment où l'auditeur tend à taper du pied et les temps de la mesure sont indiqués sur la partition ou pensés par le compositeur et les musiciens. Le premier aura un tempo proche d'un battement par seconde et sera souvent un multiple du second. De la même manière, le tempo perceptif est différent de l'écart entre deux temps consécutifs de la mesure. Là aussi, le tempo perceptif est bien souvent un multiple de cet écart, mais il est rapide si les événements sonores s'enchainent rapidement et plus lent à l'inverse. Considérons d'abord une phrase mélodique qui est jouée accompagnée d'une batterie à une vitesse donnée, puis la même phrase mélodique mais avec un accompagnement à la batterie deux fois plus rapide. L'auditeur percevra probablement une battue identique pour les deux extraits mais un tempo deux fois plus rapide dans le deuxième cas. La différence est ici le nombre d'événements sonores. Cette notion est définie par le concept de *tatum*.

Le tatum, dérivé de "temporal atom", désigne le plus petit niveau métrique. C'est le plus le plus petit découpage régulier qui coïncide le mieux avec tous les événements sonores. Les autres niveaux métriques sont donc des multiples de ce tatum. Il a une grande importance pour la perception du tempo et pour l'analyse rythmique a court terme [Bil93].

Les deux autres niveaux métriques qui nous intéressent sont le *tactus*, correspondant à la pulsation du tempo perceptif, et la *mesure*, correspondant à la périodicité des mesures

	Musique écrite	Notion perceptive
Tempo	Écart entre deux temps consécutifs	Vitesse ressentie du morceau, tactus
Temps	Temps de la mesure	Battue

TABLE 1.1 – Comparaison entre la notion absolue, liée à la musique écrite, et la notion perceptive du tempo et de la position des temps. On se place dans la définition de la musique écrite dans la suite du rapport sauf précision supplémentaire. Par simplification, *temps* sera équivalent à *position du/des temps* dans la suite du rapport sauf précision supplémentaire. Si l'on parle de *tatum/tactus* sans précision supplémentaire, il s'agit de périodicité et non de phase.

du morceau. La figure 1.1 illustre le concept de tatum, tactus et mesure et le tableau 1.1 compare la définition des temps et tempo perceptifs de celle de la musique écrite.

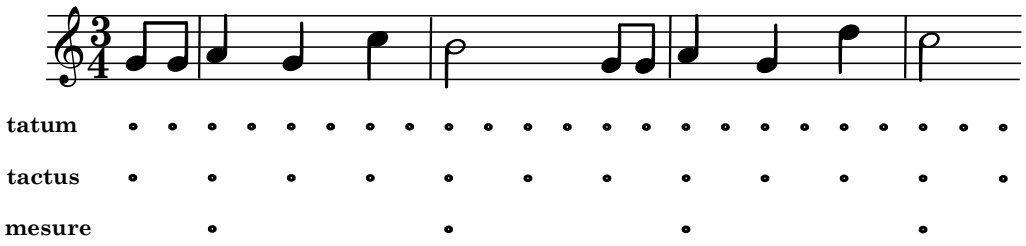


FIGURE 1.1 – Illustration du concept de niveau métrique. Ici, la phase du tatum, du tactus et de la mesure est représentée par des points noirs sous un extrait de partition. Les trois niveaux métriques sont réguliers. Les rapports entre le tatum et le tactus, et entre le tactus et la mesure sont entiers mais pas forcément identiques comme on peut le voir.

1.2.3 Ambiguïté de la position du premier temps

Nous avons vu que certains temps étaient plus accentués que d'autres. Cependant, le type de temps accentué n'est pas toujours le même et est lié au genre musical. La musique classique, avant le XX$^{\text{ème}}$ siècle, accentue principalement le premier temps alors que ce sont souvent les deuxièmes et quatrièmes temps qui sont mis en avant dans le Jazz, le Blues ou le Rock. Il n'est donc pas possible d'estimer la position des premiers temps uniquement avec la position des accents.

Les premiers temps ne sont pas situés aux mêmes endroits suivant l'intention du compositeur. Prenons exemple en musique contemporaine. Dans le cas de rythmes complexes et non réguliers, Messiaen privilégie un chiffrage constant et stable tout en accentuant les motifs rythmiques [Mes44] tandis que Stravinsky opte pour un chiffrage plus local et directement lié à ses intentions musicales comme l'illustre la figure 1.2. Il sera important pour estimer le premier temps d'établir plusieurs critères et de les hiérarchiser. Daoust propose de choisir la structure métrique la plus régulière, puis celle jouée par le plus d'instruments, enfin celle qui permettra de regrouper les parties les plus longues, afin d'éviter les liaisons [Dao08].

Enfin, la perception du rythme est subjective et propre à l'auditeur. Le rapport de Master d'Ugo Marchand, encadré par Geoffroy Peeters traite ce sujet [MP13]. Cette différence

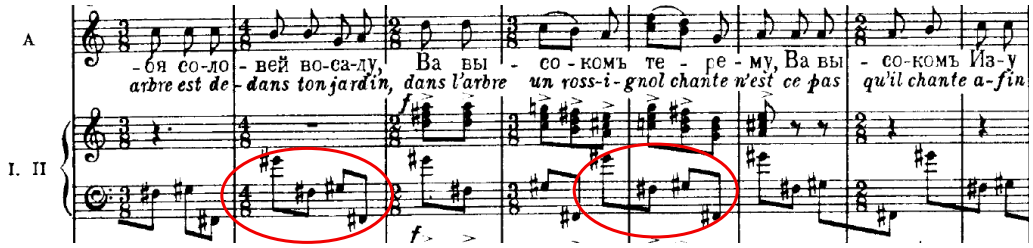


FIGURE 1.2 – Igor Stravinsky, *Les Noces*, 1922. Premier tableau, *La Tresse*. Partie 12. Illustration du chiffrage variable de la polyrythmie. Le chant passe de 3/8 à 4/8 et 2/8 pendant que l'accompagnement est un ostinato en 4/8 comme le montrent les cercles rouges. La métrique suit la partie chantée.

pose problème au moment de l'évaluation de la méthode d'estimation des temps et des premiers temps. Il faut en effet confronter l'estimation à la vérité terrain, qui consiste le plus souvent en une annotation par différents participants qui ne sont pas toujours d'accord entre eux. L'annotation finale n'est d'ailleurs pas toujours cohérente [3].

1.3 Estimer les temps et les premiers temps à partir du signal audio

1.3.1 Principes généraux de l'analyse rythmique automatique

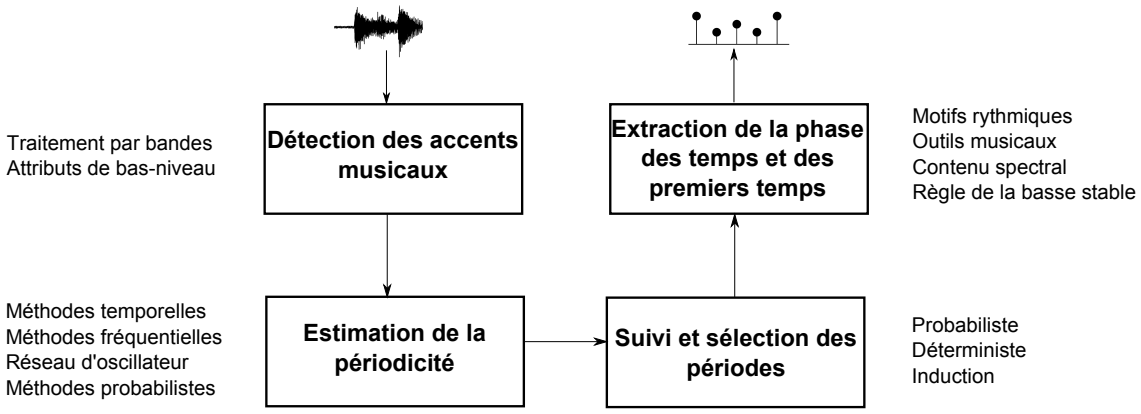


FIGURE 1.3 – Résumé d'une méthode générale d'estimation de la position des temps et des premiers temps.

L'estimation de la position des temps et des premiers temps d'un morceau de musique à partir du signal audio s'effectue le plus souvent en quatre étapes consécutives. Il faut dans un premier temps déterminer les accents rythmiques du morceau, c'est à dire les attaques des événements sonores. La deuxième partie est l'estimation des périodicités qui permet

3. On peut citer en exemple trois morceaux de Jazz de la base de données utilisée par Klapuri. L'extrait n°3 est de tempo intermédiaire, l'extrait n°204 est de tempo relatif légèrement inférieur et l'extrait n°270 de tempo relatif légèrement supérieur. De plus, le genre, les instruments et l'orchestration sont identiques. Pourtant, la pulsation estimée des extraits n°204 et 270 est deux fois plus rapide que pour l'extrait n°3. De plus, certaines fois, comme pour l'extrait 270 par exemple, la vérité terrain n'est pas en accord avec la partition.

de faire ressortir les répétitions du morceau à différents niveaux métriques. Les périodes principales doivent être sélectionnées et suivies dans le temps pour obtenir le tempo. Enfin, on choisit les instants écartés des périodicités trouvées précédemment qui représentent le mieux le signal pour obtenir la position des temps et des premiers temps comme le montre la figure 1.3.

1.3.2 Du signal audio aux accents musicaux

La première étape après la lecture du signal est de calculer ses accents. Cela correspond dans notre cas aux attaques des événements sonores permettant d'estimer le tempo puis les temps et les premiers temps du morceau.

Représentation temps/fréquence

La première étape est l'obtention d'une représentation temps fréquence du signal. Pour cela, la plupart des méthodes calculent la transformée de Fourier à court terme discrète $TFCT(x(n)) = \sum_{n=-\infty}^{\infty} x(n)w(n_m)e^{-j\omega n}$ avec $x(n)$ le signal audio à l'échantillon n, ω la pulsation et w la fenêtre d'analyse. La valeur absolue du résultat est mise en carré pour obtenir le spectrogramme qui permet d'obtenir des informations temporelles et fréquentielles utiles à l'obtention du rythme. Certaines méthodes utilisent un spectrogramme réassigné, ce qui consiste à réassigner l'énergie de chaque canal fréquentiel et temporel au canal le plus proche de la vrai région du support du signal analysé [KFROch] [Pee06] [FACM+03]. Cela permet de distinguer plus finement certaines attaques comme le montre la figure 1.4.

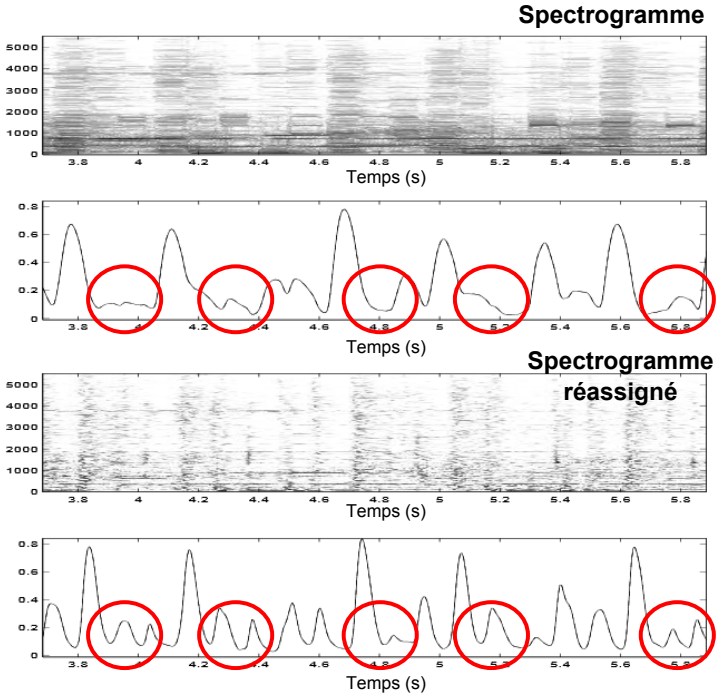


FIGURE 1.4 – Comparaison des représentation des accents du signal en utilisant un spectrogramme et un spectrogramme réassigné. Les cercles rouges montrent les attaques que l'on arrive à percevoir en appliquant la méthode du réassignement spectral. D'après [Pee05].

Séparation des composantes du signal

Il est également possible à ce moment d'effectuer une séparation entre certaines composantes du signal. Alonso propose une séparation entre les harmoniques et le bruit [AA06], Tsunoo entre les percussions et l'harmonie [TTOS11] et Zapata une séparation de la partie chantée [ZG13].

Répartition de l'énergie :

À ce moment, la plupart des méthodes de la littérature vont sommer l'énergie sur les bandes [Pee06], [Sch98], [GM99], [KEA06] ou utiliser des attributs de bas niveau [G$^+$05]. L'approche par bandes à l'intérêt de séparer les informations localisées en fréquences, de réduire la complexité et de permettre l'utilisation de bandes perceptives pour filtrer le signal [Kla99].

1.3.3 Estimation de la périodicité

Après avoir obtenu un profil d'accents musicaux, on s'aperçoit que les attaques des notes ne sont toujours proéminentes. On ne peut pas directement les détecter pour estimer la position des temps et il faut avant estimer leur périodicité afin d'obtenir les tempo de différents niveaux métriques du morceau, le tatum, le tactus et la mesure par exemple. On cherche une méthode qui permettra de bien discriminer les différents niveaux métriques et qui ne se contentera pas de donner tous les multiples de la périodicité le plus faible.

Méthodes temporelles

Les méthodes temporelles pour l'estimation de la périodicité sont utilisées par un grand nombre d'auteurs par leur simplicité et leur adéquation au problème [Sep01] [Dix01] [GM99] [Sch98]. La première méthode temporelle présentée ici est la fonction d'autocorrélation r. Elle estime la corrélation du signal x avec lui même $r_x(l) = \frac{1}{N-l}\sum_{n=0}^{N-1-l} x(n)x(n+l)$, avec N la longueur de l'intervalle sur lequel est calculée l'autocorrélation et l la périodicité que l'on souhaite étudier. Si des répétitions se produisent une valeur élevée de r apparaitra.

On peut également utiliser une méthode par banc de filtres résonants a été mise en place par Scheirer et perfectionnée par Klapuri. Il s'agit d'utiliser un banc de filtres résonants de la forme suivante :

$$y_c(\tau, n) = \alpha_\tau y_c(\tau, n-\tau) + (1-\alpha_\tau)x_c(n), \tag{1.1}$$

Avec n le numéro de la trame, τ le retard, x_c le signal des accents musicaux et $\alpha_\tau = 0.5^{\tau/T_0}$ le gain de retour sur T_0 secondes. Un temps T_0 long permettra de prendre en compte des périodicités élevées, au prix d'un temps de résonance du filtre plus important.

Méthodes fréquentielles

Il est possible également d'estimer les répétitions temporelles d'un signal à l'aide de méthodes fréquentielles comme la transformée de Fourier à temps discret ou DFT. Cette transformée mettra en avant les répétitions du signal à la manière d'une estimation de fréquence fondamentale. La figure 1.5 présente les représentations obtenues par la méthode par autocorrélation et la méthode par transformée de Fourier à temps discret pour le même type de signal. On peut remarquer que la fonction ACF a tendance à estimer des répétitions

plus lentes que le tempo alors que la DFT des répétitions plus rapides. Peeters propose alors de combiner les deux pour obtenir une répétition proche de celle du tempo [Pee06].

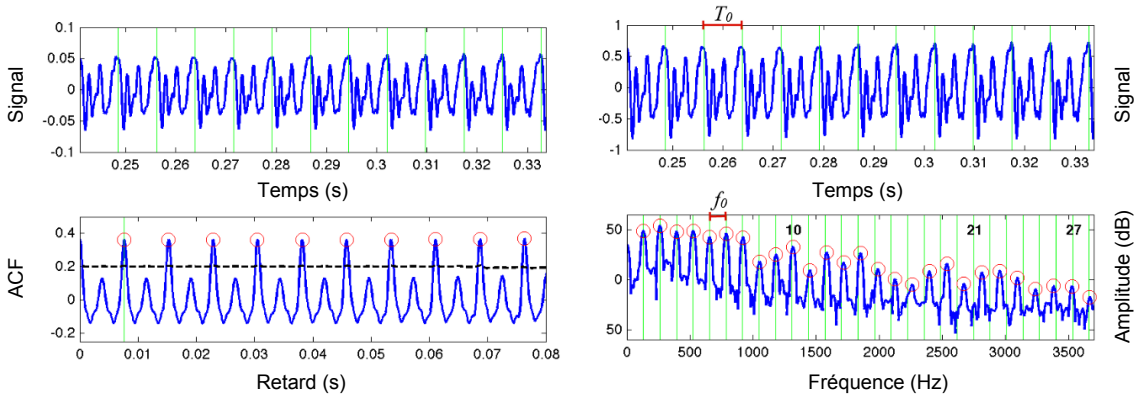


FIGURE 1.5 – Comparaison des représentation des périodicités obtenues par la méthode d'autocorrélation à gauche et par la transformée de Fourier discrète à droite. D'après [BP05].

1.3.4 Suivi et sélection des périodes

Le suivi des périodicités consiste à lier les observations de la partie précédente afin d'avoir un niveau métrique continu. Ce peut être fait par programmation dynamique. Cela consiste à donner un poids à chaque périodicité possible à tout instant puis à relier chaque valeur de la périodicité dans le temps afin qu'elle varie peu et soit bien mise en avant [AA06] [DP07]. On utilise pour cela une méthode de Viterbi qui est présentée à l'annexe A.

La sélection des périodicités consiste à conserver la plupart du temps 3 périodicités, le tatum, le tactus et la mesure qui ont été présenté à la partie précédente. Il faut alors vérifier que ces périodicités sont bien des multiples les unes des autres et qu'elles ne soient pas trop rapides ou trop lentes [KEA06].

1.3.5 Extraction de la position des temps

Les méthodes pour estimer la position des temps utilisent principalement deux informations : la durée entre deux temps consécutifs qui est égale au tempo, et les accents du signal qui seront importants au niveau de la position du temps.

On peut par exemple utiliser un modèle de temps. Ce modèle possède des valeurs élevées au niveau des temps qui sont espacés de la période du tactus estimé auparavant. On compare le modèle avec les accents du signal obtenus ci-dessus et on conserve les résultats les plus élevés comme étant la position des temps. Ces modèles de temps peuvent être simples ou permettant une forte discrimination entre les temps et les contre-temps grâce à une analyse linéaire discriminante [PP11].

D'autres méthodes vont créer une attente du temps futur qui va être de plus en plus forte au fur et à mesure que l'on s'éloigne du temps précédent d'une durée égale à une périodicité. De cette manière on obtient une pulsation régulière [DP07] [KEA06] [Ell07].

1.3.6 Extraction de la position des premiers temps

Extraire la position des temps est plus complexe car ils ne sont pas visibles directement dans le signal. Il faut alors essayer de détecter des phénomènes se passant au niveau des premiers temps pour les trouver indirectement.

Klapuri utilise un modèle de premier temps qui donne un a priori sur la forme que devra prendre les accents du signal aussi bien au niveau fréquentiel que temporel. On ajoute une dimension au modèle de temps [KEA06]. L'observation du comportement spectral du signal est d'ailleurs un bon indice pour estimer les premiers temps. Lorsque la batterie est présente sur le morceau, bien souvent le batteur donnera des coups de grosse caisse (basse fréquence) sur les premiers et troisièmes temps et des coups de caisse claire (haute fréquence) sur les deuxièmes et quatrièmes temps. Détecter une oscillation entre les basses et les hautes fréquences nous donnera des informations sur le type de temps détecté [Got01] [PP11]. Il est également possible de coupler plusieurs descripteurs de premiers temps ensemble comme le suggère Jain [JDM00]. Par exemple, Peeters utilise l'observation du comportement spectral de la batterie couplé à l'instant où un changement d'accord se produit pour en déduire la position du premier temps [PP11].

Khadkevich utilise un modèle de langage [KFROch]. Il essaye de reconnaitre des phrases composées des mots "temps" et "premiers temps" qui seront reconnus lorsqu'une séquence particulière du signal se produit à l'aide d'un apprentissage. Par exemple, le mot "temps" est détecté lorsqu'une pré-attaque, puis une attaque et enfin un silence se produisent comme le montre la figure 1.6.

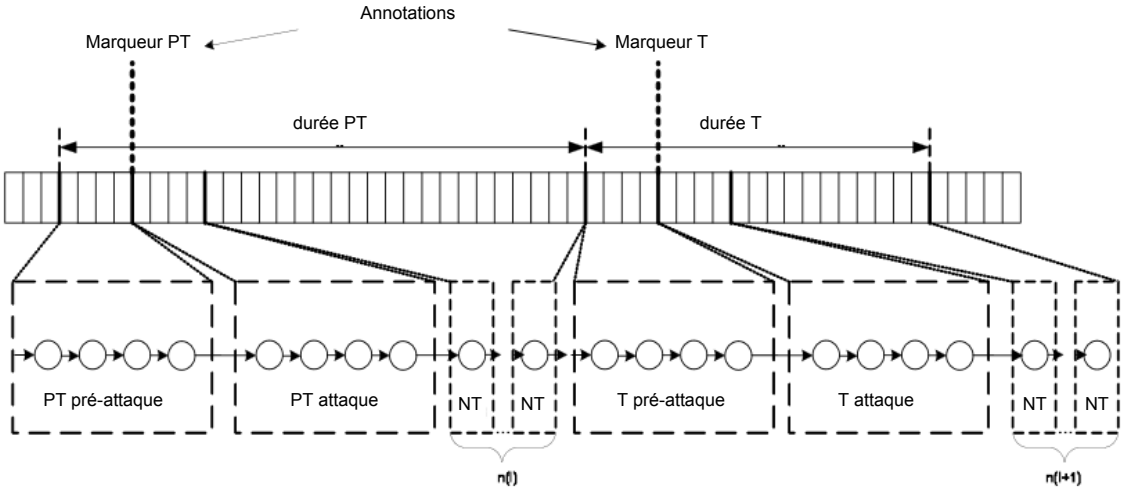


FIGURE 1.6 – Type d'événements qui forment les mots du modèle de langage de Khadkevich pour l'estimation des premiers temps. D'après [KFROch]. Sur la figure, PT signifie premier temps, T signifie temps et NT signifie non temps, ou absence de temps. Par exemple, la partie nommée "PT pré-attaque" correspond à la partie qui se situe juste avant l'attaque d'un premier temps.

Ce modèle de langage est appris dans le cadre d'un modèle de Markov caché [4] qui est souvent utilisé pour l'estimation du tempo et de la position des temps mais qui montre des limites dans le cas où les signaux présentent une grande variabilité. Gillet qui travaillait sur

4. Ce type de modèle est présenté par Rabiner et décrit à l'annexe A [Rab89].

la transcription de batterie a signalé ce problème et constate qu'utiliser plusieurs modèles de batterie améliore les performances [GR04] [GR08]. Il vaut donc mieux apprendre plusieurs modèles et utiliser celui qui correspond le mieux au signal à analyser.

Chapitre 2

Méthodes d'estimation de la position des temps et des premiers temps développées pendant ce stage

Ce chapitre regroupe les trois grandes méthodes qui ont été développées pendant ce stage. La première est une implémentation d'un article scientifique avec quelques modifications tandis que les deux autres ont été développées dans le cadre de ce stage. La méthode tirée de l'état de l'art permet d'estimer conjointement le tatum, le tactus et la mesure puis de trouver la position des temps et des premiers temps. La deuxième méthode permet de trouver le tatum, le tactus et d'estimer la position des temps dans le cas où ceux-ci sont régulièrement espacés les uns des autres. La troisième méthode prend en entrée la position des temps et en déduit la position des premiers temps.

2.1 Estimation conjointe de la période et de la phase du tatum, du tactus et de la mesure

Cette méthode s'inspire en grande partie de l'article d'Anssi Klapuri dans les transactions Speech et Audio Processing de l'IEEE [KEA06]. Cette méthode à l'avantage d'être complète, de l'audio à la phase des premiers temps, et d'offrir des résultats performants [1]. Le schéma général de cette méthode est disponible à la figure 2.1. Il reprend les différentes notations que nous verrons par la suite.

2.1.1 Calcul des accents du signal

Après la normalisation signal, échantillonné à 44100 Hz, il s'agit d'en calculer la transformée de Fourier à court terme. Pour cela, des fenêtres de Hanning de 0.023 seconde avec un recouvrement de 75 % sont utilisées. Une fenêtre de petite taille permettra d'obtenir une bonne position temporelle tandis qu'une fenêtre de taille plus grande permettra une meilleure résolution fréquentielle, ce qui est utile dans le cas de musiques non percussives où il faut détecter des changements de notes à énergie globale constante ou des attaques

1. Bien que non présent dans les dernières évaluations MIREX, cet algorithme est classé premier par les tests effectués par Holzapfel dans son article publié en novembre 2012 sur 16 beat trackers et une base de données de 217 extraits musicaux [HDZ+12].

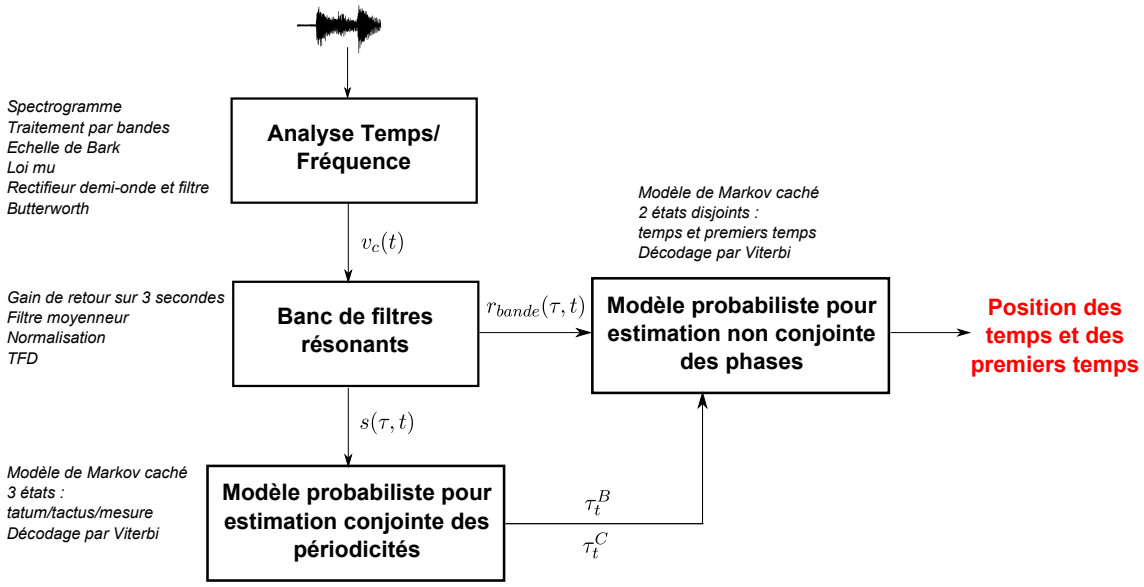


FIGURE 2.1 – Fonctionnement général de la méthode d'estimation conjointe de la période et de la phase du tatum, du tactus et de la mesure. On peut voir les entrées et les sorties des blocs entre ceux-ci. Le texte en italique à côté de chaque bloc présente les points importants de chaque étape. Chaque point est expliqué à la partie 2.1.

douces. Le spectrogramme x est ensuite réparti sur 36 bandes fréquentielles comme le propose Goto : $x_b(t)$ avec $b = 1, 2, ..., 36$ le numéro de la bande et t le numéro de la trame. Nous proposons d'utiliser une échelle uniforme de Bark pour les bandes fréquentielles. Elle se rapproche de notre perception auditive.

$$z = \frac{26.81}{1 + \frac{1960}{f}} - 0.53 \tag{2.1}$$

Avec f la fréquence et z l'échelle de Bark.

L'étape suivante est la mesure du degré de changement de l'enveloppe spectrale, afin de détecter les attaques. La loi μ est utilisée pour cela.

$$y_b(t) = \frac{\ln(1 + \mu x_b(t))}{\ln(1 + \mu)} \tag{2.2}$$

Avec $\mu = 100$. C'est une loi qui se rapproche de la perception humaine. La perception d'un changement d'intensité est proportionnelle à l'intensité elle même. Cette loi va linéariser le signal aux alentours de zéro et le compresser de manière logarithmique ailleurs afin de travailler sur des variations d'énergie relatives au niveau énergétique, comme l'oreille. Cette linéarisation présente cependant un défaut pour la transition d'un silence à une absence de silence. Cette transition va être trop fortement marquée. Nous proposons alors d'utiliser une fenêtre de Hanning pour adoucir la transition.

L'enveloppe du signal sera ensuite calculée au moyen d'un filtre passe bas de Butterworth [2] d'ordre 6 coupant à 10 Hz. $z_b(t) = \text{Butterworth}(y_b(t))$ avec Butterworth le filtre de

2. Ce filtre a pour gain $G(\omega) = \frac{1}{\sqrt{1+\omega^{2n}}}$ avec n l'ordre du filtre et ω la pulsation. Il permet d'avoir un gain très plat dans la bande passante.

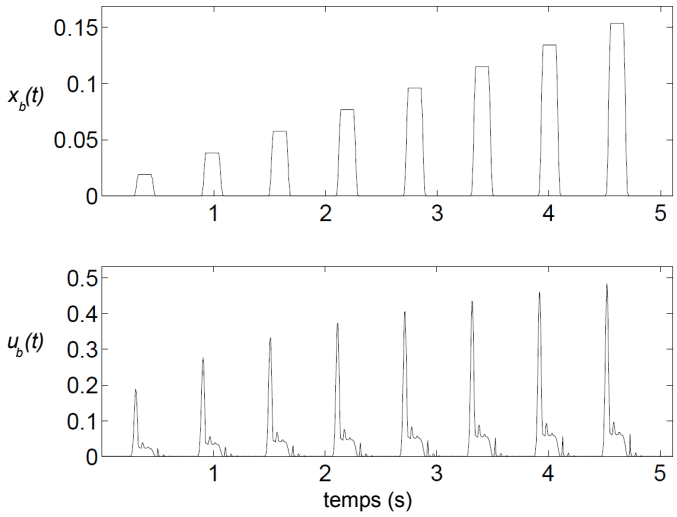


FIGURE 2.2 – Compression dynamique et différenciation permettant de faire ressortir les attaques. On remarque $u_b(t)$ n'est pas nul avec avoir atteint un maximum local grâce à la présence du terme z_b à l'équation 2.4. D'après [KEA06].

Butterworth.

On calcule $z'_b(t)$ la rectification demi-onde de la différence de $z_b(t)$:

$$z'_b(t) = \max\left(z_b(t) - z_b(t-1), 0\right) \tag{2.3}$$

La fonction de détection des accents du signal $u_b(t)$ sur 36 bandes est obtenue en effectuant une somme pondérée de z_b et z'_b :

$$u_b(t) = (1-\lambda)z_b(t) + \lambda\frac{f_e}{f_{but}}z'_b(t) \tag{2.4}$$

Avec f_e la fréquence d'échantillonnage, $f_{but} = 10$ Hz la fréquence de coupure du filtre de Butterworth et $\lambda = 0.8$ la pondération de la somme pondérée obtenu de façon empirique par Klapuri. La plupart du temps [3] $\lambda = 1$, mais la prise en compte de z_b permet de donner un peu de poids à l'enveloppe spectrale et pas uniquement à sa différenciation La figure 2.2 illustre la fonction la fonction de détection sur 1 des 36 bandes. On remarque la différenciation mais également le poids donnée par z_b.

Les 36 bandes sont alors sommées linéairement pour obtenir 4 bandes fréquentielles. On obtient la fonction de détection v_{bande} :

$$v_{bande}(t) = \sum_{b=9*(bande-1)+1}^{9*bande} u_b(t), \quad bande = \{1,2,3,4\} \tag{2.5}$$

Ce résultat a été trouvé empiriquement et permet un compromis entre une information trop segmentée avec un grand nombre de bandes gardées et une information trop diluée avec un petit nombre de bandes conservées.

3. Chez Scheirer par exemple [Sch98].

Le signal utile

La partie suivante n'est pas présente dans l'article de Klapuri et concerne toujours le calcul des accents du signal. L'étude d'exemples musicaux sur lesquels les algorithmes de l'état de l'art sont en difficultés par Holzapfel a montré que les extraits où la voix est fortement mise en avant posaient souvent problème [HDZ+12]. Holzapfel ajoute que les performances humaines sur ces extraits sont bonnes. Il est probablement plus important de s'attarder dans un premier temps sur les cas où les humains sont performants mais pas les algorithmes plutôt que sur les cas où le rythme est difficilement estimable pour tout le monde [4]. L'estimation de rythme à partir d'un enregistrement audio est une tâche complexe et plusieurs processus doivent être mis correctement bout à bout afin de parvenir à un résultat satisfaisant, mais il revient qu'un des points les plus bloquants est l'estimation précise d'accents musicaux [AA06]. Cette étape, en début de chaine conditionnera le reste et une erreur à ce niveau se propagera par la suite.

Rosenthal a estimé qu'une des principales caractéristiques de l'habileté du cerveau humain est sa capacité à se focaliser sur certaines zones fréquentielles plus faciles à estimer rythmiquement [Ros92]. Il apparait donc qu'une sélection d'une ou plusieurs bandes fréquentielles utile(s) rythmiquement peut améliorer les résultats. Sur un exemple bruité par une voix, sélectionner les bandes fréquentielles d'accompagnement peut être intéressant.

Comment définir les bandes d'accompagnement ou les bandes utiles ? L'utilité rythmique est la capacité à faire ressortir un petit nombre répétitions claires et marquées. Si l'on dispose de répétitions, il est possible d'établir un rythme. De plus, si une répétition est clairement marquée par rapport au reste, elle aura plus de chance d'être significative. Une première façon de mesurer cette utilité est d'utiliser l'autocorrélation sur le signal normalisé. La taille de la fenêtre d'autocorrélation est importante. Une petite fenêtre ne permettra pas de détecter des périodicités élevées tandis qu'une grande fenêtre ne permettra pas de considérer des variations légères de tempo. On choisit alors une petite fenêtre pour l'estimation du tatum et du tactus et une fenêtre plus longue pour l'estimation de la mesure. À chaque instant et sur chacune des 36 bandes, le signal est normalisé et l'autocorrélation a_b est calculée avec b le numéro de la bande fréquentielle. On somme les contributions qui sont multiples les unes des autres sur un intervalle non nul pour permettre les variations de tempo. Afin de conserver les motifs dont la répétition est marquée, on garde les cas où au moins trois répétitions ont eu lieu et on divise le résultat par le nombre de contributions. La valeur maximale A_b sur chaque bande est conservée. On conserve les bandes les plus significatives, c'est à dire telles que $A_b > s * \max_b(A_b)$ avec $s < 1$ un seuil.

Une étude de l'état de l'art sur le sujet a montré que cette méthode était de proche de celle proposée par Rafii [RP13]. Les efforts n'ont pas été poussés sur le sujet car l'apport n'est pas significatif sur tous types de signaux. Ce travail a cependant permis d'en apprendre plus sur les indices permettant d'extraire un rythme dans un environnement ou plusieurs contributions sont présentes.

2.1.2 Estimation des périodicités

Klapuri propose ici d'utiliser un banc de filtres résonants dont la formule a été présentée à l'équation 1.1, avec T_0 le temps de retour égal à 3 secondes, ce qui permet de prendre des

4. Dans le cas de musiques avec timing expressif comme la musique romantique par exemple.

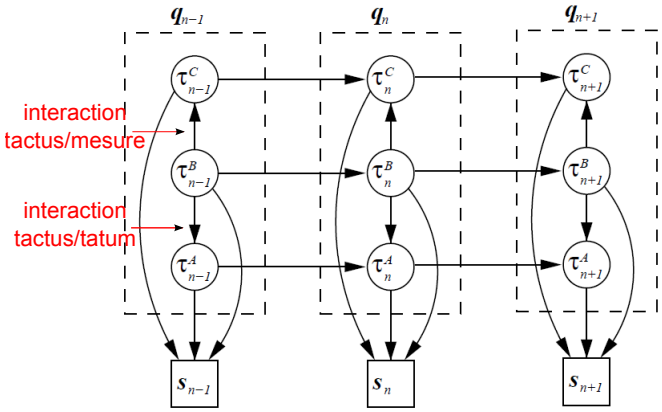


FIGURE 2.3 – Modèle de markov caché pour l'évolution temporelle du tatum, du tactus et de la mesure. D'après [KEA06]. n est l'indice temporel, s l'observation et q l'ensemble des 3 états cachés conjoints. On constate une interaction entre le tactus et le tatum et une interaction entre le tactus et la mesure

périodicités qui peuvent faire jusqu'à 4 secondes de durée. La sortie $r_{bande}(\tau, t)$ du filtre avec τ le retard et t le temps [5] est normalisée. Elle est ensuite sommée sur 4 bandes. Elle sert de descripteur au tactus et à la mesure et est appelée $s(\tau, t)$.

$$s(\tau, t) = \sum_{bande=1}^{4} r_{bande}(\tau, t) \tag{2.6}$$

À un instant t donné, un retard τ tel que $s(\tau, t)$ soit élevé a de grande chance d'être une périodicité du signal. La périodicité du tatum est estimée par une transformée de Fourier à temps discret qui met en avant la périodicité la plus rapide se répercutant dans d'autres niveaux métriques plus lents comme le tactus ou la mesure. Le descripteur est nommé $S(f, t)$. Étant donné que S est une transformée de Fourier, à un instant t donné, un retard $\frac{1}{f}$ tel que $S(f, t)$ est élevé a de grande chance d'être un tatum.

2.1.3 Suivi par estimation conjointe du tatum, tactus et de la mesure et a priori musical

L'estimation conjointe de ces trois niveaux de métrique est effectuée à l'aide d'un modèle de Markov caché. Plus d'informations sur ce type de modèle est disponible à l'annexe A. Le modèle retenu est d'ordre 1, discret, gauche-droite et avec trois états cachés conjoints pour une observation comme le montre la figure 2.3. À chaque instant, on cherche à obtenir conjointement les périodicités τ_t^A du tatum, τ_t^B du tactus et τ_t^C de la mesure. Il reste à définir les probabilités de transition et d'émission ainsi que les observations. L'estimation conjointe est réalisée en multipliant les contributions des trois niveaux métriques et en utilisant deux fonctions d'interactions qui sont décrites ci-dessous. Pour simplifier cette partie, nous expliquerons uniquement comment les probabilités d'émission et de transition du tactus ainsi que fonctions d'interaction entre les niveaux métriques sont calculées. Le raisonnement est identique pour le tatum et la mesure.

5. Ici le numéro de la trame.

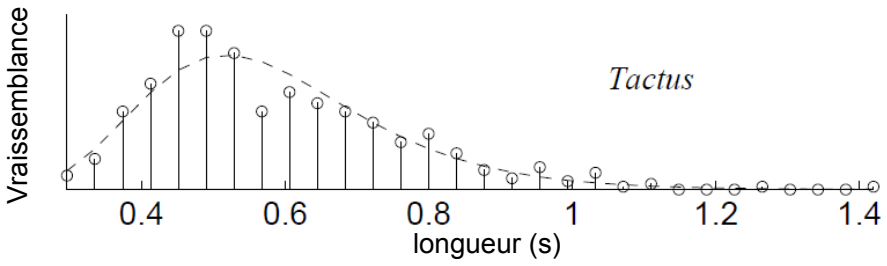


FIGURE 2.4 – Représentation de H_{tactus}. Histogramme représentant la répartition de la valeur du tactus sur la base de données de Klapuri. H_{tactus} correspond à la gaussienne qui décrit le mieux la tendance de l'histogramme, ici en pointillés. D'après [KEA06].

La séquence d'observation :

L'observation à l'instant t est la sortie du filtre résonant normalisé $s(\tau, t)$ multipliée par un histogramme H_{tactus} de la répartition du tactus dans la base de données de Klapuri. Cet histogramme agit comme un filtre qui isole les périodicités liées au tactus et est illustré à la figure 2.4. Pour des raisons de coût en calcul, les retards conduisant aux 5 valeurs maximales de ce produit sont conservées. $Obs_{\tau_1} = \arg\max_{\tau}(s(\tau) * H_{tactus})$ pour le maximum, puis on obtient le deuxième maximum $Obs_{\tau_2} = \arg\max_{\tau}(s2(\tau) * H_{tactus})$ avec $s2(\tau) = s(\tau)\ \forall \tau \neq Obs_{\tau_1}$ et $s2(Obs_{\tau_1}) = 0$, et ainsi de suite jusqu'à obtenir le cinquième maximum. À chaque instant il y a donc 5 observations liées au tactus. Étant donné que le procédé est identique pour la tatum et la mesure et que l'on garde toutes les possibilités, il y a à chaque instant $5^3 = 125$ observations.

La probabilité d'émission :

La probabilité d'émission E du tactus est la valeur du descripteur s au niveau du retard retenu comme étant une observation $E_t = s(Obs_\tau, t)$.

La probabilité de transition :

La probabilité de transition T regroupe deux contributions différentes $\{T_1, T_2\}$, qui sont multipliées entres elles : $T = T_1 * T_2$.

La première contribution est l'histogramme de répartition du tactus H_{tactus}. La deuxième contribution est une gaussienne en échelle logarithmique centrée sur l'unité et d'écart type tel que les doublements de la valeur du tactus soient peu probables comme le montre la figure 2.5. Le raisonnement est que pour passer de τ^B_{t-1} à τ^B_t, il faut que τ^B_t soit un tactus (première contribution) et qu'il soit suffisamment proche de τ^B_{t-1} (deuxième contribution).

L'interaction entre les niveaux métriques :

Cette interaction est regroupée dans la probabilité de transition car elle concerne la transition entre deux états conjoints consécutifs. Elle traduit le fait que les niveaux métriques doivent être des multiples entre eux et que certains multiples sont favorisés. Par exemple, une mesure contiendra plus souvent 4 temps que 7 temps. L'interaction concerne uniquement les niveaux métriques adjacents : le tatum/tactus et le tactus/mesure. La fonction

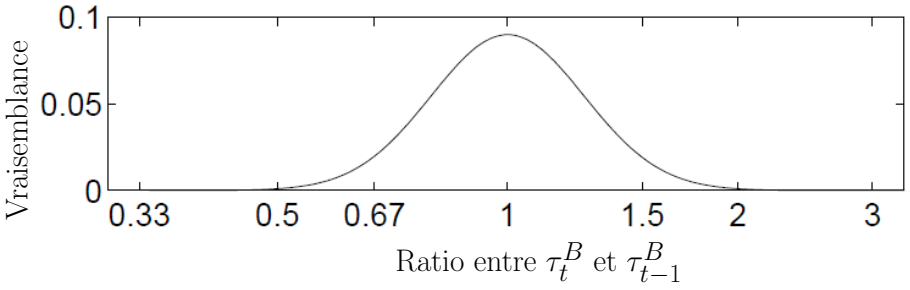


FIGURE 2.5 – Fonction de transition entre deux tactus consécutifs.

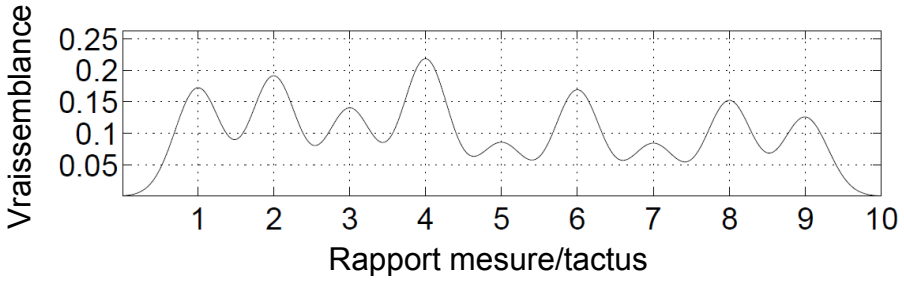


FIGURE 2.6 – Fonction d'interaction entre le tactus et la mesure. D'après [KEA06].

d'interaction qui prend comme entrée le rapport entre ces niveaux métriques a donc la forme d'une somme pondérée de gaussiennes comme le montre la figure 2.6. Klapuri utilise les même fonctions d'interaction entre les tactus/tatum et mesure/tactus. Nous proposons une formulation différente pour le rapport tactus/tatum donnant plus de poids aux rythmes ternaires : $p(\frac{\tau^B}{\tau^A} = 3) > p(\frac{\tau^C}{\tau^B} = 3)$.

Séquence optimale d'états cachés :

La séquence optimale d'états est obtenue à l'aide d'un algorithme de Viterbi. Pour résumer, à chaque instant, 125 combinaisons de valeurs de tactus, tatum et mesure sont calculées à l'aide du filtre résonant des périodicités et d'un histogramme de représentation de ces niveaux métriques. La transition d'un de ces états à l'un des 125 de l'instant suivant est obtenue en privilégiant une continuité du tempo, un rapport entier et binaire entre les niveaux métriques adjacents et en respectant une répartition des niveaux métriques proches de ce que l'on observe le plus souvent. On obtient au final à chaque instant [6] les valeurs du tatum, du tactus et de la mesure.

La figure 2.7 résume la procédure globale de l'estimation des trois niveaux métriques.

6. Plus précisément tous les 172ème de seconde.

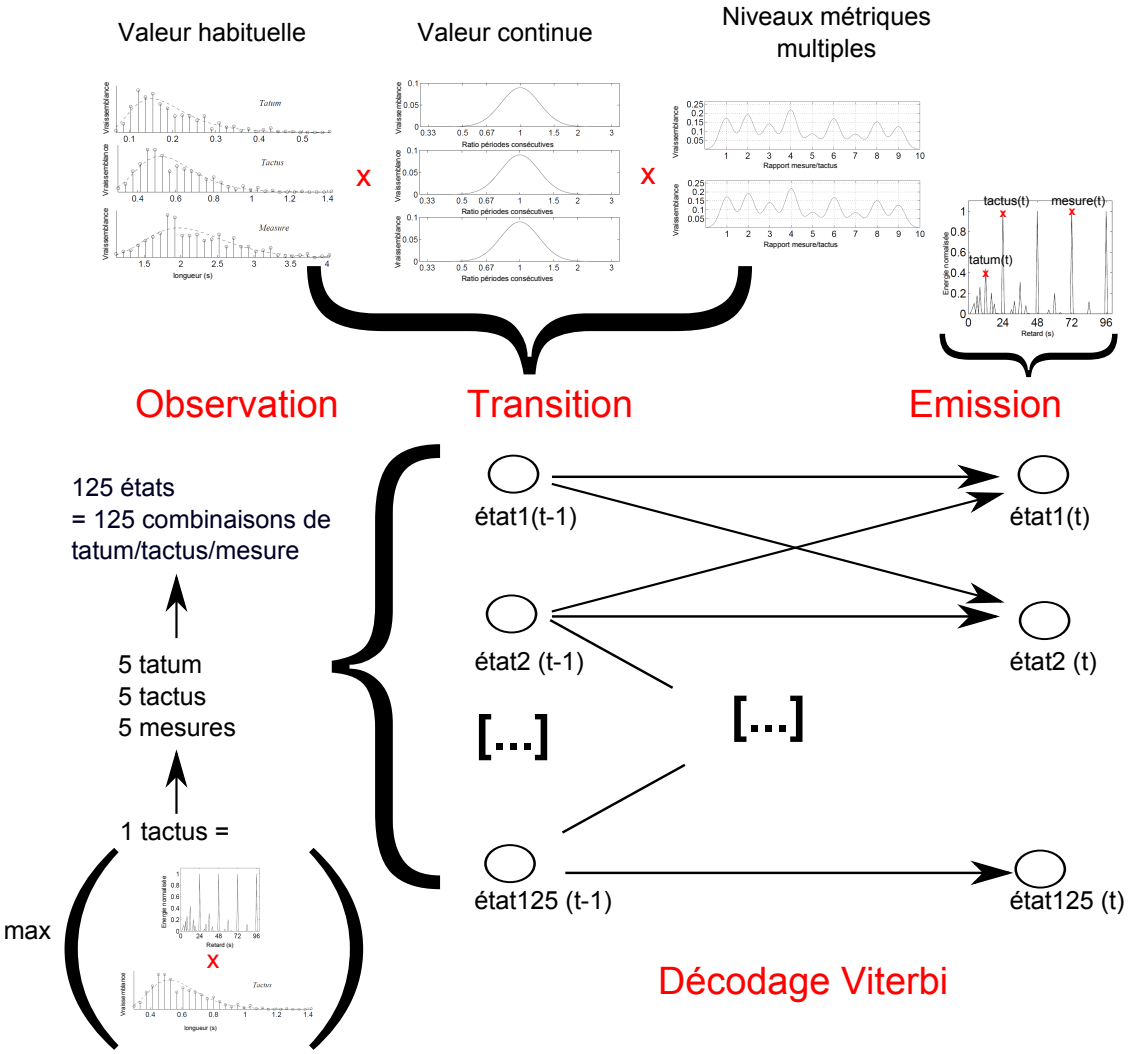


FIGURE 2.7 – Résumé de l'estimation du tatum, du tactus et de la mesure. Les croix rouges au niveau de la partie sur la transition correspondent aux multiplications des fonctions affichées et celles au niveau de la partie sur l'émission correspondent à la probabilité d'émission en fonction de l'abscisse, c'est à dire la valeur des niveaux métriques. Les fonctions sont tirées de [KEA06]. À chaque instant, 125 états représentant chacun une valeur de tatum/tactus/mesure sont possibles. Un algorithme de Viterbi permet de trouver l'enchainement optimal de ces combinaisons de trois états à l'aide des probabilités de transitions et d'émission présentées ci-dessus.

2.1.4 L'estimation de la position des temps, des premiers temps

L'estimation de la position des temps et des premiers temps s'effectue ici également à l'aide d'un modèle de Markov caché. On a ici deux modèles disjoints, un pour les temps et l'autre pour les premiers temps. De plus, comme il s'agit de la phase des temps et non de périodicités, le problème n'est pas de savoir si l'on se situe à chaque instant sur un temps ou non mais de savoir à chaque instant où se situe le temps précédent.

La probabilité de transition :

On considère à chaque instant l'écart avec dernier temps estimé :

$$T = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e^2}{2\sigma^2}\right) \tag{2.7}$$

Avec T la probabilité de transition d'un temps au suivant, $\sigma = 0.1$ l'écart type de l'oscillation de tempo et :

$$e = \frac{1}{\tau_t^n}\left\{\left[\left(|\phi_t^n - \phi_{t-1}^n| + \frac{\tau_t^n}{2}\right) \mod \tau_t^n\right] - \frac{\tau_t^n}{2}\right\} \tag{2.8}$$

Avec τ_t^n et $n = \{B, C\}$ le tactus ou la mesure à la trame t et ϕ_t^n la position du temps ou du premier temps précédent la trame t.

On a une matrice de transition qui donne une probabilité maximale lorsque le temps suivant est situé une périodicité plus loin que le temps actuel comme le montre la figure 2.8.

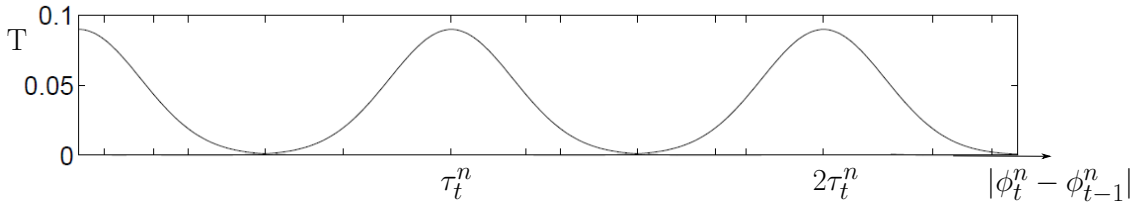


FIGURE 2.8 – Probabilité de transition T en fonction de $|\phi_t^n - \phi_{t-1}^n|$ avec $n = \{B, C\}$ les phases du tactus et de la mesure.

La probabilité d'émission :

La matrice d'émission est obtenue à partir du descripteur $r_{bande}(\tau, t)$, avec $bande = \{1, 2, 3, 4\}$ qui est égal au descripteur du tactus et de la mesure avant la somme sur les bandes. Ainsi, on conserve une information différente sur les basses, moyennes et hautes fréquences. On considère la valeur de r au niveau du niveau métrique correspondant $R_{bande}^{tactus}(t) = r_{bande}(\tau_t^B, t)$.

Pour l'estimation de la position des temps, on effectue une somme pondérée sur les différentes bandes de R^{tactus} en donnant plus de poids aux basses fréquences qui se révèlent plus stables et robustes : $E_{tactus}(t) \propto \sum_{i=1}^{4}(4 - i + 2)R_{bande_i}^{tactus}(t)$.

L'estimation de la position des premiers est plus complexe car elle nécessite des informations de plus haut niveau. Ici, la probabilité d'une trame d'être un premier temps E_{mesure} est liée à la corrélation entre les accents du signal sur les quatre bandes de R^{mesure} et des motifs d'une durée de quatre temps sur chacune des bandes et échantillonné en quatre points seulement. À chaque instant t, on regarde la valeur de $R^{mesure}_{bande}(t) = r_{bande}(\tau^C_t, t)$ à $\{t - \tau^B_t, t - 2\tau^B_t, t - 3\tau^B_t, t - 4\tau^B_t\}$ au niveau des 4 bandes fréquentielles. Si ces valeurs correspondent à celles des motifs, alors E_{mesure} est élevée. Ces motifs ont été définis de manière empirique et Klapuri a trouvé que les deux motifs suivants donnaient les meilleurs résultats :

$$\begin{bmatrix} 12 & 1.0 & 0 & 5.7 \\ 0 & 2.0 & 0 & 2.0 \\ 0 & 3.0 & 0 & 3.0 \\ 0 & 4.0 & 0 & 4.0 \end{bmatrix} \text{ et } \begin{bmatrix} 10 & 0 & 1.4 & 1.3 \\ 0 & 0 & 2.8 & 0.8 \\ 0 & 0 & 4.3 & 1.2 \\ 0 & 0 & 5.8 & 1.5 \end{bmatrix}$$

Les lignes représentent les bandes de fréquence et les colonnes les temps précédents la trame. On peut voir une certaine balance harmonique dans le premier cas et une structure binaire dans les deux cas. L'estimation sera en effet mauvaise pour des mesures en multiple de 3 temps, qui sont peu représentées dans les bases de données. Nous proposons alors un motif de mesure en trois temps :

$$\begin{bmatrix} 12 & 1.0 & 1.0 \\ 0 & 2.0 & 2.0 \\ 0 & 3.0 & 3.0 \end{bmatrix}$$

Ce motif est pris en compte lorsque $\frac{\tau^C_t}{\tau^B_t} \mod 3 < s$, avec un seuil $s = 0.3$.

Séquence optimale d'états cachés :

À chaque instant t, 15 états sont conservées. Ce sont les 15 plus grandes valeurs de $E_{mesure}([t - \tau^C_t, t])$ pour la mesure et les 15 plus grandes valeurs de $E_{tactus}([t - \tau^B_t, t])$ pour le tactus. La séquence optimale d'états cachés est obtenue à l'aide d'un algorithme de Viterbi.

La figure 2.9 résume la procédure de l'estimation de la position des temps.

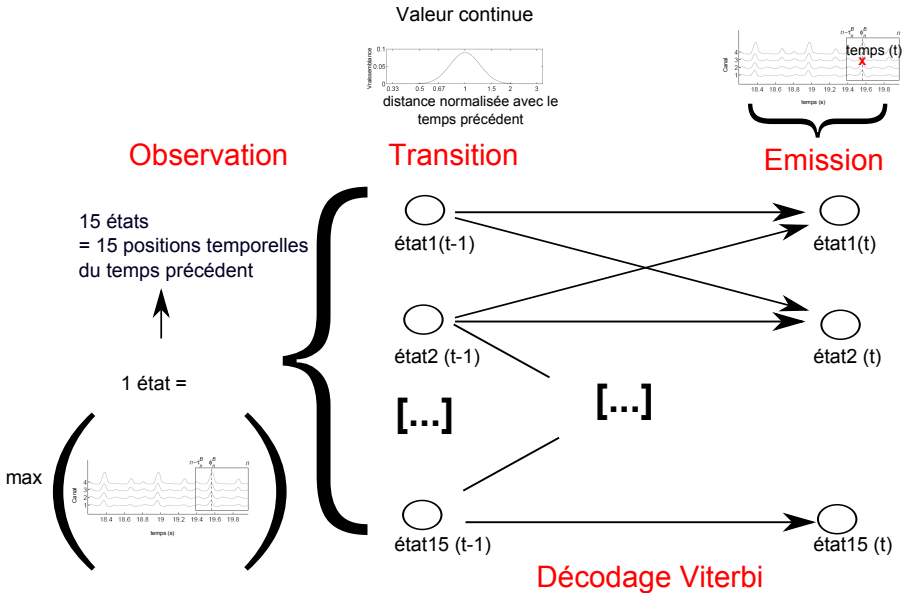


FIGURE 2.9 – Résumé de l'estimation de la position des temps. La croix rouges au niveau de la partie sur l'émission correspond à la probabilité d'émission en fonction de l'abscisse, c'est à dire la position du temps précédent. Les fonctions sont tirées de [KEA06]. Á chaque instant, 15 états représentant chacun la position temporelle du temps précédent sont possibles. Un algorithme de Viterbi permet de trouver l'enchainement optimal de ces positions à l'aide des probabilités de transitions et d'émission présentées ci-dessus.

On obtient au final la tatum, la tactus, la mesure ainsi que la position des temps et des premiers temps.

2.1.5 Différence avec la méthode originale de Klapuri

Voici les modifications que nous avons effectuées par rapport à la version originale de la méthode de Klapuri :

1) utilisation de l'échelle de Bark,
2) lissage du signal après un silence,
3) définition du signal utile,
4) pondération différente du rapport tactus/tatum et mesure/tactus,
5) ajouts d'un motif de mesure.

Malheureusement, ces modifications n'ont pas apporté d'améliorations significatives des résultats. Par exemple, les points 4) et 5) permettent une meilleure adaptation aux mesures à trois temps, qui sont très peu présentes dans la base de données et dont l'influence est donc difficile à évaluer. Cependant, l'objectif principal du stage n'était pas d'améliorer les méthodes de l'état de l'art, mais de s'inspirer des méthodes actuelles afin d'établir des algorithmes d'estimation de la position des temps et des premiers temps. C'est ce qui est présenté dans les parties suivantes.

2.2 Estimation du tempo et de la phase des temps par système d'arbre

Cette méthode mise en place dans le cadre de ce stage permet d'obtenir le tatum, le tactus et la phase des temps. La méthode ne prend pas les changements de tempo en compte [7]. Le fonctionnement général de la méthode est illustré à la figure 2.10.

Le principe général de cette méthode est de choisir de multiples candidats pour le tempo et la phase des temps, puis de choisir ceux qui représenteront le mieux le contenu rythmique du signal ainsi qu'une stabilité dans le tempo.

2.2.1 Calcul des accents du signal

Les accents du signal sont calculés de la même façon que dans la méthode développée par Klapuri à la partie 2.1. Même si cette étape est importante, il a été choisi de se concentrer sur l'établissement de périodicités et de phases de temps.

2.2.2 Calcul de la périodicité potentielle

La première étape du calcul de la périodicité est le calcul du tatum. Ce calcul est effectué à l'aide de l'autocorrélation du descripteur de périodicité d'un intervalle de 4 secondes à chaque instant permettant un compromis entre une mesure locale et globale. On oberve alors plusieurs pics dont certains sont des multiples les uns des autres. Étant donné que le tatum est la plus petite périodicité significative du signal, on considère chaque série de tatums multiples les uns des autres et on garde, dans un intervalle raisonnable de tatum [8] :

- Le tatum le plus rapide de la série
- Le tatum d'amplitude maximale de la série

On obtient alors plusieurs tatum potentiels. Leur valeur est affinée en procédant à une moyenne de l'écart des multiples de ce tatum, pour éviter les erreurs d'arrondi.

On calcule alors le poids des tactus associés à ces tatums. Les tactus étant des multiples des tatums, on considère plusieurs multiples plausibles de la périodicité la plus faible du signal musical. Le poids des tactus est calculé à l'aide de deux méthodes complémentaires.

- La présence absolue du tactus dans la périodicité, obtenue à l'aide d'une somme spectrale
- La présence relative du tactus dans la périodicité, obtenue en regardant la position relative de chaque multiple du tatum par rapport aux autres sur une période de tactus

Ce choix se justifie par le fait que le descripteur de périodicité a tendance à surestimer les périodicités élevées. Ainsi, la mesure absolue basée sur la somme spectrale aura tendance à favoriser les tactus lents et la mesure relative aura tendance à favoriser les tactus rapides. Ces deux mesures se compensent et leur produit permettra d'avoir une estimation plus robuste.

7. Pour contourner cette limitation, le signal audio pourrait être découpé en plusieurs parties sur lesquelles l'hypothèse de tempo constant serait plus facilement vérifiée. D'autres développements pourraient être pris en compte afin de prendre en compte les changements de tempo mais ce n'a pas été le but ici.

8. Cet intervalle est obtenu en considérant la distribution des valeurs du tatum sur plusieurs centaines de morceaux de genres différents obtenue par Klapuri dans [KEA06] et en considérant les valeurs significatives comme celles situées dans un intervalle de confiance de 95 %

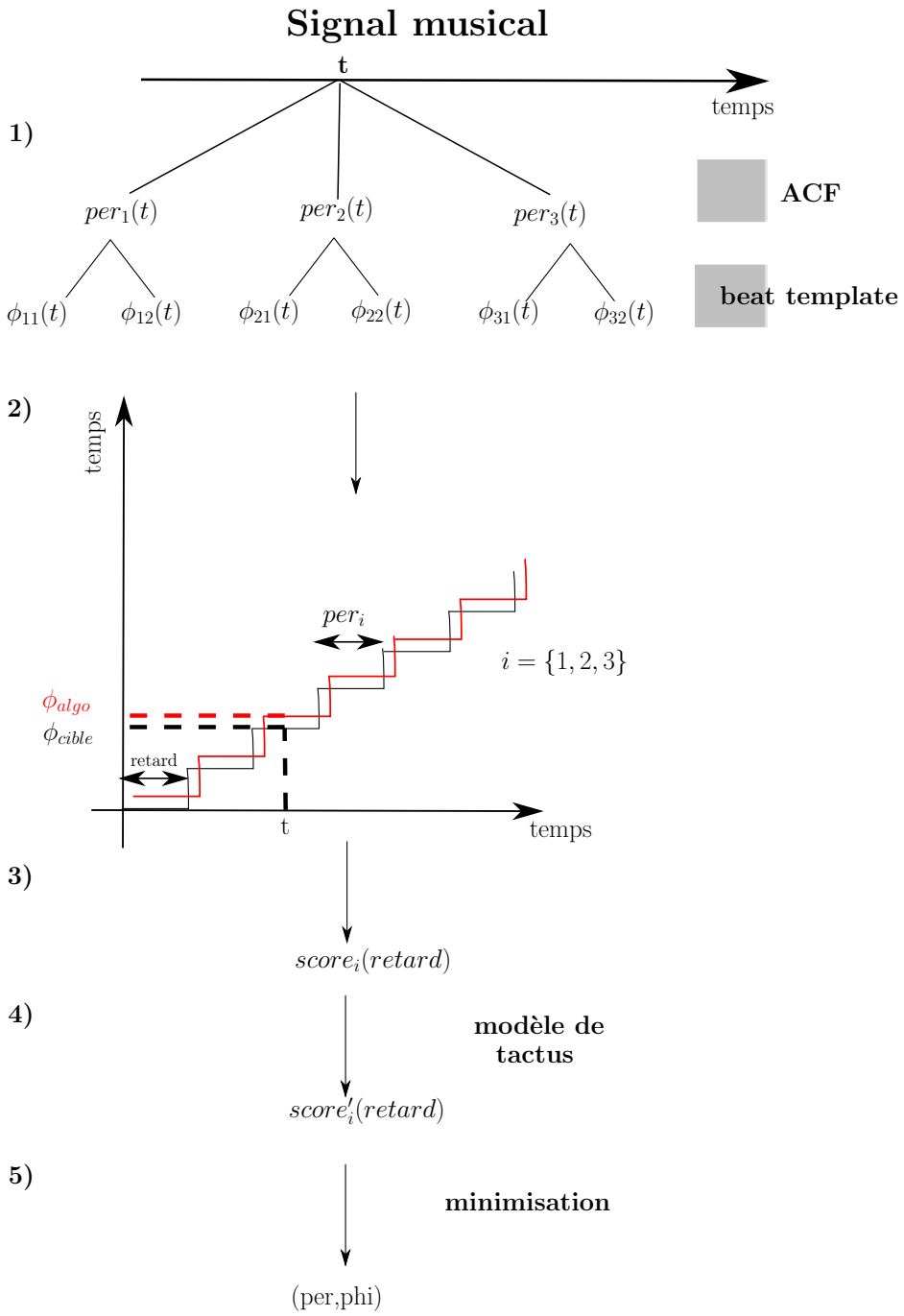


FIGURE 2.10 – Schéma général de la méthode d'estimation du tempo et de la phase des temps par système d'arbre. *per* est la périodicité et ϕ la phase du temps précédent le plus proche. Dans la première partie, à chaque instant trois périodes sont estimées à l'aide d'une fonction d'autocorrélation et d'un système de décision. Pour chaque période, deux phases de temps sont obtenues à l'aide d'un beat template. Dans la deuxième partie, un écart est calculé entre les phases obtenues par l'algorithme (en rouge) et les phases obtenues par différentes pulsations, espacées entre elles d'une durée per_i et espacées de l'origine d'une durée *retard* (en noir). Dans la troisième partie un score correspondant à cet écart est calculé, puis un modèle de tactus est appliqué pour obtenir un score plus robuste. Enfin, une minimisation de ce score est effectuée à la dernière partie pour obtenir le tempo *per* et la position des temps ϕ.

Les poids des tactus proches sont sommés avec une pénalité afin de favoriser les tactus qui émergent pour différents tatums et qui expliquent donc bien le signal. Les trois tactus

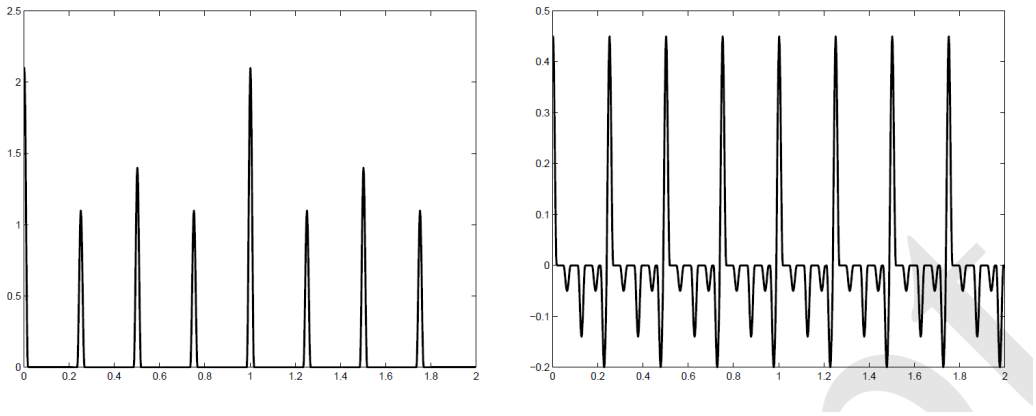


FIGURE 2.11 – Comparaison entre un beat-template classique à gauche et un beat-template obtenue par analyse discriminante linéaire à droite. En ordonnée le poids et en abscisse le temps, en seconde. Les maxima correspondent à la position des temps, espacés les uns des autres de 0.5 secondes. On observe à droite des valeurs négatives entre les temps qui permettent de discriminer au maximum les temps du reste lors d'une corrélation avec une fonction de détection par exemple. D'après [PP11].

qui ont le poids le plus élevé sont considérés comme étant les trois périodicités les plus probables du signal à cet instant. Si plusieurs tatums leur sont associés, on choisit celui qui permet le score maximal.

2.2.3 Calcul des phases des temps potentielles

À Chaque instant, trois périodicités de tatum et de tactus sont obtenues. Pour chaque couple {tatum,tactus}, deux phases de temps sont calculés de la façon suivante.

Le descripteur de temps est celui calculé par Klapuri, basé sur la règle de la basse stable [9]. On calcule alors la corrélation avec un beat template. Le beat template obtenu par Peeters et Papadopoulos par analyse discriminante linéaire est intéressant car en possédant des valeurs négatives entre les temps il permet de mieux éviter les erreur d'estimation des contre-temps et de maximiser la discrimination entre les temps et le reste. Les deux scores les plus élevés sont conservés et permettent d'obtenir les deux phases de temps les plus probables précédents l'instant considéré pour une périodicité donnée.

2.2.4 Minimisation des résultats par rapport à un modèle cible

Afin d'obtenir la position des temps la plus probable sur tout le morceau et en considérant toujours un tempo fixe, il va s'agir ici de minimiser l'écart entre des temps espacés de manière égale et les 6 phases de temps obtenues à chaque instant à la partie précédente.

On cherche à obtenir à chaque instant la position du temps précédent. Ainsi, on obtient sur un axe xy {x=instant du morceau, y=position du temps précédent l'instant du morceau} une fonction en escalier avec un certain retard à l'origine et dont les marches sont espacées d'une périodicité. On peut voir cette fonction à la partie 2) de la figure 2.10. La périodicité choisie per_j avec $j = \{1, 2, 3\}$ est la médiane, plus robuste que la moyenne,

9. On somme une fonction d'attaque par bande en donnant plus de poids aux bandes basse fréquence

des valeurs prises par tactus $per_j(t)$ à chaque instant t. $per_j = med_t(per_j(t))$.

Pour un retard allant de zéro à une fois et demie la valeur de la périodicité et pour chacune des trois périodicités obtenues précédemment, on calcule l'écart quadratique entre la fonction en escalier cible ou idéale et la valeur de phase la plus proche parmi les six obtenues.

$$S_r^j = \sum_t \min_i \left(\phi_i(t) - C_r^j(t)\right)^2 \tag{2.9}$$

Avec $r = 0...1.5per_j$ le retard à l'origine, per_j la périodicité $j = \{1, 2, 3\}$, $i = \{1, 2, 3, 4, 5, 6\}$ le numéro de la phase ϕ_i, C_r^j la fonction cible qui représente des temps espacés de per_j et dont la position du premier temps est égale à r.

Le score $S^j = \max_r(S_r^j)$ obtenu pour chacune des trois périodicités est modifié suivant la probabilité d'obtenir une telle périodicité [10].

$$S'^j = S^j H_{tactus}(per_j) \tag{2.10}$$

Avec H_{tactus} défini à la figure 2.4. Le minimum global des scores des trois périodicités permettra d'obtenir le tatum, le tactus et la phase des temps du morceau.

2.3 Estimation des premiers temps : introduction de concepts musicaux

Un des points clés de ce stage était l'estimation de la position des premiers temps. En effet, cette tâche délicate qui requiert des informations de haut niveau a assez peu été traité dans la littérature. Étant donné que les premiers temps ne sont pas directement observables dans le contenu audio mais qu'il sont assez bien perçus par l'Homme, l'angle d'approche a été de choisir plusieurs aspects musicaux qui sont présents au même instant que les premiers temps plutôt que de procéder à une approche d'apprentissage aveugle ou aléatoire. L'idéal est d'avoir des descripteurs fiables pour une grande variabilité de signaux musicaux, complémentaires les uns avec les autres et peu coûteux en calcul. Si certains descripteurs sont fiables sur une partie du domaine, on en tiendra compte lorsque l'intervalle de confiance est suffisamment élevé, c'est à dire lorsqu'un temps est fortement mis en avant.

Une fois qu'un descripteur semblait approprié, une implémentation a été mise en place et sa validité a été testée. Pour avoir une estimation des premiers temps fiable et robuste, il faut au préalable estimer le nombre de temps par mesure puis sélectionner le résultat majoritaire des différents descripteurs.

Quatre descripteurs de premier temps ont été choisis, l'instant de changement d'accord, la balance harmonique, le profil mélodique, et la distance entre patterns musicaux. Le premier est directement repris de l'état de l'art bien que son utilisation soit différente, le deuxième est repris sous une forme différente et les deux derniers ont peu été considérés par la littérature dans ce but à notre connaissance. Les méthodes suivantes supposent la connaissance préalable de la position des temps. De plus, on considère le chiffrage constant sur un intervalle donné que l'on peut étendre par simplification à la durée totale de l'extrait. Le principe général est illustré à la figure 2.12.

10. Cette probabilité a été obtenue par Klapuri en calculant l'histogramme des tactus de centaines de morceaux de différents genres annoté en tactus. Elle est illustrée à la figure 2.4.

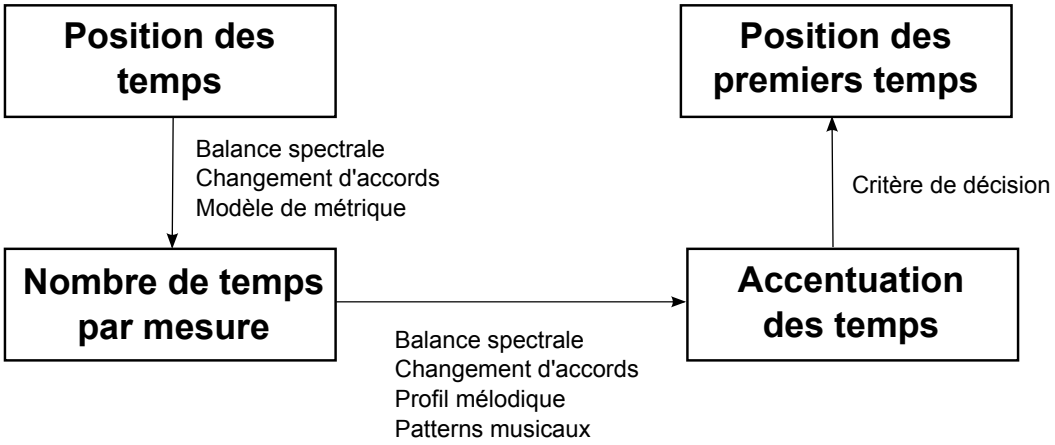


FIGURE 2.12 – Schéma général de la méthode d'estimation de la position des premiers temps à partir de la position des temps.

2.3.1 Changements d'accords - Ac

L'hypothèse ici est qu'un changement d'accord à plus de chance de se produire autour du premier temps qu'autour d'un autre temps de la mesure. Cette hypothèse est utilisée par plusieurs auteurs comme Goto ou Peeters afin d'estimer les premiers temps [Got01] [PP11]. Fujishima a été l'un des premiers à proposer une méthode efficace pour estimer en temps réel les changements d'accords à partir du contenu audio [Fuj99]. Il s'agit donc d'estimer les instants de changement d'accord, et de considérer le temps le plus proche comme ayant plus de chance d'être un premier temps. Nous reprenons ici pour l'estimation des instants de changements d'accord la méthode proposée par Oudre [OGF09].

Lorsqu'un changement d'accord se produit, les contenus harmoniques antérieurs et postérieurs à ce changement sont assez différents. L'estimation de changement d'accord peut alors se calculer rapidement à l'aide d'un chromagramme, qui représente l'évolution des chromas en fonction du temps. Les vecteurs de chroma sont traditionnellement des vecteurs à 12 dimensions où chaque dimension correspond à l'intensité associée à l'une des 12 classes de hauteur de la gamme chromatique de la musique tonale occidentale.

Le chromagramme est calculé comme le propose Bello à l'aide de la transformée à Q constant X_{Qc} [BP05]. Un calcul efficace de la transformée à Q constant a été proposé par Brown en 1991 pour des applications aux signaux musicaux [Bro91]. Cette transformée permet d'espacer les canaux fréquentiels lors de l'analyse spectrale de manière logarithmique, à la manière de l'oreille humaine :

$$X_{Qc}(k) = \sum_{n=0}^{N(k)-1} w(n,k)x(n)e^{-j2\pi f_k n} \tag{2.11}$$

Avec w la fenêtre d'analyse de longueur N dépendant de la position k du canal, x le signal temporel et f_k la fréquence centrale du $k^{\text{ème}}$ canal définie de la façon suivante :

$$f_k = 2^{\frac{k}{\beta}} f_{min} \tag{2.12}$$

Avec β le nombre de canaux par octave et f_{min} la fréquence minimale de l'analyse fréquentielle.

La transformée à Q constant permet d'espacer les harmoniques d'une note de façon constante et de calculer les chroma simplement :

$$Chroma(b) = \sum_{m=0}^{M} |X_{Qc}(b + m\beta)| \tag{2.13}$$

Avec b le numéro du chroma et M le nombre d'octaves de l'analyse fréquentielle. Le chromagramme est calculé sur 36 canaux et un filtre médian vient lisser les résultats. Il est enfin réduit à 12 canaux. La figure 2.13 vient illustrer les étapes d'obtention du chromagramme.

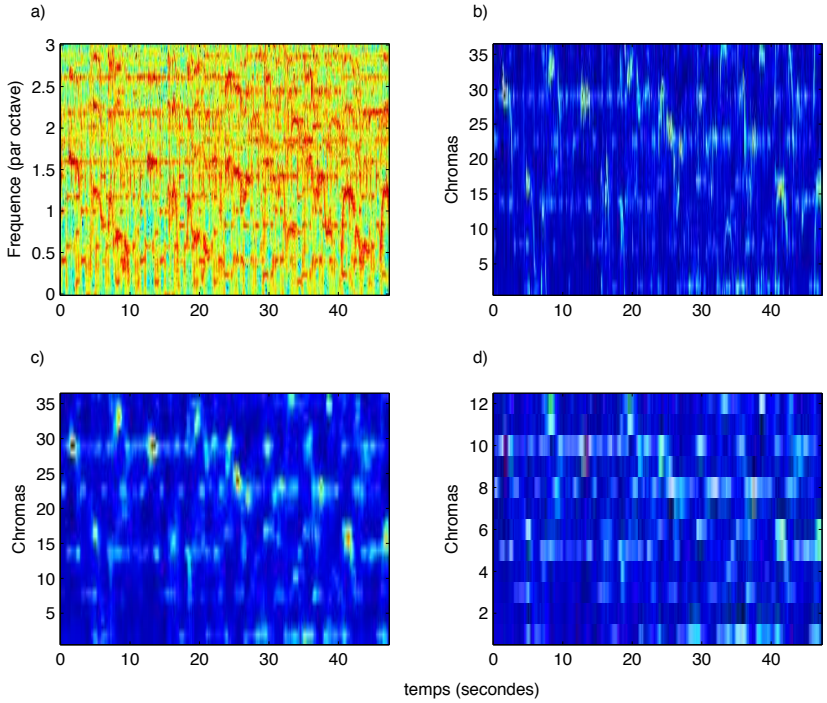


FIGURE 2.13 – Illustration de quatre étapes permettant l'obtention d'un chromagramme. Le temps est en abscisse des 4 figures. a) La transformée à Q constant. b) Les chromas à 36 dimensions obtenus. c) Filtrage médian du chromagramme. d) Chromagramme à 12 dimensions liées aux 12 notes de la gamme chromatique. Sur un signal extrait de *13 Bar Blues* de *Pelle Miljoona* de la base de données.

Une fois le chromagramme calculé, on compare la répartition harmonique à chaque instant à celle obtenue dans le cas d'accords précis. Cette comparaison est effectuée par la divergence Itakura Saito D_{IS}.

$$D_{IS}(x|y) = \sum_i \frac{x_i}{y_i} - \log\left(\frac{x_i}{y_i}\right) - 1 \tag{2.14}$$

On considère uniquement tous les accords parfaits majeurs et mineurs. Il vaut mieux ne pas estimer tous les changements d'accord mais obtenir des résultats fiables car on ne cherche pas à retrouver tous les accords mais uniquement une tendance permettant d'obtenir la position des premiers temps. L'accord $A(t)$ associé à chaque instant t est celui dont la distance avec le chromagramme $C(t)$ est la plus faible.

$$A(t) = \arg\min_A \{D_{IS}(C(t)|d_A)\} \tag{2.15}$$

Avec d_A le dictionnaire contenant la répartition spectrale des différents accords mappée sur 12 canaux à la manière du chromagramme. Il contient pour chaque accord la fondamentale, la tierce et la quinte ainsi que les deux premières harmoniques de ces 3 notes, d'amplitude décroissante d'un facteur 0.6 pour chaque harmonique. La figure 2.14 illustre la représentation de l'accord de do majeur dans le dictionnaire pour un nombre croissant d'harmoniques. La distance entre le chromagramme et le dictionnaire d'accords est lissée en fonction du temps par un filtre moyenneur d'une durée de 2 fois le tempo afin de ne pas permettre des changements d'accords trop rapides.

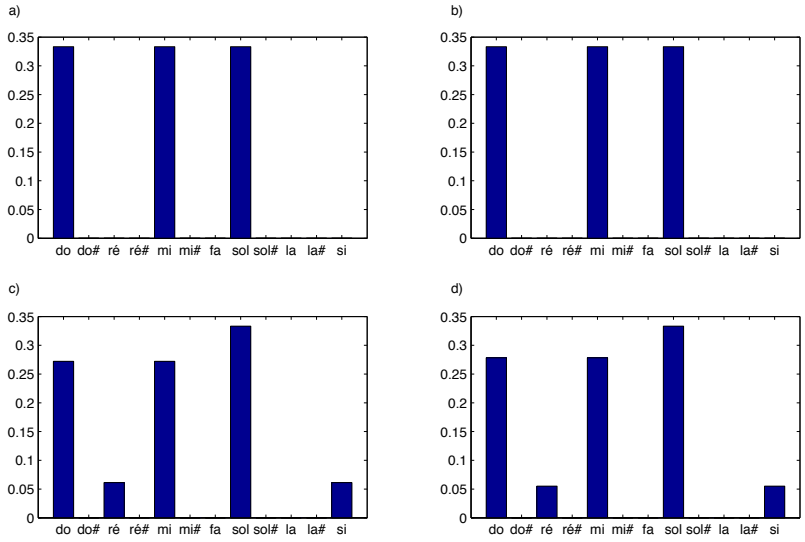


FIGURE 2.14 – Représentation de l'accord de do majeur dans le dictionnaire d'accords pour l'estimation des accords en fonction du temps. Le nom des notes est en abscisse et l'intensité spectrale normalisée est en ordonnée. Représentation pour a) 1 harmonique, b) 2 harmoniques, c) 3 harmoniques, d) 4 harmoniques

Une fois qu'un accord est associé à chaque instant. Les instants de changements d'accord sont les instants ou deux accords différents s'enchainent. Les deux accords doivent avoir une durée supérieure à une durée minimale. Cette durée minimale est fixée à deux fois le tempo.

L'utilisation d'un modèle de Markov caché qui permet de prendre en compte l'aspect temporel de la musique avec les probabilités de transitions pourrait être une variation à l'utilisation d'un filtre moyen et au choix d'une durée minimale pour un accord. Cependant, la durée de calcul est ici plus faible. Une limitation des chromas est la présence de sons inharmoniques. Les notes étant mappées dans les 12 dimensions du chromagramme, si une des composantes du spectre de la note n'est pas un multiple de la fondamentale, on obtiendra une note qui ne devrait pas être présente, comme l'a remarqué Papadopoulos dans sa thèse [Pap10].

2.3.2 Balance harmonique - BH

Goto a été l'un des premiers à lier l'estimation de la position des temps avec l'alternance de coups de grosse caisse et de caisse claire [GM94]. Les batteurs ont tendance à marquer les premiers et troisièmes temps avec un coup de grosse caisse (basse fréquence) et

les deuxièmes et quatrièmes temps avec un coup de caisse claire (haute fréquence) lorsque les mesures sont composées de quatre temps dans la musique populaire. Ainsi, les temps ont plus de chance de se trouver sur les attaques présentant une alternance entre les basses et les hautes fréquences. Cette oscillation entre les basses et les hautes fréquences est appelée balance harmonique. Peeters et Papadopoulos ont appliqué ce concept, illustré à la figure 2.15, à la détection des premiers temps en le couplant avec un détecteur de changement d'accords [PP11]. Tsunoo a utilisé une estimation conjointe des patterns de batterie et de la position des premiers temps par programmation dynamique et algorithme des k-moyennes [11] [TTOS11].

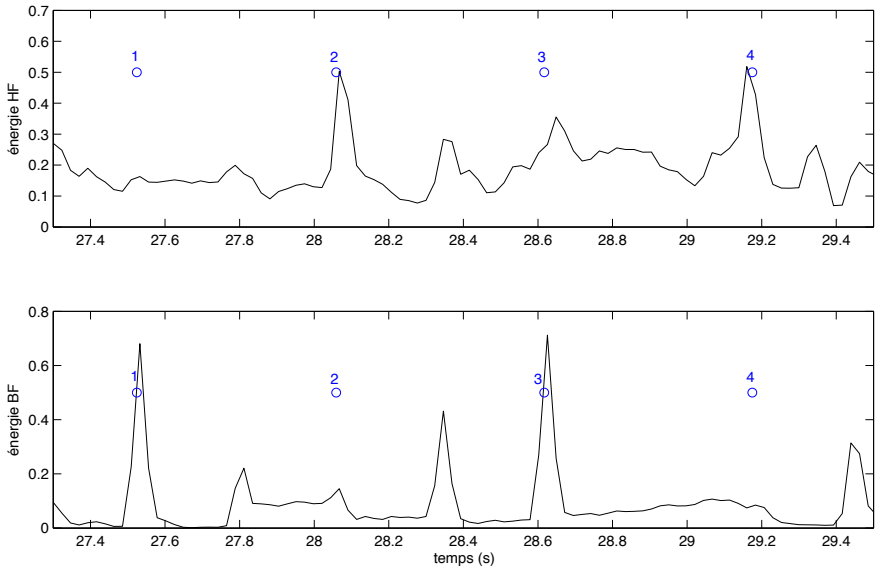


FIGURE 2.15 – Répartition de l'énergie produite par la caisse claire (en haut) et la grosse caisse (en bas) pour une mesure du morceau Alone des Bee Gees, disponible sur la base de données accompagnant l'article [KEA06]. La position des temps est marquée par les ronds bleus et leur numéro dans la mesure par les indices {1,2,3,4}. On observe une accentuation des temps 2 et 4 par la caisse claire et des temps 1 et 3 par la grosse caisse.

Une première remarque est que ce descripteur de premiers temps est moins généralisable que les autres. En effet, les changements de patterns musicaux, d'accords ou le profil mélodique sont vérifiés dans la plupart des genres de musique occidentale mais la balance basse fréquence/haute fréquence est surtout vérifiée pour les morceaux contenant une batterie ou un instrument proche, pour les mesures à 4 temps et pour certains genres musicaux. Néanmoins, une grande partie de la musique populaire suit ce schéma et il peut être intéressant d'en tenir compte. Nous essaierons de contourner ce problème en établissant un critère de décision pour tenir compte ou non de ce descripteur dans le morceau étudié. Ainsi, on procède à l'obtention du descripteur de balance harmonique, puis à une classification de son suivi ou non par le morceau. Enfin, dans le cas échéant on analyse cette

11. L'algorithme des k-moyennes proposé par MacQueen en 1967 est une méthode utilisée pour partitionner un ensemble de données en k groupes. On commence par sélectionner k centres de gravité initiaux de regroupement de données (clusters) et on affine la position de ces centres de gravité par itération de la façon suivante : 1) Chaque donnée est assignée à la partition la plus proche. 2) Chaque centre de gravité d'une partition est mis à jour pour être la moyenne des données qui lui sont attribuées [M$^+$67].

balance harmonique afin d'obtenir des informations sur la position des premiers temps. La figure 2.16 illustre le fonctionnement général de cette méthode.

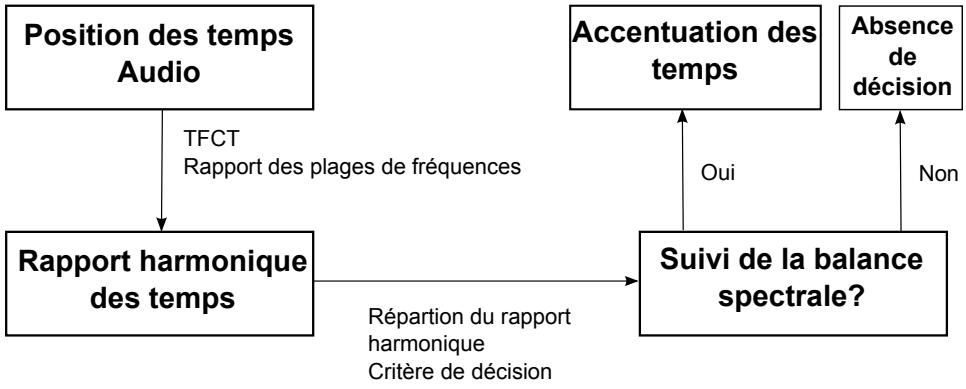


FIGURE 2.16 – Fonctionnement général de la méthode de la balance harmonique utilisée dans ce stage.

La première étape est le calcul du spectrogramme $S(w,t) = |STFT(w,t)|^2$ avec $STFT$ la transformée de Fourier à court terme du signal d'axe temporel t et fréquentiel w. Le deuxième étape est établissement d'un rapport entre l'énergie liée aux coups de caisse claire et celle liée aux coups de grosse caisse. La dernière étape est l'analyse des résultats obtenus afin de prendre une décision ou non. Plusieurs façons d'effectuer les deux dernières étapes sont présentées ici et testées par la suite.

Calcul du rapport harmonique

Le rapport harmonique nous intéressant est la partie du spectrogramme au niveau des temps (1), liée à la caisse claire et/ou la grosse caisse (2), sur une plage de fréquence donnée (3) et sous forme de rapport général ou local (4), absolu ou relatif (5). Chacun des cinq points suivants est présenté dans cette partie.

Extraire les valeurs du spectrogramme au niveau des temps n'est pas trivial pour conserver une certaine robustesse car les temps estimés ne coïncident pas toujours exactement aux coups de grosse caisse ou de caisse claire lorsqu'il y en a. Peeters a proposé de sommer les contributions sur un intervalle de la durée d'un demi tactus autour du temps estimé [PP11]. Nous testerons également une méthode plus restrictive en considérant non pas la somme mais le maximum sur un intervalle de temps de la taille d'un dixième du tactus, afin de considérer les attaques brèves et précises liées à la grosse caisse et à la caisse claire.

La balance harmonique part du principe que les coups de grosse caisse et de caisse claire sont importants et que c'est le balancement de l'un à l'autre qui va être caractéristique. Cependant, cet enchainement est peut être trop précis. Considérons également la prise en compte unique des coups de grosse caisse ou de caisse claire. D'après Goto, les coups de caisse claire sur les deuxièmes et quatrièmes temps sont plus généralisés que les coups de grosse caisse qui peuvent être syncopés [GM99]. De plus, cela permettrait de ne pas avoir à prendre en compte deux phénomènes opposés et donc de ne pas avoir à utiliser un rapport qui est souvent instable par l'effet du dénominateur.

La plage de fréquence utilisée par Peeters dans [PP11] pour extraire l'énergie des coups de caisse claire va de 150 Hz à $\frac{Fe}{4}$ Hz avec Fe la fréquence d'échantillonnage. Cette formulation n'est peut être pas tout à fait adaptée à l'extraction de l'énergie produite par un coup de caisse claire qui se situe bien au delà de 150 Hz. Goto propose un intervalle qui va de 1400 à 7500 Hz qui sera également testé comme le montre la figure 2.17 et permet de ne pas considérer l'intervalle entre 150 et 1400 Hz qui pourrait perturber les résultats [Got01].

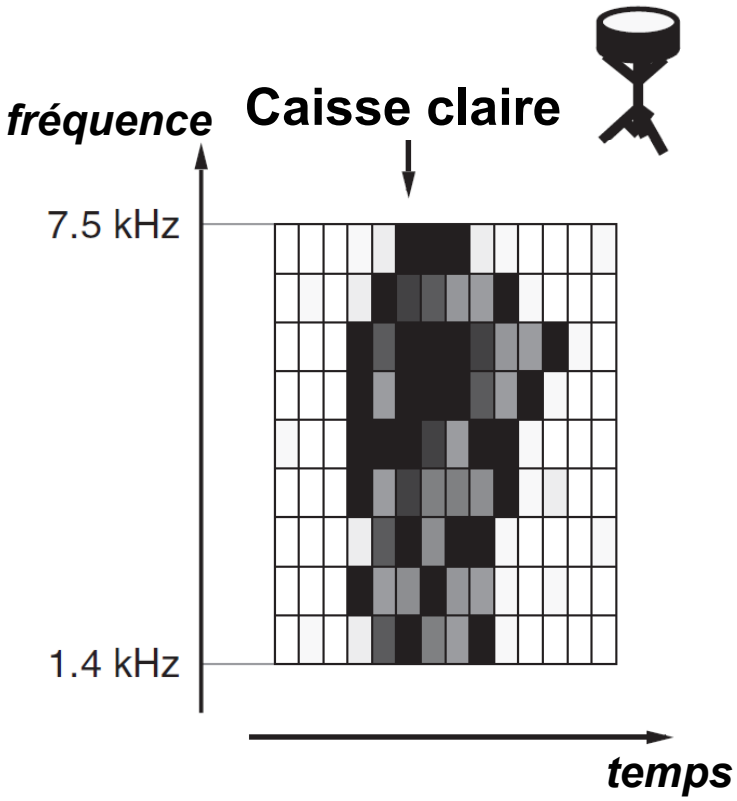


FIGURE 2.17 – Répartition de l'énergie produite par un coup de caisse claire. D'après [Got01].

Le point suivant est lié aux faiblesses intrinsèques du calcul d'un rapport de deux composantes. Si l'énergie basse fréquence autour de la position d'un temps est très faible, sa position au dénominateur du rapport harmonique va donner un résultat très élevé qui peut détériorer les mesures. Par exemple, s'il n'y a pas de coup de grosse caisse sur un premier ou un troisième temps et que l'énergie basse fréquence est également faible à un moment donné, le calcul général de la balance harmonique risque d'être assez perturbé. Ainsi, il est intéressant de regrouper toute l'énergie liée aux coups de caisse claire, puis celle liée aux coups de grosse caisse et enfin d'effectuer le rapport des quantités totales obtenues, après avoir normalisé le signal localement par un filtre médian.

Le dernier point concerne la manière de concevoir un balancement harmonique. Si l'on procède de façon absolue comme c'est le cas généralement, le résultat final sera la somme des résultats intermédiaires autour de chaque temps. Comme pour le point précédent, une valeur très élevée est gênante. De plus, les valeurs intermédiaires seront hétérogènes car le contenu musical varie avec le temps. Ainsi, on considèrera également le balancement relatif

entre 3 temps consécutifs avec le système de poids suivant :

$$\begin{aligned} rh_{rel}(b) &= 1 \quad \text{si} \quad rh_{abs}(b) > rh_{abs}(b \pm 1) \\ rh_{rel}(b) &= 0 \quad \text{si} \quad rh_{abs}(b) < rh_{abs}(b \pm 1) \\ rh_{rel}(b) &= 0.5 \quad \text{sinon} \end{aligned} \tag{2.16}$$

Avec $rh_{abs}(b)$ et $rh_{rel}(b)$ les valeurs absolues et relatives du rapport harmonique autour du temps b. Cela revient à procéder à de l'estimation de pics [12].

Pour résumer, le descripteur de la balance harmonique peut être défini suivant les choix considérés de l'une des deux façons suivantes :

$$BH(k) = \left(\sum_{b \in k} e_{snare}(b) \right)^{p_{snare}} \left(\sum_{b \in k} e_{kick}(b) \right)^{p_{kick}} \tag{2.17}$$

ou

$$BH(k) = \sum_{b \in k} rh(b) \tag{2.18}$$

Avec $BH(k)$ la balance harmonique liée au temps $k = \{1, 2, 3, 4\}$ d'une mesure théorique à quatre temps, b un des temps annotés, $p_{snare} = \{0, 1\}$ la présence ou non de la caisse claire, $p_{kick} = \{-1, 0, 1\}$ la présence ou non de la grosse caisse, $e_{snare} = \{e_{abs_{snare}}, e_{rel_{snare}}\}$ et $e_{kick} = \{e_{abs_{kick}}, e_{rel_{kick}}\}$ les énergies produites par les coups de caisse claire et de grosse caisse de façon absolue ou relative :

$$e_{abs_{type}}(b) = M_t \left(\sum_{i=B_{type}}^{H_{type}} S(w_i, t) \right), \quad t \in [b - L, b + L] \tag{2.19}$$

Avec $type = \{snare, kick\}$ la caisse claire ou la grosse caisse, $M = \{\max, \sum\}$ le type de mesure utilisé, B et H les limites basses et hautes de l'intervalle de fréquence considéré, S le spectrogramme et L la longueur de l'intervalle temporel autour du temps.

- $e_{rel_{type}}(b) = 1$ si $e_{abs_{type}}(b) > e_{abs_{type}}(b \pm 1)$
- $e_{reltype}(b) = 0$ si $e_{abs_{type}}(b) < e_{abs_{type}}(b \pm 1)$
- $e_{reltype}(b) = 0.5$ sinon

Et enfin $rh(b) = \{rh_{abs}(b), rh_{rel}(b)\}$ les rapports harmoniques absolus ou relatifs, avec :

$$rh_{abs}(b) = \left[\frac{e_{snare}(b)}{e_{kick}(b)} \right] \tag{2.20}$$

et rh_{rel} défini à l'équation 2.16.

Suivi ou non de la balance harmonique

Si un morceau suit le schéma basse fréquence sur les premiers et troisièmes temps et haute fréquence sur les deuxièmes et quatrièmes temps, le descripteur de balance harmonique contiendra des valeurs élevées un temps sur deux [13]. Dans le cas contraire, il y a

12. Ou du peak picking en anglais
13. Par exemple, $BH = \{a, b, c, d\}$ avec a et c supérieurs à b et d

plus de chance que ses valeurs soit réparties autrement. De plus, si on morceau suit cette balance harmonique, il est fort probable qu'une mesure contienne 4 temps, moyennement probable qu'une mesure contienne 2 ou 8 temps et improbable qu'une mesure ne contienne pas un multiple de 2 temps. Cette notion sera utile par la suite pour établir le chiffrage du morceau et choisir les premiers temps.

Pour obtenir le critère de suivi de la balance harmonique, les quatre valeurs de BH sont comparées de trois façons différentes.

La première façon est la plus contraignante : on veut que l'énergie du signal soit similaire lors des coups de caisse claire et qu'elle soit différente entre les coups de caisse claire et de grosse caisse. Cela se traduit par ce qui suit. Si deux temps non adjacents sont α fois supérieurs à chacun des deux autres temps et qu'ils sont moins de β fois différents l'un de l'autre, alors on estime que la balance harmonique est suivie [14].

La deuxième façon prend en compte uniquement les différences d'énergie et non plus les similarités. Si deux temps non adjacents sont α fois supérieurs à chacun des deux autres temps alors on estime que la balance harmonique est suivie.

La dernière comparaison est moins précise mais plus générale. On considère uniquement le maximum de deux temps non adjacents qui doit être α fois supérieur au maximum des deux autres temps.

Résumé des choix effectués

De nombreuses façons de calculer le rapport harmonique ont été présentées, voici le résumé des tests effectués :

- découpler le rapport entre les basses et les hautes fréquences améliore grandement les résultats. Cela permet de ne pas être perturbé par une absence locale d'énergie basse fréquence et d'obtenir un résultat plus caractéristique de l'extrait global,
- l'intervalle de fréquence de la caisse claire tel que le présente Goto améliore les résultats de façon significative. On peut supposer alors que l'énergie situé entre 150 et 1400 Hz n'appartenant pas à la plage de fréquence de la caisse claire va perturber les résultats,
- cependant, cet intervalle de la caisse claire est vaste et contient de nombreuses autres contributions comme le chant, la mélodie etc. L'information contenu uniquement à ce niveau de fréquence n'est pas suffisante et il faut ajouter des informations basse fréquence. On peut même la considérer perturbatrice car la prise en compte unique de l'énergie sur les plages de fréquences de la grosse caisse donne des résultats équivalents à la définition plus classique de la balance harmonique. Il faudrait donc travailler à une extraction plus précise de l'énergie produite par la caisse claire car l'écoute des extraits audio montre bien sa présence majoritaire sur les deuxièmes et quatrième temps,
- les autres paramètres présentent une influence non significative.

14. On aurait pu procéder à une classification du descripteur de balance harmonique sur les quatre temps, ainsi qu'à la détection de la présence ou non de la batterie dans le signal musical mais cela n'a pas pu être effectué dans le temps imparti du stage.

Le critère de suivi le moins restrictif a été gardé [15] car même si l'utilisation de critères plus restrictifs améliore le résultat final, la paramétrisation doit être précise et semble moins généralisable à d'autres bases de données. Cela mérite une vérification.

2.3.3 Profil mélodique - Melo

Lier le profil mélodique aux premiers temps suppose que certaines notes sont plus accentuées que d'autres et permettent de percevoir du rythme, notamment les premiers temps du morceau. Nous entendons par profil mélodique les indices aussi bien mélodiques que temporels des notes.

Comment définir l'accentuation des notes ? Pfordresher utilise des tests perceptifs. Il demande à différents auditeurs de taper dans leurs mains à l'instant où une note leur parait accentuée [Pfo03]. D'autres auteurs comme Bigand peuvent demander de taper pour se synchroniser à une pulsation régulière [Big00]. Néanmoins, des difficultés intrinsèques au caractère fuyant de la musique est qu'une fois qu'un accent mélodique est détecté, il a disparu. Il est alors délicat d'obtenir une définition de l'accent mélodique partagée par tout le monde, mais des constantes reviennent. On peut considérer dans un premier temps l'accent lié au contour de la mélodie. Lorsque l'on passe d'une mélodie ascendante à une mélodie descendante ou inversement, la note précédent ce changement de direction ou la note suivante est considérée accentuée. Un accent peut également être lié à une pause. Les notes suivant et précédant la pause seront considérées accentuées. La figure 2.18 illustre les exemples précédents.

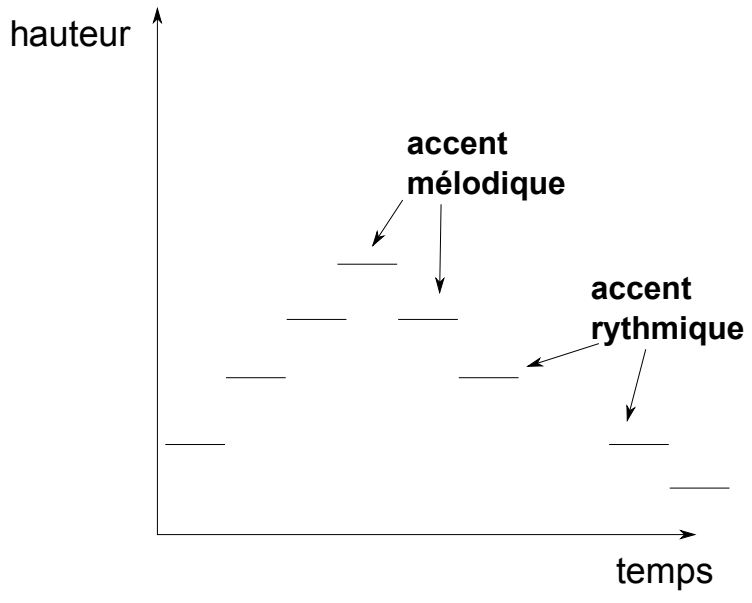


FIGURE 2.18 – Accents mélodiques et rythmiques sur un enchainement de notes représentées dans un plan {temps,fréquence}. La note pivot est la note la plus haute sur de la figure.

Selon Hannon, les indices mélodiques ne sont pas suffisants pour prédire le rythme car

15. Le maximum de deux temps non adjacents doit être $\alpha = 1.1$ fois supérieur au maximum du reste

ils se contredisent parfois [16]. Cependant l'utilisation d'indices mélodiques et rythmiques des notes améliore les résultats à l'utilisation seule d'indice rythmiques [HSEK04]. Nous prendrons donc en compte des indices rythmiques et mélodiques des notes. De plus, Hannon suggère que l'utilisation de nombreux indices sur le profil mélodique permet d'obtenir une estimation fiable de la métrique du morceau grâce à un critère de décision [17].

Nous procéderons de la façon suivante. La mélodie sera estimée à partir du signal audio. Son profil mélodique sera calculé. Enfin, un critère de décision permettra de prendre en compte le profil mélodique ou non, comme le montre la figure 2.19.

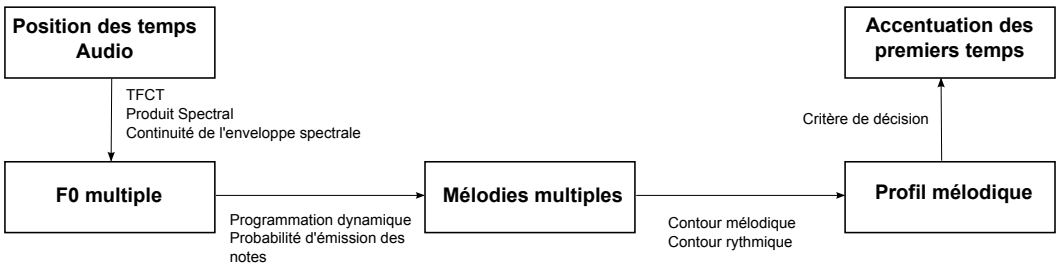


FIGURE 2.19 – Fonctionnement général de la méthode du profil mélodique utilisée dans ce stage.

Estimation de la mélodie :

On estime dans un premier temps les fréquences fondamentales multiples du signal audio. L'approche proposée est inspirée par les travaux de Klapuri [Kla03] et [Kla01]. Cette approche simple consiste à estimer dans un premier temps la fréquence fondamentale dominante à chaque instant, à en déduire la position et l'amplitude de ses harmoniques sur le spectre puis à supprimer les harmoniques avec l'application du principe de la continuité de l'enveloppe spectrale [18]. Cette étape est ensuite itérée afin d'extraire l'ensemble des fréquences fondamentales du son initial. Le produit spectral est calculé pour $H = 4$ versions compressées.

La mélodie est ensuite estimée en utilisant la programmation dynamique. Les notes sont définies par leur numéro MIDI. La probabilité d'émission de chacune des notes est liée à la valeur de son produit spectral. La probabilité de transition est représentée par une gaussienne centrée sur le numéro MIDI de la note et d'écart type $\sigma = 5$ permettant de décrire une tendance de faible variation de la hauteur des notes. Aucun état initial n'est privilégié. On utilise un algorithme de type Viterbi pour déterminer le contour optimal de la mélodie. Une fois qu'un chemin a été trouvé. Toutes les notes par lequel il est passé ainsi les deux adjacentes sont supprimées [19] et un nouveau chemin est estimé. On extrait ainsi trois mélodies.

16. Sauf si les notes sont isochrones.
17. Quand les indices concordent la bonne décision est prise et quand les indices sont contradictoires aucune décision n'est prise. C'est également le résultat de tests perceptifs qui ont montré que les auditeurs avaient tendance à ignorer les indices mélodiques et temporels lorsque ceux-ci étaient contradictoires.
18. Ou spectral smoothness.
19. Leur probabilité d'émission est mise à zéro.

Un post-traitement est utilisé. Le silence et la répétition de notes identiques ne sont pas considérés. Chaque note dure jusqu'à l'apparition de la suivante. Si une note à une durée inférieure à un quart du tempo, elle est supprimée. Si la note qui l'a suit est identique à la note qui l'a précède, ces deux notes sont alors liées.

Calcul du profil mélodique :

Le calcul du profil mélodique est obtenu selon deux critères bien qu'il aurait été intéressant d'en ajouter d'autres : le contour mélodique $M(n)$ et rythmique $R(n)$ des notes n. Le contour mélodique est obtenu d'après le travail de Thomassen qui attribue des poids aux deux dernières notes d'un trio en fonction de leur hauteur relative comme le montre la figure 2.20 [Tho82]. Le contour rythmique des notes est obtenu en attribuant des poids aux notes en fonction de leur durée. Plus une note est longue, plus son poids va être élevé via une relation quadratique :

$$R(n) = \left(\frac{t_{n+1} - t_n}{\max(t_{i+1} - t_i)} \right)^2, \quad i = [1..N-1] \tag{2.21}$$

Avec t_n l'instant d'apparition de la note n et N le nombre de notes. Les contributions rythmiques et mélodiques sont multipliées l'une avec l'autre pour obtenir le profil mélodique P de la note : $P(n) = M(n)R(n)$.

C(n+1)	C(n+2)	M(n+1)	M(n+2)
= 0	= 0	0.00	0.00
≠0	= 0	1.00	0.00
= 0	≠0	0.00	1.00
> 0	< 0	0.83	0.17
< 0	> 0	0.71	0.29
> 0	> 0	0.33	0.67
< 0	< 0	0.50	0.50

FIGURE 2.20 – Poids $M(n+1)$ et $M(n+2)$ des notes $n+1$ et $n+2$ d'un trio de notes $\{n, n+1, n+2\}$ établi par Thomassen afin de définir un accent mélodique. Le poids est obtenu en fonction de $C(n+1)$ la position relative de la note $n+1$ avec la note n. À la quatrième ligne, on a $C(n+1) > 0$ et $C(n+2) < 0$ ce qui signifie que la mélodie formée des notes n, $n+1$ et $n+2$ est ascendante puis descendante. Dans le cas, on voit que le poids de la note $n+1$ est de 0.83 et celui de la note $n+2$ est de 0.17. D'après [Tho82].

Critère de décision :

Le poids relatif à chaque temps est calculé et sommé tous les N temps, avec N obtenu avec l'estimation du chiffrage du morceau. Le temps accentué par le profil mélodique sera conservé si son poids est α fois supérieur au maximum des autres temps et supérieur à un seuil s. Le paramètre α permet de s'assurer que l'estimation est cohérente et le paramètre s qu'elle se base sur une quantité d'information suffisante.

2.3.4 Distance entre patterns musicaux - P

Un pattern musical est défini ici comme une partie répétitive ou non d'un morceau caractérisée par son ambiance musicale. Le refrain et le couplet peuvent être deux patterns différents mais cela peut être plus précis avec par exemple l'arrivée d'un nouvel instrument, un long silence ou un changement d'orchestration. L'hypothèse ici est que l'enchainement entre deux patterns a plus de chance de s'effectuer au niveau d'un premier temps qu'ailleurs dans la mesure.

La difficulté est de trouver l'instant où deux patterns s'enchainent. Cela peut être fait de manière absolue mais également de manière relative en comparant les enchainements entre les patterns pour différents placements de premiers temps et en gardant les premiers temps qui permettent l'enchainement le plus propre possible.

Nous présenterons trois méthodes. La première estime les patterns comme on peut le faire pour obtenir le résumé audio d'une chanson. La deuxième identifie les instants de façon absolue où le contenu musical du morceau change de façon significative. La dernière méthode est relative et compare différents enchainements de premiers temps pour choisir l'enchainement le plus adéquat. La première méthode est largement inspirée de la littérature, même si l'objectif est différent tandis que les deux suivantes sont plus originales.

Le point commun de ces méthodes est qu'elles utilisent la notion de ressemblance ou de similarité entre des contenus musicaux. Nous utiliserons pour cela une matrice de similarité, ou d'auto similarité S du morceau, issue des travaux de Foote en 1999 [Foo99]. Chaque coefficient $s_{ij} = s(t_i, t_j)$ représentera la similarité entre les instants t_i et t_j du signal. On notera $\underline{\underline{S}} = s_{ij} \quad \forall i, j$. Une valeur élevée du coefficient s_{ij} représente une similarité importante entre les instants t_i et t_j. Si une séquence de temps t_i, t_{i+1}, t_{i+2}, ... est similaire à une séquence de temps t_j, t_{j+1}, t_{j+2}, ... nous observerons une diagonale supérieure et inférieure avec des coefficients élevés. De plus, si $t_i \simeq t_{i+1} \simeq t_{i+2}$... nous observerons un bloc. L'inspection de ces lignes et de ces blocs nous renseignera sur la structure du morceau.

La similarité est liée à la distance entre des observations faites du signal. Sa forme sera donc différente suivant le type de distance ou d'observation choisie. La distance choisie peut être euclidienne : $D_e(t_i, t_j) = \sqrt{\sum_k \left(t_k^i - t_k^j\right)^2}$ ou cosinusoidale : $D_c(t_i, t_j) = \frac{\sum_k (t_k^i . t_k^j)}{\sqrt{\sum_k (t_k^i)^2}\sqrt{\sum_k (t_k^j)^2}}$ entre autres. Les résultats dépendent de la mesure choisie. Par exemple, on ne tient pas compte de l'intensité du signal avec la mesure cosinusoidale. Les observations sont également variées mais nous nous concentrerons sur les coefficients cepstraux en échelle de Mel (MFCC), qui nous permettent d'estimer l'ambiance musicale ou le timbre du morceau de musique. Il sont obtenus en cinq étapes :

- Le calcul de la transformée en amplitude de Fourier du signal, -
- La conversion du signal en échelle de Mel $M = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$ [20], avec f la fréquence en Hertz avant conversion,
- Le passage en échelle logarithmique à chacune des fréquences de mel,
- Le calcul de la transformée en cosinus discrète inverse
- La sélection des coefficients du spectre résonant proches de zéro.

20. Il existe plusieurs formulations possibles.

Les coefficients obtenus seront les coefficients cepstraux en échelle de Mel que l'on recherche. Les trois méthodes ci-dessous permettent d'analyser la matrice de similarité de façon adaptée à notre problème d'estimation des premiers temps.

Classification des différentes parties musicales

Cette méthode est inspirée des travaux de Peeters et Cooper visant à obtenir une résumé audio d'une chanson. [PLBR02] et [CF02]. Il s'agit d'obtenir les différentes parties du morceau comme le couplet, le refrain et le pont dans la musique populaire par exemple, ainsi que l'enchainement temporel de ces parties. L'instant où deux parties vont s'enchainer aura une grande chance d'être un premier temps et sera assimilé comme tel.

La matrice de similarité $\underline{\underline{S}}$ est calculée à partir de la distance euclidienne entre 12 coefficients MFCC. Les instants de plus grande variation de $\underline{\underline{S}}$ obtenus en trois étapes permettant d'affiner à chaque fois le résultat. Le principe est repris de l'article de Peeters et quelques points sont présentés ici.

La première étape est une simple étude des variations de la diagonale supérieure de $\underline{\underline{S}}$. Les coefficients sur cette diagonale représente la similarité de deux instants consécutifs. Si la similarité est rompue, on observera un MFCC élevé. Ceci nous permet de repérer des segments où l'ambiance varie peu. Les MFCC sont moyennés sur ces segments. Deux segments dont la distance euclidienne est inférieure à un seuil sont regroupés ensemble. On obtient alors différentes classes représentant différentes ambiances du signal.

Ces classes sont envoyées en entrée un algorithme des k-moyennes qui permet de lier chaque instant du signal avec la classe la plus proche. Pour cela, sur N itérations :

- Les observations, sous forme de 12 MFCC à chaque instant, sont groupées à la classe la plus proche selon une distance euclidienne,
- Chaque classe est mise à jour en moyennant les observations qui lui sont associées.

Enfin, on utilise un modèle de Markov caché afin de tenir compte de l'aspect temporel de la musique et que chaque partie du morceau possède une durée non négligeable. Les états cachés sont les classes que l'on vient d'obtenir à la sortie de l'algorithme des k-moyennes. La probabilité d'être dans l'état E sachant l'observation O est égale à l'inverse de la distance entre l'observation et la l'état donnée. La matrice de transition est équivalente pour chaque état et pénalise les changements d'état. Les paramètres du modèle de Markov caché sont entrainés par un algorithme de Baum Welsh et la séquence d'état la plus probable est obtenue à l'aide d'un algorithme de Viterbi.

À chaque changement d'état, on enregistre le temps annoté le plus proche et connaissant le nombre de temps par mesure, le premier temps sera celui qui correspond le plus aux changements d'état. On peut noter que cette méthode ne sera pas gardée par la suite pour des raisons expliquées à la partie 3.3.2.

Changement absolu de pattern musical - P_{abs}

Cette deuxième méthode ne s'attache pas à retrouver les différentes parties d'un mor-

ceau mais uniquement à estimer les instants où l'ambiance musicale change. Pour cela on intègre une procédure P (que l'on décrira ci-dessous) permettant de trouver l'instant de changement entre deux patterns au schéma suivant :

- On initialise le temps t à zéro et on effectue les trois points suivants jusqu'à ce que t soit proche à la durée totale du morceau :
- On effectue P avec une précision temporelle faible, de t à la fin du signal,
- On choisit l'instant le plus significatif supérieur à une distance minimale,
- On effectue P avec une précision temporelle élevée, autour de l'instant estimé au point précédent
- On stocke l'instant le plus significatif et on met à jour t avec la valeur obtenue.

La procédure P est la suivante :

- On calcule la matrice de similarité $\underline{\underline{S}}$ en utilisant une distance euclidienne et 12 MFCC. Les MFCC sont obtenus sur des trames courtes si l'on veut une précision temporelle élevée et longue si l'on veut une précision temporelle faible,
- On considère que $\underline{\underline{S}}$ est proche d'une matrice par bloc et on souhaite identifier l'instant de changement entre le premier et le deuxième bloc. On note B_i le bloc de $\underline{\underline{S}}$ contenant toutes les interactions entre les instants antérieurs à l'instant i. $B_i = \underline{\underline{S}}(t_k, t_l)$, $\forall (k,l) \leq i$. On calcule la gaussienne g_i ayant la même moyenne et le même écart type que B_i.
- La première mesure de distance entre les patterns est la distance euclidienne entre g_i et g_{i+1}. La deuxième mesure est la distance euclidienne entre g_i et g_{Di} la gaussienne ayant la même moyenne et le même écart type que $B_{i+1} - B_i$. La deuxième mesure est plus précise temporellement que la première mais elle est également moins robuste car il suffit qu'un seul instant soit peu lié aux instants précédents pour qu'un changement soit détecté.
- On somme les deux mesures précédentes et on obtient une mesure de l'instant de changement de pattern.

On obtient alors en sortie les différents instants de changement de pattern que l'on traite de la même manière que dans la méthode précédente.

Cette méthode est plus efficace que la précédente car elle se concentre uniquement sur la distance entre deux patterns musicaux. Cependant, elle reste imprécise temporellement. Une solution à ce problème est de procéder de manière relative en comparant plusieurs segmentations du signal et en déterminant la meilleure segmentation. La méthode est présentée à la figure 2.21

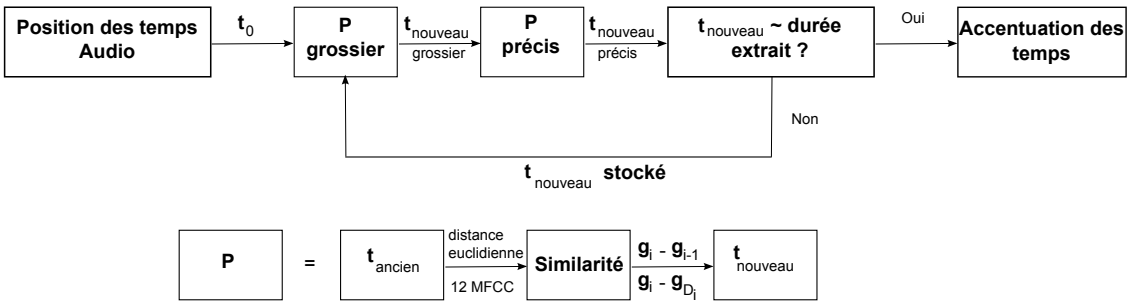


FIGURE 2.21 – Fonctionnement général de la méthode du changement de pattern absolu. t_0 est l'instant initial. Le terme "grossier" signifie imprécis. $t_{nouveau}$ est le nouvel instant de changement de pattern qui permet l'accentuation des premiers temps. Les explications sont fournies dans le corps du texte.

Changement relatif de pattern musical - P_{rel} et P_{discri}

Soit un morceau contenant des mesures à N temps et dont on connait la position des temps. Soit t_1 la position temporelle du temps initial [21]. I_1^1 est le signal contenu entre t_1 et t_{N+1}. Si l'on continue la segmentation, on obtient I_2^1 le signal contenu entre t_{N+1} et t_{2N+1} et I_i^1 le signal contenu entre $t_{(i-1)*N+1}$ et t_{i*N+1}. On peut effectuer la même procédure en partant du $j^{\text{ème}}$ temps annoté ou estimé avec $j = [1 : N]$ et l'on obtient I_i^j le signal contenu entre $t_{(i-1)*N+j}$ et t_{i*N+j} comme le montre la figure 2.22. Il s'agit ici de trouver j tel que I_i^j corresponde au signal contenu dans les mesures du morceau. On en déduit alors les premiers temps.

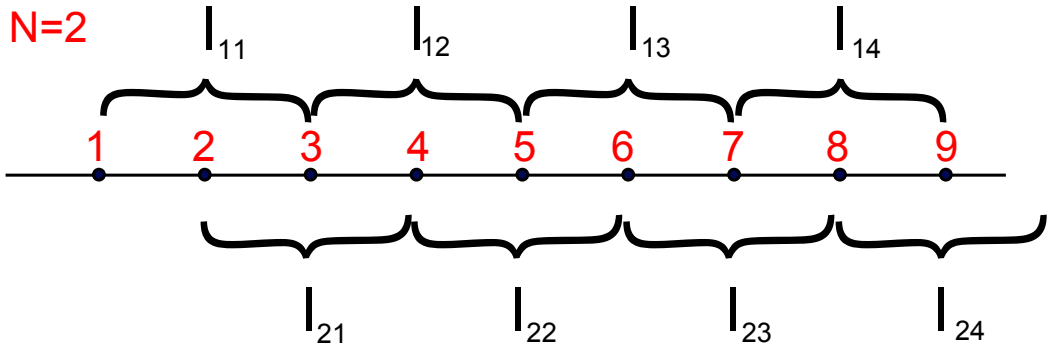


FIGURE 2.22 – Segmentation d'un morceau à $N = 2$ temps. Les chiffres en rouge correspondent au numéro du temps annoté ou estimé.

L'hypothèse est que si l'on prend j tel que I_i^j, $\forall i$ corresponde au signal contenu dans les mesures du morceau, l'enchainement entre les différentes parties du morceau sera plus marqué. Cela veut dire que I_i^j sera similaire à I_{i-1}^j et différent de I_{i+1}^j si I_{i+1}^j est la première mesure de la nouvelle partie du morceau. la figure 2.23 illustre cela sur un exemple concret.

21. Attention, ce n'est pas forcément un premier temps au sens musical du terme, uniquement le premier temps qui a été annoté ou estimé.

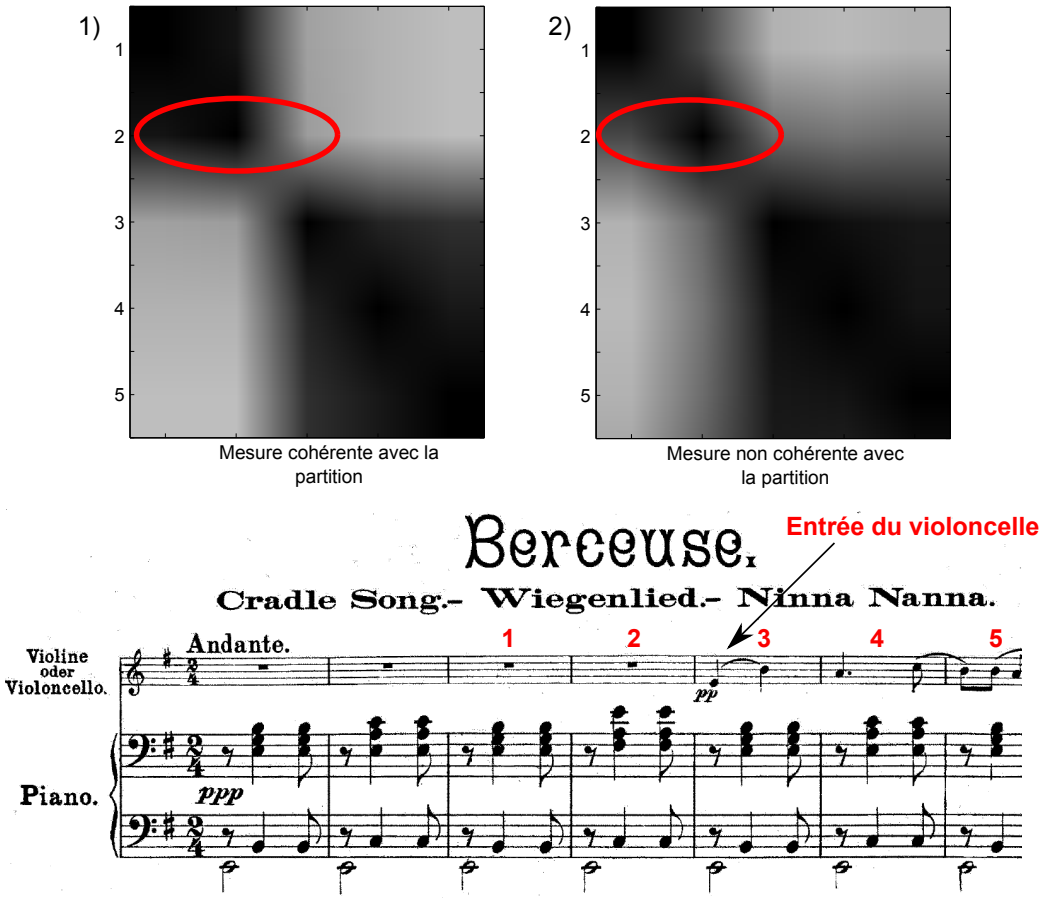


FIGURE 2.23 – *Berceuse* pour piano et violoncelle d'Armas Järnefelt, obtenue sur *imslp.org*. La partie supérieure contient la similarité entre 5 mesures du morceau, indiquées en rouge sur la partition. Des mesures similaires ont une teinte foncée l'une par rapport à l'autre et des mesures non similaires une teinte claire. Par exemple, dans le cas 1), les mesures 1 et 2 sont très similaires. Dans le cas 1) les mesures ont été correctement estimées et dans le cas 2) elles ont été décalées d'un temps. Quand le violoncelle arrive à la troisième mesure, le contenu musical change et la similarité entre les mesures 2 et 3 change radicalement dans le cas 1). Dans le cas 2), l'entrée du violoncelle se produit au milieu de la mesure 2 et la similarité avec la mesure suivante ne change pas radicalement. Pour les mêmes raisons, les mesures 1 et 2 sont plus similaires dans le cas 1) que dans le cas 2).

On commence là aussi par calculer la matrice de similarité $\underline{\underline{S}}^j$. Cette fois-ci, chaque observation est faite sur I_i^j. S_{kl}^j représente la similarité entre I_k^j et I_l^j. De la même manière que précédemment on trouve l'instant t où l'on change de pattern musical et l'on calcule le score de changement de pattern $Score_t^j = S_{t+1,t}^j - S_{t-1,t}^j$. Si j correspond à un premier temps, on a vu que $S_{t+1,t}^j$ est grand (car peu similaire) et $S_{t-1,t}^j$ est petit (car similaire). Le premier temps p est assimilé à l'indice j donnant le score le plus élevé : $p = \arg\max_j \left(\sum_t Score_t^j\right)$.

Cette méthode est intéressante car elle utilise toutes les informations dont on dispose. Elle évite également l'imprécision temporelle liée à l'utilisation de matrices de similarité et est robuste aux différents genres présents dans la base de données. Plusieurs améliorations sont possibles, aussi bien pour définir le type d'observation le plus adéquat à une mesure de similarité que pour calculer le score de changement de pattern. Elle est également utile pour discriminer le temps qui à le moins de chance d'être un premier temps. Le tableau 3.2 présenté à la partie suivante montrera que cette discrimination est performante. La méthode est illustrée par la figure 2.24.

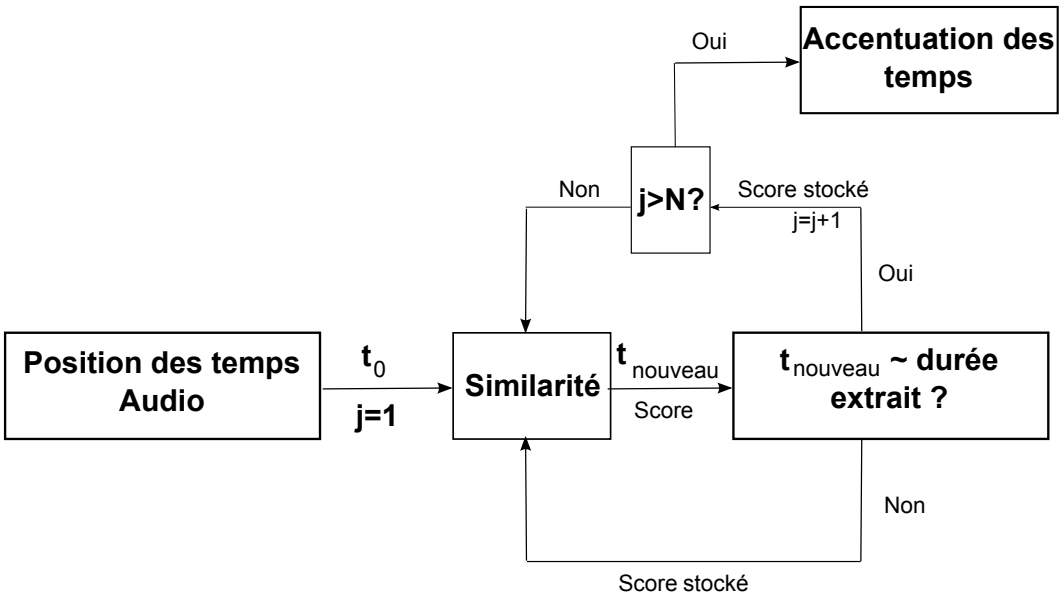


FIGURE 2.24 – Fonctionnement général de la méthode du changement de pattern relatif. t_0 est l'instant initial. Les explications sont fournies dans le corps du texte. N est le nombre de temps dans une mesure. $t_{nouveau}$ est l'instant de changement de pattern relatif.

2.3.5 Des descripteurs précédents à l'estimation du premier temps

Afin d'estimer les premiers temps, on estime préalablement le chiffrage [22], on calcule les descripteurs présentés ci-dessus pour un chiffrage donné et on conclue grâce à un critère de décision par agents multiples.

22. Par soucis de simplicité, le chiffrage est assimilé par la suite au nombre de temps par mesure.

Estimation du chiffrage

On considère qu'une mesure peut contenir N temps, avec $N = \{2, 3, 4\}$. On fait une hypothèse de chiffrage constant sur l'extrait. S_N sera la probabilité que l'extrait contienne des mesures à N temps. On conclue sur le chiffrage avec l'argument du maximum de S_N. Le calcul se déroule en deux étapes.

Premièrement, on calcule le descripteur lié à la balance harmonique. Si la condition de suivi de la balance harmonique est remplie, un associe un poids P_N à chaque chiffrage. P_n est associé à la probabilité qu'un extrait suivant une balance harmonique contienne des mesures à N temps. P_4 est élevé, P_2 est faible et P_3 est très faible.

Deuxièmement, on estime les instants de changement d'accord et on considère les temps n les plus proches de chaque instant de changement d'accord. On appelle $T(n)$ le vecteur obtenu. $T(n) = 1$ si le temps n est le plus proche d'un instant de changement d'accord. $T(n) = 0$ sinon. On multiplie T par une fonction peigne F_{iN} représentant les différentes configurations $i = \{1 : N\}$ d'une mesure à N temps. $F_{iN}(n) = 1$ si $n = i + kN$ avec k entier [23]. L'intérêt est que pour chaque temps n, $F_{iN}(n) = 1$ si n est un premier temps et $F_{iN}(n) = 0$ sinon. On normalise le résultat suivant la somme de $F_{iN}(n)$ et on garde le maximum selon i, ce qui correspond à la configuration de premiers temps la plus probable pour un chiffrage donné. En multipliant ce résultat pour le poids P_n obtenue précédemment, on obtient S_N :

$$S_N = max_i \left(\frac{\sum_n F_{iN}(n)T(n)}{\sum_n F_{iN}(n)} \right) P_N \tag{2.22}$$

Le chiffrage C_e de l'extrait e est l'argument du maximum de S_N.

Cette formulation permet de favoriser les mesures à 4 temps, majoritaires dans la plupart des bases de données tout en prenant en compte les mesures à 2 et 3 temps.

Fusion des descripteurs par agent multiple

Chaque descripteur présenté ci-dessus donne une information qui est soit positive, en indiquant quel temps est un premier temps (changement d'accord), soit négative en indiquant quel temps n'est pas un premier temps (balance spectrale). Lorsque l'information est positive, le score général du temps indiqué augmente et lorsque l'information est négative le score général du temps indiqué diminue. Le temps qui obtient le score maximal après l'utilisation de tous les descripteurs est assimilé à un premier temps. La figure 2.25 résume le fonctionnement de notre estimation de la position des premiers temps. La figure 2.26 présente les paramètres choisis pour les différents descripteurs.

23. Par exemple $F_{23}(1 : 10) = [0100100100]$ et $F_{14}(1 : 10) = [1000100010]$

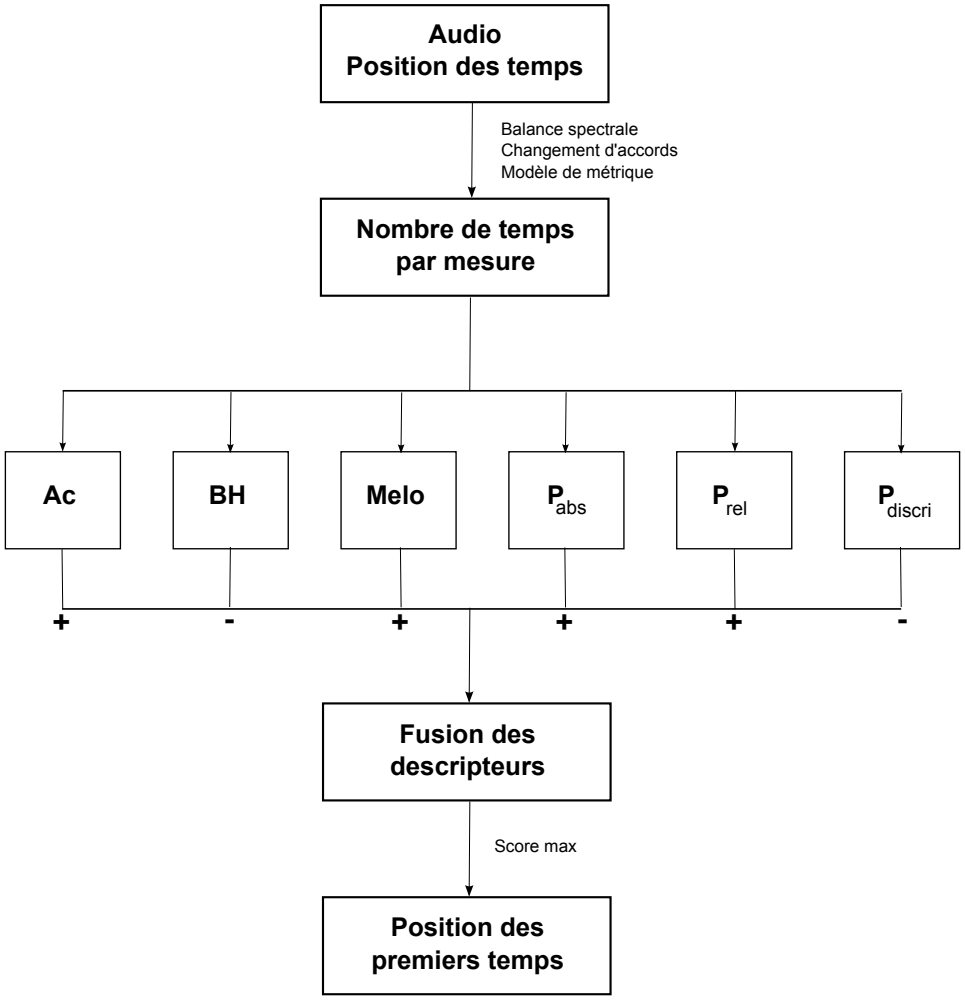


FIGURE 2.25 – Présentation générale de la méthode d'estimation de la position des premiers temps permettant de voir quels descripteurs ont été utilisés. Les + et les - en dessous d'un descripteur signifient que le score général du temps indiqué par ce descripteur augmente ou diminue.

Ac

Choix par rapport à état de l'art :
- Accords majeurs et mineurs
- Distance Itakura Saito entre le dictionnaire et les vecteurs de chroma à 12 dimensions
- Dictionnaire avec 2 harmoniques
- Durée minimale : 2 fois le temps
- max(temps=changement accord) unique

BH

Choix par rapport à état de l'art :
- Grosse Caisse + Caisse Claire
- Poids absolu
- Fréquence CC = [1400,7500]
- Suivi par max(temps)>alpha*max(temps adjacents), alpha=1.1
- Discrimine le temps choisi

Melo

Non présent état de l'art :
- 4 versions compressées Somme Spectrale
- 3 mélodies
- Contour * Durée notes
- max(temps)>alpha*max(temps restants) et max(temps)>s
- alpha=1.05; s=1.05

P_{abs}

Non présent état de l'art :
- 12 MFCC
- Distance euclidienne entre observations
- Intervalle grossier : 0.5 seconde
- Intervalle précis : 0.05 seconde

P_{rel}

Non présent état de l'art :
- 12 MFCC
- Distance euclidienne entre observations
- Suivi si maximum unique

P_{discri}

Non présent état de l'art :
- 12 MFCC
- Distance euclidienne entre observations
- Suivi si minimum unique

FIGURE 2.26 – Choix effectués pour les différentes méthodes. CC signifie Caisse Claire.

Chapitre 3

Performance du système : méthodes, résultats et discussion

Les performances de nos algorithmes sont évaluées à l'aide de mesures issues de la littérature afférente [DDP09] sur une base de données de dizaines d'extraits musicaux et de plusieurs genres différents que nous avons annotés en premier temps. Nous comparerons les performances de nos méthodes avec cinq algorithmes de l'état de l'art et analyserons les résultats obtenus.

3.1 Bases de données

Quarante extraits musicaux de 30 à 60 secondes équitablement répartis selon quatre genres musicaux : le Classique, le Jazz et le Blues et la Dance/Electro ont été annotés en premier temps. L'annotation en temps était déjà disponible. Leur description est disponible à l'annexe B. Ces extraits ont été choisis aléatoirement avant le développement de nos méthodes d'estimation de la position des temps et des premiers temps avec pour seule contrainte un chiffrage constant [1]. Ces genres ont l'intérêt d'être assez variés et de poser plus de difficultés à l'état de l'art que le Hip Hop/rap, le Pop/rock et la Country par exemple [2].

- La musique classique possède des attaques ou des transitions peu marquées. Le tempo est assez expressif et les percussions sont souvent absentes [3] ou ne soulignent quasiment jamais la pulsation comme c'est le cas dans le Rock ou le Jazz,
- dans le Jazz, la présence de swing (décalage des croches vers des positions ternaires, propre à l'interprète) perturbe le côté régulier et la précision temporelle de la phase. La faible qualité de l'enregistrement de certains extraits assez anciens rend parfois leur analyse délicate,
- le chant sur le Blues, dont les extraits ici sont assez anciens, est assez expressif et vient perturber l'estimation des temps et des changements de mesure,

1. Il ne peut pas y avoir de mesure à 2 temps puis de mesure à 3 temps au sein du même extrait.

2. Les performances pour la position des temps pour le Rock/Pop, Hip Hip/Rap et Soul/RnB/Funk sont en moyenne 55% plus élevées que pour le Jazz, la Dance/Electro et la musique classique chez Klapuri [KEA06]. En ce qui concerne la position des premiers temps chez le même auteur, elle n'a pas été tentée pour la musique classique et les performances pour le Rock/Pop, Hip Hip/Rap et Soul/RnB/Funk sont en moyenne 60% plus élevées que pour le Jazz et la Dance/Electro.

3. Aucune percussion membranophone ou idiophone n'est présente dans la base de données.

– la Dance/Electro possède une grande variabilité, de longs silences, le signal est très compressé et la charleston possède une très grande énergie qui a tendance à marquer les contre-temps. De plus, l'estimation des premiers temps est délicate car peu d'indices permettent leur identification (peu de changement d'accords, des basses sur tous les temps, peu de changements de sections musicales, un signal très compressé au niveau de l'intensité).

Ces extraits viennent de la base de données mise en place par Klapuri [4]. Malheureusement la position des premiers temps n'était pas disponible dans la version présente. Étant donné l'importance du travail d'annotation, nous nous sommes limités à 40 extraits.

La position des premiers temps est relativement aisée à estimer non automatiquement, mais souvent à un facteur près. Il arrive donc plusieurs fois que les annotateurs hésitent entre une mesure à deux ou quatre temps ou qu'ils ne soient pas d'accord entre eux. Dans notre cas, les extraits délicats ont été tranchés en concertation avec plusieurs personnes diplômées en musicologie. Des ambiguïtés peuvent subsister, mais comme cela a été dit précédemment, les efforts ont d'abord été fournis afin d'améliorer les cas où les algorithmes ont des difficultés alors que les annotateurs humains sont d'accords.

3.2 Mesures d'évaluation

Il est important de pouvoir comparer ses résultats avec ceux d'autres chercheurs. Pour cela, des bases de données sont mises à disposition, mais également des mesures unifiées. On présentera et évaluera nos méthodes avec les mesures ci-dessous qui sont souvent utilisés dans la littérature.

F-mesure

La F-mesure est l'évaluation générique la plus souvent utilisée en recherche d'informations musicale. On la retrouve dans la plupart des évaluations du Music Information Retrieval EXchange [5] par exemple. Pour l'estimation de la position des temps, cette mesure est calculée à l'aide de trois paramètres : A le nombre de temps annotés, E le nombre de temps estimés et $CE(fp)$ le nombre de temps correctement estimés dans l'intervalle d'une fenêtre de précision fp. On définit alors la précision p et le rappel r :

$$p(fp) = \frac{CE(fp)}{E} \tag{3.1}$$

$$r(fp) = \frac{CE(fp)}{A} \tag{3.2}$$

Ils nous permettent de calculer la F-mesure F :

$$F(fp) = \frac{2 * r(fp) * p(fp)}{r(fp) + p(fp)} \tag{3.3}$$

La F-mesure, exprimée en pourcentage, favorise donc les résultats avec une bonne précision et un bon rappel. La fenêtre de précision sera définie comme un pourcentage du tempo afin de ne pas discriminer les morceaux lents et favoriser les morceaux rapides. On peut remarquer qu'une estimation à contre-temps obtiendra un score de 0% alors qu'une

4. Les détails de cette base de données sont disponibles ici : http ://www.cs.tut.fi/∼klap/iiro/meter
5. www.music-ir.org/mirex/

estimation deux fois plus rapide ou plus lente obtiendra un score bien supérieur d'environ 67%. L'estimation de la position des premiers temps est identique bien que dans notre cas, la fenêtre de précision ne présente plus d'intérêt.

Évaluation basée sur la continuité

La classification est ici non plus binaire comme dans le cas de la F-mesure mais basée sur la mesure de régions dans lesquelles les temps sont correctement estimés.

La continuité se fait dans une fenêtre de tolérance θ en pourcentage du tempo autour de chaque annotation a_j. Le temps le plus proche γ_t de chaque annotation est considéré correct si il se situe dans l'intervalle de la fenêtre de tolérance et si le temps précédent se situe également dans l'intervalle de la fenêtre de tolérance. Les conditions de continuité prennent également en compte le tempo annoté Δ_j et le tempo estimé Δ_t comme le montrent les trois formules suivantes :

(i) $a_j - \theta\Delta_j < \gamma_t < a_j + \theta\Delta_j$
(ii) $a_{j-1} - \theta\Delta_{j-1} < \gamma_{t-1} < a_{j-1} + \theta\Delta_{j-1}$
(iii) $(1-\theta)\Delta_j < \Delta_t < (1+\theta)\Delta_j$

En comparant chaque temps γ_t à chaque annotation a_j d'après (i) et (iii), on peut estimer le nombre de temps corrects sur chaque segment de continuité correct Υ_m comme le montre la figure 3.1. Cela nous permet de déterminer la première mesure de performance de continuité : la rapport du segment de continuité correct le plus long avec la longueur de l'entrée. La mesure CML_c ou niveaux métriques corrects continus permet d'indiquer la proportion de temps au niveau métrique correct avec la continuité requise :

$$CML_c = \frac{\max(\Upsilon_m)}{A} * 100\% \tag{3.4}$$

Avec A le nombre d'annotations. CML_c tient simplement compte du segment d'estimation correcte le plus long. Si un seul temps est mal estimé au milieu du morceau, $CML_c = 50\%$ alors que si le temps est mal estimé au début ou à la fin du morceau $CML_c \simeq 100\%$.

CML_t ou niveaux métriques corrects totaux permet d'éviter cette dépendance à la position des temps mal estimés avec une mesure moins restrictive :

$$CML_t = \frac{\sum_m \Upsilon_m}{A} * 100\% \tag{3.5}$$

Les deux mesures précédentes sont recalculées pour des tempo deux fois plus rapides et deux fois plus lents[6] et le meilleure score est gardé. On appelle ces mesures AML_c et AML_t ou niveaux métriques admis (continus ou totaux). Goto et Hainsworth sont à l'origine de ces mesures basées sur la continuité [GM97] [Hai03].

Le lecteur souhaitant plus d'informations sur les mesures adaptées à l'évaluation de méthodes d'estimation de la position des temps et des premiers temps de contenu musical est invité à consulter l'article de Davies sur le sujet [DDP09].

6. Ces écarts sont admis car il arrive plusieurs fois que le tempo soit perceptivement perçu à un facteur deux ou un demi.

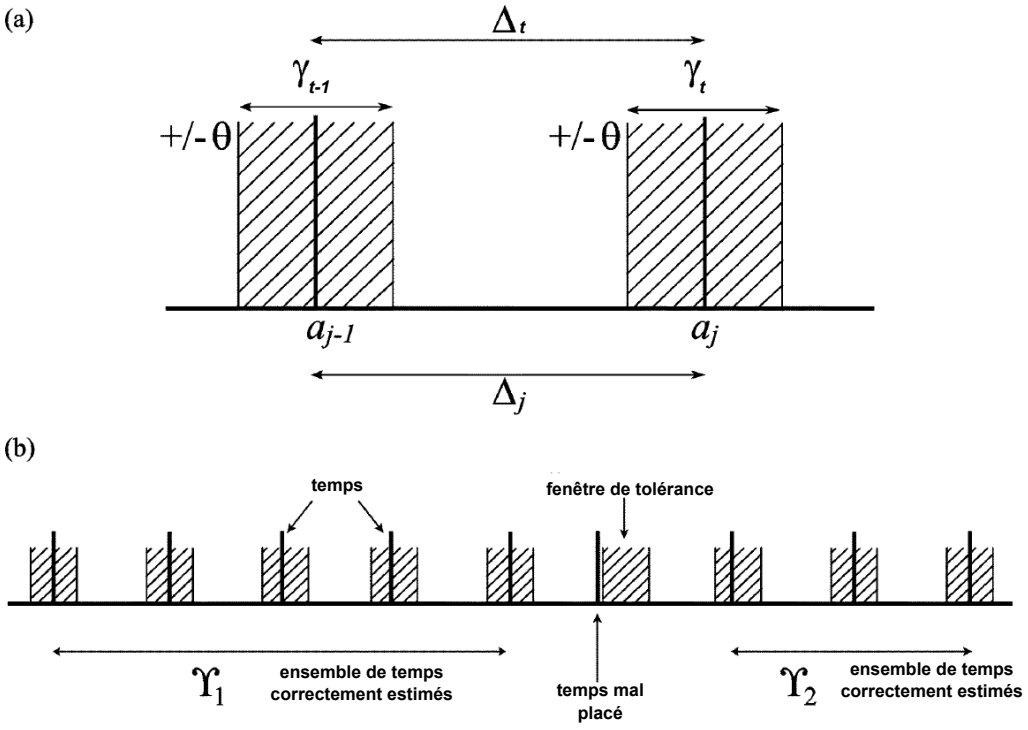


FIGURE 3.1 – (a) Mesure basée sur la continuité pour l'estimation de la position des temps. L'acceptation des temps γ_t dépend de leur proximité avec l'annotation a_j et de la position du temps précédent. θ correspond à la fenêtre de tolérance autour de chaque annotation. (b) Résumé de la méthode sur plusieurs temps consécutifs. Les notations sont les mêmes que dans le corps du document. Traduit de Davies [DP07].

3.3 Résultats obtenus

3.3.1 Estimation du tempo et de la position des temps

La F-mesure et les évaluations basées sur la continuité de notre méthode d'estimation du tempo et de la position des temps que nous nommons *Arbres* seront comparées à 4 algorithmes de l'état de l'art :

- Ellis utilise la programmation dynamique pour estimer la position des temps au sein d'une méthode que nous appellerons *Ellis*.
- La méthode développée par Klapuri que nous avons présentée au chapitre précédent sera évaluée sous le nom de *Klapuri*. Il s'agit de la version implémentée pendant ce stage à partir de l'article de Klapuri [KEA06].
- Thomas Fillon, Cyril Joder et Slim Essid utilisent le cadre probabiliste des champs conditionnels aléatoires [7] pour estimer la position des temps. Nous appellerons cette méthode *FJE*.
- Davies a développé une méthode que nous appellerons *Davies* qui utilise un modèle à deux état afin de s'adapter aux changements de tempo tout en conservant une continuité avec l'estimation du temps précédent.

7. Ils permettent de considérer des dépendances globales au problème. Voir [LMP01] pour plus d'informations sur le sujet.

Les résultats sont disponibles au tableau 3.1. Les résultats de la F-mesure sont plus stables que ceux basés sur la continuité. En effet, on a vu qu'il suffisait d'un temps mal estimé pour que la mesure de continuité chute de moitié. La méthode d'Ellis obtient des performances bien meilleures lorsque tous les niveau métriques sont autorisés. En effet, elle à tendance à surestimer le tactus et ses performances globales en sont affectées. L'algorithme de Klapuri obtient une bonne F-mesure mais n'est pas adapté à la mesure de continuité. Son utilisation des modèles de Markov cachés fait qu'un temps peut être localement sur ou sous-estimé pour obtenir une meilleure estimation globale ce qui est rédhibitoire pour les mesures de continuité. La méthode *FJE* obtient de très bonnes performances aussi bien au niveau de la F-mesure que de la mesure de continuité. Quelques rares erreurs sont l'estimation en contre-temps d'extraits de Jazz et de Dance et l'estimation du tempo à un facteur 2 ou un demi parfois (on voit une forte hausse des performance lorsque les erreurs de tempo sont autorisées). L'algorithme de Davies obtient des performances proches de la méthode précédente et les erreurs se ressemblent également. Enfin, notre algorithme obtient des résultats intéressants avec le meilleur score pour deux évaluations sur cinq dont la F-mesure qui constitue la mesure la plus répandue sur le sujet. Néanmoins, les performances des quatre meilleures méthodes sont proches.

On peut ajouter que la méthode par Arbres est plus sensible que les autres méthodes à la taille de fenêtre de précision. On observe une diminution de 12% de la valeur de la F-mesure pour une fenêtre 33% plus étroite, alors que cette diminution n'est que de 5 % pour les autres algorithmes en moyenne [8]. Ceci s'explique par la façon dont les systèmes ont été construits. La méthode par Arbres retourne des temps régulièrement espacés les uns des autres, qui seront légèrement décalés avec le signal réel si le morceau connait des variations de tempo. Cela arrive moins avec les autres méthodes qui sont moins rigides. Cela se voit sur les résultats. Les performances sur la Dance qui possède un tempo stable sont plus élevées pour la méthode par Arbres qu'ailleurs, alors que pour la musique classique beaucoup plus dynamique les performances chutent plus qu'ailleurs. Par contre, la méthode par Arbre permet la meilleure estimation du tempo, et est performante pour trouver l'estimation de la position des temps à un tempo donné. De plus, les performances pour le Jazz, le Blues et la Dance sont comprises dans un intervalle de $\pm 7.5\%$ ce qui montre une certaine robustesse.

En complément des résultats, on peut noter que la complexité des méthodes d'Ellis et Davies est beaucoup plus faible que celle des autres méthodes. La position des temps d'un extrait de 60 secondes est estimée en deux secondes en moyenne chez Ellis et Davies, avec un processeur Intel(R) Core(TM)2 Duo 3 Ghz. Il faut respectivement 45, 50 et 90 secondes pour faire la même chose avec l'algorithme FJE, la méthode proposée par Klapuri et la procédure par Arbres. L'optimisation du temps de calcul n'a pas été la préoccupation principale de ce stage.

8. L'augmentation des performances pour une fenêtre plus grande est également plus importante pour la méthode par Arbres que pour les autres méthodes.

Nom	**Fmesure**		CMLc		CMLt		AMLc		AMLt	
Ellis	72.52	82.80	30.35	56.09	36.46	56.16	60.57	85.98	80.75	95.78
		66.86		23.08		41.48		52.93		77.15
		71.13		12.24		18.22		52.69		89.07
		69.29		30.00		30.00		50.68		61.00
Klapuri	78.93	90.75	38.57	76.63	62.29	89.13	46.36	77.39	77.39	96.18
		75.84		24.90		58.41		40.45		77.62
		74.36		28.83		54.50		32.58		74.80
		74.77		23.93		47.12		27.58		60.93
FJE	79.87	66.73	**66.98**	60.06	68.84	60.13	**92.59**	99.82	**94.66**	99.82
		86.07		80.00		80.00		100		100
		77.06		54.89		58.50		88.23		92.03
		89.34		71.77		75.70		81.86		86.52
Davies	79.90	73.81	60.85	70.00	65.73	70.00	83.70	95.88	90.22	99.03
		79.27		55.14		59.14		94.97		98.97
		75.58		41.17		52.95		65.36		79.06
		91.75		77.07		80.83		78.61		83.83
Arbres	**80.83**	86.50	60.92	79.39	**74.11**	79.81	71.44	99.23	86.23	99.64
		83.27		63.62		78.16		70.76		86.84
		66.62		33.15		55.49		39.92		72.04
		86.90		67.51		83.00		75.86		86.43

TABLE 3.1 – Résultats de l'évaluation de la méthode d'estimation de la position des temps, nommée *Arbres*. Comparaison avec des algorithmes de l'état de l'art de Davies, Ellis, Klapuri ainsi que Fillon, Joder et Essid [DP07], [Ell07], [KEA06]. Les quatre valeurs regroupées les unes sous les autres au sein d'une même case représentent les résultats pour la Dance/Electro, le Jazz, le Classique et le Blues, de haut en bas. La fenêtre de précision de la F-mesure est de 25% et la fenêtre de tolérance de la mesure de continuité est de 30%.

3.3.2 Estimation des premiers temps à partir de la position des temps

Nous présenterons d'abord la performance de chacun des descripteurs utilisée dans notre méthode d'estimation de la position des premiers temps à partir de la position des temps pour s'assurer qu'ils apportent chacun une valeur ajouté. Nous comparerons ensuite les résultats obtenus en utilisant plusieurs combinaisons de descripteurs avec ceux obtenus par le logiciel Ircambeat.

Résultat de chaque descripteur

Pour évaluer la performances de chaque descripteur, nous n'utiliserons pas la F-mesure qui requiert un méthode d'estimation complète mais plutôt le nombre d'estimations correctes lorsque le chiffrage est bien estimé. Les résultats sont disponibles au tableau 3.2.

Les changements d'accords :

Ce descripteur est performant avec plus de deux tiers de résultats corrects contre un hasard de 26.32% et influent avec 85% de prise de décision. Un inconvénient est la tendance à relativement surestimer le troisième temps qui est difficile à discriminer du premier temps. Les erreurs se produisent surtout pour les extraits de Blues et de Dance. En effet un accord est souvent tenu longtemps sur ces genres dans la base de données, ce qui entraine des erreurs de l'algorithme d'estimation d'accord qui va considérer des changements qui n'ont pas lieu par exemple.

La balance harmonique :

La balance harmonique est prise en compte sur presque autant d'extraits et prend des décisions correctes encore plus souvent que le descripteur de changement d'accord. Cependant, elle ne peut que discriminer les temps 2 et 4 et sera utile en conjonction avec d'autres descripteurs.

Le profil mélodique :

Le profil mélodique offre une performance intéressante pour une prise de décision inférieure aux descripteurs précédents mais non négligeable. La prise de décision est effectuée en prenant le poids suivant aux paramètres définis à la partie 2.3.3 : $\alpha = 1.05$ et $s = 1.05$.

Les changements de pattern :

La première remarque est de la méthode reprenant les principes des algorithmes de détection des différentes parties d'un morceau ne présente pas de bons résultats et n'a pas été conservée pour la suite. L'implémentation n'a peut être pas bien été effectuée mais il est également possible que cette méthode ne soit pas adaptée à l'estimation de la position des premiers temps. La durée des extraits est limitée et il y a peu d'enchainements de couplets et de refrains ou autres. Ainsi, le faible nombre d'instants à estimer affecte la robustesse de la mesure. De plus, la précision temporelle est insuffisante ce qui affecte également la robustesse.

La deuxième méthode, nommée P_{abs}, est plus adaptée à notre problème car on se

Nom	Prise de décision (%)	Décision correcte (%)	Autres temps estimés (%)
BH	82.5	90.9	x
Ac	85.0	67.7	8.8 14.17 8.8
Melo	45.0	72.2	11.1 5.6 11.1
P_{abs}	52.5	42.86	33.3 4.8 19.1
P_{rel}	97.5	43.6	12.8 23.1 20.5
P_{discri}	95.0	94.7	47.4 26.3 21.1

TABLE 3.2 – Performance des descripteurs isolés. BH : Balance hamonique, Ac : Changement d'accords, Melo : Profil mélodique, P_{abs} Changement de pattern absolu, P_{rel} : Changement de pattern relatif, P_{discri} : Changement de pattern relatif discriminant. les autres temps estimés sont de haut en bas les temps 2, 3 et 4. Il y a 26.32% de deuxième temps, 24.34% de troisième temps et 23.0% de quatrième temps dans la base de données après normalisation du nombre de mesures par morceau.

situe à une échelle temporelle plus petite et l'on considère tous les changements de patterns musicaux possibles. De plus, on ne cherche pas à retrouver un refrain ou un couplet dans le morceau mais uniquement à détecter les instants de changement d'ambiance. On peut ainsi obtenir plus de données qui seront également plus précises temporellement. Les résultats semblent moins performants qu'ailleurs mais la discrimination du troisième temps se révèlera utile par la suite.

La dernière méthode est assez intéressante et différente des deux autres. On utilise ici toutes les informations à notre disposition à savoir la position des temps et le chiffrage afin de conclure. Il serait utile de perfectionner cette méthode ainsi que les deux précédentes mais les résultats obtenus sont déjà intéressants. On s'aperçoit notamment que le rejet des mesures les moins vraisemblables est très performant : seulement 5,26% des résultats bons sont considérés faux alors que les hasard est de 26,32%.

La fusion :

Nous étudierons l'effet d'un descripteur à la différence entre la F-mesure avec et sans ce descripteur. Si un descripteur n'est pas pris en compte, le temps qu'il indique n'est ni sommé ni soustrait pour choisir le premier temps. On peut remarquer que le changement de pattern absolu possède une bonne interaction avec la balance harmonique. L'effet du changement de pattern absolu sans la balance harmonique est plus faible que l'effet du changement de pattern absolu sans n'importe quel autre descripteur. L'inverse est également vérifiée. L'effet de la balance harmonique sans le changement de pattern absolu est moindre que l'effet de la balance harmonique sans n'importe quel autre descripteur. La table 3.3.2 illustre cela. Une explication possible est que le changement de pattern absolu manque

Effet	Sans le complémentaire	Sans chacun des autres
P_{abs}	**0.75**	moy = 1.72 min=0.97
BH	**5.02**	moy = 6.76 min=5.56

TABLE 3.3 – Comparaison des effets de la balance harmonique BH et du pattern absolu P_{abs}. Les valeurs affichées sont les différences de score en pourcentage. Le terme complémentaire signifie BH à la première ligne et P_{abs} à la deuxième. On voit les effets moindres de ces deux descripteurs l'un sans l'autre.

Algorithme	Fmesure
Ac +BH	60.92
Ircambeat	**61.63**
Ac + BH + P_{abs}	67.17
Ac + BH + P_{abs} + P_{discri}	72.80
Ac + BH + P_{abs} + P_{discri} + $Melo$	**76.33**
Tout	75.3

TABLE 3.4 – Comparaison de la Fmesure de la méthode développée dans le cadre de ce stage pour plusieurs configurations différentes et du logiciel Ircambeat au niveau de l'estimation des premiers temps à partir de la position des temps. La comparaison est effectuée sur la base de données présentée à la partie 3.1 et à l'annexe B

parfois de précision temporelle. Que ce soit à cause de la performance des musiciens ou de l'estimation du descripteur, il existe une période floue autour du changement de pattern qui peut alors être estimé légèrement trop tôt ou trop tard. Ainsi, les temps adjacents au premier temps risquent d'être estimés à tort. Cependant, la balance harmonique est efficace pour discriminer les temps adjacents au premier temps. L'interaction des deux descripteurs est donc positive. D'une manière générale, il est plus difficile de discriminer le premier et le troisième temps dans un morceau de musique à quatre temps. Le changement de pattern absolu est un des seuls descripteurs présents à pouvoir le faire.

Résultat de l'ensemble des descripteurs et comparaison avec le logiciel Ircambeat

Afin de savoir si les performances de notre algorithme sont intéressantes, nous allons les comparer avec celles du logiciel Ircambeat, dont une version exécutable nous a aimablement été transmise par Hélène Papadopoulos [PP11]. Cette version estime la position des premiers temps et à l'avantage de prendre en entrée la position temporelle des temps. D'autres algorithmes, comme celui développé par Klapuri par exemple, doivent estimer eux même le tempo et la position des temps avant de trouver les premiers temps. Ainsi, des risques d'erreurs en cascade viendraient les désavantager et biaiser la comparaison.

La table 3.3.2 présente les résultats obtenus. On remarque des performances similaires en utilisant les même types de descripteurs, à savoir la balance harmonique et le changement d'accord, avec une Fmesure aux environs de 61%. L'ajout du pattern absolu qui interagit avec les descripteurs précédent permet ensuite une amélioration des performances de plus de 6%. Ensuite, la prise en compte des patterns relatifs discriminants puis du profil

mélodique permet de faire monter les performances de 4% environs à chaque fois pour arriver à une Fmesure finale de 76.33 %. La prise en compte de tous les descripteurs donne une Fmesure de 75.3%, proche de la valeur maximale et montrant que la prise de décision par des descripteurs multiples est assez performante. Les résultats sont intéressants car les performances globales sont bonnes et le principe d'utilisation de plusieurs descripteurs est vérifié par l'augmentation progressive de la performance en fonction du nombre de descripteurs pris en compte. On peut noter qu'il existe un biais de comparaison avec le logiciel Ircambeat car notre algorithme a été optimisé avec la base de données. Cependant, un effort a été fourni pour ne pas trop dépendre de cette base de données en utilisant des valeurs de paramètres stables [9].

Un point encourageant est que les erreurs d'estimation sont surtout liées à un manque de consensus et non à une estimation globalement erronée. En effet, lorsque le premier temps n'est pas correctement estimé, seuls un ou deux descripteurs sur les 5 positifs s'accordent entre eux. Lorsque les descripteurs s'accordent mieux entre eux, qu'il y a une majorité absolue, le premier est toujours correctement estimé sur la base de données.

9. Cela veut dire que si on note p le paramètre et $R(p)$ les résultats obtenus en fonction du paramètre, on a $R(p+\epsilon) \simeq R(p)$ pour ϵ inférieur à un seuil.

Conclusion

Durant ce stage de cinq mois, il s'agissait de se familiariser avec les méthodes d'estimation du tempo, de la position des temps et des premiers temps. Pour cela un état de l'art a été effectué et plusieurs méthodes ont été implémentés. Une base de données de quarante extraits de trente à soixante secondes a été annotée en premiers temps. Cependant, la valeur ajoutée de notre travail a été l'établissement de méthodes d'estimation de la position des temps et surtout des premiers temps car la littérature sur le sujet n'était pas très développée. Il n'existe par exemple pas d'évaluation MIREX ou autres sur les premiers temps à notre connaissance.

L'estimation du tempo et de la position des temps a été faite à l'aide d'une méthode par arbres à choix multiples. À chaque instant, 3 tempo et 6 temps étaient calculés à l'aide d'une fonction d'autocorrélation à décisions multiples et a priori musicaux et d'un modèle de temps. La position retenue du temps parmi les 6 possibles était celle qui minimisait la distance quadratique entre le signal accentué et une fonction cible. Cette méthode a été comparée avec 4 algorithmes performants de l'état de l'art et à obtenue de bons résultats même si sa rigidité aux variations de tempo reste un handicap.

Pour l'estimation des premiers temps, l'étude de méthodes performantes, des limitations des méthodes actuelles, de bases de données musicales et des phénomènes qui rentrent en jeu dans notre perception des premiers temps à permis d'établir plusieurs descripteurs qui se devaient d'être suffisamment généraux. Les changements d'accords, le profil mélodique, la répartition des basses fréquences ou la balance harmonique et l'enchainement de patterns musicaux ont été retenus. Une base de données de 40 extraits musicaux de genres différents a été annotée en premier temps afin d'évaluer la méthode mise en place. Les résultats, basés sur la F-mesure et les mesures de continuité, ont été comparés au logiciel Ircambeat. Les résultats obtenus sont positifs et montrent que l'utilisation de nombreux descripteurs musicaux permet d'augmenter la qualité de l'estimation de la position des premiers temps.

Perspectives

Plusieurs améliorations sont possibles pour l'estimation de la position des temps et des premiers temps. L'estimation des notes utilisant le produit spectral, la continuité de l'enveloppe spectrale et la programmation dynamique pourrait être plus performante et la mesure de l'accent mélodique plus fournie. La précision temporelle du changement de pattern absolu doit également être améliorée et la mesure de "propreté" des changements de patterns relatifs mérite d'être affinée. Le critère de suivi ou non d'un descripteur pourrait faire l'objet d'une classification plus approfondie. Une base de données plus importante et comportant plus de morceaux dont les mesures ne sont pas composées de quatre temps permettrait de rendre notre estimation plus robuste. Il faut également vérifier la validité des résultats sur des bases de données non connues à l'avance.

Une autre amélioration possible pour la musique non uniquement instrumentale serait la prise en compte du texte. La principale difficulté serait d'extraire le texte d'un morceau de musique, environnement bruité et complexe. Cependant, cela pourrait être en partie contourné en utilisant un fichier comprenant les paroles, très facile à obtenir sur internet par exemple, et en synchronisant les paroles avec les trajectoires formantiques détectées sur le signal comme peuvent le faire Fujihara, Mauch ou Mesaros [FGO+06], [MFG10] et [MV08] et en liant paroles et contenu musical de la façon suivante :

- Une contrainte syllabique importante est fortement liée à un pic mélodique et inversement,
- Une contrainte syllabique importante est fortement liée à une durée de note élevée et inversement,
- Les mots vides [10] ont beaucoup moins de chance d'être liés à un pic mélodique que les autres mots,
- On associera plus facilement une note courte à une voyelle courte qu'à une voyelle longue. De plus, on associera plus facilement une note longue à une diphtongue qu'à une voyelle longue.

Une fois cela fait, on pourrait utiliser le lien entre les paroles et la métrique, comme cela est décrit chez Nichols pour renforcer ou non la probabilité d'être sur un premier temps ou sur un temps [NMBR09]. Voici par exemple quelques liens entre la métrique et les paroles :

- Une contrainte syllabique importante est fortement liée à une position métrique forte et inversement,
- Les mots vides, contenant peu de poids sémantique, sont fortement liés aux positions métriques faibles.

10. Aussi appelés *stopwords*, ce sont des mots communs non significatifs comme les prépositions, les pronoms ou les articles comme *le*, *la*, *des*, *ce*... Non significatif sous-entend que leur distribution statistique dans une collection de textes est uniforme.

Annexe A

Les chaines de Markov cachées adaptées au problème d'estimation conjointe du tatum, du tactus et de la mesure

A.1 Les chaines de Markov

Une chaine de Markov est un processus stochastique discret possédant la propriété de Markov : toute l'information utile pour la prédiction du futur est contenue dans l'état présent du processus. Cela signifie que la prédiction du futur à partir du présent n'est pas rendue plus précise par des éléments d'information supplémentaires concernant le passé :

$$P(q_t = E_j | q_0 = E_{i_0}, q_1 = E_{i_1}, ..., q_{t-1} = E_i) = P(q_t = E_j | q_{t-1} = E_i) \tag{A.1}$$

Avec $E_1, E_2, ..., E_N$ les N états du système, et q_t l'état à l'instant t. L'égalité précédente est appelée la probabilité de transition car elle mesure la probabilité d'être dans un état E_j à l'instant t en étant dans l'état E_i à l'instant précédent. On peut faire l'hypothèse que cette probabilité de transition est indépendante du temps. Dans le cas présent, l'état de sortie est directement observable. Ce pourrait être par exemple une note de musique que l'on entend :

- État 1, E_1 : fa
- État 2, E_2 : sol
- État 3, E_3 : si

On peut postuler que l'on entend une seule note à un instant t et que l'on passe d'une note à l'autre en fonction de leur éloignement sur le clavier. On obtiendrait un matrice de transition T_{ij} de la forme :

$$P(q_t = E_j | q_{t-1} = E_i) = T_{ij} = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.3 & 0.55 & 0.15 \\ 0.1 & 0.15 & 0.75 \end{pmatrix}$$

Étant donné que la première note est un fa, quelle est la probabilité que l'on entende dans cet ordre {fa, fa, si, sol, sol, fa} ? Il faut calculer pour cela la probabilité de la séquence d'observation $O = \{E_1, E_1, E_3, E_2, E_2, E_1\}$ connaissant le modèle (la première note est un

fa et la probabilité de transition entre deux notes consécutives est définie ci-dessus) :

$$P(O|\text{Modèle}) = (P(E_1, E_1, E_3, E_2, E_2, E_1|\text{Modèle}) \quad (A.2)$$
$$= P(E_1).P(E_1|E_1) * P(E_3|E_1) * P(E_2|E_3) * P(E_2|E_2) * P(E_1|E_2) \quad (A.3)$$
$$= 1 * 0.6 * 0.1 * 0.15 * 0.55 * 0.3 \quad (A.4)$$
$$= 0.0015 \quad (A.5)$$

A.2 Les modèles de Markov cachés

Dans les modèles de Markov cachés, le processus de sortie n'est pas un événement physique observable. Dans notre cas, on n'observe pas de temps ou de premier temps mais une partie de l'énergie du signal audio dont il faut déduire de quel état caché elle découle. L'observation est donc ici une fonction probabiliste de l'état. La figure A.1 illustre ce modèle sur un exemple concret. On peut voir sur cette figure les différents éléments des modèles de Markov cachés.

- Le nombre d'états du modèle. Il y a deux états dans notre exemple qui sont les deux dés utilisés.
- Le nombre d'observations différentes possibles pour chaque état. Elles sont au nombre de 6 dans notre exemple, les 6 faces d'un dé.
- La distribution de probabilité de transition d'états. Elle correspond à la probabilité de passer d'un état à un autre ou à lui même. Dans notre exemple, c'est la probabilité que le lanceur passer du premier dé au deuxième ou non et inversement. $p(o_{t+1} = e_j|o_t = e_i)$ est la probabilité d'être dans l'état "j" à l'instant "t+1" après avoir été dans l'état "i" à l'instant "t". Notre exemple permet de se convaincre que la somme selon "j" de cette probabilité est égale à 1 car après avoir lancé un dé, on ne peut que le relancer le même dé ou l'autre disponible.
- La probabilité d'émission. Elle correspond à la probabilité d'un état d'émettre telle observation. Dans notre cas, le premier dé a 40% de chance de donner un 6 par exemple. De même que précédemment, un état à 100% de chance de donner une de ces observations possible, (1,2,3,4,5,6) dans notre cas.
- L'état initial. Ce dernier élément permet de savoir quel est le premier état du système. Dans notre exemple l'état initial est tiré de façon aléatoire et chaque dé à 50% de chance d'être lancé initialement. Là aussi, on a 100 % de chance de lancer un des deux dés.

Les modèles de Markov d'ordre 1, en outre d'avoir une complexité réduite et une résolution assez peu coûteuse en calculs car chaque état dépend uniquement du précédent, ont également l'intérêt d'être génératifs. Il est alors possible de générer une séquence similaire à celle observée une fois les 5 éléments précédents connus.

Rabiner montre et résout les 3 problèmes permettant de résoudre les modèles de Markov cachés et le lecteur est invité à s'y référer pour de plus amples détails [Rab89]. Il vont cependant être présentés ici avec quelques remarques additionnelles.

- Le premier problème est de calculer, pour l'optimiser et en déduire la séquence d'état, $p(O|\lambda)$ avec O les observations et λ le modèle (les transitions, l'émission et la distribution initiale). Il faut pour cela utiliser une variante de la formule des probabilités totales et ce que l'on sait sur la séquence d'état E. En utilisant les propriétés des

probabilités conjointes et conditionnelles, on a : $p(O, E) = p(O|E)p(E)$ et $p(O) = \sum_E p(O, E)$. Ainsi, $p(O) = \sum_E p(O|E)p(E)$ et $p(O|\lambda) = \sum_E p(O|E, \lambda)p(E|\lambda)$. Or on peut montrer que la probabilité de la séquence d'états connaissant le modèle (deuxième terme de la somme) est égale aux probabilités de transition des états de la séquence et que la probabilité de la séquence d'observation sachant la séquence d'état et le modèle (premier terme de la somme) est égale aux probabilités d'émission des observations de la séquence d'état. Les probabilités d'émission, de transition et d'état initial sont donc suffisantes pour calculer $p(O|\lambda)$. Le résultat direct étant une imbrication de sommes, ce qui implique un nombre élevé de calculs [1]. Heureusement, il est possible de résoudre se problème par récursion et de réduire grandement le coût en calcul [2] à l'aide de la procédure backward-forward. Ainsi, la probabilité d'observer la séquence d'observation $O_1...O_t$ jusqu'au temps t connaissant le modèle dépend uniquement de la probabilité d'observer la séquence d'observation $O_1...O_{t-1}$ au temps précédent, de la probabilité d'émission au temps t et de la probabilité de transition entre les temps $t-1$ et t. On retrouve l'hypothèse d'ordre 1 de dépendances du modèle de Markov.

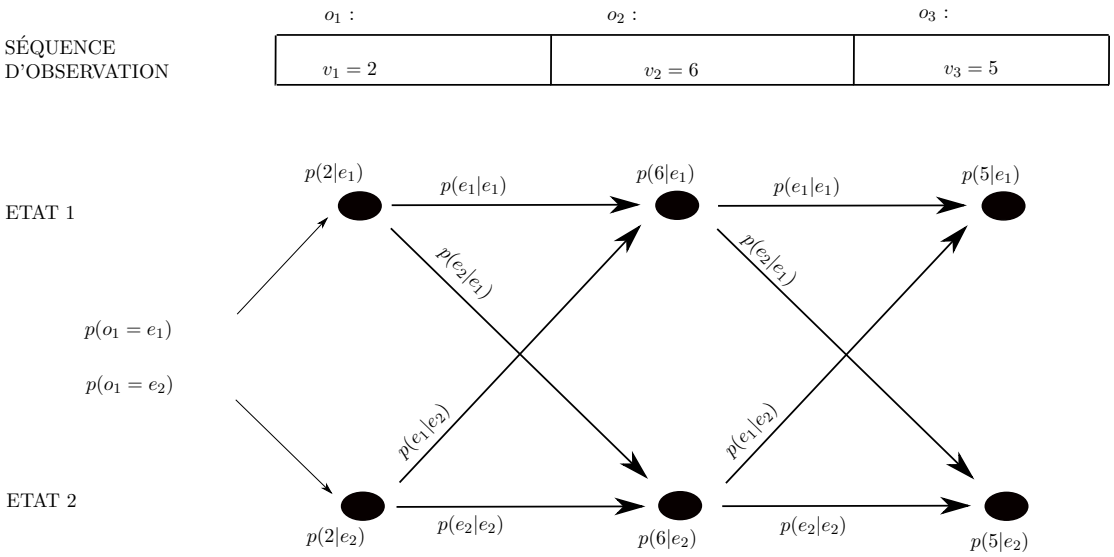
– Le deuxième problème est de trouver la séquence d'état qui explique le mieux les observation. Une solution efficace de ce problème est l'algorithme de Viterbi. La méthode est également récursive et il s'agit de calculer à chaque instant de la séquence d'observation le chemin le plus probable pour arriver dans un état donné. Ce chemin dépend uniquement du chemin le plus probable pour arriver dans un état donné à l'instant précédent. Le prolongement du chemin optimal pour arriver dans chaque état est alors effectué de proche en proche jusqu'à arriver à la fin de la séquence d'observation à l'état final. Le chemin ayant la probabilité la plus élevée est gardé et les états par lequel il est passé sont ceux qui expliquent le mieux la séquence d'observation. La figure A.2 illustre cette méthode.

– Le dernier problème est d'ajuster les paramètres du modèle ou de choisir le meilleur modèle pour maximiser $p(O|\lambda)$. Dans le cadre de ce stage, cela pourrait correspondre à trouver des templates de mesure permettant d'expliquer le mieux les observations. De plus, ces templates doivent être discriminants, c'est à dire dans notre cas que l'on ne puisse pas obtenir un résultat similaire en partant du premier temps ou d'un autre temps. L'algorithme de Baum-Welch qui reprend les principes de maximisation et attentes peut être utilisé pour cela. Il s'agit d'apprendre et de mettre à jour les paramètres du modèle (probabilité d'émission, de transition et d'état initial) jusqu'à convergence à l'aide de la séquence d'observation connue et d'une initialisation du modèle. Pour cela on calcule le rapport entre :
  · la probabilité d'observer le paramètre à estimer (une transition, une émission, un état initial) sachant la séquence d'observation et le modèle,
  · la probabilité compatible avec le paramètre à estimer d'observer la ou les séquences connues sachant le modèle.

  Ce rapport donne la nouvelle valeur du paramètre à estimer. On met à jour le modèle et on recommence jusqu'à convergence. Si on reprend l'exemple précédent du lancer de dés, la mise à jour de l'estimation de la probabilité de transition du premier dé vers le deuxième dé est égale au rapport entre, la probabilité qu'il y ait au moins une fois une transition entre le premier et le deuxième dé connaissant le modèle et

1. Plus précisément, $2T.N^T$ calculs avec T la longueur de la séquence observée et N me nombre d'états.
2. À N^2T calculs

la séquence $\{v1 = 2, v2 = 6, v3 = 5\}$, et la probabilité qu'il y ait au moins un lancer du premier dé connaissant le modèle et la séquence $\{v1 = 2, v2 = 6, v3 = 5\}$. Il est à noter que cette optimisation n'est pas globale mais seulement locale et dépendra donc du choix initial du modèle.



EMISSION(ij)= $p(i|e_j) = p(v_t = i|o_t = e_j)$

=

0.05	0.05	0.05	0.05	0.4	0.4
0.4	0.4	0.05	0.05	0.05	0.05

TRANSITION(ij)= $p(e_j|e_i) = p(o_{t+1} = e_j|o_t = e_i)$

=

0.7	0.3
0.2	0.8

INITIALISATION(i)= $p(o_1|e_i)$

=

0.5	0.5

CHEMIN AVEC LE MAXIMUM DE VRAISSEMBLANCE

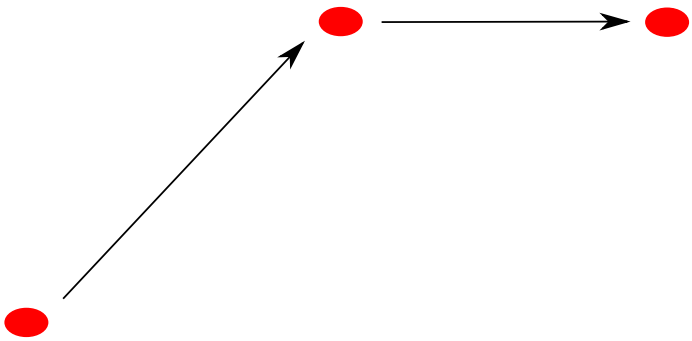


FIGURE A.1 – Exemple illustrant les modèles de Markov cachés. On considère trois lancers de deux dés (deux états cachés) pipés (probabilité d'émission non uniforme) et dont on observe uniquement la valeur obtenue (séquence d'observation) après chaque lancer : 2, 6 puis 5. Il s'agit de savoir quel dé a été utilisé à chaque instant. Dans l'exemple présent, une fois qu'un dé a été utilisé il est plus probable qu'il soit réutilisé pour le lancer suivant (probabilité de transition non uniforme). Les probabilités d'émission et de transition sont connues. Le chemin le plus probable, qui explique le mieux les observations, est affiché en rouge en bas de la figure et sa log-vraissemblance est -5.41. Il est à noter que les modèles de Markov marchent le mieux avec des processus infinis.

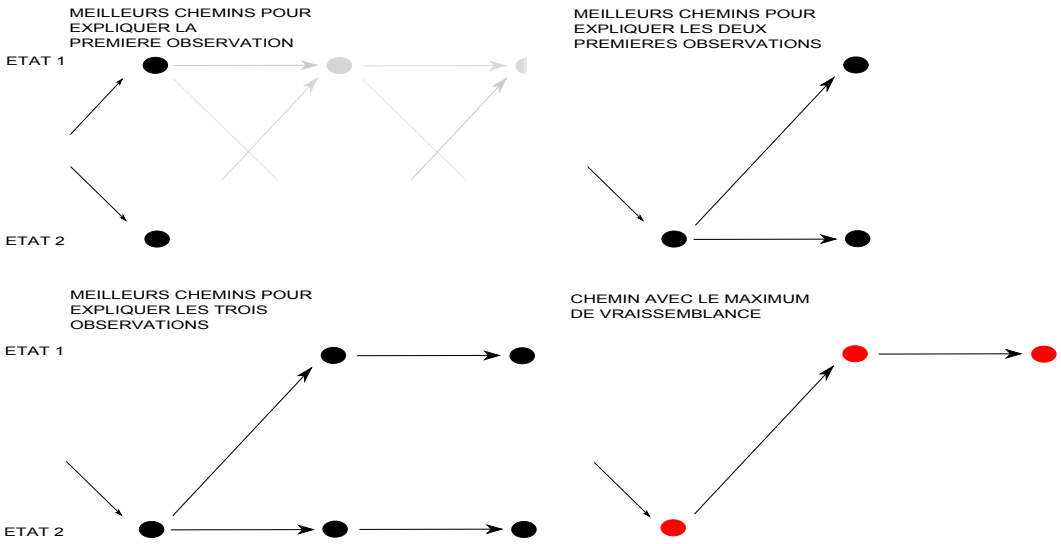


FIGURE A.2 – Choix des états expliquant le mieux la séquence d'observation présentée à la figure A.1. En noir les chemins menant deux états cachés du problème expliquant le mieux les observations. En rouge les états cachés expliquant le mieux les observations.

Annexe B

Contenu de la base de données

Voici le titre et le genre dont est tiré chaque extrait de la base de données ainsi que l'artiste ou le compositeur. Le numéro de l'extrait de la base de données mis au point par Klapuri est également précisé.

Artiste	**Titre de l'extrait**	**Numéro de l'extrait dans la base de données originale**	**Genre**
Ahmad Jamal	Autumn In New York	20	Jazz
Frank Sinatra	Bad Bad Leroy Brown	25	
Abraham Laboriel	Dear Friends	89	
Glenn Miller	In The Mood	204	
Alex Welsh	Maple Leaf Rag	270	
Glenn Miller	Over The Rainbow	329	
Lars Edegran	Panama	331	
Miles Davis	Round Midnight	363	
Bob Wilber	Lester	368	
Bob Wilber	Moonsong	369	
Pet Shop Boys	Always On My Mind	11	Dance
Armand van Helden	Alienz	13	
Armand van Helden	Boogiemonster	14	
Armand van Helden	Motherearth	15	
Artful dodger	Outrageous	16	
Artful dodger	Rerewind	17	
Dune	Raving	103	
New Order	Everything's Gone Green	120	
Daft Punk	Harder Better Faster Stronger	167	
Moodymann	Long Hot Sexy Nights	253	
Pelle Miljoona	13 Bar Blues	1	Blues
Muddy Waters	Baby Please Don't Go	22	
Pepe Ahlqvist	Bad Bad Whiskey	26	
Muska Babitzin	Cry Your Heart Out	85	
B.B King	How Blue Can You Get	182	
Billy Boy Arnold	How Long Can This Go On	185	
Gary Moore	I Loved Another Woman	198	
Billy Boy Arnold	Lowdown Thing Or Two	262	
Joe Turner	Sweet Sixteen	355	
Johnny Adams	Room With A View	362	
Beethoven	Fidelio	34	Classique
Beethoven	Symphonie n°5	36	
Järnefelt	Berceuse	38	
Biber	Sonate	39	
Bach	Brandebourgeois, Allegro	48	
Busoni	Elegie	58	
Chopin	Etude op25 no9	117	
Handel	La Paix	125	
Susato	Pavane, La dona	131	
Saint-Saëns	Le Carnaval Des Animaux	140	

TABLE B.1 – Les détails sur la base de données originale peuvent être trouvés à l'adresse suivante : http ://www.cs.tut.fi/∼klap/iiro/meter

Bibliographie

[AA06] Miguel A. Alonso-Arevalo. *Extraction d'information rythmique à partir d'enregistrements musicaux*. PhD thesis, École Nationale Supérieure des Télécommunications, 2006.

[AD90] Paul E Allen and Roger B Dannenberg. Tracking musical beats in real time. In *Proceedings of the International Computer Music Conference*, volume 1990, pages 140–3, 1990.

[Big00] Emmanuel Bigand. Tapping in time with mechanically and expressively-performed music. 2000.

[Bil93] Jeffrey Alan Bilmes. Techniques to foster drum machine expressivity. In *Proceedings of the International Computer Music Conference*, pages 276–276. International Computer Music Association, 1993.

[BP05] Juan Pablo Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signals. In *ISMIR*, volume 5, pages 304–311, 2005.

[Bro91] Judith C Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89 :425, 1991.

[CF02] Matthew L Cooper and Jonathan Foote. Automatic music summarization via similarity analysis. In *ISMIR*, 2002.

[Dao08] T.J. Daoust. *Polymeter in Twentieth-century Music : A Study in Notational Methods*. The University of North Carolina at Greensboro, 2008.

[DDP09] Matthew EP Davies, Norberto Degara, and Mark D Plumbley. Evaluation methods for musical audio beat tracking algorithms. *Queen Mary University, Centre for Digital Music, Tech. Rep. C4DM-TR-09-06*, 2009.

[Dix01] Simon Dixon. An empirical comparison of tempo trackers. In *Proceedings of the 8th Brazilian Symposium on Computer Music*, pages 832–840, 2001.

[DP07] Matthew EP Davies and Mark D Plumbley. Context-dependent beat tracking of musical audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3) :1009–1020, 2007.

[Ell07] Daniel PW Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1) :51–60, 2007.

[FACM+03] P Flandrin, F Auger, E Chassande-Mottin, et al. Time-frequency reassignment : from principles to algorithms. *Applications in Time-Frequency Signal Processing*, 5 :179–203, 2003.

[FGO+06] Hiromasa Fujihara, Masataka Goto, Jun Ogata, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals. In *Multimedia, 2006. ISM'06. Eighth IEEE International Symposium on*, pages 257–264. IEEE, 2006.

[Foo99] Jonathan Foote. Visualizing music and audio using self-similarity. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 77–80. ACM, 1999.

[Fuj99] Takuya Fujishima. Realtime chord recognition of musical sound : a system using common lisp music. In *Proc. ICMC, 1999*, pages 464–467, 1999.

[G+05] Fabien Gouyon et al. A computational approach to rhythm description-audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing. 2005.

[GM94] Masataka Goto and Yoichi Muraoka. A beat tracking system for acoustic signals of music. In *Proceedings of the second ACM international conference on Multimedia*, pages 365–372. ACM, 1994.

[GM97] Masataka Goto and Yoichi Muraoka. Issues in evaluating beat tracking systems. In *Working Notes of the IJCAI-97 Workshop on Issues in AI and Music-Evaluation and Assessment*, pages 9–16, 1997.

[GM99] Masataka Goto and Yoichi Muraoka. Real-time beat tracking for drumless audio signals : Chord change detection for musical decisions. *Speech Communication*, 27(3) :311–335, 1999.

[Got01] Masataka Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, 30(2) :159–171, 2001.

[GR04] Olivier Gillet and Gaël Richard. Automatic transcription of drum loops. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 4, pages iv–269. IEEE, 2004.

[GR08] Olivier Gillet and Gaël Richard. Transcription and separation of drum signals from polyphonic music. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3) :529–540, 2008.

[Hai03] Stephen Webley Hainsworth. Techniques for the automated analysis of musical audio. 2003.

[HDZ+12] A. Holzapfel, M.E.P. Davies, J.R. Zapata, J.L. Oliveira, and F. Gouyon. Selective sampling for beat tracking evaluation. *Audio, Speech, and Language Processing, IEEE Transactions*, 20(9) :2539–2548, November 2012.

[HHT06] Masatoshi Hamanaka, Keiji Hirata, and Satoshi Tojo. Implementing "a generative theory of tonal music"†. *Journal of New Music Research*, 35(4) :249–277, 2006.

[HSEK04] Erin E Hannon, Joel S Snyder, Tuomas Eerola, and Carol L Krumhansl. The role of melodic and temporal cues in perceiving musical meter. *Journal of Experimental Psychology : Human Perception and Performance*, 30(5) :956, 2004.

[HWF09] Jason A Hockman, Marcelo M Wanderley, and Ichiro Fujinaga. Real-time phase vocoder manipulation by runner's pace. In *Proc. Int. Conf. on New Interfaces for Musical Expression (NIME)*, 2009.

[JDM00] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition : A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1) :4–37, 2000.

[KEA06] Anssi P Klapuri, Antti J Eronen, and Jaakko T Astola. Analysis of the meter of acoustic musical signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1) :342–355, 2006.

[KFROch] M. Khadkevich, T. Fillon, G. Richard, and M. Omologo. A probabilistic approach to simultaneous extraction of beats and downbeats. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 445–448, March.

[Kla99] Anssi Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 6, pages 3089–3092. IEEE, 1999.

[Kla01] Anssi P Klapuri. Multipitch estimation and sound separation by the spectral smoothness principle. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 5, pages 3381–3384. IEEE, 2001.

[Kla03] Anssi P Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *Speech and Audio Processing, IEEE Transactions on*, 11(6) :804–816, 2003.

[LHS71] H Christopher Longuet-Higgins and Mark J Steedman. On interpreting bach. *Machine intelligence*, 6 :221–241, 1971.

[LJS85] F. Lerdahl, Ray. Jackendoff, and W. Slawson. A reply to peel and slawson's review of "a generative theory of tonal music". *Journal of Music Theory*, 29(1) :145–160, 1985.

[LK94] Edward W Large and John F Kolen. Resonance and the perception of musical meter. *Connection science*, 6(2-3) :177–208, 1994.

[LMP01] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICLM*, 2001.

[M+67] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[Mes44] O. Messiaen. *Technique de mon langage musical : texte avec exemples musicaux*. Leduc, 1944.

[MFG10] Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. Lyrics-to-audio alignment and phrase-level segmentation using incomplete internet-style chord annotations. In *Proceedings of the 7th Sound and Music Computing Conference (SMC)*, 2010.

[MP13] Ugo Marchand and Geoffroy Peeters. Estimation des tempi perçus en fonction du contenu audio et du profils utilisateurs. Master's thesis, Université Pierre et Marie Curie - Paris VI, Master ATIAM, 2013.

[MV08] Annamaria Mesaros and Tuomas Virtanen. Automatic alignment of music audio and lyrics. In *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*. Citeseer, 2008.

[NMBR09] Eric Nichols, Dan Morris, Sumit Basu, and Christopher Raphael. Relationships between lyrics and melody in popular music. In *Proceedings of the 10'th International Conference on Music Information Retrieval (ISMIR)*. Citeseer, 2009.

[OGF09] Laurent Oudre, Yves Grenier, and Cédric Févotte. Chord recognition using measures of fit, chord templates and filtering methods. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pages 9–12. IEEE, 2009.

[Pap10] Hélène Papadopoulos. *Estimation conjointe d'information de contenu musical d'un signal audio*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2010.

[Par94] Richard Parncutt. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, pages 409–464, 1994.

[PE85] Dirk-Jan Povel and Peter Essens. Perception of temporal patterns. *Music Perception*, pages 411–440, 1985.

[Pee05] Geoffroy Peeters. Time variable tempo detection and beat marking. In *Proc. ICMC*, 2005.

[Pee06] Geoffroy Peeters. Template-based estimation of time-varying tempo. *EURASIP Journal on Advances in Signal Processing*, 2007, 2006.

[Pfo03] Peter Q Pfordresher. The role of melodic and rhythmic accents in musical structure. *Music Perception*, 20(4) :431–464, 2003.

[PLBR02] Geoffroy Peeters, Amaury La Burthe, and Xavier Rodet. Toward automatic music audio summary generation from signal analysis. In *ISMIR*, volume 2, pages 94–100, 2002.

[PP11] G. Peeters and H. Papadopoulos. Simultaneous beat and downbeat-tracking using a probabilistic framework : Theory and large-scale evaluation. *Audio, Speech, and Language Processing, IEEE Transactions*, 19(6), August 2011.

[Rab89] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–285, February 1989.

[Rap97] Christopher Raphael. Modeling the perception of meter with competing subharmonic oscillators. In *Proceedings of the third Triennial ESCOM Conference*, 1997.

[Rap01] Christopher Raphael. Automated rhythm transcription. In *ISMIR*, 2001.

[Ros92] David Rosenthal. Emulation of human rhythm perception. *Computer Music Journal*, 16(1) :64–76, 1992.

[RP13] Zafar Rafii and Bryan Pardo. Repeating pattern extraction technique (repet) : A simple method for music/voice separation. 2013.

[Sch98] E.D. Scheirer. Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, 103 :588, 1998.

[Sep01] Jarno Seppänen. Computational models of musical meter recognition. Master's thesis, Tampere University of Technology, 2001.

[Tho82] Joseph M Thomassen. Melodic accent : Experiments and a tentative model. *The Journal of the Acoustical Society of America*, 71 :1596, 1982.

[TTOS11] E. Tsunoo, G. Tzanetakis, N. Ono, and S. Sagayama. Beyond timbral statistics : Improving music classification using percussive patterns and bass lines. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4) :1003–1014, 2011.

[ZG13] Jose R Zapata and Emilia Gomez. Using voice suppression algorithms to improve beat tracking in the presence of highly predominant vocals. 2013.