Université Pierre et Marie Curie – Paris VI

Mémoire de stage de Master 2 ATIAM

Estimation du tempo perceptif et réduction des erreurs d'octave du tempo

Joachim FLOCON-CHOLET

Sous la direction de Geoffroy Peeters

1er mars 2012 – 31 juillet 2012

Institut de Recherche et Coordination Acoustique/Musique
4, place Igor Stravinsky 75004 Paris

Remerciements

Je tiens tout d'abord à remercier Geoffroy Peeters pour m'avoir permis de faire ce stage et de m'avoir pleinement impliqué dans ce projet. Ses explications, ses conseils et ses directions de recherche ont fait que le stage s'est révélé très instructif.

Je remercie également Axel Roebel et l'équipe Analyse Synthèse pour leur accueil et l'ambiance générale, et en particulier Florian Kaiser et Johan Pauwels, mes collègues de bureau pendant ces cinq mois.

Enfin, je souhaite remercier tous mes camarades ATIAM pour les très bons moments passés en leur compagnie durant toute cette année mais aussi pour les discussions très enrichissantes que j'ai pu avoir concernant mes recherches pendant le stage.

Ce travail a été en partie financé par le programme Quaero financé par l'Oseo, agence nationale française pour l'innovation.

Table des matières

Notations

Symboles, fonctions et opérateurs mathématiques

$\boldsymbol{x}$	vecteur
$\boldsymbol{X}$	matrice
$p(x)$	Densité de probabilité de la variable aléatoire x
$\mathbb{E}[.]$	Espérance mathématique
$\mathbb{E}[x\|y]$	Espérance conditionnelle de x sachant y
$\mathcal{N}(\boldsymbol{z}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Densité de probabilité de la loi normale ayant pour paramètres la moyenne $\boldsymbol{\mu}$ et la matrice de covariance $\boldsymbol{\Sigma}$

Noms de variables et acronymes

T_e	Tempo estimé par un algorithme d'estimation du tempo
T_a	Tempo annoté
BPM	Battements par minute
$MFCC$	Mel-Frequency Cepstral Coefficients
GMM	Gaussian Mixture Model

Introduction

Généralités

Cette recherche s'inscrit dans la problématique de l'estimation automatique du tempo d'un extrait musical. On définit le tempo comme étant l'allure ou la vitesse d'exécution d'une œuvre musicale. Celui-ci est donné de manière précise en battements par minute (noté BPM), à la différence des indications de tempo comme *Andante*, *Allegro ma non troppo* qui peuvent dans ce cas donner une indication sur la vitesse d'exécution mais suggèrent également une intention de jeu de la pièce à interpréter.

Le tempo est un attribut fondamental et très descriptif d'un extrait musical. Sa perception est à la fois intuitive et quasi immédiate dans la plupart des cas. Bon nombre de personnes peuvent battre la mesure d'un morceau après seulement quelques secondes d'écoute. Toutefois, bien que la perception du tempo pour l'humain soit relativement aisée, l'élaboration d'un système permettant d'extraire automatiquement le tempo d'un extrait donné reste quelque peu délicat.

En effet, les algorithmes actuels permettant d'effectuer cette tâche donnent généralement de bons résultats si on néglige un défaut majeur, commun à tous ces algorithmes : les erreurs d'octave du tempo. Les erreurs d'octave de tempo apparaissent lorsque par exemple, l'algorithme de détection identifie le tempo d'un extrait musical à 120 BPM au lieu d'un tempo à 60 BPM. Ou à l'inverse, l'algorithme estime un tempo à 70 BPM pour un tempo réel à 140 BPM. Les erreurs d'octave du tempo correspondent donc à une surestimation ou une sous-estimation du tempo original, d'un facteur 2 ou 3 généralement. Même si cette estimation peut s'avérer correcte vis-à-vis de la pyramide rythmique d'un morceau, elle ne correspond plus au tempo perçu, et donc, à l'atmosphère générale que l'on se fait de la musique.

Il est à noter que malgré la définition que nous venons de donner, on peut trouver de nombreux exemples pour lesquels on pourrait penser la mesure de plusieurs manières différentes. Du fait de cette ambiguïté, il est parfois difficile de parler d'erreur d'octave du tempo. Nous introduisons alors la notion de tempo perceptif, comme étant le tempo qu'une majorité de personnes associeraient à un extrait musical, à la différence d'un tempo théorique déduit d'une partition. Le tempo perceptif nous permet donc d'avoir un tempo de référence lié à la sensation de l'auditeur.

Démarche

La majorité des algorithmes d'estimation du tempo existants fonctionnent en deux étapes. En premier lieu, ils extraient des informations de bas-niveau comme la détection d'attaques de notes. La seconde étape consiste à réaliser une analyse de périodicité afin de dégager la périodicité la plus significative qui donnera le tempo estimé.

Bien que l'étude de la périodicité dans une œuvre musicale soit une piste fondamentale pour estimer le tempo, on se propose d'y associer d'autres notions musicales. L'idée principale de ce projet est de mettre en place une méthode d'estimation du tempo utilisant de nouveaux indices acoustiques. Les descripteurs utilisés sont la **balance spectrale**, liée au motif rythmique d'une batterie ou d'autres événements sonores variant en fréquence au court du temps, les **variations de Chroma**, relatives aux changements d'accords ou de tonalités dans une pièce de musique, la **similarité acoustique**, liée aux variations musicales à court terme dans un morceau de musique et enfin la **fonction d'énergie** qui permet d'estimer la périodicité globale d'un extrait audio.

Le principe de cette démarche se rapproche finalement du schéma de pensée de l'être humain : lorsqu'on écoute un extrait musical, on ne se base pas sur l'interprétation d'un seul paramètre (comme c'est généralement le cas avec les algorithmes d'estimation du tempo se basant principalement sur la périodicité de l'énergie du signal), mais sur un ensemble d'informations. Par exemple, pour un extrait de musique simple, il suffira d'écouter seulement deux temps pour estimer le tempo. En revanche, dans des cas plus complexes, on préfèrera attendre un premier temps ou un changement d'accord pour avoir un avis définitif. On peut également se concentrer sur le motif mélodique d'une des parties instrumentales ou bien se focaliser sur la séquence rythmique de la batterie, ou autre. Le choix du critère d'écoute, c'est à dire décider sur quel critère on va baser notre jugement dépend bien évidement du contexte musical, mais dans notre cas, nous supposons qu'intégrer les différentes notions de changements harmoniques, de motif rythmique, de similarité et de périodicité permettra d'améliorer l'estimation du tempo.

Dans le cadre de ce projet, nous nous appuyons sur le logiciel **Ircambeat** (Peeters 2007b), développé à l'Ircam, qui servira de base à l'élaboration de nouveaux algorithmes.

Pour exposer notre étude concernant l'estimation du tempo et la réduction des erreurs d'octave du tempo, nous proposons le plan suivant. Le chapitre **1** sera consacré à l'état de l'art dans lequel nous étudierons les techniques d'estimation du tempo intégrant l'ajout de connaissance musicale à des algorithmes d'estimation du tempo existants. Dans le chapitre **2**, nous analyserons en détail le corpus d'étude Last.fm qui nous servira tout au long de notre projet et nous observerons le comportement d'Ircambeat vis-à-vis des erreurs d'octave du tempo. Le chapitre **3** présentera les descripteurs audio que nous utiliserons pour inférer le tempo et enfin, dans le chapitre **4**, nous expliquerons quelles approches nous avons retenues pour estimer le tempo à l'aide de nos descripteurs.

Pour clore cette étude, nous dresserons un bilan des méthodes développées dans ce document avant de proposer quelques perspectives.

Chapitre 1

État de l'art

Bien que le travail présenté dans ce document porte sur l'estimation du tempo, nous ne ferons pas ici d'étude bibliographique sur les algorithmes d'estimation du tempo, sujet très vaste qui n'est pas tout à fait le propos ici. Nous préférerons en revanche, détailler les études portant sur l'estimation du tempo par croisement de données, point central de notre projet. Cette présentation permettra de voir quelles sont les techniques utilisées mais aussi sur quelles considérations musicales les auteurs se basent pour établir leurs systèmes.

Avant cela, nous reviendrons sur la perception du tempo pour justifier quelle notion sera préférée pour cette étude.

1.1 Perception du tempo

Comme nous l'avons déjà évoqué en introduction, la notion de tempo perceptif peut donner lieu à plusieurs interprétations. On présente ici les différentes définitions que l'on peut trouver dans les études liées à l'estimation du tempo et nous argumenterons le choix de notre définition.

Dans (**Chua & Lu 2005**), Chua interprète le tempo selon trois points de vue. Il définit tout d'abord le *Score tempo*, qui est le tempo noté sur une partition. Le musicien qui exécutera la pièce se basera alors sur ce tempo. Chua décrit également le *Foot-tapping tempo* qui serait le tempo que l'on penserait inconsciemment lorsqu'on écoute une musique. L'auteur fait remarquer que ce tempo est généralement limité à une plage de tempo centrée autour de 80-100 BPM. Par conséquent, pour une œuvre musicale dont le tempo serait très rapide, le *foot-tapping tempo* serait deux fois moins élevé que le tempo réel. Enfin, il définit le *Perceptual tempo* comme le tempo intégrant la sensation de vitesse d'un morceau.

La figure 1.1 donne un exemple de différence entre les différentes définitions du tempo. Si on considère les deux mélodies indépendamment, elles ont le même *score tempo* et *foot-tapping tempo* mais pas le même *perceptual tempo*. Selon Chua, le tempo perceptif de la mélodie du haut sera plus élevé.

Si on s'attache désormais aux expériences menées sur la perception du tempo, on peut retenir l'article (**Zhu & Lu 2005**) qui propose une étude sur la visualisation symbolique de la musique à partir des informations de tempo et de timbre. Afin de mieux appréhender l'utilisation de ces deux paramètres, les auteurs réalisent une étude perceptive du timbre et du tempo. Sur une base de 600 titres audio regroupant divers styles musciaux, 20 sujets doivent estimer le tempo et le timbre d'un extrait audio à partir d'échantillons de référence. Les résultats de ces expériences montrent une convergence très nette des estimations en tempo.

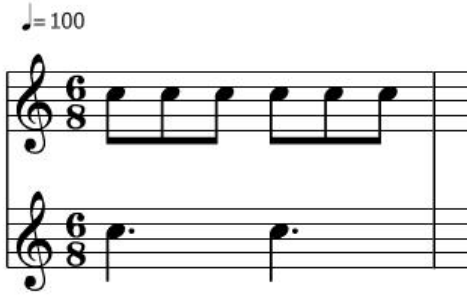

FIGURE 1.1 – Exemple donné par Chua pour illustrer la notion du *Perceptual tempo*. Les deux mélodies ont le même *Score tempo* et *foor-tapping tempo* mais la mélodie du haut a un tempo perceptif plus élevé que la mélodie du bas.

Dans un but différent, d'autres expériences sur l'estimation du tempo ont été menées dans (**Moelants & McKinney 2004**) montrant une fois encore la concordance des estimations faites par les sujets.

Lorsque l'on parle d'estimation du tempo, on peut donc faire référence à plusieurs notions de tempo comme le montre Chua. Il est alors difficile de parler d'erreur d'octave du tempo si on considère ces différentes définitions. Une estimation du tempo peut apparaître juste selon une certaine définition mais peut être considérée comme une erreur d'octave du tempo si on se réfère à une autre notion du tempo.

En revanche, si on s'appuie sur les tests perceptifs, il est clair que, bien que subjective, cette estimation semble converger vers une même valeur. Dans cette étude, étant donné que nous utilisons un corpus d'étude dont les titres ont été annotés par plusieurs personnes, nous pouvons donc faire l'hypothèse que les annotations correspondront au tempo perceptif.

Comme nous le verrons par la suite, on peut toujours trouver des cas pour lesquels les estimations ne concordent pas. Ces cas de figure seront traités particulièrement.

1.2 Croisement de données

On peut distinguer plusieurs approches dans l'état de l'art de l'estimation du tempo par croisement de données.

Tout d'abord, on peut trouver des méthodes permettant d'estimer le tempo sans utiliser d'algorithme classique. C'est le cas dans (**Seyerlehner et al. 2007**) où les auteurs font l'hypothèse que des titres audio ayant des motifs rythmiques similaires auront de fortes chances d'avoir le même tempo. Le motif rythmique est ici l'information relative à la périodicité du signal. Les deux méthodes retenues pour décrire le motif rythmique sont l'autocorrélation et la *Fluctuation Patterns*, fonction décrivant les variations de l'énergie dans 20 bandes de fréquences.

À partir d'une base de données dont chaque titre est annoté en tempo, on peut faire la correspondance entre les motifs rythmiques d'un extrait musical et son tempo associé. Pour un titre inconnu, on peut alors chercher dans la base de données des extraits audio ayant des motifs rythmiques similaires par un simple *K-NN* (K-nearest neighbour). Le tempo estimé pour le titre inconnu sera alors le tempo des extraits audio jugés similaires d'un point de vue du motif rythmique. Avec cette méthode, les auteurs ont obtenus des résultats au moins aussi bons que les algorithmes d'estimation du tempo de l'état de l'art.

(**Xiao et al. 2008**) propose une approche statistique permettant de combiner le timbre et le tempo perceptif. Les auteurs font remarquer que les extraits audio auxquels on associe un tempo élevé sont généralement perçus comme *bruités*, alors que ceux dont les tempi associés sont faibles induisent une sensation de calme. Pour cette méthode, on fait donc l'hypothèse que le timbre influe sur la perception du tempo et qu'un titre ayant un timbre global bruité renforcera la sensation de vitesse alors qu'au contraire, un timbre induisant une sensation de calme laissera deviner un tempo plutôt lent.

Pour caractériser l'information du timbre, les auteurs choisissent d'utiliser les 12 coefficients MFCC (Mel-Frequency Cepstral Coefficients). La méthode consiste à créer un modèle de mélange de Gaussiennes (GMM) dont les vecteurs d'observation sont constitués des 12 coefficients MFCC moyens de chaque titre, auxquels on ajoute le tempo estimé par un algorithme d'estimation du tempo, présenté dans (**Ellis 2007**).

Pour la prédiction du tempo d'un titre inconnu, les auteurs calculent tout d'abord les coefficients MFCC, noté M, ansi que le tempo T grâce à l'algorithme d'estimation du tempo. Quatre tempi sont ensuite générés en multipliant T par un facteur 0.33, 0.5, 2 et 3. Pour un titre inconnu, nous avons donc cinq tempi candidats, noté T_i, $i = 1, ..., 5$ et et les coefficients MFCC M. On détermine enfin la probabilité de la combinaison de T_i et de M avec le modèle GMM. Le tempo final T' sera alors :

$$T' = \arg\max_{T_i} p(T_i|M)$$

où $p(T_i|M)$ représente la probabilité conditionnelle de T_i sachant M. Cette méthode permet donc, à partir de l'information du timbre d'une musique, de retrouver la bonne octave du tempo.

Dans (**Hockman & Fujinaga 2010**) les auteurs s'intéressent à la classification des titres audio en deux classes : les titres dont le tempo serait apparenté à un tempo *lent*, et ceux à un tempo *rapide*. Pour cela, une base de données est constituée de titres étant reconnus comme *lent* ou *rapide*. Ensuite, pour chaque titre plus de 80 descripteurs relatifs au timbre, à l'intensité, à la hauteur spectrale etc., sont calculés. Sur ces données, les auteurs évaluent six méthodes de classification : k-NN, Machines à Vecteur de Support (SVM), Naive Bayes, C4.5 Decision Trees, AdaBoost avec C4.5 et Bagging avec C4.45. Les meilleurs résultats sont obtenus avec les techniques AdaBoost et Machines à Vecteurs de Support (SVM) dont les taux de bonne classification s'élèvent à 99.44 %.

Dans (**Chen et al. 2009**), est proposé un système qui utilise l'information d'ambiance musicale pour créer un modèle statistique de classe de tempo perceptif. Ce modèle est ensuite utilisé pour corriger les estimations de n'importe quel algorithme d'estimation de tempo.

Les auteurs considèrent que la notion d'ambiance d'une musique est un paramètre porteur de sens qui peut être relié à la perception du tempo. Par exemple, une musique dont on pourrait caractériser l'ambiance comme 'agressive' ou 'effrénée', peut suggérer que le tempo perçu sera *rapide*. À l'inverse, une musique qualifiée de 'romantique' ou 'sentimentale' induira un tempo perceptif *lent*.

Pour la classification des ambiances musicales, 90 descripteurs sont utilisés pour modéliser 101 humeurs musicales grâce à un GMM. Quatre classes de tempo perceptif sont ensuite considérées : les tempi *Très lent*, *Plutôt lent*, *Plutôt rapide* et *Très rapide*. Chaque titre de la base de données a été annoté selon l'une de ces quatre classes. Ensuite, un SVM est utilisé pour modéliser ces quatre classes en utilisant les descripteurs d'ambiance musicale comme vecteurs d'observation.

Le système détermine donc l'ambiance générale grâce au calcul des 90 descripteurs, puis en fonction de ce résultat, affecte le titre audio analysé à l'une des quatre classes de tempo perceptif. Pour la correction du tempo, les auteurs définissent un ensemble de règles heuristiques comme

par exemple : si un titre analysé est classé dans la classe de tempo perceptif *très lent*, et que le tempo estimé par un algorithme d'estimation du tempo est supérieur à 90 BPM, alors il faut diviser le tempo estimé par 2.

L'évaluation de cette méthode montre une amélioration systématique des résultats des algorithmes d'estimation du tempo.

Les techniques utilisées dans l'état de l'art se démarquent les unes des autres à plusieurs niveaux. D'une part, elles divergent par leurs objectifs : les méthodes utilisées permettent soit d'estimer le tempo de manière directe, soit de corriger un algorithme d'estimation du tempo en faisant une classification de la classe de tempo. D'autre part, on remarque des différences sur les descripteurs utilisés et sur les hypothèses musicales faites pour déduire le tempo.

Invariablement, on observe une amélioration des résultats de l'estimation du tempo lorsqu'on ajoute de l'information musicale aux algorithmes d'estimation existants.

1.3 Ircambeat

Étant donné que l'algorithme Ircambeat nous sert de base dans cette étude, il apparaît nécessaire de donner un aperçu de son fonctionnement général.

L'algorithme Ircambeat fonctionne en trois étapes. Tout d'abord, le logiciel calcule une fonction du flux d'énergie spectrale réassigné, notée y. Celle-ci mesure les variations du spectrogramme au cours du temps et permet de localiser les événements musicaux. Cette fonction sert de base pour l'estimation de la périodicité.

Pour cela, l'algorithme calcule séparément la transformée de Fourier discrète ($TF(y)$) et l'autocorrélation ramenée dans le domaine fréquentiel ($FM\text{-}ACF(y)$) de la fonction d'énergie. La fonction de périodicité est obtenue par le produit $TF(y) \times FM\text{-}ACF(y)$. Cette fonction permet d'observer des motifs caractéristiques selon le tempo et la métrique du signal audio analysé.

Enfin, l'algorithme de Viterbi permet de déterminer au cours du temps la séquence de tempi les plus probables à partir des motifs de la fonction de périodicité.

Le synoptique suivant regroupe les différentes étapes de fonctionnement d'Ircambeat.

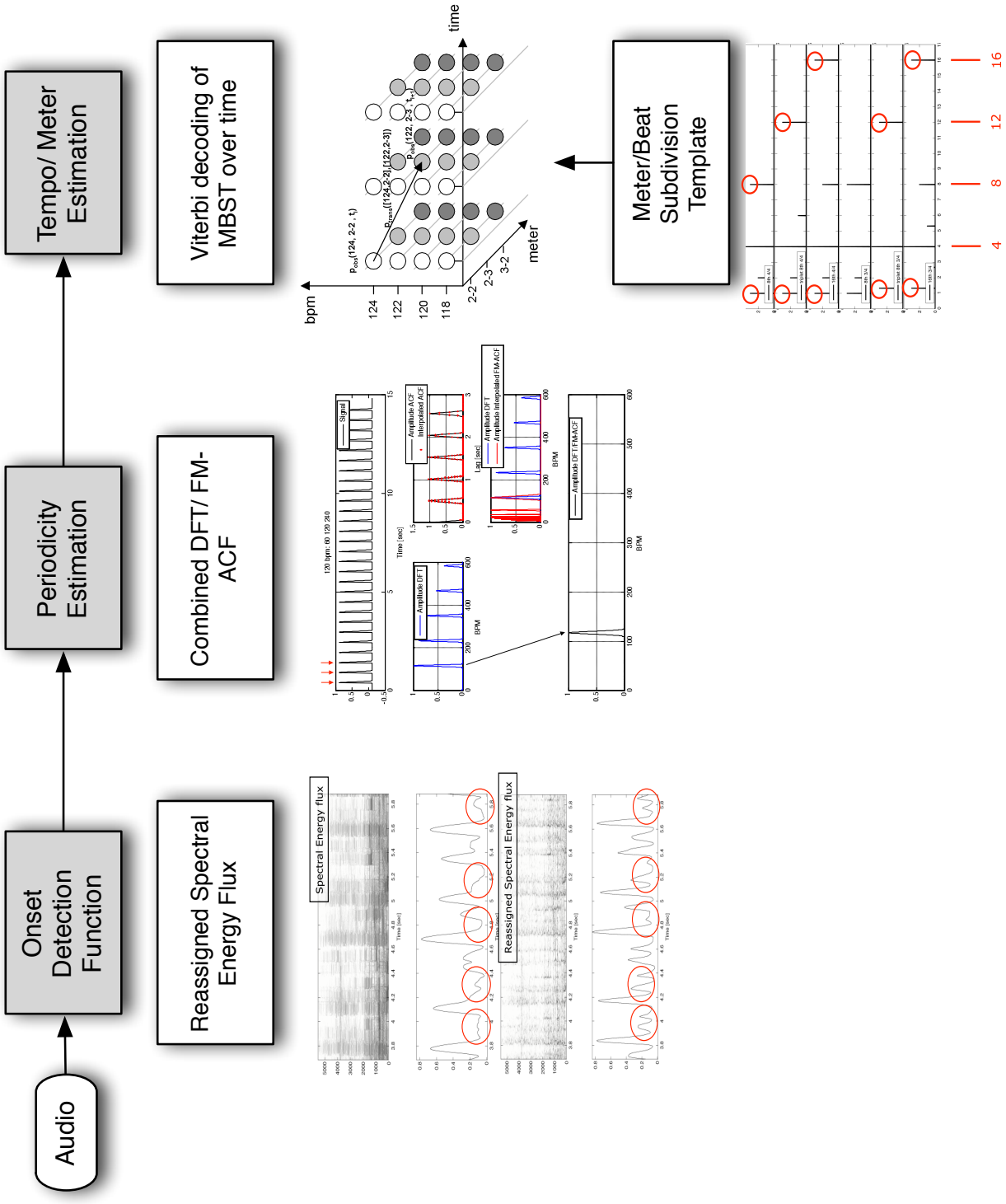

FIGURE 1.2 – Synoptique de l'algorithme **Ircambeat**, d'après la conférence donnée par Geoffroy Peeters à l'UPF. Octobre 2010

Chapitre 2

Analyse de la base de données Last.fm

La première étape de cette étude consiste à étudier le comportement d'Ircambeat vis-à-vis des erreurs d'octave du tempo. Pour ce faire, nous disposons du corpus Last.fm regroupant un grand nombre d'extraits audio annotés en tempo.

Dans ce chapitre, nous détaillerons tout d'abord la manière dont a été créée la base de données Last.fm et sur quels critères nous nous baserons pour sélectionner les titres et leurs annotations servant à l'étude finale. Ensuite, nous procéderons à l'évaluation d'Ircambeat pour mieux situer le problème d'erreur d'octave du tempo.

2.1 Présentation du corpus Last.fm

Last.fm est une webradio et un site internet proposant un système de collection de statistiques et de recommandations de musique. Outre les statistiques d'écoute, chaque utilisateur a la possibilité d'associer un mot-clé (on parle généralement de *tag*) à un artiste, un album ou une chanson, ce qui permet d'avoir de nombreuses informations concernant une musique précise.

Pour ses recherches sur l'estimation du tempo, Mark Levy (**Levy 2011**) a mis en place ce que nous appelons le corpus d'étude Last.fm : un ensemble d'extraits audio issus du fond musical Lasf.fm dont chaque titre a été annoté selon des caractéristiques liées au tempo.

2.1.1 Contenu musical

Dans un premier temps, il est intéressant de savoir quel type de musique, ou quels genres musciaux sont représentés dans notre corpus d'étude. Pour cela, nous utilisons l'API (Application Programming Interface) de Last.fm afin d'obtenir les informations en genre renseignées par les internautes, pour chaque titre. Ceci montre que plus des trois quarts des titres dans la base de données peuvent être expliqués par les grands courants musicaux de la musique populaire. La figure 2.1 représente la répartition des titres dans notre base de données selon le genre musical.

2.1.2 Annotations

La constitution des annotations pour la base de données Last.fm s'est faite sous la forme d'un formulaire sur une page web. Pour chaque page web, correspondant à un titre musical, l'internaute écoute un extrait audio d'une durée de trente secondes à une minute, et doit répondre à plusieurs questions.

Tout d'abord, celui-ci doit indiquer la classe de tempo de l'extrait qu'il écoute : *Fast*, noté

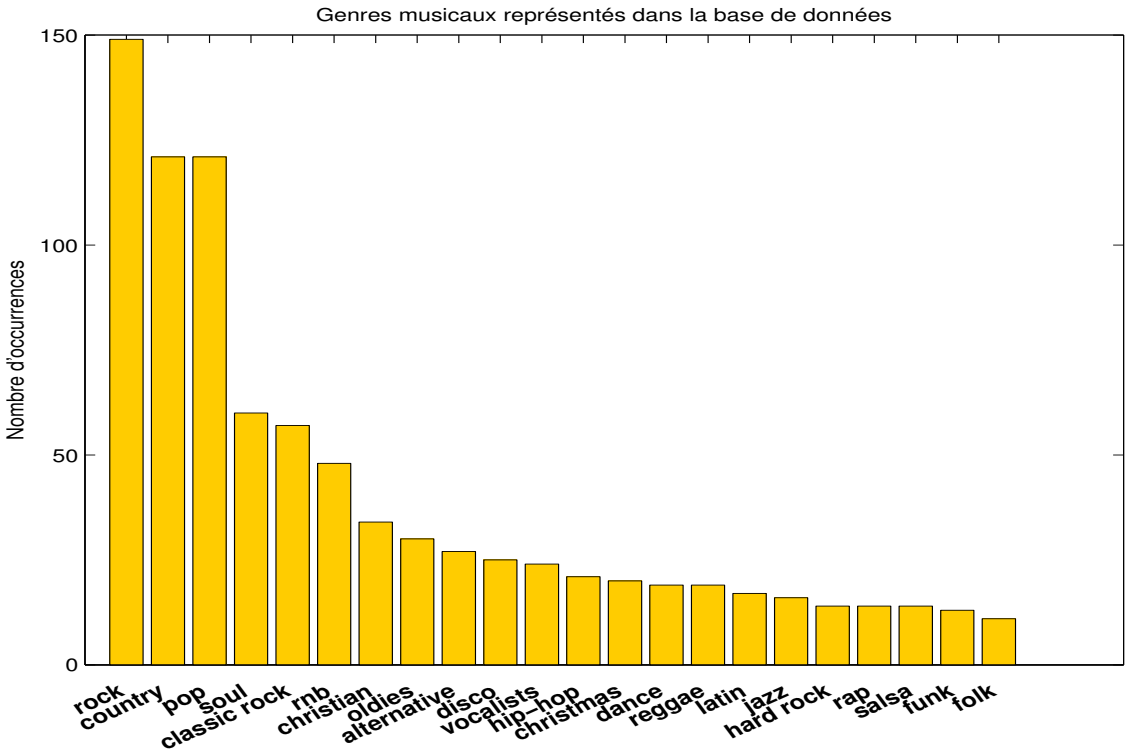

FIGURE 2.1 – Genres musicaux représentés dans la base de données

classe '3', *In between*, noté classe '2', *Slow*, noté classe '1' et *Hard to say*, noté classe '0'. Ensuite l'utilisateur doit marquer les temps du morceau qu'il écoute en appuyant sur la barre Espace de son clavier. Seuls les tempi entre 30 et 300 BPM sont conservés. Enfin, l'annotateur écoute un deuxième titre et doit comparer le tempo des deux extraits entendus. Dans notre étude, nous ne traitons pas ce type d'informations.

Les annotations de tempo et les classes de tempo indiqués par les internautes pour chaque titre constituent nos références pour tester Ircambeat.

Dans la suite de ce document, on parlera de *tempo estimé* pour faire référence au tempo estimé par Ircambeat, de *tempo perceptif* pour le tempo annoté par les internautes et de *classe de tempo* pour les indications de vitesse d'un morceau, également indiquées par les internautes dans la base de données Last.fm.

L'auteur de l'article expliquant la constitution de cette base de données est conscient du fait que la fiabilité des données n'est pas toujours garantie. En effet, un utilisateur peut annoter un titre sans même l'avoir écouté. C'est pour cette raison que nous serons vigilants quant à l'utilisation des annotations.

Les détails de la création du corpus Last.fm sont donnés par Levy dans (**Levy 2011**).

2.2 Exploitation des données

Compte tenu du fait que tous les titres ne sont pas annotés de manière fiable, nous devons définir des critères de sélection afin de ne retenir que des données exploitables.

On peut par exemple observer des titres n'ayant qu'une seule annotation, ou alors deux annotations mais non concordantes. De plus, étant donné que les internautes ne vont pas tous marquer les temps de la même manière, même si on possède plusieurs annotations, il faut réussir à dégager un tempo de référence parmi celles-ci. Par exemple, on peut observer pour un titre donné la séquence d'annotations : '137', '127', '135', '75', '133', '136'. Il faut alors définir une méthode pour déterminer le tempo qui nous servira de référence.

Afin d'obtenir des données fiables, nous avons décidé de ne conserver que les titres ayant au minimum trois annotations en tempo et en classe de tempo. Ensuite, pour déterminer le tempo annoté qui nous servira de référence, nous avons introduit l'estimation du tempo majoritaire. À partir de la séquence d'annotations, on construit un clustering hiérarchique ascendant en autorisant la création d'un cluster si la différence entre les moyennes des deux clusters est suffisamment faible.

$$|\bar{x_r} - \bar{x_s}| < 6\% \max\{\bar{x_r}; \bar{x_s}\}$$

où $\bar{x_r}$ et $\bar{x_s}$ représentent les moyennes des clusters r et s respectivement.

$$\bar{x_r} = \frac{1}{n_r} \sum_{i=1}^{n} x_{ri}$$

Notre algorithme de clustering utilise donc une distance euclidienne appliquée aux centroïdes de nos clusters (*centroid linkage*).

Cette méthode nous permet de regrouper les annotations proches. Si un cluster dominant se dégage, c'est-à-dire si celui-ci regroupe un maximum d'annotations, c'est le tempo moyen de ce cluster qui est choisi comme tempo de référence. Sinon, le titre est rejeté car il n'y a pas de tempo majoritaire. On réalise cette même opération pour déterminer la classe de tempo majoritaire.

Le tempo mis en évidence est celui qu'une majorité d'annotateurs ont estimé et correspond à notre définition du tempo perceptif. Cette étape est primordiale car elle permet de renforcer la certitude du tempo de référence.

2.3 Analyse avec Ircambeat

Après cette étude préalable, on peut désormais étendre l'estimation du tempo par Ircambeat sur chaque titre de la base de données. On compare ensuite les résultats de cette estimation avec les tempi annotés. On utilise pour cela une une fenêtre de tolérance de 6% du tempo annoté Ta. On peut trouver dans d'autres études l'utilisation d'une fenêtre de tolérance plutôt de 4%. Cependant, comme nous avons affaire à des annotations dont la précision n'est pas garantie, nous avons jugé qu'une fenêtre de 4% était trop discriminante.

Pour plus de lisibilité, on regroupe les résultats obtenus en plusieurs catégories. On note Te, le tempo estimé par Ircambeat et Ta, le tempo annoté. Les catégories choisies pour expliquer le comportement d'Ircambeat sont les suivantes :

– Classe Estimation correcte

– Classe Erreurs d'octave $Te = kTa$ (le tempo estimé est un multiple du tempo annoté. $k \in \{1/3; 1/2; 2/3; 2; 3\}$)
– Classe Incertitude-erreur (erreurs ne correspondant pas à des erreurs d'octave)

Le tableau 2.1 indique quels types d'erreur sont faits par Ircambeat et la figure 2.2 donne une représentation plus globale des erreurs d'estimation commises par Ircambeat.

	Nombre de titre	%
Estimation correcte	912	67.26
Erreur d'octave $T_e = 2T_a$	280	20.65
Erreur d'octave $T_e = T_a/2$	38	2.80
Erreur d'octave $T_e = 2/3T_a$	27	1.99
Erreur d'octave $Te = 3/2T_a$	3	0.22
Erreur d'octave $T_e = 3T_a$	18	1.33
Erreur d'octave $T_e = T_a/3$	0	0
Reste / Incertitude-erreur	78	5.75
Total	1356	100

TABLE 2.1 – Résultats de l'analyse d'Ircambeat sur la base de données, après tri dans la base de données. La tolérance pour qu'une estimation soit juste est ±6% du tempo annoté

On se sert également de l'information de classe de tempo pour déterminer dans quelle classe les erreurs d'octave sont les plus fréquentes. Ces résultats sont donnés dans le tableau 2.2. De la même manière que le figure 2.2, les représentations des estimations d'Ircambeat en fonction des annotations pour chaque classe de tempo se trouvent en annexes **B**.

	Classe *Slow*		Classe *Medium*		Classe *Fast*	
Type d'estimation	N titre	%	N titre	%	N titre	%
Estimation correcte	283	65.21	424	65.43	186	67.88
Erreur d'octave $T_e = 2T_a$	108	24.28	138	16.66	52	18.97
Erreur d'octave $T_e = T_a/2$	7	1.61	20	3.08	8	2.91
Erreur d'octave $T_e = 3T_a$	4	0.92	4	0.61	5	1.82
Erreur d'octave $T_e = T_a/3$	0	0	0	0	0	0
Total	434	100	648	100	274	100

TABLE 2.2 – Analyse des estimations d'Ircambeat, selon la classe de tempo annoté

2.4 Conclusions

Cette étude s'est révélée très instructive sur le comportement général d'Ircambeat vis-à-vis des erreurs d'octave, à la fois sur le plan qualitatif et quantitatif. Tout d'abord, on remarque que les erreurs d'octave du tempo interviennent majoritairement pour des extraits audio relativement lents, comme le montre le taux d'erreurs pour la classe tempo *Slow*. Ensuite, les erreurs d'octave du tempo sont à 83.5% des *surestimations* du tempo orignal ce qui va dans le sens de la première remarque. Enfin, les erreurs d'octave représentent un peu plus de 20 % des erreurs totales. Résoudre ce problème permettrait de gagner grandement en qualité d'estimation.

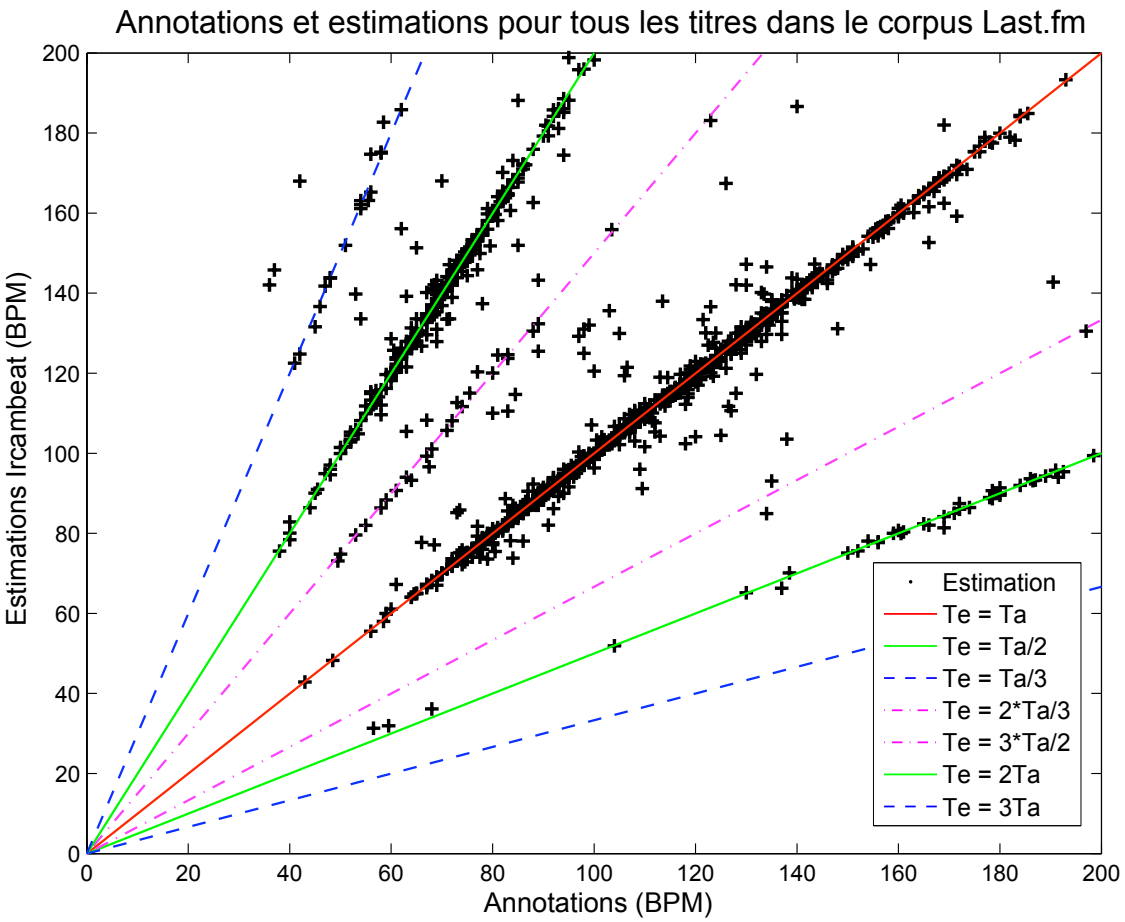

FIGURE 2.2 – Observation des estimations d'Ircambeat en fonction des annotations.

Chapitre 3

Étude des indices acoustiques

La deuxième partie de notre recherche s'attache à l'étude d'indices acoustiques permettant de donner de nouvelles informations relatives au tempo. Les descripteurs choisis sont les *variations de Chroma*, la *balance spectrale*, la *similarité acoustique* et la *fonction d'énergie*.

3.1 Variations de Chroma

Avec ce descripteur, on fait l'hypothèse que les changements harmoniques au cours d'un extrait musical peuvent être utilisés pour estimer le tempo. Le descripteur *variations de Chroma* permet de faire une telle estimation en suivant deux phases. Dans un premier temps on calcule les vecteurs de Chroma qui permettent de nous donner une indication sur le contenu tonal du signal. Dans un second temps, on cherche les variations du contenu harmonique au cours du temps, mettant ainsi en évidence les changements d'accords.

Le principe de ce descripteur est donné dans (**Peeters & Papadopoulos 2011**).

3.1.1 Vecteurs de Chroma

Les vecteurs de Chroma, introduits par Wakefield dans (**Wakefield 1999**), également proposés par Fujishima (**Fujishima 1999**) sous le nom de Pitch Class Profile, sont des vecteurs représentant, pour un signal audio à un instant t, l'intensité de chacun des douze demi-tons d'une gamme diatonique.

Pour éviter toute confusion par la suite, nous rappelons la distinction entre la hauteur tonale et les Chroma. La hauteur tonale décrit la hauteur d'un son en fonction de sa fréquence. Dans le cas général, si la fréquence fondamentale d'un son augmente, la hauteur tonale de ce son augmentera également. Les Chroma cependant, décrivent une représentation cyclique des sons en regroupant tous les demi-tons d'une gamme diatonique quelque soit leur octave. Ainsi, deux sons séparés par un nombre entier d'octave, auront la même valeur de Chroma mais pas la même hauteur tonale.

La méthode sur laquelle nous nous basons pour calculer les vecteurs de Chroma est présentée dans (**Peeters 2006**). Dans cet article, les vecteurs de Chroma sont obtenus après calcul de la transformée de Fourier à court terme du signal $S(f_k, t)$, puis en faisant la correspondance entre les fréquences f_k d'une trame et son équivalent dans la gamme diatonique. La correspondance entre la fréquence f_k à l'instant t_i et l'un des douze demi-tons $l \in [0, 11]$, où l est un entier, est définie comme :

$$n(f_k, t_i) = 12 \log_2 \left(\frac{f_k}{f_{ref}} \right) \mod 12$$

avec f_{ref} est la fréquence de référence. On obtient le vecteur de Chroma pour l'instant t_i en sommant toutes les contributions de chaque fréquence associée à un des demi-tons :

$$\boldsymbol{c}(l, t_i) = \sum_{f_k \text{ tel que } n(f_k, t_i)=l} |S(f_k, t_i)|^2 \; l = 0, 1, ..., 11$$

Nous obtenons ainsi un vecteur à douze dimensions. En opérant ainsi ce calcul sur tout le signal, on obtient la matrice $\boldsymbol{C}(l, t)$ donnant une représentation du signal en fonction de son contenu harmonique.

3.1.2 Variations des vecteurs de Chroma

Les vecteurs de Chroma nous fournissent une représentation du signal selon une échelle tonale. Pour mettre en évidence les changements d'accords, nous allons maintenant calculer les variations du contenu tonal au cours du temps. Si un changement d'accord survient à un instant t, le contenu harmonique sera très différent à gauche et à droite de t. Nous comparons donc les valeurs dans $\boldsymbol{C}(l, t)$ à la gauche de t_i et à sa droite en utilisant une fenêtre d'observation notée L pour la fenêtre temporelle à gauche de t_i et R la fenêtre d'observation à droite de t_i.

Dans notre cas, on fait l'hypothèse que les changements d'accords sont plus susceptibles de survenir tous les quatre temps. Nous considérons également que le contenu harmonique est homogène sur toute la mesure. Nous devons donc définir nos fenêtres L et R en fonction de ces hypothèses. Étant donné que nous ne connaissons pas le tempo, nous calculons les variations des vecteurs de Chroma pour plusieurs hypothèses de tempo T_h, où $T_h \in [30, 200]$.

On note ainsi : $L = [t_i - \alpha T_h, t_i]$ et $R = [t_i, t_i + \alpha T_h]$. Les fenêtres d'observation sont de cette manière d'une durée multiple du tempo T_h. Nous faisons l'hypothèse que les accords changent tous les quatre temps (une fois par mesure dans le cas d'une mesure à 4/4). Ceci nous amène à choisir $\alpha = 4$.

La comparaison entre le contenu harmonique à droite et à gauche de t_i se fait par calcul d'une distance cosinusoidale entre les vecteurs moyennes $\boldsymbol{\mu}_L$ de $\boldsymbol{C}(l, L)$ et $\boldsymbol{\mu}_R$ $\boldsymbol{C}(l, R)$. Si $\boldsymbol{\mu}_L$ et $\boldsymbol{\mu}_R$ sont orthogonaux, la distance vaut 1. On définit alors :

$$d(L, R) = 1 - \frac{\boldsymbol{\mu}_L \cdot \boldsymbol{\mu}_R}{||\boldsymbol{\mu}_L|| \cdot ||\boldsymbol{\mu}_R||}$$

La figure 3.1 schématise le principe de détection des changements d'accords par comparaison du contenu harmonique.

En faisant cela pour tout le morceau, on obtient en fonction du temps et de chaque hypothèse de tempo, les changements d'accords de l'œuvre musicale. Par une transformée de Fourier, on peut alors mettre en évidence la fréquence de variation harmonique et donc estimer le tempo. Il est à noter que du fait de l'hypothèse d'un changement d'accord tous les quatre temps, la fréquence (en BPM) résultant du descripteur Chroma sera le quart du tempo réel (toujours dans la cas où cette hypothèse est vérifiée).

La figure 3.2 montre les changements d'accords mis en évidence au cours du temps ainsi que la fréquence des variations harmoniques.

Par la suite on notera $\boldsymbol{C}_i(\omega, T_h)$ ou $\boldsymbol{C}_i$ la matrice issue du descripteur des variations de Chroma pour un titre i représentant le fréquence de changement d'accords ω en fonction des hypothèses de tempo T_h.

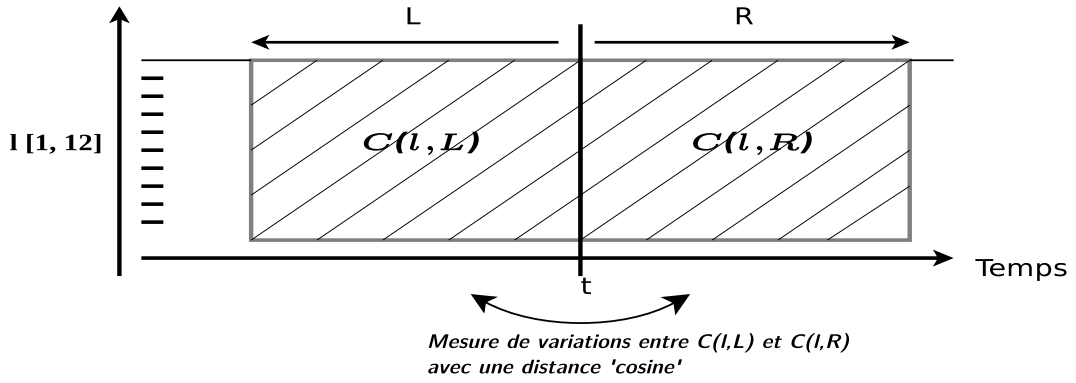

FIGURE 3.1 – Schéma représentant le calcul de la variation de Chroma

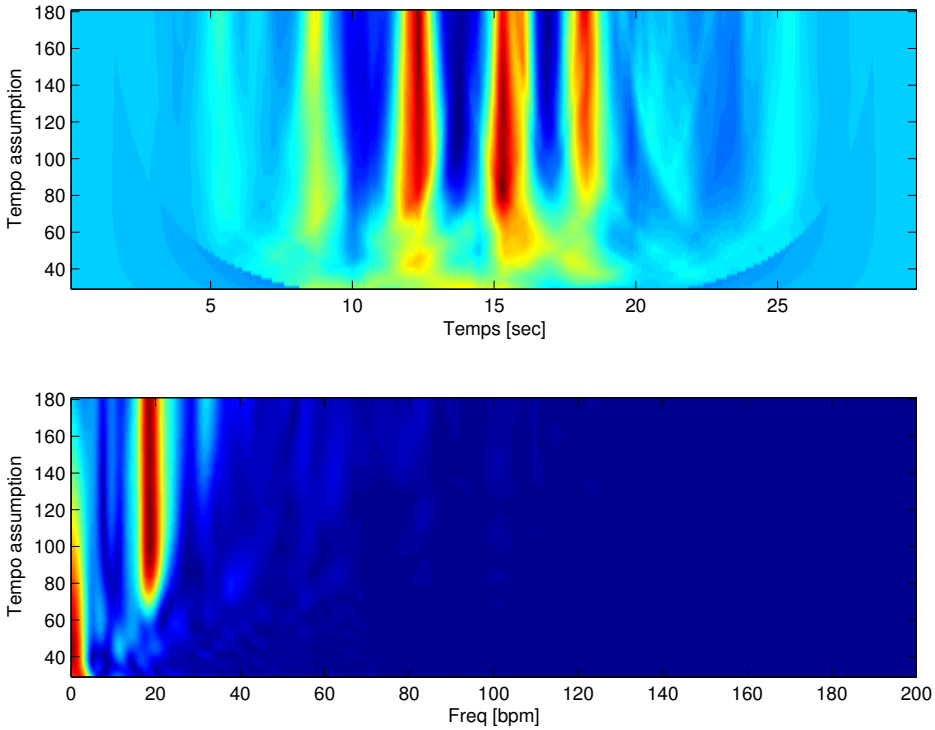

FIGURE 3.2 – En haut : variations des vecteurs de Chroma au cours du temps, selon chaque hypothèse de tempo T_h. En bas : transformée de Fourier des variations de Chroma selon T_h. Tempo annoté : 73 BPM. La période estimée est est 18.47 BPM donc le tempo estimé sera quatre fois cette valeur : 73.88 BPM. Titre : Adam Lambert - Time for miracles

3.2 Balance spectrale

La balance spectrale met en évidence les variations de concentration d'énergie spectrale. En musique populaire, on remarque qu'on a souvent un motif rythmique particulier à la batterie d'alternance grosse caisse, caisse claire. D'un point de vue spectral, ceci implique que l'énergie

spectrale sera concentrée en basse fréquence sur les premiers et troisièmes temps, puis dans les hautes fréquences sur les deuxièmes et quatrièmes temps (dans le cas d'une mesure à 4/4). La mesure des variations de concentration d'énergie spectrale nous permet donc d'identifier la vitesse du motif rythmique de la batterie et ainsi d'estimer le tempo.

Le principe de ce descripteur est donné dans (**Peeters & Papadopoulos 2011**).

À chaque instant t_i et pour une hypothèse de tempo $T_h \in [30; 200]$, on calcule le rapport des énergies spectrales entre les hautes et les basses fréquences. On utilise pour cela une fenêtre temporelle centrée sur t_i d'une largeur L et une fréquence de coupure $kmax$:

$$r(t_i) = \frac{\sum_{t=t_i-L/2}^{t_i+L/2} \sum_{k=kmax}^{N/2} |S(\omega_k, t)|^2}{\sum_{t=t_i-L/2}^{t_i+L/2} \sum_{k=1}^{kmax} |S(\omega_k, t)|^2}$$

Où N est le nombre de points de la transformée de Fourrier. L correspond à $T_h/2$ et $kmax$ à 150Hz.

On cherche ensuite à déterminer si l'instant t_i correspond à un premier, deuxième, troisième ou quatrième temps en observant les valeurs de r aux instants multiples de l'hypothèse de tempo T_h. Formellement, on calcule :

$$r'_j(t_i) = r(t_i - (j-1)T_h)$$

où $j \in \{1; 2; 3; 4\}$. On normalise les valeurs de $r'_j(t_i)$ de manière à avoir $\sum_j r'_j(t_i) = 1$, et comme on souhaite avoir des grandes valeurs sur les premiers et troisièmes temps, on impose $r''_j(t_i) = 1 - r'_j(t_i)$. On calcule enfin :

$$r_{tot}(t_i) = r''_{j=1;3}(t_i) - r''_{j=2;4}(t_i)$$

De cette manière, si l'hypothèse de tempo T_h est correcte, et que t_i correspond à un premier temps, alors $r_{tot}(t_i)$ sera grand, puis faible pour $r_{tot}(t_i + T_h)$ et ainsi de suite.

Ce calcul nous permet d'obtenir une matrice $\boldsymbol{B}(t_i, T_h)$ représentant les variations de $r_{tot}(t_i)$ au cours du temps et pour chaque hypothèse de tempo.

Ensuite, en calculant la transformée de Fourier de r_{tot} pour chaque T_h, on obtient la matrice $\boldsymbol{B}(\omega, T_h)$ indiquant la fréquence de variation de la concentration de l'énergie spectrale. (w en BPM). La figure 3.3 donne un exemple des matrices $\boldsymbol{B}(t_i, T_h)$ et $\boldsymbol{B}(w, T_h)$.

Il est à noter que ce descripteur permet d'estimer la période entre les temps 1 et 3 (ou 2 et 4) et que par conséquent, la fréquence de variation de l'énergie spectrale sera la moitié du tempo estimé.

Par la suite on notera $\boldsymbol{B}_i(\omega, T_h)$ ou $\boldsymbol{B}_i$ la matrice issue du descripteur de balance spectrale pour un titre i représentant la fréquence d'alternance de l'énergie spectrale ω en fonction des hypothèses de tempo T_h.

3.3 Similarité acoustique

La perception du tempo peut être liée aux répétitions à court terme d'événements musicaux, comme par exemple les répétitions de sons ayant la même hauteur ou le même timbre. Le descripteur de similarité acoustique permet de mettre en évidence ce genre de répétitions et nous servira pour inférer le tempo.

Le descripteur de similarité acoustique repose sur le calcul de la matrice d'auto-similarité du signal audio.

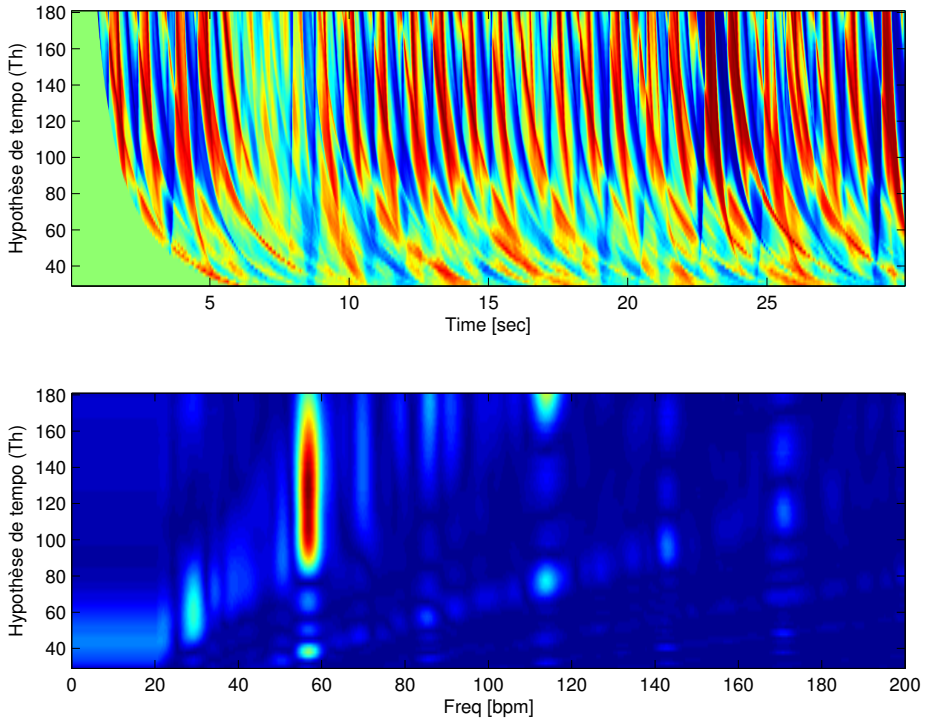

FIGURE 3.3 – En haut : variations temporelles de r selon chaque hypothèse de tempo T_h. En bas : transformée de Fourier de r selon T_h. Tempo annoté = 114 BPM. La période estimée est 56.89 donc le tempo sera deux fois cette valeur : 113.78 BPM. Titre : Third Eye Blind - Never let you go

3.3.1 Matrice d'auto-similarité

La matrice d'auto-similarité nous permet d'avoir une représentation en deux dimensions des événements répétitifs au cours d'un extrait audio. Pour construire cette matrice, nous nous basons sur la méthode proposée dans (**Peeters 2007a**).

En premier lieu, on calcule pour chaque instant du signal t_i les descripteurs permettant de caractériser un événement sonore. Nous utilisons ici les 12 coefficients MFCC et le *spectral contrast* (proposé dans (**Jiang et al. 2002**)) pour capter les informations de timbre ainsi que les Chroma pour obtenir les informations harmoniques.

Pour chaque descripteur, on construit ensuite la matrice d'auto-similarité $\boldsymbol{S}(t_i, t_j)$ en effectuant une mesure de distance entre les vecteurs d'observation de deux segments audio aux instants t_i et t_j.

Soit $\boldsymbol{v}_{t_i}$ le vecteur d'observation issu d'un des descripteurs à l'instant t_i, on définit la matrice d'auto-similarité par :

$$\boldsymbol{S}(t_i, t_j) = d(\boldsymbol{v}_{t_i}, \boldsymbol{v}_{t_j})$$

La distance utilisée dans notre cas pour la mesure de similarité est une distance euclidienne. On calcule ainsi trois matrices d'auto-similarité pour chacun des descripteurs. Ces matrices sont ensuite normalisées pour prendre leurs valeurs entre 0 et 1 puis sont sommées afin de n'obtenir qu'une seule matrice d'auto-similarité.

3.3.2 Matrice de retard et déduction du tempo

On convertit ensuite la matrice d'auto-similarité $\boldsymbol{S}(t_i, t_j)$ en matrice de retard $\boldsymbol{L}(t_i, l_j)$ où $l_j = t_j - t_i$ représente le retard entre les répétitions. Dans la matrice de retard, une grande valeur dans la ligne l_j signifie que des répétitions surviennent régulièrement avec un retard l_j.

Enfin, on somme la matrice de retard à travers chaque temps t_i pour obtenir un vecteur représentant la quantité de répétitions selon le retard. On calcule ensuite la transformée de Fourier de ce vecteur pour mettre en évidence la fréquence de répétition dans le signal.

La figure 3.4 illustre les différentes étapes de calcul du descripteur similarité acoustique avec la matrice d'auto-similarité, la matrice de retard et enfin la fréquence de répétition.

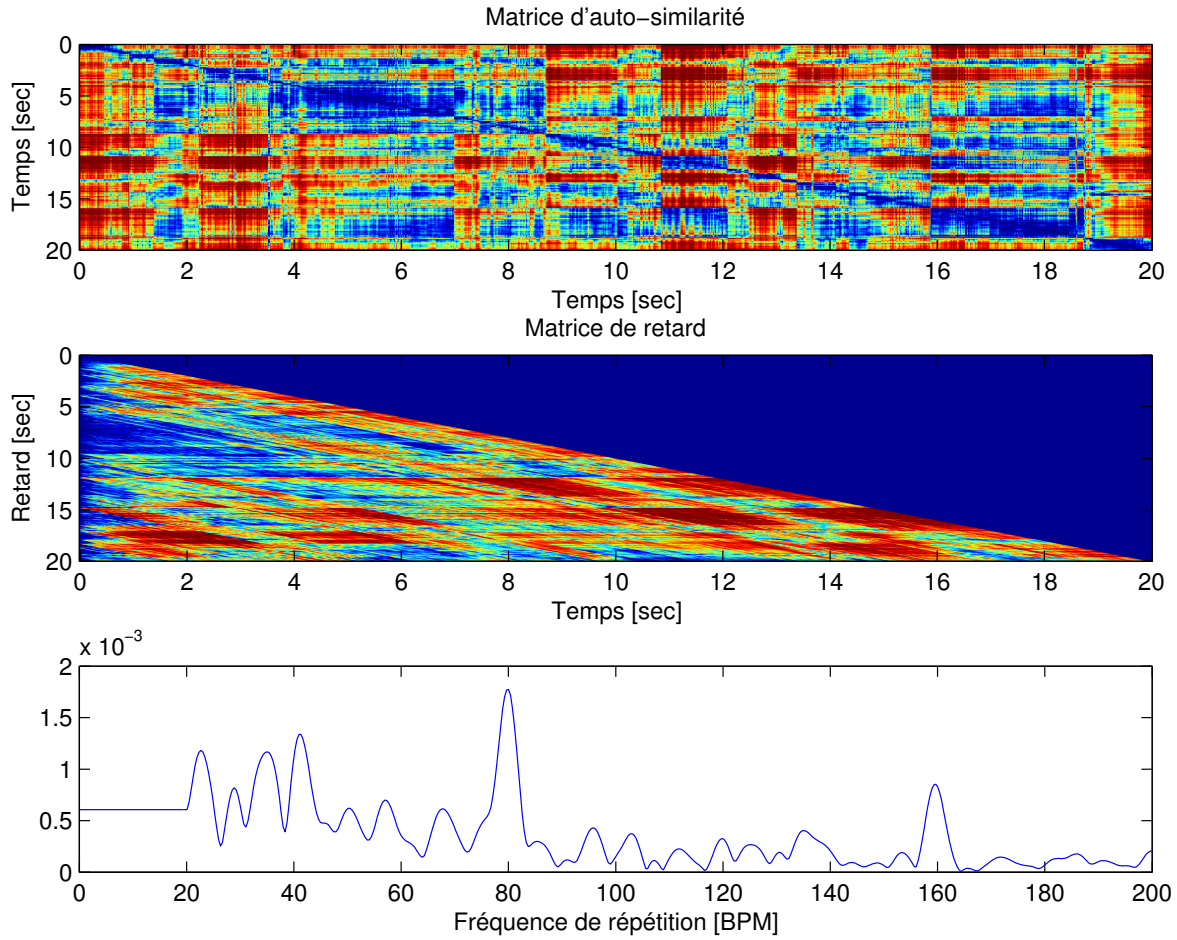

FIGURE 3.4 – Haut : matrice d'auto-similarité. Milieu : matrice de retard. Bas : Fréquence de répétition. Tempo annoté : 81 BPM. Le tempo estimé par le descripteur est 79.8 BPM. Titre : Abba - Lay all your lovin on me

3.4 Fonction d'énergie

La fonction d'énergie qui nous sert de descripteur ici est en réalité déjà présente dans Ircambeat et est présentée dans (**Peeters 2007b**). Toutefois, nous l'utilisons comme descripteur afin d'avoir une mesure de la périodicité du signal. Les trois autres descripteurs nous apportent une indication du tempo grâce aux changements d'accords (*variations de Chroma*), au motif rythmique particulier de la batterie (*balance spectrale*) et aux répétitions à court terme d'événements musicaux, basés sur la hauteur et le timbre (*similarité acoustique*). Cette fonction d'énergie nous permettra d'obtenir une indication du tempo à partir d'une autre caractéristique musicale, la périodicité globale d'un morceau.

Brièvement, le calcul de cette fonction d'énergie est basé sur le calcul du flux d'énergie spectrale proposé par Laroche dans (**Laroche 2003**). Le flux d'énergie spectrale met en évidence les variations fréquentielles au cours du temps et peut s'écrire de la manière suivante :

$$E(i) = \sum_{f} |X(f, t_i)| - |X(f, t_{i-1})|$$

Où $X(f, t_i)$ représente la transformée de Fourier à court terme du signal x à l'instant t_i. Bien que cette fonction mette bien en évidence les attaques de notes, elle souffre d'un problème souvent rencontré en traitement du signal, le problème de résolution fréquentielle contre la résolution temporelle. En effet, pour calculer $E(i)$, on aurait tendance à utiliser une fenêtre d'analyse assez grande pour avoir une résolution fréquentielle suffisante, mais ce serait au détriment de la résolution temporelle, qui est cruciale dans notre cas.

La fonction d'énergie que nous utilisons limite en partie l'influence du paradoxe résolution temporelle contre résolution fréquentielle en définissant le calcul du flux d'énergie spectrale réassigné. Celui-ci se base sur le calcul du flux d'énergie spectrale à partir d'un spectrogramme dont les fréquences ont été repositionnées. Il en résulte une meilleure localisation des attaques de notes.

La fonction d'énergie est donc une fonction continue qui indique les attaques des notes. À partir de cette fonction, on peut estimer la fréquence la plus significative concernant l'apparition des notes et donc déduire le tempo. La figure 3.5 donne un exemple de la fonction d'énergie calculée pour un signal audio ainsi que le tempo estimé à partir de cette fonction.

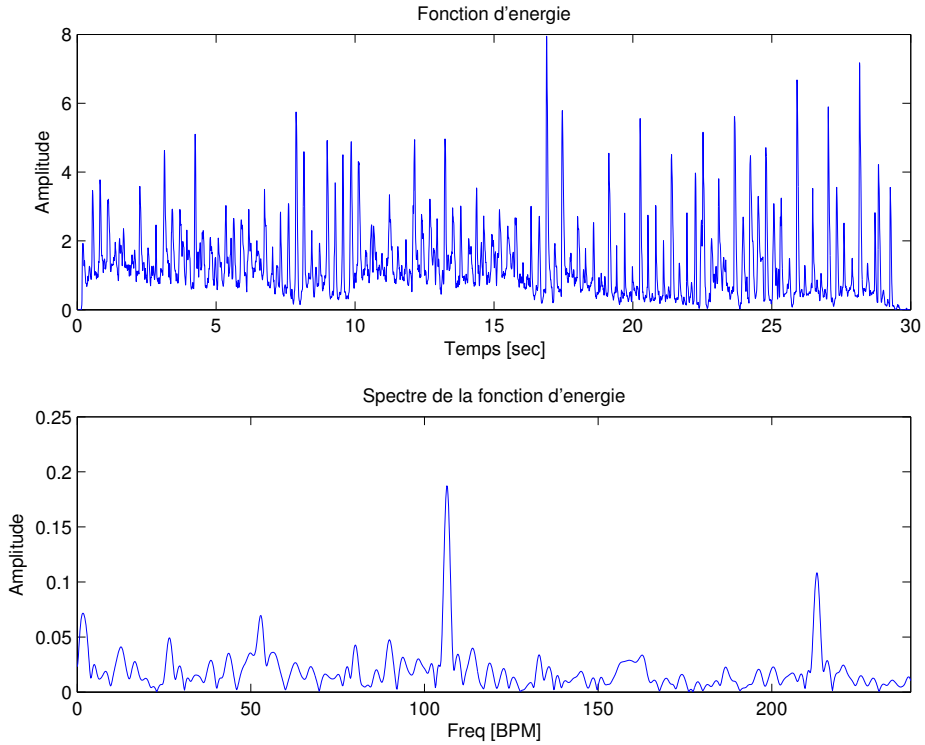

FIGURE 3.5 – Extrait audio : 'Til Tuesday - Voices Carry. BPM annoté 106. BPM estimé 106.3

Chapitre 4

Combinaison des données

Les descripteurs audio que nous avons présentés nous permettent d'avoir une estimation du tempo en se basant sur des propriétés musicales, à savoir que les changements d'accords, le motif rythmique de la batterie, la répétition à court terme et la périodicité sont autant d'éléments utiles pour inférer le tempo. Nous sommes toutefois conscients du fait que ces descripteurs ne sont pas infaillibles -auquel cas le problème de l'estimation du tempo serait résolu- et que dans certains cas leurs estimations peuvent être biaisées.

Cependant, nous pensons qu'utiliser ces descripteurs en plus de l'estimation d'Ircambeat pourrait améliorer la qualité de l'estimation du tempo.

Le problème qui se présente alors devant nous est de définir une méthode capable d'allier le résultat des descripteurs que nous venons de décrire avec l'estimation d'Ircambeat.

La technique retenue est une approche statistique basée sur un modèle de mélange de Gaussiennes (GMM) et plus précisément sur la régression par modèle de Gaussiennes. On peut retrouver cette technique en traitement de la parole comme dans (**En-Najjary et al. 2003**) où les auteurs arrivent à estimer la hauteur d'un son uniquement à partir de l'enveloppe spectrale du signal. Le principe est de construire un modèle de mélange de Gaussiennes sur des données composées d'observations et de valeurs cibles (c'est-à-dire les données qu'on cherche à estimer). Dans l'article, les vecteurs d'observation sont les 12 coefficients MFCC et les valeurs cibles sont les fréquences en Hz.

Dans notre cas, les observations seront constituées de l'estimation d'Ircambeat et des résultats des descripteurs, et l'information que l'on cherchera à atteindre sera le tempo annoté.

Dans ce chapitre nous présenterons le cadre de travail des modèles de mélange de Gaussiennes et plus particulièrement la méthode de régression par GMM. Nous détaillerons également la manière dont nous définissons nos vecteurs d'observation, point-clé pour l'utilisation de cette technique. Pendant cette étude, nous avons eu l'occasion de tester plusieurs approches. Nous ne présenterons ici que celles ayant les résultats les plus intéressants.

4.1 Modèle de mélange de Gaussiennes (GMM)

On définit la densité de probabilité d'une variable $\boldsymbol{z}$, où $\boldsymbol{z}$ est un vecteur de dimension D, suivant un modèle GMM d'ordre K par :

$$p(\boldsymbol{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Chaque densité $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ est appelée *composante du mélange*. $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ est la densité de probabilité de la loi normale donnée par :

$$\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_k) \right\}$$

où $\boldsymbol{\mu}_k$ est le vecteur moyenne de dimension D, $\boldsymbol{\Sigma}_k$ la matrice de covariance de dimension $D \times D$ et $|\boldsymbol{\Sigma}|$ est le déteminant de $\boldsymbol{\Sigma}$.

Les π_k sont les *coefficients de mélange*. π_k est la probabilité a priori que $\boldsymbol{z}$ soit généré par la k^{eme} composante du modèle, avec :

$$\sum_{k=1}^{K} \pi_k = 1, \text{ et } \pi_k \geqslant 0$$

Pour cette étude, nous nous appuyons sur la toolbox Matlab proposée par Sylvain Calinon (**Calinon 2009**).

4.2 Régression par modèle de mélange de Gaussiennes

Soit $\boldsymbol{x} = [x_1 x_2 ... x_N]$ un vecteur contenant les données cibles et $\boldsymbol{Y} = [\boldsymbol{y}_1 \boldsymbol{y}_2 ... \boldsymbol{y}_N]$ les vecteurs d'observation associés aux valeurs cibles. On définit le vecteur $\boldsymbol{z}$ comme étant la combinaison des valeurs x_i et des vecteurs $\boldsymbol{y}_i$:

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{y}_i \\ x_i \end{bmatrix}$$

On estime les paramètres du modèle de Gaussiennes (π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$) de $p(\boldsymbol{z})$ (qui est la densité jointe $p(x_i, \boldsymbol{y}_i)$), grâce à l'algorithme EM (Expectation/Maximization). Avec ces paramètres, on peut déduire la valeur x à partir d'un vecteur d'observation $\boldsymbol{y}$ grâce à la fonction de prédiction :

$$F(\boldsymbol{y}) = \mathbb{E}[x|\boldsymbol{y}] = \sum_{k=1}^{K} h_k \left[\boldsymbol{\mu}_k^x + \boldsymbol{\Sigma}_k^{xy} (\boldsymbol{\Sigma}_k^{yy})^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_k^y) \right]$$

avec

$$h_k(\boldsymbol{y}) = \frac{\pi_k \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_k^y, \boldsymbol{\Sigma}_k^{yy})}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_k^y, \boldsymbol{\Sigma}_k^{yy})}$$

qui est la probabilité a posteriori qu'un vecteur $\boldsymbol{y}$ appartienne à la k^{eme} composante du mélange. La matrice de covariance et le vecteur moyenne sont de la forme :

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^{yy} & \boldsymbol{\Sigma}_k^{yx} \\ \boldsymbol{\Sigma}_k^{xy} & \boldsymbol{\Sigma}_k^{xx} \end{bmatrix} \qquad \text{et} \qquad \boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_k^y \\ \boldsymbol{\mu}_k^x \end{bmatrix}$$

La démonstration de cette fonction se trouve en annexe **A**, page 40.

Dans notre cas, le vecteur $\boldsymbol{x}$ correspond aux annotations de tempo. Ce sont les données que l'on cherche à atteindre. Notre vecteur $\boldsymbol{y}$ sera composé des résultats de nos descripteurs audio.

4.3 Définition des vecteurs d'observation

4.3.1 Plusieurs approches

La manière dont nous définissons les vecteurs d'observation à partir du résultat des descripteurs est d'une grande importance pour l'utilisation de notre méthode. Cela va influer sur la quantité et le type d'informations que nous allons fournir à notre modèle. C'est aussi lors de cette étape que nous effectuons la phase nécessaire de réduction de dimension des vecteurs d'observation. En effet, les méthodes du type GMM sont sensibles aux espaces à grand nombre de dimensions. En résumé, pendant cette étape nous devons donc définir quelle(s) information(s) retenir de nos descripteurs.

Deux approches ont été envisagées pour la création de nos observations. La première est de considérer l'information telle qu'elle nous apparaît à l'issue du descripteur, en échantillonnant le vecteur à l'aide d'un banc de filtres. Cette méthode a été utilisée dans (**Peeters & Flocon-Cholet 2012**) où les vecteurs sont échantillonnés grâce à vingt filtres espacés de manière logarithmique entre les tempi 32 et 208. De cette manière, on réalise un sous-échantillonnage du vecteur et on conserve une information réduite de l'estimation du tempo faite par le descripteur.

La deuxième approche que nous avons étudiée, et qui sera retenue par la suite, a cette fois pour but d'exprimer une corrélation entre l'estimation faite par le descripteur et l'estimation d'Ircambeat. En cherchant à traduire la notion d'accord ou de désaccord entre le résultat des descripteurs et Ircambeat, on se rapproche finalement d'une indication d'erreur d'octave du tempo de la part d'Ircambeat.

La première approche pourrait être qualifiée d'approche *absolue* dans le sens où l'on considère l'information de manière brute, et la deuxième approche pourrait être qualifiée de *relative* car dans ce cas, on essaie d'obtenir une information de corrélation entre l'estimation faite par Ircambeat et celle des descripteurs. Nous reviendrons un peu plus loin sur les différences entre les deux méthodes.

4.3.2 Valeurs cibles x

On rappelle que pour chaque titre m_i présent dans notre corpus d'étude, nous avons les informations associées :

$$m_i = \begin{cases} Ta_i & \text{Tempo annoté} \\ Te_i & \text{Tempo estimé par Ircambeat} \end{cases}$$

Nous reprenons les notations que nous avons utilisées plus haut pour définir les données partitionnées $\boldsymbol{z} = [\boldsymbol{y}\ x]^T$. Les valeurs que l'on cherche à estimer, nos valeurs cibles x, correspondent aux tempi annotés. Ces tempi sont eux-mêmes issus de la phase d'analyse des annotations du corpus Last.fm vue dans le chapitre **2.**

4.3.3 Vecteurs d'observation $\boldsymbol{y}$

Les vecteurs d'observation $\boldsymbol{y}$ sont créés à partir du résultat des descripteurs selon l'approche que nous avons définie plus haut. Cette méthode qui nous permet d'obtenir une corrélation entre l'estimation d'Ircambeat et l'estimation d'un descripteur fonctionne en deux étapes : un changement d'échelle de l'axe des fréquences puis un échantillonnage des vecteurs aux points caractéristiques.

On note pour un titre i, $\boldsymbol{C}_i$, la matrice obtenue par le descripteur Chroma, $\boldsymbol{B}_i$ la matrice issue du descripteur balance spectrale, $\boldsymbol{s}_i$ le vecteur issu du descripteur similarité et $\boldsymbol{e}_i$ le vecteur issu du descripteur énergie. Pour chaque $\boldsymbol{C}_i$, $\boldsymbol{B}_i$, $\boldsymbol{s}_i$, $\boldsymbol{e}_i$, on divise l'axe des fréquences (en BPM) par Te_i, le tempo estimé par Ircambeat. De cette manière, chaque descripteur est exprimé en rapport au tempo d'Ircambeat. Ensuite, $\boldsymbol{C}_i$, $\boldsymbol{B}_i$, $\boldsymbol{s}_i$, $\boldsymbol{e}_i$ sont échantillonnés aux points :

$$\boldsymbol{k} = \left\{ \frac{1}{4}; \frac{1}{3}; \frac{1}{2}; \frac{2}{3}; \frac{3}{4}; 1; 1.25; 1.33; ...; 2 \right\}$$

de la même manière que dans (**Peeters 2011**). Ces points correspondent aux rapports caractéristiques entre le tempo estimé par un descripteur et le tempo d'Ircambeat. Pour l'échantillonnage des matrices $\boldsymbol{C}_i$ et $\boldsymbol{B}_i$, on fait une simple recherche de maximum à travers Th pour un k donné. Pour chaque descripteur, le vecteur échantillonné est donc de dimension $D = 12$. Si on utilise tous les descripteurs, on arrive à des vecteurs d'observation de dimension $D = 48$.

On note également qu'avant le processus d'échantillonnage, chaque descripteur est normalisé pour avoir leur maximum à 1 afin d'avoir des données du même ordre de grandeur.

La figure 4.1 illustre la méthode d'échantillonnage utilisée sur nos descripteurs. Dans cet exemple, on reprend le vecteur issu du descripteur d'énergie, vu au chapitre précédent. On peut voir le changement de l'axe des fréquences qui passe de T à $\frac{T}{Te}$, puis l'échantillonage aux points k.

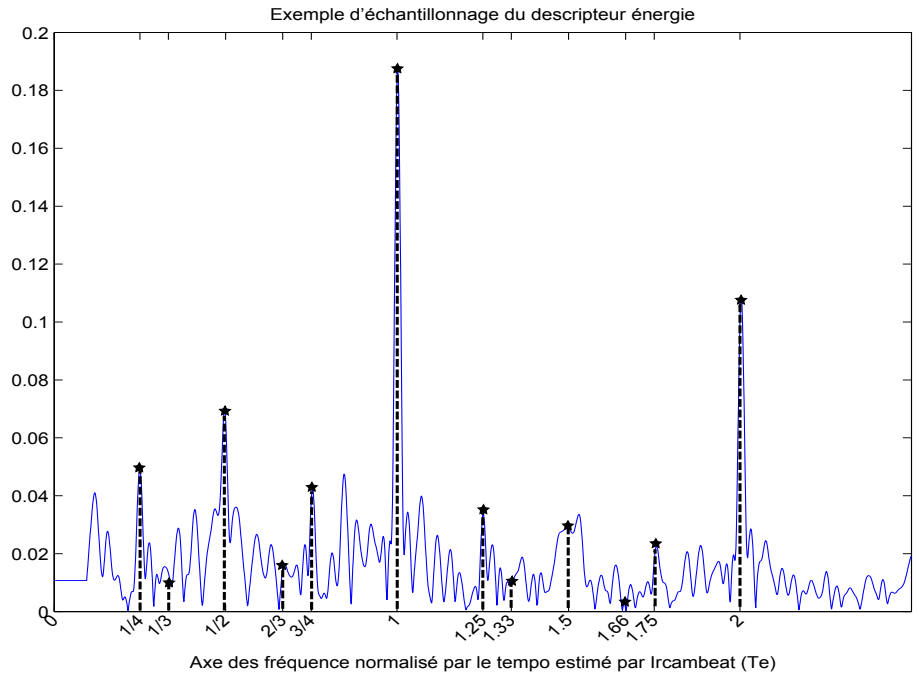

FIGURE 4.1 – Échantillonnage du vecteur issu du descripteur d'énergie. L'axe des fréquences est divisé par Te où $Te = 106.3$ BPM. Extrait audio : 'Til Tuesday - Voices Carry.

Nos vecteurs de descriptions finaux sont la concaténation du résultat de chaque descripteur échantillonné $\boldsymbol{c'}_i$, $\boldsymbol{b'}_i$, $\boldsymbol{s'}_i$ et $\boldsymbol{e'}_i$:

$$\boldsymbol{y}_i = \begin{bmatrix} \boldsymbol{c'}_i \\ \boldsymbol{b'}_i \\ \boldsymbol{s'}_i \\ \boldsymbol{e'}_i \end{bmatrix}$$

Les vecteurs d'observation globaux seront donc de la forme :

$$\boldsymbol{z}_i = \begin{bmatrix} \boldsymbol{y}_i \\ x_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{c'}_i \\ \boldsymbol{b'}_i \\ \boldsymbol{s'}_i \\ \boldsymbol{e'}_i \\ Ta_i \end{bmatrix}$$

4.3.4 Comparaison entre les deux méthodes d'échantillonnage

Pour clore cette présentation de la méthode d'échantillonnage, nous faisons une comparaison entre les deux approches que nous avions envisagées, à savoir la méthode dite *relative* (que nous venons d'expliquer) et la méthode *absolue*. Nous illustrons les distinctions entre ces deux approches avec la figure 4.2 représentant les profiles d'un descripteur échantillonné selon les deux techniques. Dans (a) et (b), on peut voir les profils obtenus grâce aux deux méthodes, pour un titre ayant un tempo de 80 BPM. De même, dans (c) et (d), nous avons les profils pour un titre dont le tempo est de 130 BPM. On considère ici que le tempo estimé par Ircambeat (Te) et par les descripteurs correspondent au tempo annoté.

Ainsi, dans (a) et (c), on peut voir un pic prépondérant pour $\frac{T}{Te} = 1$ ce qui indique que le descripteur a bien estimé le même tempo d'Ircambeat. On remarque que les profils (a) et (c) ne sont pas très différents car ils expriment tous deux leur accord avec le tempo estimé Te. En revanche, dans (b) et (d), on peut voir que les profils évoluent en fonction du tempo estimé. Ceci est la conséquence de l'échantillonnage par bancs de filtres.

4.4 Protocole d'évaluation

Dans la partie suivante, nous présenterons les méthodes d'estimation du tempo par mélange de modèle de Gaussiennes. Pour attester de leurs performances, à la suite de leurs présentations, nous procéderons aux évaluations. On donne donc ici le protocole d'évaluation qui sera le même pour toutes les techniques utilisées.

La méthode d'évaluation est la validation par plis croisés à dix plis. On décompose notre corpus d'étude en dix plis ayant la même distribution en terme de titres correspondant aux différentes catégories que nous avons définies lors de l'analyse de la base de données (Estimation correcte $Te = Ta$, Erreur d'octave $Te = 2Ta$, etc.). Sur les dix plis, neuf servent à l'apprentissage de notre modèle et le pli restant sert à l'évaluation. Les résultats présentés sont les valeurs moyennes sur les dix plis.

Lors de la création de nos modèles GMM, nous estimons les paramètres de chaque gaussienne grâce à l'agorithme EM, lui-même initialisé par un clustering *k-means*.

Afin d'avoir des résultats suffisamment reproductibles, le modèle GMM retenu lors de la phase d'apprentissage sera celui maximisant la log-vraisemblance sur nos données d'apprentissage après 50 itérations. On précise également que le critère d'arrêt de l'algorithme EM est

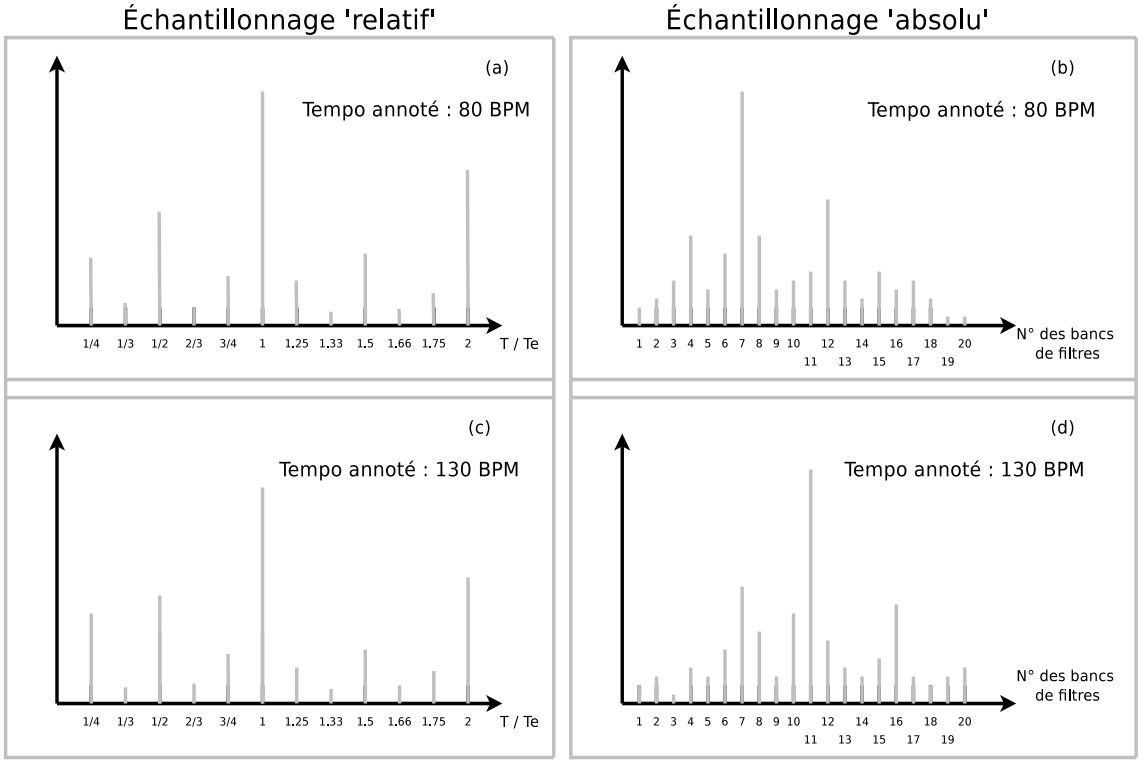

FIGURE 4.2 – Représentation des profils d'un descripteur échantillonné selon la méthode dite *relative* (a) et (c) et selon la méthode *absolue* (b) et (d), pour deux tempi différents.

basé sur la convergence de la log-vraisemblance du modèle sur les données d'apprentissage. Soit $loglikelihood_N$ la log-vraisemblance à l'étape N, le critère d'arrêt est :

$$\frac{loglikelihood_N}{loglikelihood_{N-1}} - 1 < 1^{-10}$$

4.5 Régression du tempo annoté T_a

4.5.1 Présentation

Pour notre première approche, nous cherchons à faire une estimation directe du tempo grâce à la régression par modèle de Gaussiennes. Les données d'apprentissage sont constituées de l'estimation d'Ircambeat (Te), du résultat de nos descripteurs échantillonnés et du tempo annoté (Ta). Nous faisons donc une régression sur Ta à partir de Te et du calcul des descripteurs audio, comme le montre la figure 4.3.

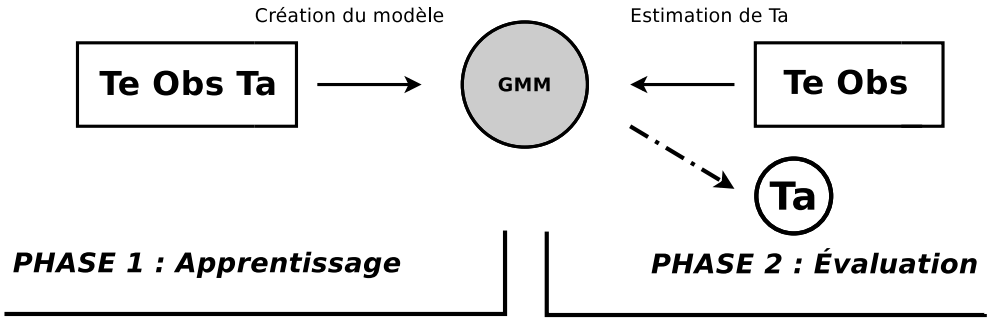

FIGURE 4.3 – Schéma de fonctionnement de la régression sur T_a. T_e désigne le tempo estimé par Ircambeat, *Obs* le vecteur d'observation composé des descripteurs audio et T_a le tempo annoté.

4.5.2 Évaluation

On donne dans les tableaux suivants les résultats de l'évaluation de cette méthode. Afin d'obtenir les meilleurs scores, nous faisons varier le nombre de composantes K de notre modèle et nous testons différentes combinaisons des descripteurs audio servant de vecteurs d'observation.

On donne à chaque fois le taux d'estimation correcte avec une fenêtre de tolérance de 6% du tempo annoté Ta. On donne également le gain de la méthode (i.e, la différence du taux d'estimation correcte entre Ircambeat seul et la méthode présente).

Le tableau 4.1 montre l'influence du nombre de composantes K du modèle de mélange de Gaussiennes. Ces résultats ont été obtenus en utilisant tous les descripteurs audio (variations de Chroma, balance spectrale, similarité acoustique et fonction d'énergie). Le choix des descripteurs utilisés ne changent pas la conclusion de ces résultats : les meilleurs performances ont été obtenues pour un nombre de composantes $K = 20$. Nous retiendrons par la suite ce paramètre.

	$K = 4$	$K = 8$	$K = 12$	$K = 16$	$\boldsymbol{K = 20}$	$K = 24$
Moyenne sur 10 plis	50.34	64.77	68.14	68.41	**68.77**	67.37

TABLE 4.1 – Influence du nombre de composantes K de notre modèle. Les vecteurs d'observation sont composés des variations de Chroma, de la balance spectrale, de la similarité acoustique et de la fonction d'énergie.

Le tableau 4.2 indique cette fois-ci les performances de la méthode en utilisant un nombre de composantes $K = 20$ et en utilisant différentes combinaisons des descripteurs audio. Par souci de lisibilité, chaque descripteur sera identifié comme suit : (**1**) pour les variations de Chroma,

(**2**) pour la balance spectrale, (**3**) pour la similarité acoustique et (**4**) pour la fonction d'énergie. On donne aussi l'écart-type σ des résultats de chacun des plis.

Descripteurs	Moyenne sur 10 plis	Gain
3 4	63.40 ($\sigma = 1.45$)	-4.1
2 3 4	**68.40** ($\sigma = 2.96$)	**+0.9**
1 3 4	67.90 ($\sigma = 3.50$)	$+0.4$
1 2 4	66.37 ($\sigma = 2.31$)	-1.13
1 2 3 4	67.14 ($\sigma = 1.91$)	-0.36
1 2 3	65.45 ($\sigma = 2.33$)	-2.05

TABLE 4.2 – Résultats de l'estimation du tempo annoté Ta utilisant la régression par GMM en utilisant différentes combinaisons de descripteurs audio.

Les résultats indiquent que l'amélioration des estimations du tempo par cette méthode est très faible (moins de 1%) comparé à l'estimation seule d'Ircambeat. Le modèle créé à partir de nos données d'apprentissage n'est pour le moment pas suffisant pour faire de l'estimation directe.

4.6 Estimation du facteur de correction α

4.6.1 Présentation

Au vu des résultats précédents, nous décidons de modifier notre approche. Au lieu d'inférer directement le tempo comme c'était le cas auparavant, nous cherchons désormais à estimer un facteur correctif α tel que :

$$\alpha Te = Ta$$

Le tempo estimé par cette méthode sera alors $Te' = \hat{\alpha} Te$, où $\hat{\alpha}$ correspond à la valeur estimée de α par le modèle. Après avoir défini pour chaque titre le facteur $\alpha = \frac{Te}{Ta}$, nous créons le modèle GMM de la même manière que la méthode précédente, c'est-à-dire en utilisant l'estimation Ircambeat Te, les descripteurs audio et le facteur correctif α, comme le montre le synoptique 4.4.

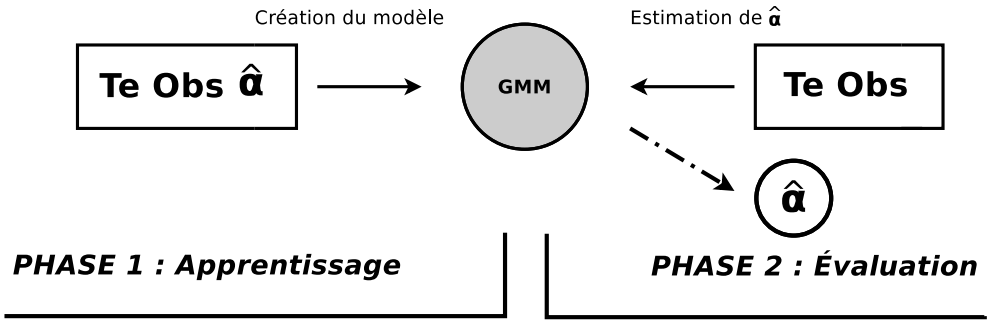

FIGURE 4.4 – Schéma de fonctionnement de la régression sur α

Nous pourrions utiliser le facteur α estimé par le modèle GMM sans a priori. Cependant, nous allons nous servir des observations faites lors de l'étude di'Ircambeat.

En effet, d'après l'analyse du comportement d'Ircambeat, nous avons pu voir que les erreurs du tempo les plus fréquentes étaient les erreurs du type $Te = 2Ta$ et $Te = \frac{Ta}{2}$, ce qui correspond à un facteur correctif $\alpha = \frac{1}{2}$ et $\alpha = 2$ respectivement. Si on ajoute à cela les estimations correctes

d'Ircambeat ($\alpha = 1$), on peut donc faire l'hypothèse que les estimations du modèle seront en majorité $\alpha \in \left\{\frac{1}{2}; 1; 2\right\}$.

Ainsi, au lieu de prendre en compte la valeur de α donnée par la méthode de régression, on peut effectuer un seuillage autour des points $\alpha \in \{1/2; 1; 2\}$. Nous définissons une fenêtre de tolérance $\alpha \in \{1/2; 1; 2\} \pm 0.2$. Par exemple, si le modèle estime $\alpha = 0.65$, on peut supposer que le facteur correctif le plus probable dans ce cas sera $\alpha = 0.5$.

Finalement, contrairement à la méthode précédente, on réduit l'ensemble des possibilités des valeurs estimées par la régression. En faisant la régression de Ta, on pouvait supposer que $Ta \in [30, 200]$, ce qui est un espace trop grand pour faire une quelconque hypothèse sur la valeur estimée. On se retrouve à présent avec un espace à trois dimensions, ce qui nous permet d'avoir un a priori relativement fort sur les valeurs estimées.

En terme de démarche, on se rapproche d'une certaine manière des méthodes vues dans l'état de l'art, notamment celle proposée par Chen dans (**Chen et al. 2009**), dans le sens où on cherche à faire une correction de l'estimation d'Ircambeat. En d'autres termes, on cherche à identifier si l'estimation d'Ircambeat correspond à une erreur d'octave du tempo.

4.6.2 Évaluation

De la même manière que précédemment, nous évaluons cette méthode en faisant varier le nombre de composantes K du modèle de mélange de Gaussiennes et en testant plusieurs combinaisons des descripteurs audio. Le tableau 4.3 résume les résultats obtenus.

	$K = 2$		$K = 4$		$K = 8$	
Descripteurs	% 10 Plis (σ)	Gain	%10 Plis (σ)	Gain	% 10 Plis (σ)	Gain
1(chroma)	68.77 (1.79)	+1.64	66.84 (2.36)	-0.29	66.47 (2.22)	-0.66
2(spec)	74.17 (3.27)	+7.03	74.16 (3.39)	+7.02	74.02 (2.38)	+7.89
3(simi)	73.57 (1.79)	+6.43	73.65 (3.69)	+6.51	72.91 (1.94)	+5.78
4(ener)	73.65 (2.24)	+6.51	73.28 (3.78)	+6.14	72.61 (3.26)	+5.48
1 2	74.39 (2.16)	+7.25	73.34 (2.29)	+6.21	72.53 (3.00)	+5.40
1 2 3	75.12 (2.34)	+7.98	74.53 (2.03)	+7.39	73.27 (2.65)	+6.14
1 2 3 4	77.61 (2.33)	+10.47	76.89 (3.53)	+9.75	74.39 (2.95)	+7.25
1 2 4	76.61 (3.11)	+9.47	75.34 (1.72)	+8.20	74.46 (3.16)	+7.33
2 3	75.86 (2.52)	+8.73	74.97 (2.51)	+7.84	73.21 (2.63)	+6.08
1 3 4	77.42 (2.31)	+10.28	75.64 (1.99)	+8.51	73.79 (2.82)	+6.66
3 4	75.50 (1.43)	+8.36	75.86 (2.01)	+8.73	76.01 (2.01)	+8.87
2 3 4	**77.79** (2.59)	**+10.65**	75.79 (2.86)	+8.65	75.05 (2.84)	+7.92

TABLE 4.3 – Résultats de l'estimation du tempo par la méthode de l'estimation du facteur correctif α

Les résultats montrent clairement l'amélioration de l'estimation du tempo par cette méthode.

On fait d'une certaine manière de la classification : on cherche à savoir si un titre donné correspond à la catégorie *bonne estimation* ($\alpha = 1$), à la catégorie *doublement du tempo* ($\alpha = 1/2$) ou bien à la catégorie *sous-estimation de moitié* ($\alpha = 2$). La dernière méthode présentée fera donc l'objet d'une classification.

4.7 Approche GMM Classification

4.7.1 Présentation

Pour cette troisième méthode, nous étudions une approche de type classification. On crée pour cela trois modèles de mélange de Gaussiennes correspondant à trois classes : la classe *Estimation correcte* ($\alpha = 1$), la classe *Erreur d'octave* $Te = 2Ta$ ($\alpha = 1/2$) et la classe *Erreur d'octave* $Te = Ta/2$ ($\alpha = 2$).

La classification se fait ensuite en calculant la probabilité qu'a un vecteur d'observation d'un titre inconnu d'appartenir à l'un des trois modèles définis. Pour un modèle donné, le calcul de cette probabilité se fait de la manière suivante :

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Où $\boldsymbol{y}$ est le vecteur d'observation pour un titre inconnu, K, le nombre de composantes du modèle et $\boldsymbol{\mu}_k$ et $\boldsymbol{\Sigma}_k$ sa moyenne et sa matrice de covariance.

La figure 4.5 illustre le principe général de cette méthode.

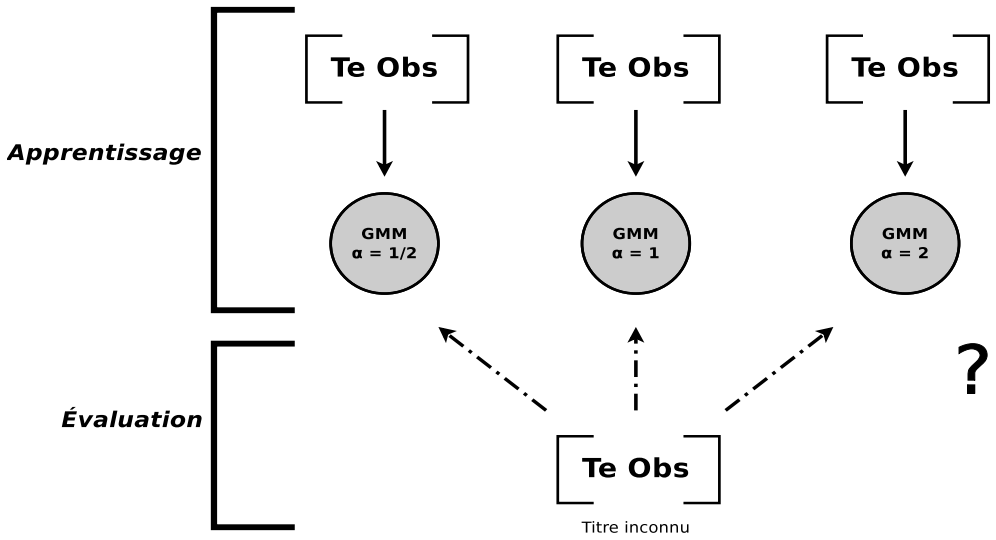

FIGURE 4.5 – Schéma de fonctionnement de la classification

4.7.2 Évaluation

Les résultats obtenus par cette méthode sont donnés dans le tableau 4.4. On peut constater que les performances sont similaires à la technique précédente. Encore une fois les meilleurs résultats sont obtenus en utilisant les descripteurs de balance spectrale, de similarité acoustique et la fonction d'énergie.

	$K = 2$		$K = 4$	
Descripteurs	% 10 Plis	Gain	% 10 Plis	Gain
1(chroma)	61.42 ($\sigma = 3.22$)	-5.71	64.24 ($\sigma = 3.08$)	-2.90
2(spec)	70.61 ($\sigma = 4.03$)	+3.47	71.21 ($\sigma = 2.71$)	+4.07
3(simi)	69.05 ($\sigma = 2.53$)	+ 1.91	68.39 ($\sigma = 2.29$)	+1.25
4(ener)	71.19 ($\sigma = 2.97$)	+4.06	70.97 ($\sigma = 3.03$)	+3.84
1 2	71.41 ($\sigma = 3.70$)	+4.28	71.28 ($\sigma = 3.90$)	+4.14
1 2 3	74.98 ($\sigma = 3.08$)	+7.84	73.94 ($\sigma = 2.23$)	+6.81
1 2 3 4	76.91 ($\sigma = 2.83$)	+9.77	73.14 ($\sigma = 2.46$)	+6.00
1 2 4	75.05 ($\sigma = 3.04$)	+7.91	74.31 ($\sigma = 2.39$)	+7.17
2 3	75.13 ($\sigma = 2.72$)	+ 7.99	73.94 ($\sigma = 1.63$)	+6.80
1 3 4	75.19 ($\sigma = 3.56$)	+7.69	74.82 ($\sigma = 2.30$)	+7.69
3 4	72.24 ($\sigma = 2.28$)	+5.09	74.30 ($\sigma = 2.90$)	+7.16
2 3 4	**77.34** ($\sigma = 3.82$)	**+10.21**	76.31 ($\sigma = 2.76$)	+9.18

TABLE 4.4 – Résultats obtenus pour la méthode de classification

4.8 Réflexions sur l'utilisation des descripteurs

4.8.1 Vecteurs d'observation et performance des méthodes

Les deux dernières méthodes proposées reposent principalement sur l'identification des erreurs d'octave du tempo de la part d'Ircambeat, plutôt que sur l'estimation du tempo comme c'était le cas dans la première méthode.

On peut supposer que les bonnes performances de ces deux dernières techniques sont liées à la manière dont nous avons créé les vecteurs d'observation à partir des descripteurs audio. Comme nous l'avons déjà évoqué dans la section **4.3.1**, plusieurs approches étaient envisageables mais nous avons retenu l'approche que nous avions qualifiée de *relative* qui avait pour but de rendre compte de la corrélation entre l'estimation faite par le descripteur et celle faite par Ircambeat. Au vu des résultats, il apparaît donc logique d'obtenir de meilleures performances pour les techniques cherchant à identifier les erreurs d'octave du tempo étant donné que nos vecteurs d'observation permettent de rendre compte de ce type d'informations.

À l'inverse, dans (**Peeters & Flocon-Cholet 2012**), de bons résultats ont été obtenus en utilisant la régression sur Ta alors que les vecteurs d'observation ont été créés en utilisant la méthode *absolue*.

4.8.2 Utilisation combinée des descripteurs

Les résultats montrent que l'utilisation des descripteurs audio améliore de manière significative l'estimation du tempo. Les meilleurs résultats sont obtenus avec l'utilisation combinée de la balance spectrale, de la similarité acoustique et de la fonction d'énergie. En revanche, les variations de Chroma paraissent ne pas apporter suffisamment d'informations pour améliorer la qualité de l'estimation du tempo. Cette conclusion est également apparue dans (**Peeters & Flocon-Cholet 2012**).

La faiblesse de ce descripteur réside peut-être dans l'hypothèse faite pour calculer les variations harmoniques, qui est que les accords changent tous les quatre temps. Cette hypothèse peut s'avérer vraie dans un grand nombre d'œuvres musicales, mais ce n'est pas une caractéristique suffisamment générale.

Conclusion et perspectives

Les travaux menés au cours de cette étude ont permis d'apporter quelques améliorations concernant l'estimation du tempo d'un extrait musical. Le but de ce projet était de travailler à partir de l'algorithme Ircambeat, afin de réduire le problème très fréquent des algorithmes d'estimation du tempo : les erreurs d'octave du tempo. Pour cela, nous avons choisi d'associer à Ircambeat des descripteurs audio donnant une information relative au tempo. Cette étude s'est appuyée sur le corpus Last.fm nous donnant pour chaque titre le tempo perceptif, ce qui nous permet par la suite de parler d'erreur d'octave du tempo sans ambiguïté.

Dans un premier temps, l'étude de notre corpus d'étude Last.fm nous a permis de mieux comprendre le comportement d'Ircambeat vis-à-vis des erreurs d'octave du tempo. En particulier, nous avons mis en évidence la forte propension de l'algorithme à faire des erreurs du type $Te = 2Ta$ et $Te = Ta/2$.

Ensuite, nous avons défini les descripteurs qui nous serviront pour raffiner l'estimation du tempo. Les descripteurs retenus sont les *variations de Chroma*, qui permettent d'obtenir une information de tempo à partir du taux de changement d'accords d'un morceau de musique, la *balance spectrale*, qui mesure les variations de concentration de l'énergie spectrale, la *similarité acoustique*, basée sur la répétition d'événements musicaux à court terme et enfin une *fonction d'énergie* qui mesure la périodicité globale de la variation d'énergie d'un extrait musical.

Pour inférer le tempo à partir de l'estimation d'Ircambeat et du résultat de nos descripteurs, nous avons eu recours à une approche statistique utilisant la régression par modèle de Gaussiennes. Pour pouvoir utiliser cette technique, nous avons dû définir nos vecteurs d'observation et donc choisir quelles informations extraire de nos descripteurs. Pour cela, nous avons pris le parti d'avoir des vecteurs d'observation rendant compte de la corrélation entre l'estimation d'Ircambeat et le résultats des descripteurs.

Finalement, nous avons présenté et évalué trois méthodes permettant d'estimer le tempo. La première repose sur l'estimation directe du tempo alors que les deux autres sont basées sur l'estimation d'un facteur correctif α, tel que $\alpha Te = Ta$. L'évaluation de ces deux dernières méthodes montre une amélioration significative des performances de l'estimation du tempo, avec près de 10% d'estimations correctes en plus par rapport à Ircambeat seul.

Ce projet montre que l'utilisation de descripteurs musicaux, liée à un algorithme d'estimation de tempo permet d'augmenter la qualité de l'estimation du tempo. On notera toutefois que l'utilisation du descripteur *variations de Chroma* n'a pas toujours amélioré les performances de nos méthodes. À l'avenir, il faudrait donc définir une méthode plus robuste pour déduire le tempo à partir des changements d'accords. Dans cette étude, nous n'avons pas utilisé l'information de classe de tempo, pourtant disponible dans les annotations de notre corpus Last.fm. Par la suite, on pourrait envisager d'intégrer cette donnée dans les méthodes d'estimation du tempo.

Annexes

Annexe A

Démonstration de la fonction de prédiction

La démonstration qui suit est en partie tirée de l'ouvrage de Christopher M. Bishop *Pattern recognition and Machine learning*, Chapitre 2, section 2.3.1 Conditional Gaussian distributions, page 85, 86 et 87. (**Bishop & SpringerLink 2006**)

On rappelle que la densité de probabilité de la loi normale donnée par :

$$\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_k)^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_k)\right\}$$

où $\boldsymbol{\mu}_k$ est le vecteur moyenne de dimension D, $\boldsymbol{\Sigma}_k$ la matrice de covariance de dimension $D \times D$ et $|\boldsymbol{\Sigma}|$ est le déteminant de $\boldsymbol{\Sigma}$.

On rappelle également une propriété de la loi normale multidimensionnelle : si deux ensembles de variables suivent une loi normale, alors la loi conditionnelle d'une de ces deux variables suit également une loi normale.

On considère tout d'abord un vecteur $\boldsymbol{x}$ de dimension D suivant une loi normale $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. On considère également que $\boldsymbol{x}$ est partitionné en deux sous-ensembles $\boldsymbol{x}_a$ et $\boldsymbol{x}_b$, avec $\boldsymbol{x}_a$ comprenant les M premières valeurs de $\boldsymbol{x}$ et $\boldsymbol{x}_b$ comprenant les $M - D$ valeurs restantes. On a donc :

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix}$$

On définit le vecteur moyenne associé $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

ainsi que la matrice de covariance $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

On note la symétrie de cette matrice : $\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}$ implique que $\boldsymbol{\Sigma}_{aa}$ et $\boldsymbol{\Sigma}_{bb}$ sont symétriques alors que $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^T$. Il est aussi utile d'introduire la *matrice de précision*, $\boldsymbol{\Lambda}$ où :

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$$

La forme partitionnée de la matrice de précision est alors :

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab} \\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb} \end{pmatrix}$$

On remarque que $\mathbf{\Lambda}$ a les mêmes propriétés de symétrie que $\mathbf{\Sigma}$. Il faut cependant souligner que $\mathbf{\Lambda}_{aa}$ n'est pas simplement l'inverse de $\mathbf{\Sigma}_{aa}$. Cette relation sera étudiée un peu plus loin.
En développant l'expression présente dans l'exponentielle de la densité de la loi normale, dans le cas de données partitionnées, on obtient :

$$\begin{aligned} -\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= -\tfrac{1}{2}(\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T\Lambda_{aa}(\boldsymbol{x}_a-\boldsymbol{\mu}_a) - \tfrac{1}{2}(\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T\Lambda_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \\ &\quad -\tfrac{1}{2}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)^T\Lambda_{ba}(\boldsymbol{x}_a-\boldsymbol{\mu}_a) - \tfrac{1}{2}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)^T\Lambda_{bb}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \end{aligned}$$

On remarque que l'exponentielle de la distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu})$ peut s'écrire sous la forme :

$$-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = -\frac{1}{2}\boldsymbol{x}^T\mathbf{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu} + const$$

où 'const' correspond aux termes indépendants de $\boldsymbol{x}$, soit $\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$. On voit donc qu'en exprimant la forme quadratique de l'exponentielle sous cette forme, on peut alors assimiler l'inverse de la matrice de covariance $\mathbf{\Sigma}^{-1}$ à la matrice présente dans l'opération des termes de x au second ordre. De même, les coefficients linéaires à x peuvent être assimilés à $\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$.

Nous allons maintenant appliquer cette méthode à la distribution conditionnelle gaussienne $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$, dont la forme quadratique a été donnée plus haut. On appelle la moyenne et la matrice de covariance de cette distribution $\boldsymbol{\mu}_{a|b}$ et $\mathbf{\Sigma}_{a|b}$ respectivement. Nous observons la dépendance de x_a en supposant que x_b est constant. En prenant alors tous les termes de x_a au second ordre, on a :

$$-\frac{1}{2}\boldsymbol{x}_a^T\Lambda_{aa}\boldsymbol{x}_a$$

d'où on peut conclure que la matrice de covariance de $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ est :

$$\mathbf{\Sigma}_{a|b} = \mathbf{\Lambda}_{aa}^{-1}$$

En prenant à présent tous les termes linéaires à x_a :

$$\boldsymbol{x}_a^T\big\{\mathbf{\Lambda}_{aa}\boldsymbol{\mu}_a - \mathbf{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b\big\}$$

En utilisant le fait que $\mathbf{\Lambda}_{ba}^T = \mathbf{\Lambda}_{ab}$. Comme vu précédement, le coefficient de $\boldsymbol{x}_a$ doit être égal à $\mathbf{\Sigma}_{a|b}^{-1}\boldsymbol{\mu}_{a|b}$. On a donc :

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \mathbf{\Sigma}_{a|b}\big\{\mathbf{\Lambda}_{aa}\boldsymbol{\mu}_a - \mathbf{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)\big\} \\ &= \boldsymbol{\mu}_a - \mathbf{\Lambda}_{aa}^{-1}\mathbf{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \end{aligned}$$

Pour le moment, $\boldsymbol{\mu}_{a|b}$ et de $\mathbf{\Sigma}_{a|b}$ sont exprimés en fonction de la forme partitionnée de la matrice de précision $\mathbf{\Lambda}$. Pour revenir à des expressions faisant intervenir la matrice de covariance, on utilise la propriété matricielle suivante :

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{M} & -\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \end{pmatrix}$$

Avec $\boldsymbol{M} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}$.

$\boldsymbol{M}^{-1}$ est connu sous le nom de *complément de Schur* de la matrice de gauche. En utilisant la définition que nous avons déjà établie, on peut écrire :

$$\begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

En appliquant ensuite la propriété matricielle vue à l'instant, on a :

$$\begin{aligned} \boldsymbol{\Lambda}_{aa} &= (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1} \\ \boldsymbol{\Lambda}_{ab} &= -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1} \end{aligned}$$

D'où on obtient les expressions de la moyenne et de la covariance de la distribution conditionnelle $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \end{aligned}$$

On remarque alors que $\boldsymbol{\mu}_{a|b}$ représente le second terme de la fonction de prédiction. On rappelle que la fonction de prédiction est :

$$F(\boldsymbol{y}) = \mathbb{E}[x|\boldsymbol{y}] = \sum_{k=1}^{K} h_k \left[\boldsymbol{\mu}_k^x + \boldsymbol{\Sigma}_k^{xy}(\boldsymbol{\Sigma}_k^{yy})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_k^y)\right]$$

avec

$$h_k(\boldsymbol{y}) = \frac{\pi_k N(\boldsymbol{y}|\boldsymbol{\mu}_k^y, \boldsymbol{\Sigma}_k^{yy})}{\sum_{k=1}^{K} \pi_k N(\boldsymbol{y}|\boldsymbol{\mu}_k^y, \boldsymbol{\Sigma}_k^{yy})}$$

Dans le cas d'un modèle de mélange de Gaussienne à une composante, le fonction de prédiction devient simplement :

$$F(\boldsymbol{y}) = \mathbb{E}[x|\boldsymbol{y}] = \boldsymbol{\mu}_{x|\boldsymbol{y}}$$

Le terme $h_k(\boldsymbol{y})$, représentant la probabilité qu'un vecteur de données appartienne à une certaine composante du mélange, permet simplement d'étendre la fonction de prédiction à un modèle de mélange de Gaussiennes à plusieurs composantes.
On peut retrouver cette explication dans (**Stylianou et al. 1998**).

> [...] In the jointly Gaussian case, the optimal conversion function is thus a simple linear transformation given by $[\boldsymbol{\mu}_{x|y}]$. It was decided to extend this result to the GMM by weighting terms that are analogous to the Gaussian conditional expectation. These weighting terms were chosen to be the conditional probabilities that the vector $\boldsymbol{x}_i$ belongs to the different classes $\boldsymbol{C}_i$

Annexe B

Erreurs d'octave par classe de tempo

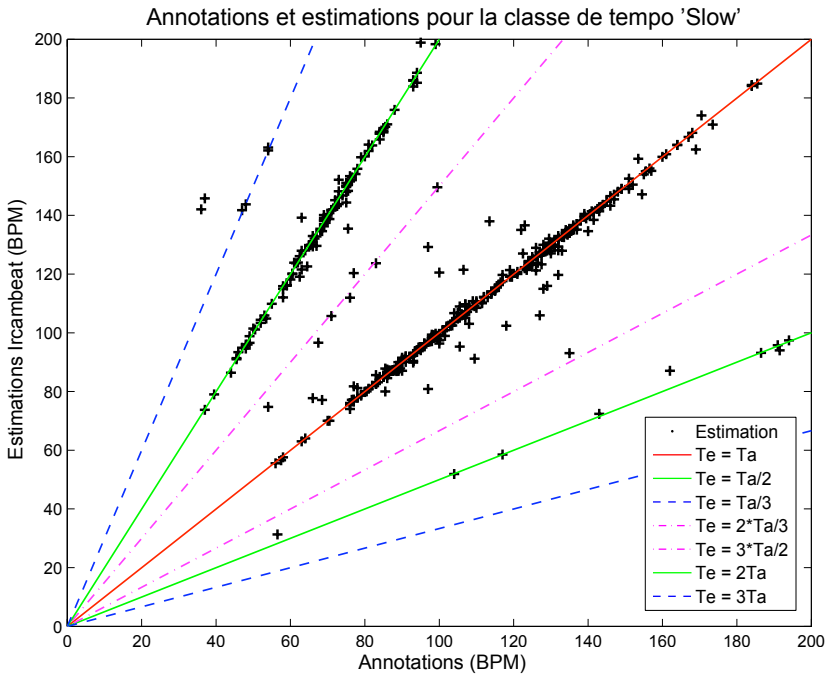

FIGURE B.1 – Répartition des estimations d'Ircambeat par rapport au tempo perceptif, pour la classe de tempo 'Slow'

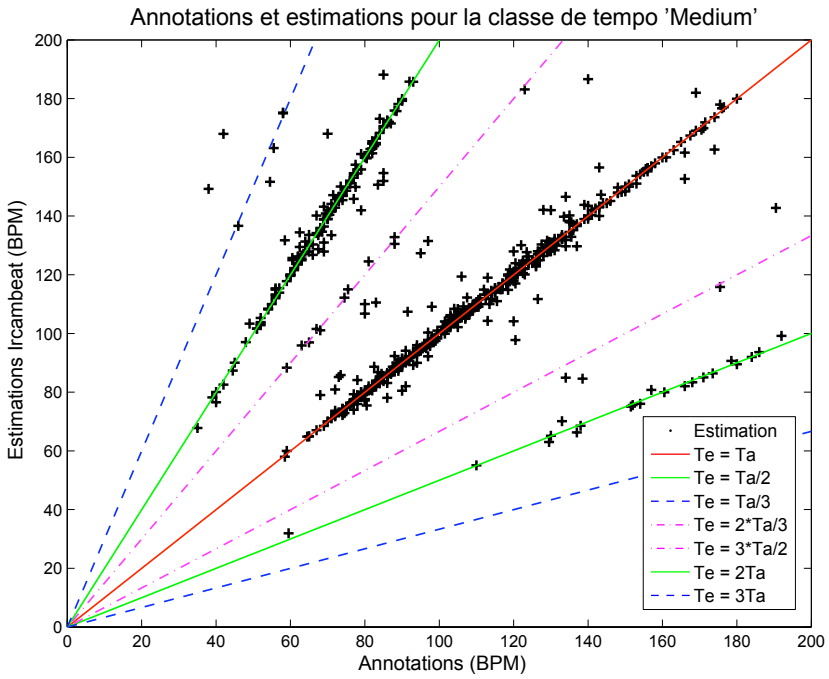

FIGURE B.2 – Répartition des estimations d'Ircambeat par rapport au tempo perceptif, pour la classe de tempo 'Medium'

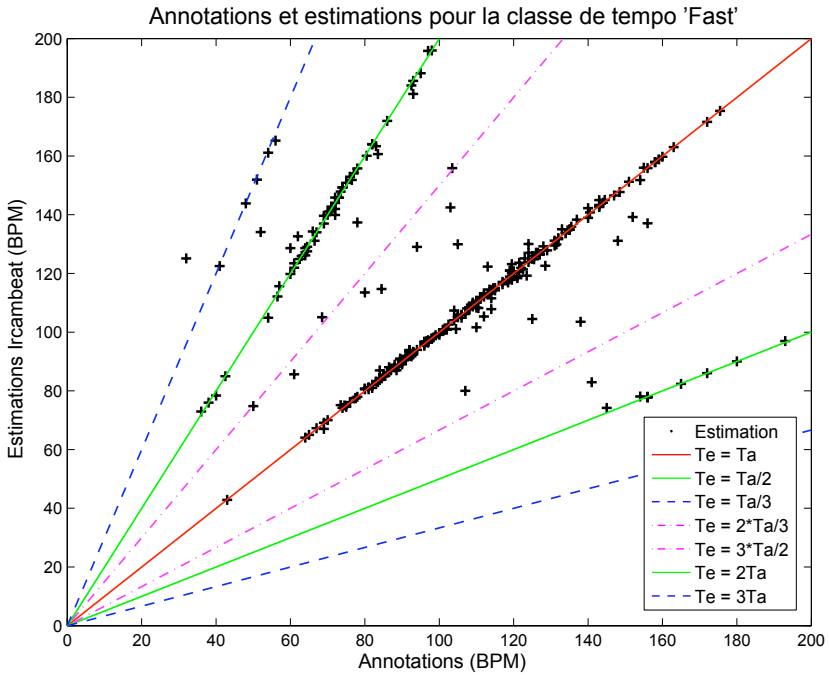

FIGURE B.3 – Répartition des estimations d'Ircambeat par rapport au tempo perceptif, pour la classe de tempo 'Fast'

Bibliographie

Bishop, C. & SpringerLink (2006). *Pattern recognition and machine learning*, volume 4. springer New York.

Calinon, S. (2009). *Robot Programming by Demonstration : A Probabilistic Approach*. EPFL/CRC Press. EPFL Press ISBN 978-2-940222-31-5, CRC Press ISBN 978-1-4398-0867-2.

Chen, C., Cremer, M., Lee, K., DiMaria, P., & Wu, H. (2009). Improving perceived tempo estimation by statistical modeling of higher level musical descriptors. In *Proceedings of the 126th Audio Engineering Society Convention*.

Chua, B. & Lu, G. (2005). Determination of perceptual tempo of music. *Computer Music Modeling and Retrieval*, 61–70.

Ellis, D. (2007). Beat tracking by dynamic programming. *Journal of New Music Research*, *36*(1), 51–60.

En-Najjary, T., Rosec, O., & Chonavel, T. (2003). A new method for pitch prediction from spectral envelope and its application in voice conversion. In *Eighth European Conference on Speech Communication and Technology*.

Fujishima, T. (1999). Realtime chord recognition of musical sound : a system using common lisp music. In *Proceedings of the International Computer Music Conference*, (pp. 464–467).

Hockman, J. & Fujinaga, I. (2010). Fast vs slow : Learning tempo octaves from user data. In *Proc. ISMIR*.

Jiang, D., Lu, L., Zhang, H., Tao, J., & Cai, L. (2002). Music type classification by spectral contrast feature. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, (pp. 113–116). Ieee.

Laroche, J. (2003). Efficient tempo and beat tracking in audio recordings. *Journal-Audio Engineering Society*, *51*(4), 226–233.

Levy, M. (2011). Improving perceptual tempo estimation with crowd-sourced annotations. *12th International Society for Music Information Retrival Conference (ISMIR 2011)*.

Moelants, D. & McKinney, M. (2004). Tempo perception and musical content : What makes a piece fast, slow or temporally ambiguous. In *Proceedings of the 8th International Conference on Music Perception and Cognition*, (pp. 558–562).

Peeters, G. (2006). Chroma-based estimation of musical key from audio-signal analysis. In *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR)*, (pp. 115–120).

Peeters, G. (2007a). Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. *Proc. ISMIR, Vienna, Austria*.

Peeters, G. (2007b). Template-based estimation of time-varying tempo. *EURASIP Journal on Applied Signal Processing*, *2007*(1), 158–158.

Peeters, G. (2011). Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal. *Audio, Speech, and Language Processing, IEEE Transactions on*, (99), 1–1.

Peeters, G. & Flocon-Cholet, J. (2012). Perceptual tempo estimation using gmm regression. In *Submitted to ACM Multimedia, MIRUM Workshop*, Nara, Japan.

Peeters, G. & Papadopoulos, H. (2011). Simultaneous beat and downbeat-tracking using a probabilistic framework : theory and large-scale evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, (99), 1–1.

Seyerlehner, K., Widmer, G., & Schnitzer, D. (2007). From rhythm patterns to perceived tempo. In *Proceedings of the 8th International Conference on Music Information Retrieval*, (pp. 519–524).

Stylianou, Y., Cappé, O., & Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *Speech and Audio Processing, IEEE Transactions on*, *6*(2), 131–142.

Wakefield, G. (1999). Mathematical representation of joint time-chroma distributions. In *International Symposium on Optical Science, Engineering, and Instrumentation, SPIE*, volume 99, (pp. 18–23).

Xiao, L., Tian, A., Li, W., & Zhou, J. (2008). Using a statistic model to capture the association between timbre and perceived tempo. In *Proceedings of the International Conference on Music Information Retrieval*, (pp. 659–662).

Zhu, J. & Lu, L. (2005). Perceptual visualization of a music collection. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, (pp. 1058–1061). IEEE.