Estimation des tempi perçus en fonction du contenu audio et du profil utilisateur.

Ugo MARCHAND
(ugo.marchand@ircam.fr)

Stage de Master 2 ATIAM
Encadrant : Geoffroy PEETERS
Du 1 mars 2013 au 31 juillet 2013

IRCAM
1 Place Igor-Stravinsky
75004 Paris, FRANCE

Table des matières

Notations

$\mathcal{A}/\mathcal{D}$	Symbolisent les classes Accord et Désaccord.
IQR	Écart interquartile.
$d_i(\lambda)$	Indice acoustique, où i dénote le type d'indice (*ener*, *sim*, *specbal* ou *harmo*) et λ le retard temporel en sec. Ces indices sont décrits en détail dans la partie 3.1.
a	Dénote un extrait audio.
t	Dénote un tempo.
u	Dénote un utilisateur.

Acronymes

ACP	Analyse en Composantes Principales. Une analyse en composantes principales permet de réduire la dimension d'un ensemble de variables possiblement corrélées, en un ensemble de variables dé-corrélées. La première composante est telle qu'elle ait le maximum de variance (c'est-à-dire qu'elle représente le maximum de variabilité des données possible), la seconde telle qu'elle ait aussi le maximum de variance avec la condition supplémentaire d'être orthogonale à la première, et ainsi de suite.
BPM	Battements par minute.
GMM	Modèle de Mélange Gaussien (Gaussian Mixture Model). C'est une méthode d'apprentissage statistique qui consiste à modéliser des ensembles de données par un certain nombre de distributions gaussiennes. La classification par modèle de mélanges gaussiens est décrite dans la partie 4.1. Un exemple de régression par modèle de mélanges gaussiens se trouve dans la partie 3.2.4.
MIR	Music Information Retrieval : désigne le champs de recherche dans lequel s'inscrit l'estimation de tempo perceptif. Ce champs comprend aussi des applications telles que la séparation de sources, la reconnaissance d'instruments ou la transcription automatique de la musique.
SVM	Machines à Vecteurs Support (Support Vector Machine).

Introduction

Le tempo est un des éléments perceptifs les plus importants de la musique. Il est définit comme l'allure d'un extrait musical, exprimé en battement par minute (BPM). Bien que sa perception soit en général aisée (bon nombre de personnes peuvent taper les temps d'un extrait musical après quelques secondes seulement), son estimation automatique n'est pas une problèmatique facile.

L'estimation de tempo est un des problèmes majeur du domaine de recherche sur la récupération de données musicales (Music Information Retrieval, MIR) car ses applications sont multiples (par exemple : application synchronisées sur le battement, analyse musicologique).

Le tempo reste avant tout une donnée perceptive, et il arrive que plusieurs personnes auxquelles on demande de taper le tempo d'un extrait ne soient pas d'accord entre elles. Le tempo perceptif est alors souvent défini par le tempo majoritaire (tapé par le plus grand nombre de sujets). Si on note a un extrait audio et t son tempo, l'estimation de tempo consiste à trouver f telle que $f(a) = \hat{t} \simeq t$. Si on considère plusieurs utilisateurs notés u, la tâche revient donc à trouver f telle que $f(a, u) = \hat{t}_u \simeq t_u$. Actuellement, les efforts se concentrent surtout sur le cas où les utilisateurs sont d'accord entre eux, et cherchent à estimer le plus précisément f telle que $f(a, \forall u) = \hat{t}$. Cette estimation est indépendante de l'utilisateur.

Dans certains cas cependant, un tempo prédominant ne se dégage pas, et il devient alors difficile de parler de tempo unique. Ce mémoire s'intéresse à ce cas où les annotateurs ne sont pas d'accord entre eux et un tempo unique ne peut donc pas être défini. Le but est de pouvoir estimer le tempo en fonction de l'utilisateur et de l'audio. Cependant, avant de créer le modèle complet $f(a, u) = \hat{t}_u \simeq t_u$, on s'intéresse d'abord à prédire si la perception du tempo va être partagée ($f(a, u) = f(a, u')$) ou non ($f(a, u) \neq f(a, u')$) pour un extrait musical donné.

Pour cela, on crée un modèle de classification des extraits musicaux en deux classes notées $\mathcal{A}$ pour accord et $\mathcal{D}$ pour désaccord. Ce travail s'appuie sur le corpus créé par Levy [11], qui nous fourni une liste de titres musicaux, ainsi que les annotations de ces titres par de multiples utilisateurs. Les extraits audio sont analysés au moyen de quatre indices acoustiques : l'énergie, la similarité à court-terme, la balance spectrale et l'harmonicité ([18]). Ces indices servent ensuite à la prédiction des classes $\mathcal{A}$ et $\mathcal{D}$ à travers différents modèles.

Ce mémoire est organisé de la façon suivante :

La partie 1 propose un état de l'art sur l'ambiguïté de tempo, le désaccord entre utilisateur et l'estimation de tempo. La partie 2 présente comment nous avons obtenu notre corpus à partir de celui de Levy [11]. Ensuite, nous détaillons notre modèle de classification dans la partie 3. Enfin, dans la partie 4, nous exposons et analysons les résultats de ces modèles.

1 État de l'art

Il existe relativement peu d'articles traitant du sujet du partage de la perception d'un tempo. Les premières études sur ce sujet ont été réalisées par Moelants et McKinney, qui ont proposé dans [13, 14, 15] un modèle de résonance pour expliquer l'apparition d'un tempo préférentiel. En effet, lors de plusieurs expériences perceptives, où il était demandé aux sujets de taper les temps d'un extrait musical, il est apparu un tempo préférentiel, souvent centré autour de 120 bpm. Ce modèle est validé par des tests de perceptions dans leur laboratoire. Ils proposent aussi une explication, basée sur les accents rythmiques, lorsque les résultats s'éloignent de leur modèle de résonance.

Dans [15], ils vont plus loin et font l'hypothèse que la perception du tempo est partagée (i.e. tous les sujets tapent le même tempo) si l'extrait musical contient un niveau métrique proche de ce tempo de résonance (120 bpm). Par contre, si l'extrait contient un niveau métrique ayant deux pics de part et d'autre du tempo préférentiel, la perception du tempo risque d'être ambiguë (i.e. les sujets ne sont pas d'accord entre eux).

Une des rares autres études traitant du désaccord sur la perception du tempo entre utilisateurs a été réalisée par Zapata. Dans [23], Zapata et al. montrent que l'on peut prédire la confiance à apporter aux résultats des algorithmes d'estimation automatique de tempo, grâce à une mesure de l'accord entre annotateurs (MMA : mean mutual agreemeent). Cette mesure (MMA) permet aussi la sélection de l'annotateur de tempo le plus fiable pour un extrait musical donné. Les résultats de cette expérience montrent que les corpus classiques sur lesquels travaillent les algorithmes d'estimation de tempo sont souvent biaisés et fournissent trop d'exemples "faciles". D'après leur article précédent ([8]), ce biais est à l'origine du plafonnement des algorithmes d'estimation automatique du tempo.

Contrairement à l'ambiguïté de tempo perceptif, il existe beaucoup plus d'études portant sur l'estimation du tempo. La plupart des algorithmes actuels souffrent des fameuses erreurs d'octave (l'algorithme estime le double ou la moitié du tempo), l'objectif principal de la recherche dans ce domaine est donc de réduire ces erreurs d'octaves . Peeters [18] et Seyerlehner [20] ont une approche par apprentissage machine à partir de descripteurs bien choisis. Gkiokas [6], Hockman [7] et Chen [2] utilisent des classes de vitesse de tempo (de lent à rapide). Enfin Xiao [22] explore la relation entre timbre et tempo.

Peeters et Flocon-Chollet dans [18] proposent d'estimer le tempo perceptif au moyen de quatre indices acoustiques basés sur des considérations perceptives (variation d'énergie, similarité à court-terme, variation harmonique et alternance grave/aigu). Ensuite, une régression GMM est utilisée pour estimer le tempo perceptif. Cette méthode a permis une diminution des erreurs d'octave, et une meilleure répartition de celles-ci sur l'échelle des tempi.

Seyerlehner dans [20] propose d'utiliser la méthode d'apprentissage statistique des plus proches voisins (k-NN) pour inférer le tempo perçu. Les extraits audio sont représentés par leur motif de fluctuation (fluctuation patterns) ou leur fonction d'auto-corrélation. La distance entre deux extraits est calculée comme le coefficient de corrélation de Pearson. Enfin, pour chaque extrait, on récupère les k extraits les plus similaires ($k = 5$), et le tempo majoritaire parmi ces k extraits est affecté à l'audio inconnu.

Gkiokas et al. dans [6] proposent de réduire les erreurs d'octave en apprenant des classes de tempo. Chaque titre est représenté par un vecteur de périodicité et une classe (lent/modéré/rapide). Un modèle de machines à vecteurs support (SVM) est utilisé pour apprendre la classe de tempo. Enfin le tempo perceptif estimé est calculé comme le pic prédominant de la fonction de périodicité,

appartenant à l'intervalle de la classe inférée.

Hockman et al. dans [7] proposent aussi de réduire les erreurs d'octave en apprenant des classes de tempo. Chaque titre est représenté par une série de descripteurs générés par jAudio ([12]). L'apprentissage se fait au moyen de 6 algorithmes de classification. Enfin, la séparation en classes est utilisée pour corriger le tempo. Le point à noter de cet article est qu'ils n'utilisent pas de fonctions de périodicité pour ce problème mais une série de 80 descripteurs liés à la hauteur, à l'intensité et au timbre du morceau. L'algorithme de classification donne d'excellents résultats (plus de 96 %), mais la correction de tempo ne montre pas de résultats vraiment améliorés.

Chen dans [2] propose une méthode de correction des erreurs d'octave. Son hypothèse est que le caractère du morceau est lié à son tempo (par exemple : agressif signifiera souvent un tempo rapide, alors que romantique ou sentimental signifiera un tempo lent). Basé sur une centaine de descripteurs, un système SVM apprend 4 classes de tempo (de lent à rapide). Puis suivant ces classes, le tempo estimé est multiplié par 2, divisé par 2, ou laissé inchangé. Cette méthode permet d'améliorer les résultats de beaucoup d'algorithmes de l'état de l'art.

Xiao dans [22] émet l'hypothèse que le tempo perçu est lié au timbre de l'extrait musical. Il représente donc chaque musique par un vecteur de MFCC (Mel Frequency Cepstral Coefficents). Il utilise ensuite un GMM à 8 gaussiennes pour modéliser les MFCC extraits et le tempo annoté. Pour chaque titre musical inconnu, une première estimation de tempo T_e est faite. Le modèle GMM sert ensuite à estimer les probabilités de T_e, $\frac{T_e}{2}$, $\frac{T_e}{3}$, $2T_e$, $3T_e$. La plus grande probabilité donne le tempo perçu. Cette méthode donne des résultats similaires aux algorithmes d'estimation de tempo actuels.

2 Le corpus

Le corpus que nous utilisons est dérivé de celui issu de l'expérience effectuée par Last-FM en 2011 et publiée par Levy [11].

2.1 Présentation de l'étude menée par Mark Levy

Pour ce corpus, Levy [11] a fait appel à un grand nombre d'internautes, dans le cadre d'une expérience web perceptive. Cette expérience est toujours disponible [1]. Il était demandé aux sujets d'écouter des extraits musicaux (de 30 secondes), de les classer en trois classes *slow* (lent), *fast* (rapide) et *in-between* (entre les deux), de quantifier leur tempo (en bpm) et enfin de comparer le premier titre avec un autre extrait musical (selon les critères plus lent, la même vitesse ou plus rapide). Nous nous intéressons uniquement à la partie annotation du tempo perceptif. L'annotation de tempo se fait avec la barre d'espace. Pour qu'une estimation de tempo soit prise en compte, il faut que l'utilisateur ait tapé au moins 10 fois. S'il y a plus de 2 secondes entre les battements, le compteur est réinitialisé (cela limite le tempo minimal possible à 30 bpm). La moyenne de l'intervalle entre les battements est ensuite prise comme annotation de tempo.

Cette étude a donc permis de réaliser une base de données assez conséquente, comme en attestent les chiffres du tableau 1. Nous possédons presque 4000 titres, annotés par 2000 utilisateurs, pour un nombre total de presque 18000 annotations.

Nombre de titres	3698
Nombre d'annotateurs différents	1896
Nombre d'annotations de tempo	17884

TABLE 1 – Chiffres-clés du corpus de Levy.

La figure 1 montre la répartition des genres musicaux des titres de la base. Ces genres ont été extrait au moyen de l'API de 7-digital [2]. Pour les morceaux possédant plusieurs genres, nous n'avons pris que le genre principal. Nous possédons au final une base très majoritairement composée de musique pop/rock.

2.2 Utilisation du corpus & limitations

L'utilisation de ce corpus est néanmoins sujette à limitations.

Le corpus ne contient pas l'audio associé aux annotations, seulement le titre et l'artiste. Or, ces informations peuvent être parfois ambiguës, ou plusieurs versions d'une même chanson peuvent exister. De plus, Levy dans [11] a fait écouter un extrait de 30 secondes du titre, et nous ne savons pas lequel. Ces raisons font que nous ne sommes jamais sûr que notre audio, téléchargé au moyen de l'API de 7-digital, corresponde exactement à ce qu'ont écouté les annotateurs.

Nos extraits choisis (par 7-digital) ne correspondent pas aux 30 premières secondes du morceau mais plutôt à 30 secondes représentatives de ce titre (on note que le refrain d'un morceau pop/rock

1. `http://playground.last.fm/demo/speedo`
2. `http://developer.7digital.com/resources/api-docs`

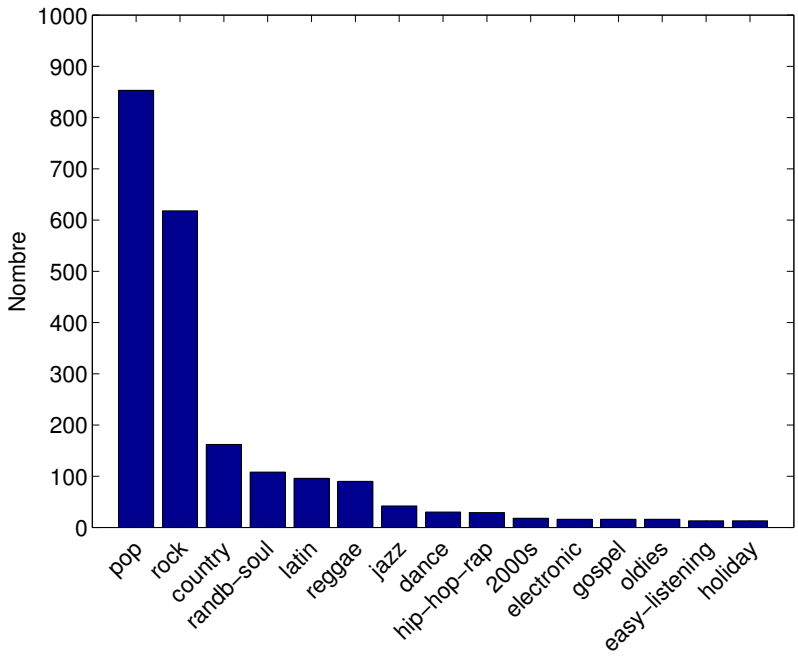


FIGURE 1 – Genres musicaux du corpus, obtenu grâce à l'API de 7digital.com.

est toujours présent). On peut espérer que Mark Levy a fait de même lorsqu'il a sélectionné les extraits pour son expérience, et que donc, notre audio correspond dans la majorité des cas à celle de l'expérience.

Comme l'indique Levy, l'environnement dans lequel se fait l'expérience est lui aussi sujet à caution. On est loin d'une expérience perceptive en laboratoire. Pour limiter les problèmes, l'expérience est restreinte aux seuls utilisateurs enregistrés. Levy a aussi mis en place un système de points et une page de classement représentant les meilleurs contributeurs, afin de motiver les gens à répondre à l'expérience. On note que, pour limiter la triche, chaque réponse est liée à son annotateur. Donc, si un problème est découvert, on peut facilement supprimer toutes les annotations associées. On peut noter, que vu le peu d'intérêt qu'a la triche sur cette expérience, les annotations sont relativement fiables.

Un autre point de l'expérience à noter, est que l'annotation se fait au moyen de la barre d'espace. La précision du tempo battu n'est donc pas très importante. C'est pourquoi, nous prenons par la suite une fenêtre de tolérance de 8%. Cette fenêtre est supérieure à celle de 4% habituellement utilisée en indexation audio, mais compte-tenu de la faible précision des résultats, et du caractère très informel de l'expérience, une fenêtre de 4 % est trop discriminante.

Pour conclure, l'expérience a été faite dans un environnement plutôt hostile à une expérience perceptive par comparaison à un test perceptif en laboratoire. On ne peut même pas assurer que

les utilisateurs aient écouté l'extrait musical avant de répondre. Mais les résultats de Peeters et Flocon-Cholet [3, 18] ont montré que ce corpus était suffisamment fiable pour tester des algorithmes d'estimation de tempo perceptif.

2.3 Nettoyage

Le corpus de Levy est donc assez fiable pour estimer un tempo perceptif lorsque les annotateurs sont d'accord entre eux. Il s'est révélé inutilisable tel quel dans le cas où les utilisateurs ne sont pas d'accord entre eux.

Nous nous intéressons au cas où les utilisateurs sont en désaccord, ce qui pose deux questions : est-ce que les utilisateurs ne sont pas d'accord entre eux parce qu'il y a une ambiguïté de tempo (ce qu'on aimerait), ou est-ce qu'ils ne sont pas d'accord parce qu'ils ont mal effectué l'expérience ? Il s'avère que dans un bon nombre de cas, c'est une erreur d'annotation qui est la cause du désaccord.

Ce problème n'est pas mis en valeur dans [18], car le fait de sélectionner les titres pour lesquels les utilisateurs sont d'accord entre eux, opére déjà un filtrage des mauvaises annotations sur le corpus. Ce filtrage implicite n'est plus possible dans notre cas, vu que l'on s'intéresse justement au désaccord. Nous avons donc mis en place un outil manuel de nettoyage du corpus.

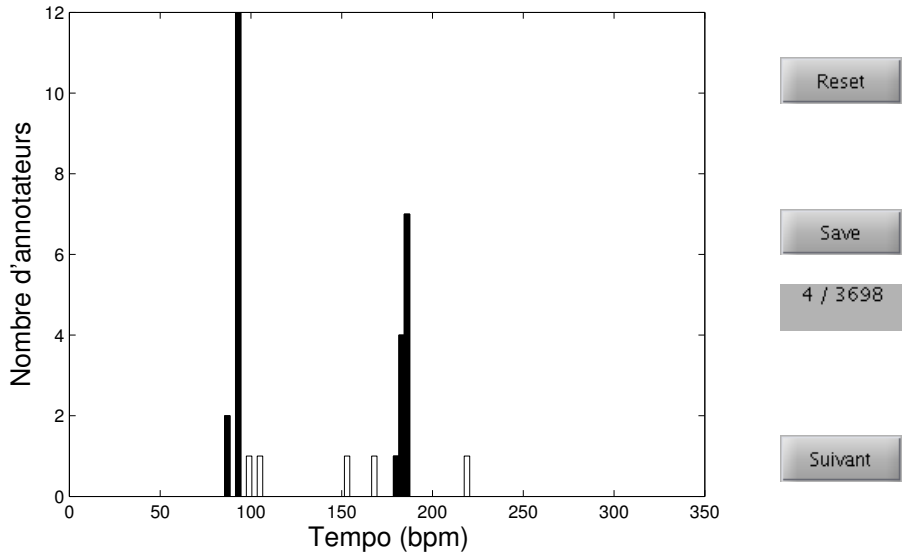


FIGURE 2 – Capture d'écran de l'interface de nettoyage du corpus. L'histogramme est celui des annotations en fonction du tempo, pour un titre donné. Les annotations en noir seront conservées, les annotations transparentes seront supprimées.

Sur la figure 2, on peut voir un exemple typique du travail que nous avons effectué pour chaque titre. Au milieu est tracé l'histogramme des annotations de tempo. Les annotations en noir sont

conservées, les annotations transparentes sont supprimées. On remarque qu'ici, on est dans le cas d'une ambiguïté de tempo, on repère deux tempi principaux vers 90 et 180 bpm. Les 5 annotations transparentes non associées à ces tempi sont des personnes qui ne sont soit pas assez précises, soit qui n'ont pas compris la consigne.

On peut questionner une telle démarche, qui pourrait être vue comme une adaptation des données pour notre confort. Cela n'en est rien car ce nettoyage consiste simplement à ne conserver que les tempi principaux (qui sont des multiples l'un de l'autre) repérés par les annotateurs, supprimer les outliers et les titres pour lesquels aucun tempo ne semble prédominant. Compte-tenu du protocole expérimental de l'expérience (web, aucun contrôle sur l'utilisateur, précision floue due à la barre d'espace), cela est parfaitement justifié.

2.4 Catégorisation

Maintenant que l'on possède un corpus propre, on peut s'intéresser à la catégorisation des titres. On va utiliser deux classes qui seront notées $\mathcal{A}$ (pour accord) et $\mathcal{D}$ (pour désaccord).

La méthode retenue pour la catégorisation utilise l'écart interquartile (ou interquartile range, noté *iqr*) des tempi. C'est une mesure de dispersion plus robuste que l'écart-type (qui est assez sensible aux valeurs extrêmes). L'*iqr* est calculé comme la différence entre le premier et le troisième quartile d'une série de mesures.

Pour chaque titre, on calcule l'IQR de tous les tempi annotés dont on a pris le log_2. On note *tempi* le vecteur de toutes les annotations d'un titre. On sépare le corpus en deux classes $\mathcal{A}$ et $\mathcal{D}$ grâce au critères suivants :

– les titres de la classe $\mathcal{A}$ satisfont :

$$iqr\left(log_2\left(tempi\right)\right) < \tau = 0.2$$

– les titres de la classe $\mathcal{D}$ sont évidemment sélectionnés suivant :

$$iqr\left(log_2\left(tempi\right)\right) \geq \tau$$

Prendre le log_2 des tempi annotés permet de mettre en évidence les erreurs d'octave. En effet, un tempo multiple de 2 d'un tempo de référence aura son log_2 supérieur de 1 : $log_2\left(2t\right) = 1 + log_2\left(t\right)$.

La figure 3 montre la distribution de l'*iqr* sur tous les titres du corpus. À gauche, avant le nettoyage du corpus (décrit au 2.3), et à droite, après. Le trait vertical rouge en pointillés représente le seuil τ de sélection de la classe. On observe bien l'effet du nettoyage. Avant, une bonne partie des titres avaient une valeur comprise entre 0.1 et 0.9, ce qui les rendait peu facile à classer. Après nettoyage, la majorité des valeurs sont concentrées autour de 0 (classe $\mathcal{A}$) et de 1 (symbolisant un désaccord d'octave : classe $\mathcal{D}$).

Le tableau 2, montre, à chaque étape le nombre de titres utilisables pour notre problème. On voit qu'il est drastiquement réduit. Malheureusement, les étapes de nettoyage et de sélection sont nécessaires pour avoir un corpus de travail fiable. Il nous reste donc 249 titres fiables, formant deux classes équilibrées : **134** titres sont dans la classe $\mathcal{A}$ et **115** titres dans la classe $\mathcal{D}$. Ces titres vont donc former notre corpus d'étude, que l'on utilisera dans toute la suite de ce rapport.

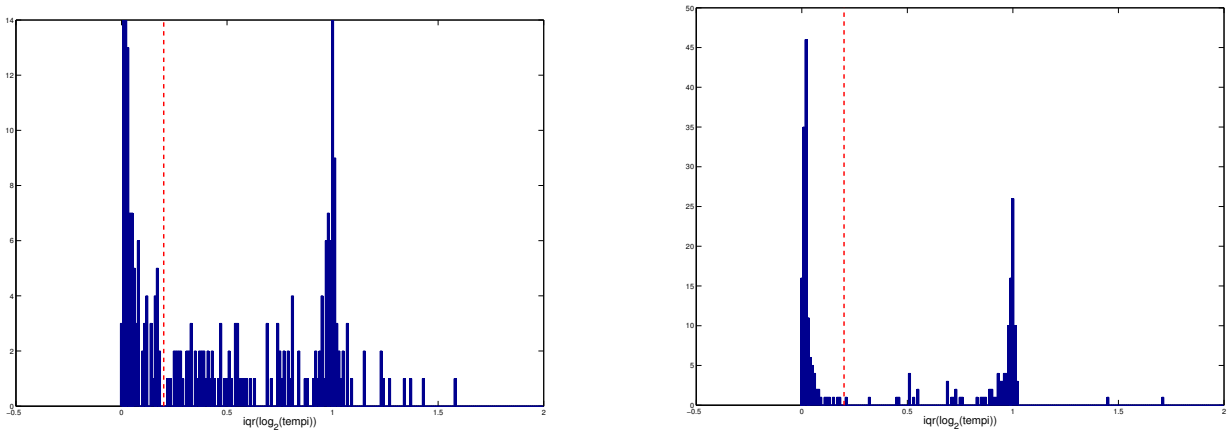


FIGURE 3 – Distribution de l'$iqr(log_2(tempo))$. [Gauche] Avant nettoyage du corpus. [Droite] Après nettoyage. La ligne rouge verticale en pointillés représente le seuil τ de sélection de classe.

		Nombre de titres
Corpus complet		3698
Après nettoyage		820
Avec 10 annotations		**249**
dont :	$\mathcal{A}$	**134**
	$\mathcal{D}$	**115**

TABLE 2 – Nombre de titres à chaque étape de préparation du corpus. En gras, le corpus qui va nous servir dans toute la suite de ce rapport.

3 Classification $\mathcal{A}$ccord/$\mathcal{D}$ésaccord

La classification $\mathcal{A}/\mathcal{D}$ est une première étape vers l'estimation de tempo perceptif prenant en compte l'utilisateur. Si on sait prédire l'accord ou le désaccord des utilisateurs sur le tempo d'un titre musical, on pourra, dans le cas du désaccord, affiner notre analyse pour prédire un tempo pour chaque utilisateur.

On cherche donc à créer un modèle de classification entre nos quatre indices acoustiques et les classes $\mathcal{A}$ccord et $\mathcal{D}$ésaccord. Le schéma général de notre expérience est présenté dans la figure 4. On extrait du corpus les annotations, et on en déduit les deux classes $\mathcal{A}$ et $\mathcal{D}$ (voir partie 2.4). En parallèle, on extrait les indices acoustiques de l'audio (partie 3.1). Ces indices sont les entrées de quatre modèles d'estimation de l'accord des annotateurs (3.2). Comme on pré-suppose que l'information $\mathcal{A}/\mathcal{D}$ est dans nos indices acoustiques, le travail va consister à trouver le bon modèle qui permette de passer des indices à l'accord inter-utilisateurs.

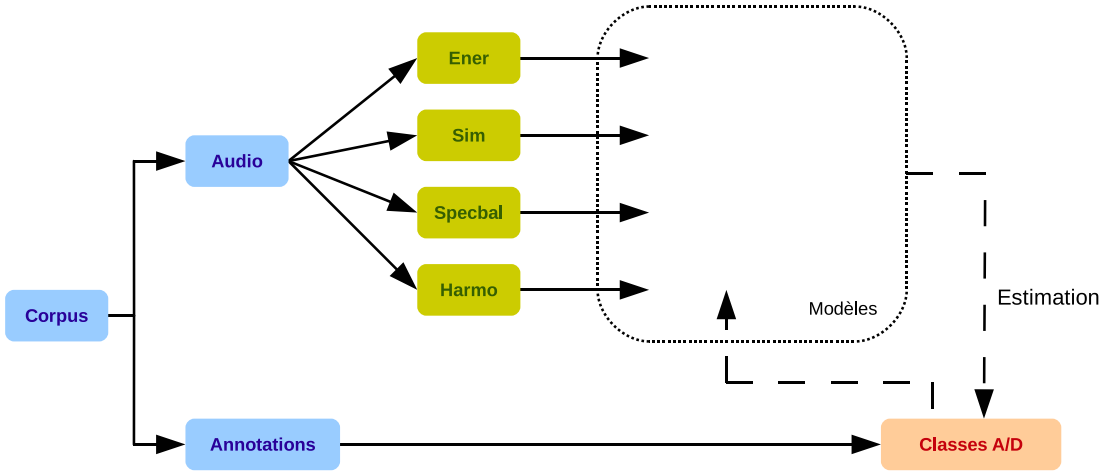


FIGURE 4 – Schéma général de la classification $\mathcal{A}/\mathcal{D}$.

3.1 Les indices acoustiques

Nos modèles d'estimation $\mathcal{A}/\mathcal{D}$ sont basés sur quatre indices acoustiques. Ces quatre indices sont choisis pour leur relation avec la perception du tempo. Ils ont déjà été utilisés plusieurs fois dans le cadre d'algorithmes d'estimation du tempo perceptif, la dernière étant [18]. Nous allons décrire brièvement dans cette partie comment ils sont calculés et en quoi ils peuvent être liés au tempo perçu.

3.1.1 Variation d'énergie $d_{ener}(\lambda)$

Cet indice met en évidence les attaques de notes. Il utilise la variation du contenu énergétique à l'intérieur de bandes de fréquences.

Cet indice a d'abord été introduit dans [10]. Il localise les changements rapides dans le domaine fréquentiel, et se calcule de la manière suivante :

$$E(i) = \sum_{f=f_{min}}^{f_{max}} G(|X(f,t_i)|) - G(|X(f,t_{i-1})|)$$

où $X(f,t_i)$ est la transformée de Fourier du signal $x(t)$, à l'instant t_i, et $G(x)$ une fonction telle que $\sqrt{x}$ ou $\arcsin(x)$, qui permet d'éviter que les composantes haute-fréquences ne soient masquées par des composantes basses-fréquences ayant une amplitude plus faible.

Malheureusement, cet indice souffre du problème de résolution fréquentielle contre résolution temporelle. En effet, pour avoir une meilleure détection des changements fréquentiels, on voudrait une résolution fréquentielle élevée, c'est-à-dire une grande fenêtre d'analyse. Or cela serait au détriment de la résolution temporelle, qui est cruciale dans notre cas (on cherche à observer les changements rapides).

Nous calculons donc notre indice acoustique $d_{ener}(\lambda)$ par une autre méthode, proposée par Peeters dans [16], qui permet d'éviter ce problème. Cette méthode utilise le spectrogramme réassigné. Cela consiste à repositionner l'énergie des points de la transformée de Fourier à court-terme à leurs centres de gravité (ω, t). $d_{ener}(\lambda)$ est enfin obtenu en calculant la fonction d'auto-corrélation de cette fonction d'énergie.

La fonction d'énergie basée sur ce spectrogramme montre une meilleure localisation des attaques. Le logiciel Ircambeat, développé par l'IRCAM [16], est entièrement basé sur cet indice et a donné de bons résultats [3].

3.1.2 Similarité à court-terme $d_{sim}(\lambda)$

Cet indice est basé sur l'hypothèse que la perception du tempo est liée au taux de répétitions à court-terme d'évènements musicaux. Il repose sur le calcul d'une matrice d'auto-similarité [4]. La méthode utilisée est décrite en détails dans [17].

Pour expliquer brièvement, on calcule 3 matrices d'auto-similarité, sur 3 aspects différents de l'audio : une représentant le timbre (MFCC), une autre l'harmonie (utilisant les chromas [5]) et une dernière le rapport bruit/harmonicité (en utilisant des coefficients de contraste spectral [9]).

On somme ensuite ces trois matrices en une seule matrice d'auto-similarité $S(t_i, t_j)$, où chaque point représente la similarité entre les temps t_i et t_j. On convertit ensuite cette matrice en matrice de retard $L(t_i, l_j)$, où $l_j = t_j - t_i$ représente le retard entre répétitions. On somme enfin cette matrice selon les t_i pour obtenir $d_{sim}(\lambda)$, où λ représente les retards.

3.1.3 Balance spectrale $d_{specbal}(\lambda)$

Ce vecteur met en évidence l'alternance d'énergie entre hautes et basses fréquences. En musique populaire, cela correspond à l'alternance grosse caisse / caisse claire. On fait une forte hypothèse

3. http ://music-ir.org/mirex/wiki/2009 :Audio_Beat_Tracking_Results

ici, à savoir que le titre est à 4 temps, et qu'on aura une alternance entre les premiers / troisièmes temps (souvent grosse caisse, donc basses fréquences) et les deuxièmes / quatrièmes temps (souvent caisse claire, donc hautes fréquences). Cette hypothèse fonctionne aussi dans le cadre des mesures à deux temps, mais exclue tous les rythmes ternaires. Elle est donc particulièrement adaptée à la musique pop/rock, principalement présente dans notre corpus (figure 1) et dans les corpus habituels d'estimation du tempo.

Le calcul de cet indice est décrit dans [19]. On cherche à déterminer ensuite si l'instant t_i correspond à un temps fort (premier et troisième temps) ou à un temps faible (deuxième et quatrième temps) On calcule le ratio $r(t_i)$ entre énergie à haute-fréquence et énergie à basse fréquence. Ensuite, pour plusieurs hypothèses de tempo T_h, on calcule les ratios de chaque temps de la mesure $r_n(t_i)$, $n = 1..4$. On calcule enfin $r_{final}(t_i) = r_{n=1,3}(t_i) - r_{n=2,4}(t_i)$. Si l'hypothèse de tempo T_h est correcte, ce ratio r_{final} sera élevé sur les premiers et troisièmes temps et faible sinon. Les variations de ce ratio calculées par auto-corrélation donnent $d_{specbal}(\lambda, T_h)$. On somme la fonction sur les T_h pour obtenir $d_{specbal}(\lambda)$.

3.1.4 Harmonicité $d_{harmo}(\lambda)$

Pour cet indice, on utilise l'idée que la musique populaire est souvent une succession de segments harmoniques homogènes (en musique populaire, on a souvent un accord par mesure). Le taux de changement d'accord est directement proportionnel au tempo. On fait donc l'hypothèse qu'il y a un accord par mesure, et que celle-ci est à quatre temps. Le taux de changement d'accord vaut donc $\frac{1}{4}$ du tempo.

Cet indice est décrit dans [18]. On calcule d'abord le vecteur de Chromas, à douze dimensions, qui donne une représentation d'un temps t, en fonction de son contenu harmonique (des douze demi-tons) : on obtient une matrice de Chromas $C(l, t)$, $l = 0, 1, ..., 11$. La variation temporelle de cette matrice est ensuite calculée, comme dans [18]. On prend enfin l'auto-corrélation de ce résultat, que le notera $d_{harmo}(\lambda)$.

On obtient donc une fonction représentant les changements d'accords de l'oeuvre en fonction du retard temporel λ entre deux instants. La variation de cette fonction est (d'après notre hypothèse) 4 fois plus rapide que le tempo (en secondes, ou 4 fois plus lente en bpm).

3.1.5 Réduction de dimension

On possède donc 4 vecteurs d'indices acoustiques notés $d_{ener}(\lambda)$, $d_{sim}(\lambda)$, $d_{specbal}(\lambda)$, et $d_{harmo}(\lambda)$, que l'on notera $d_i(\lambda)$, $i \in [1..4]$. Ces vecteurs ont une grande dimensionnalité, ce qui est trop discriminant dans le cas d'estimation de tempo. On va donc utiliser une technique de floutage pour les rendre plus invariants : on applique un banc de filtres sur leur axe temporel. On utilise 20 filtres triangulaires, logarithmiquement espacés entre 32 et 208 bpm. Chaque vecteur $d_i(\lambda)$, $i \in [1..4]$ est donc multiplié par ce banc de filtre, ce qui donne 4 vecteurs à 20 dimensions, que l'on notera $d_i(b)$, $b \in [1..20]$, $i \in [1..4]$.

Cette méthode a l'avantage aussi de réduire la dimensionnalité des vecteurs, ce qui les rend plus facilement utilisables dans un algorithme d'apprentissage statistique.

Pour réduire encore cette dimensionnalité, nous avons fait une Analyse en Composantes Principales (ACP), nous avons gardé seulement les axes composant plus de 95 % de la variance. Cela réduit ce vecteur à 34 dimensions.

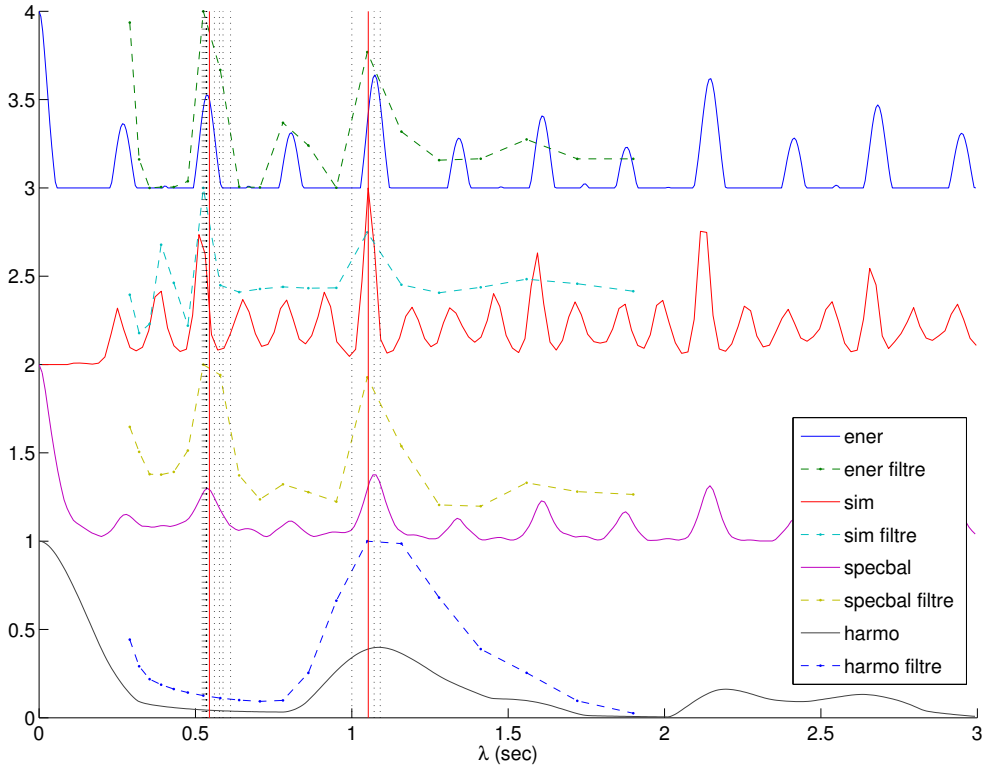


FIGURE 5 – Indices acoustiques pour le titre 'Such a Shame' de 'Talk Talk'. Les lignes pleines représentent les 4 indices acoustiques (voir légende), les lignes en pointillés le filtrage par banc de filtre de ces indices. Les lignes pointillées noires représentent les annotations des utilisateurs, et les lignes verticales rouges les tempi retenus.

La figure 5 montre les 4 indices acoustiques calculés pour un titre du corpus (lignes pleines). On y observe aussi le filtrage par banc de filtre (lignes en pointillés), ainsi que les annotations des utilisateurs (lignes verticales noires, en pointillés), et les tempi retenus (lignes verticales rouges).

3.2 Modèles de classification $\mathcal{A}/\mathcal{D}$

Nous présentons ici quatre modèles d'estimation de l'accord ou du désaccord sur le tempo perceptif, ayant pour entrées les indices acoustiques décrits précédemment (pour rappel, le schéma

général présentant la classification est donné à la figure 4). Dans tous les cas, le but est de passer de l'information se trouvant dans les indices à l'accord inter-utilisateurs.

Le **Modèle MM** est basé sur les travaux de McKinney et Moelants, il s'appuie sur l'existence d'un tempo perceptif préférentiel. Le **Modèle Feature-GMM** consiste simplement à essayer d'apprendre les classes $\mathcal{A}/\mathcal{D}$ directement à partir des indices acoustiques. Le **Modèle Infor-GMM** cherche à exploiter l'hypothèse que l'accord entre les annotateurs est lié à l'accord entre les indices acoustiques. L'accord entre indices est représenté par la corrélation de Pearson puis la divergence de Kullback-Leibler symétrisée. Enfin, le **Modèle Tempo-GMM** exploite la même hypothèse, mais en apprenant les classes $\mathcal{A}/\mathcal{D}$ à partir de quatre estimations de tempo, obtenues à partir des quatre indices acoustiques pris indépendamment.

3.2.1 Modèles MM-Ener et MM-Sim

Ce premier modèle reprend les hypothèses de McKinney et Moleants [13]. La première hypothèse est l'existence d'un tempo préférentiel autour de 120 bpm. La seconde que l'on va chercher à exploiter est décrite dans l'article comme :

Si un extrait musical contient un niveau métrique (dans notre cas, un des indices acoustiques) dont le tempo est proche du tempo résonnant (entre 110 et 170 bpm d'après les auteurs), le tempo perçu a de grandes chances de ne pas être ambigu (i.e. les annotateurs sont d'accord entre eux). Si, par contre, les tempi principaux encadrent le tempo résonnant, le tempo perçu aura tendance à se séparer en plusieurs valeurs, et les annotateurs ne pas être d'accords.

Dans ce modèle, on va donc prendre un indice acoustique, détecter ses deux pics principaux, et voir s'il l'un d'eux appartient à l'intervalle $110 - 170$ bpm. Si oui, on classifie le titre dans la catégorie $\mathcal{A}$, sinon dans la classe $\mathcal{D}$. Le principe est résumé sur la figure 6.

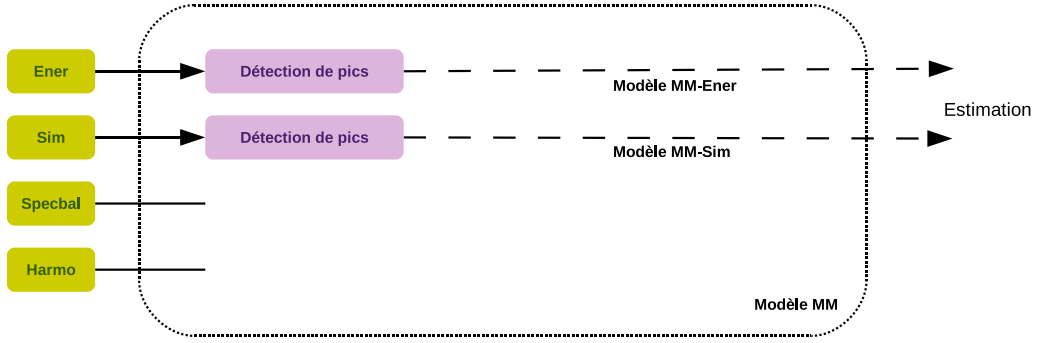


FIGURE 6 – Schéma du modèle MM.

On n'applique ce test qu'aux deux indices donnant les meilleurs résultats (d'après [18]), à savoir l'énergie et la similarité à court-terme. On appelle les deux modèles résultants **Modèle MM-Ener** et **Modèle MM-Sim**.

Il est important de noter que l'on ne travaille pas exactement sur les indices obtenus dans la partie 3.1, mais sur les indices, calculés identiquement, mais dont la dernière étape d'auto-corrélation est remplacée par une DFT. Les indices calculés ainsi donnent de très légèrement moins bons résultats

lors d'une estimation de tempo, mais ont des pics principaux bien définis et peu nombreux alors qu'avec l'auto-corrélation, on a des pics avec une forte amplitude à tous les multiples du retard minimal.

On voit dans la figure 7 un exemple de détection de pics sur la fonction d'énergie. La fonction d'énergie est en bleu. Les croix rouges représentent les pics détectés, les deux gros points rouges sont les deux pics principaux. Le trait pointillé rouge vertical correspond au tempo prédominant (120 bpm). Les traits verts pleins verticaux représentent l'intervalle $[110 - 170]$ bpm. Si un pic se trouve dans cet intervalle, on estimera qu'il y a accord ($\mathcal{A}$) entre les annotateurs. Dans le cas de la figure 7, les deux pics principaux encadrent l'intervalle, on estime donc qu'il y aura désaccord ($\mathcal{D}$).

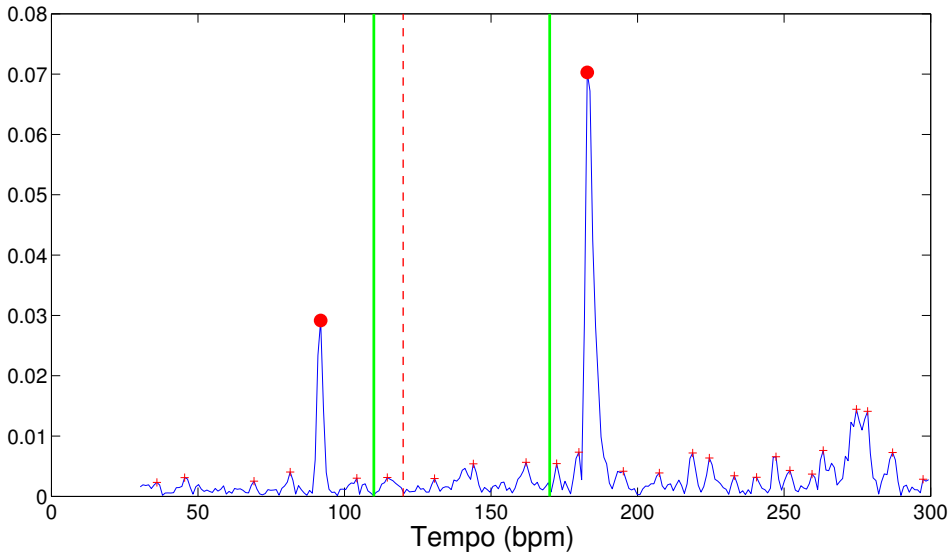


FIGURE 7 – Détection de pics. On a tracé la fonction d'énergie en bleu, les croix rouges sont les pics détectés, les gros points rouges sont les deux pics principaux retenus. Les lignes pleines verticales vertes représentent l'intervalle $[110 - 170]$ bpm, et la ligne pointillée verticale rouge le tempo de résonance 120 bpm.

3.2.2 Modèle Feature-GMM

Ce modèle paraît le plus intuitif. On possède déjà des indices acoustiques qui sont capables de prédire correctement le tempo [18], pourquoi ne seraient-ils pas capable de prédire aussi l'accord / désaccord ?

Le principe du modèle est décrit sur la figure 8. On utilise d'abord une ACP pour réduire la dimension des quatre vecteurs (voir partie 3.1.5). On passe ainsi de 80 à 34 dimensions. À partir de ce vecteur, on entraîne deux modèles GMM à 4 gaussiennes, et avec une matrice de covariance pleine : un pour la classe $\mathcal{A}$ccord, l'autre pour la classe $\mathcal{D}$ésaccord. On attribue ensuite aux titres inconnus la classe dont la probabilité à posteriori est la plus grande (cette classification par GMM

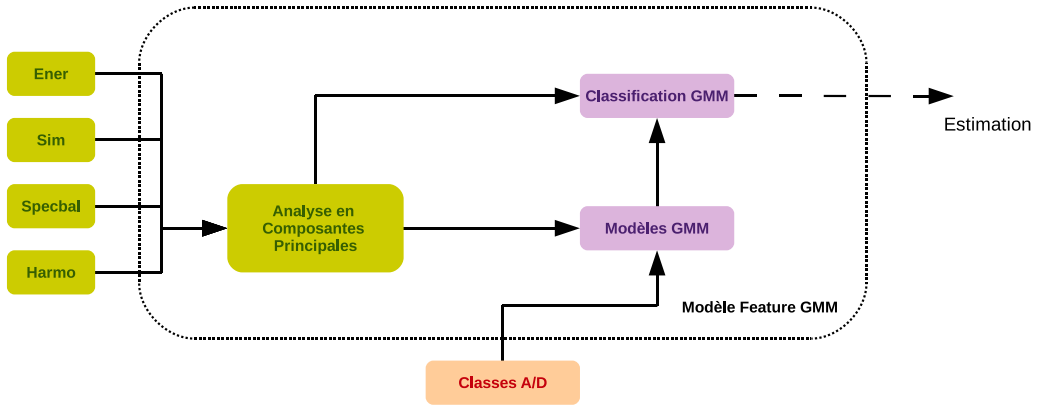
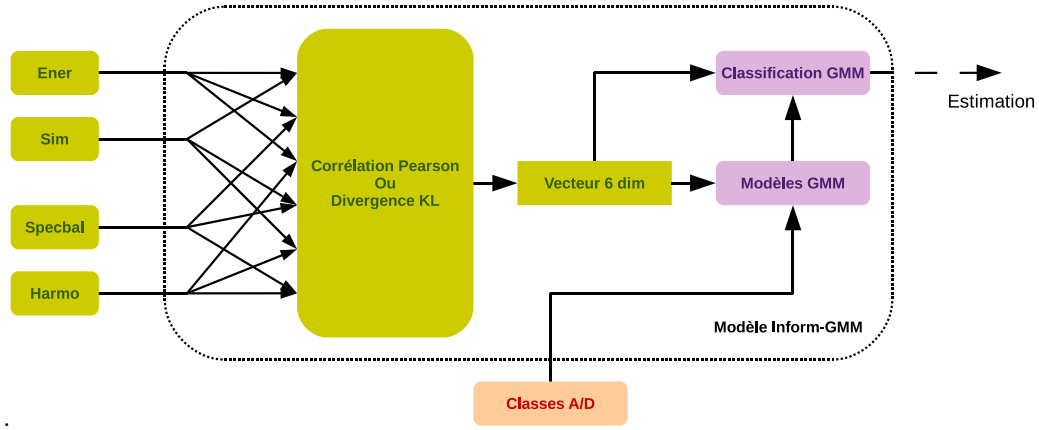


FIGURE 8 – Schéma du modèle Feature-GMM.

est décrite plus en détails dans la partie 4.1).

3.2.3 Modèle Inform-GMM



.

FIGURE 9 – Schéma du modèle Inform-GMM.

Les indices acoustiques $d_i(b)$ représentent les périodicités du signal suivant plusieurs points de vue différents. $d_{ener}(b)$ représente les variations énergétiques, $d_{sim}(b)$ les similarités structurelles à court-terme, $d_{specbal}(b)$ l'alternance hautes/basses fréquences, et $d_{harmo}(b)$ les changements harmoniques. Nous faisons l'hypothèse que si deux vecteurs $d_i(b)$ et $d_{i'}(b)$ apportent la même information de périodicité, il apportent aussi la même information de tempo perceptif.

Nous mesurons l'information partagée par les indices acoustiques $d_i(b)$:

$$\underline{C} = [c(d_{ener}, d_{sim}),\ c(d_{ener}, d_{specbal}), ...]$$

où c est une fonction qui teste le partage d'informations entre les deux indices acoustiques. Nous allons tester par la suite deux types de fonctions : la corrélation de Pearson, et la divergence de Kullback-Leibler (KL) symétrisée. On utilise ensuite deux modèles GMM (décrits partie 4.1) appris sur le vecteur $\underline{C}$ à 6 dimensions pour estimer les classes $\mathcal{A}/\mathcal{D}$. La figure 9 résume ce modèle.

3.2.4 Modèle Tempo-GMM

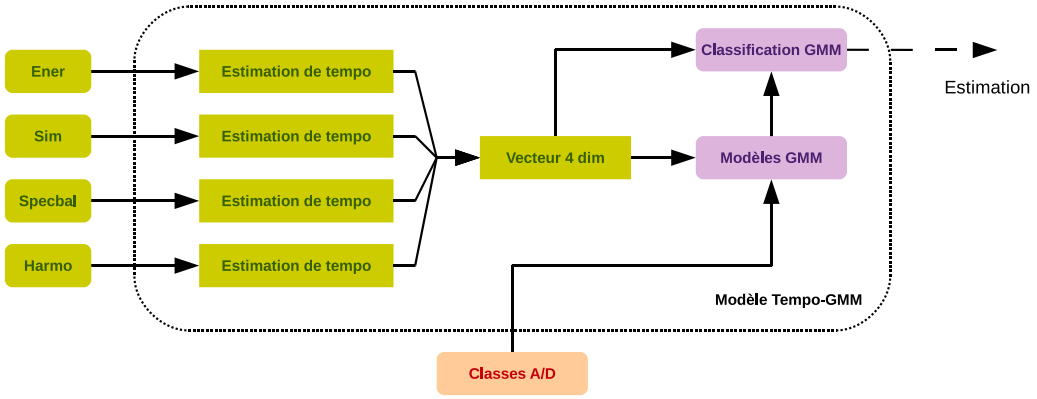


FIGURE 10 – Schéma du modèle Tempo-GMM.

Ce dernier modèle est basé sur la même hypothèse : si les indices acoustiques partagent la même information alors il y aura plutôt $\mathcal{A}$ccord sur le tempo perceptif. Mais dans ce modèle, plutôt que d'apprendre deux modèles GMM sur les indices acoustiques, on va le faire sur quatre estimations de tempo faites sur chacun des indices pris indépendamment. Le principe de ce modèle est résumé sur la figure 10.

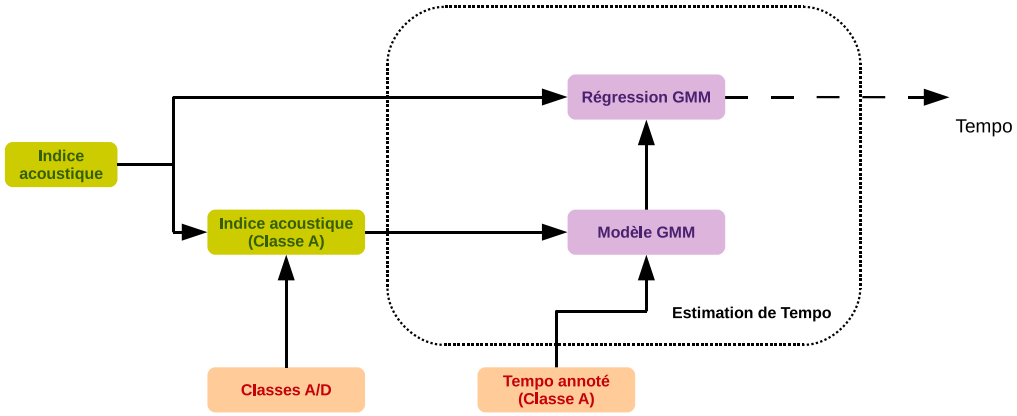


FIGURE 11 – Schéma de l'estimation de tempo sur un indice acoustique.

On estime donc d'abord quatre tempi $\hat{t}_{ener}, \hat{t}_{sim}, \hat{t}_{specbal}, \hat{t}_{harmo}$ sur les quatre indices acous-

tiques. Chaque algorithme d'estimation est une méthode de régression GMM, décrite dans [18]. La figure 11 résume cette estimation de tempo. Pour chaque indice acoustique, on sélectionne d'abord les titres appartenant à la classe $\mathcal{A}$ccord. Grâce à ces titres, on créé un modèle GMM en prenant pour paramètres $d_i(b)$ et le tempo annoté. Ensuite on estime par régression GMM le tempo sur tous les titres (classes $\mathcal{A}$ et $\mathcal{D}$) par ce modèle.

Ensuite ce vecteur $\left[\hat{t}_{ener}, \hat{t}_{sim}, \hat{t}_{specbal}, \hat{t}_{harmo}\right]$ est utilisé comme paramètre d'entrée pour la même estimation par GMM que précédemment (et décrite partie 4.1).

4 Résultats et analyse

4.1 Protocole expérimental

Sur la figure 12, nous présentons notre méthode de classification-GMM. Nous possédons un ensemble de **données**, contenant des descripteurs et les classes ($\mathcal{A}/\mathcal{D}$). Ces données sont d'abord séparées selon la méthode de la validation croisée à N plis (dans notre cas, 5 plis). Nous obtenons donc un **ensemble d'apprentissage** sur lequel on va créer les deux (un pour chaque classe) modèles de mélanges gaussiens (dans notre cas, on utilise 4 gaussiennes), et un **ensemble de test**, sur lequel nous estimons les probabilité d'appartenance à chaque classe. La probabilité la plus forte donne la classe estimée.

La **validation croisée** à 5 plis consiste à séparer l'ensemble de données en 5 fois 2 ensembles d'apprentissage et de test, qui contiendront respectivement $\frac{4}{5}$ des données et $\frac{1}{5}$ des données, de façon à ce que l'ensemble de test couvre toutes les données.

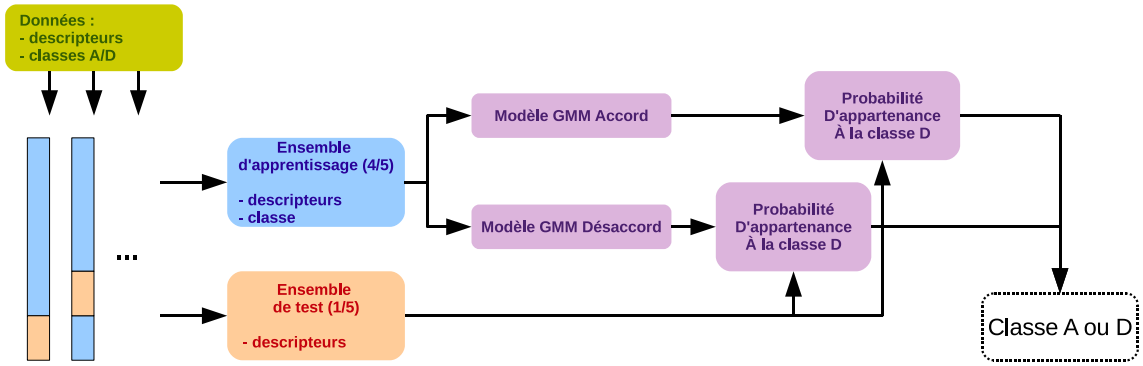


FIGURE 12 – Classification par GMM, avec validation croisée à 5 plis.

Dans notre étude, nous utilisons la toolbox Matlab développé par Sylvain Calinon [1] pour l'implémentation des modèles GMM.

4.2 Résultats des modèles

Dans cette partie, nous présentons les résultats de nos algorithmes de catégorisation à deux classes obtenus par les différents modèles proposés.

Les résultats sont présentés en terme de rappel (en anglais recall) et de rappel moyen (mean-recall). Ce dernier est préféré à la précision et la F-Mesure car il n'est pas sensible à la distribution des classes.

Rappel et rappel moyen

Prenons un exemple simple de classification à deux classes Positifs et Négatifs. Les résultats sont comme indiqués dans le tableau 3.

	Positifs estimés	Négatifs estimés
Positifs	Vrais Positifs (TP)	Faux Négatifs (FN)
Négatifs	Faux Positifs (FP)	Vrais Négatifs (TN)

TABLE 3 – Illustration du rappel : exemple d'une estimation à deux classes Positifs et Négatifs.

Le Rappel est le nombre d'éléments estimé correctement, divisé par le nombre total d'éléments de cette classe. Pour les Positifs : $\frac{TP}{TP+FN}$, pour les Négatifs ce sera $\frac{TN}{FP+TN}$. Le rappel moyen est la moyenne du rappel pour chacune de ces deux classes.

Un rappel moyen de 50 % correspond donc à une estimation aléatoire.

Résultats

	Rappel $\mathcal{A}$	Rappel $\mathcal{D}$	Rappel moyen
Modèle MM-Ener	62.69	42.61	52.65
Modèle MM-Sim	56.71	58.26	57.49
Modèle Feature-GMM	55.21	45.22	50.22
Modèle Inform-GMM (Pearson)	51.51	49.57	50.54
Modèle Inform-GMM (KL)	61.17	50.43	55.80
Modèle Tempo-GMM	**73.73**	**66.52**	**70.10**

TABLE 4 – Résultats obtenus pour la classification en deux classes $\mathcal{A}$ et $\mathcal{D}$ pour les différents modèles présentés dans la partie 3.2, en terme de rappel pour chaque classe $\mathcal{A}$ et $\mathcal{D}$ et de rappel moyen.

Dans le tableau 4, nous présentons les résultats des différentes modèles. La classification $\mathcal{A}/\mathcal{D}$ a été faite en utilisant une validation croisée à 5 plis. Les résultats sont présentés en terme de rappel pour la classe $\mathcal{A}$, rappel pour la classe $\mathcal{D}$, et rappel moyen.

Comme on peut le constater, les modèles MM-Ener, Feature-GMM et Inform-GMM (avec la corrélation de Pearson) ne catégorisent pas mieux qu'un classificateur aléatoire. Les modèles MM-Sim et Inform-GMM (utilisant KL) ont des résultats juste au-dessus de l'aléatoire. Enfin, les résultats obtenus par le modèle Tempo-GMM sont encourageants, et dépassent largement ceux des autres modèles.

Nous allons maintenant analyser en détails ces résultats et essayer d'expliquer leur sens pour chaque modèle.

4.3 Analyse du modèle MM

4.3.1 Modèle de résonance de McKinney et Moelants

McKinney et Moelants supposent qu'il existe un tempo préférentiel autour de 120 bpm. Ils modélisent donc toutes les annotations de leur corpus par une résonance [21] :

$$R = \frac{1}{\sqrt{(f_0^2 - f^2)^2 + \beta f^2}} - \frac{1}{\sqrt{{f_0}^4 - f^4}}$$

Notre premier travail a d'abord été de tester cette hypothèse. On voit sur la figure 13 l'histogramme de toutes les annotations selon leur tempo, la gaussienne rouge est le modèle de résonance décrit précédemment, dont les paramètres ont été trouvés par la méthode des moindres carrés.

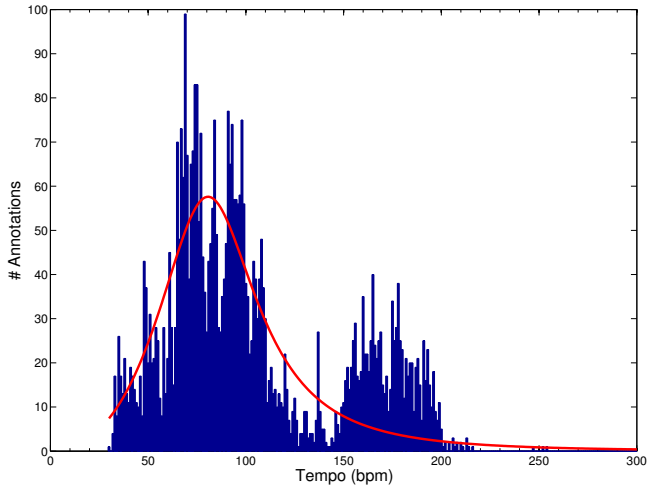


FIGURE 13 – Modèle de tempo prédominant de McKinney & Moelants. L'histogramme de toutes les annotations du corpus selon leur tempo est tracé en bleu. La gaussienne rouge représente le modèle de résonance de McKinney & Moelants.

Pour la base Last-FM, le tempo de résonance est de 80 bpm. On remarque même un vide fréquentiel autour de 120 bpm, ce qui est contraire à l'hypothèse d'un tempo préférentiel à 120 bpm. Ce désaccord avec l'hypothèse de McKinney et Moelants peut s'expliquer par trois facteurs.

- Notre **corpus** est très différent du leur comme le montre la figure 14. Le corpus de McKinney & Moelants est composé de titres également répartis en musique classique, country, dance, hip-hop, jazz, latin, reggea, rock/pop et soul. Le notre composé à plus de 50 % de musique rock/pop, de presque 10 % de country et soul, de seulement 5 % de latin et reggea, et de pas ou peu de classique, dance, hip-hop et jazz.
- Les **annotateurs** n'ont pas les mêmes profils. Dans l'expérience de McKinney & Moelants, les 33 sujets ont une éducation musicale de 7 années en moyenne. Nous estimons que, dans notre cas d'expérience web, la majorité des annotateurs n'a pas reçu d'éducation musicale.
- Enfin, les **protocoles** expérimentaux de création des corpus diffèrent énormément. Le notre est une expérience sur le web, pratiquement sans contrôle des annotateurs, tandis que le corpus de McKinney et Moelants a été créé dans de meilleures conditions.

4.3.2 MM-Ener et MM-Sim

Ces deux modèles ont un rappel moyen de 52.65 % (MM-Ener) et de 57.49 % (MM-Sim), c'est-à-dire juste au dessus de l'aléatoire pour MM-Ener, et au niveau de l'aléatoire pour MM-Sim. Ces

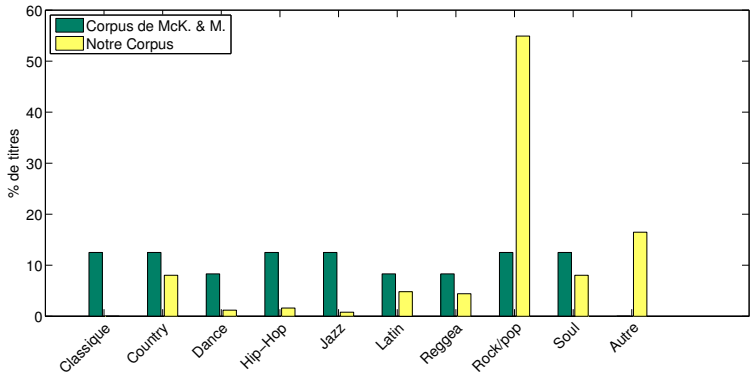


FIGURE 14 – Comparaison de notre corpus avec celui de McKinney & Moelants, en terme de genres musicaux. En vert est représenté la distribution des genres musicaux du corpus de McKinney et Moelants, en jaune la distribution des genres de notre corpus.

modèles ne fonctionnent donc pas sur notre corpus.

Dans [15], cette hypothèse fonctionnait bien sur l'une de leurs expériences, mais mal sur les deux autres. Les auteurs concluent que si un extrait possède un tempo clair autour de 120 bpm, alors il sera très probablement non ambigu, et que en revanche, ralentir ou augmenter ce tempo peut créer des ambiguïtés. Ils concluent aussi que notre perception du tempo peut être modifiée par d'autres paramètres, comme la structure du morceau ou la répartition des accents. Ce modèles ne sont donc pas utilisables sur tous les types de corpus, et pas sur le notre en particulier.

Ce non-fonctionnement peut être du à plusieurs paramètres :

- notre **corpus**, les **annotateurs**, le **protocole** (comme expliqué dans la partie précédente).
- le **tempo préférentiel de 120 bpm** ne correspond pas à notre corpus (figure 13), dont la résonance est à 80 bpm. Pour supprimer cette différence, nous avons donc essayé de changer l'intervalle signifiant l'accord de $[110 - 170]$ bpm à $[70 - 90]$ bpm. Cela n'a malheureusement pas amélioré les résultats, comme on peut le constater dans le tableau 5.
- les **indices acoustiques** ne sont peut être pas appropriés pour ce type de modélisation. Ces indices sont mieux appropriés à un estimation de tempo, car ils possèdent souvent beaucoup de pics, à des multiples du tempo principal. Cela rend l'estimation de tempo plus robuste, mais pollue un peu notre détection de pic. Il est à noter que nous n'avons pas utilisé les indices de balance spectrale et d'harmonicité, vu leur faibles performances en estimation de tempo seul.

4.4 Analyse du modèle Feature-GMM

Ce modèle donne pour résultat un rappel moyen de 50.22 %, c'est-à-dire l'équivalent d'une estimation aléatoire.

	Rappel $\mathcal{A}$	Rappel $\mathcal{D}$	Rappel moyen
Modèle MM-Ener	14.93	87.83	51.38
Modèle MM-Sim	14.93	88.70	45.09

TABLE 5 – Résultats obtenus avec les modèles MM-Ener et MM-Sim en changeant l'intervalle de résonance à $[70 - 90]$ bpm.

Nous avons essayé ce modèle car il donnait de bons résultats en estimation de tempo perceptif [3, 18]. Dans ces études, basées sur le même corpus de Last-FM, il est possible d'estimer de façon fiable le tempo perceptif à partir des quatre indices acoustiques utilisés comme paramètres d'une Régression-GMM (comme celle décrite dans 3.2.4, mais en utilisant les 4 indices simultanément comme paramètres d'entrée du modèle GMM). Notre problème, par rapport à ces articles, est la sélection des titres du corpus. Dans [3, 18], le fait de sélectionner les titres pour lesquels les gens sont d'accord supprime automatiquement les cas un peu flous du corpus. Notre méthode de sélection, plus intuitive, et qui cherche justement les cas ambigus, bénéficie moins de cette protection.

Le manque d'exemples et la trop grande dimension du vecteur de paramètres peuvent aussi être la cause du non-fonctionnement de cette méthode. On ne possède en effet que 250 titres, répartis en deux classes et un vecteur de paramètres de 34 dimensions.

L'information que l'on cherche (accord/désaccord des utilisateurs) n'est donc pas modélisable directement par les indices, c'est pourquoi nous nous sommes intéressés à l'information partagée par ces indices. On note aussi qu'il est intéressant de réduire les dimensions du vecteur de paramètres d'apprentissage, vu la faible taille de notre corpus. Les modèles suivants prennent donc en considération ces deux aspects en réduisant les descripteurs à 6 et 4 paramètres modélisants l'information partagée des indices acoustiques.

4.5 Analyse du modèle Inform-GMM

Les deux modèles ont 50.54 % (Inform-GMM (Pearson)) et 55.80 % (Inform-GMM (KL)) de rappel moyen, c'est-à-dire au niveau de l'aléatoire avec la corrélation de Pearson, et juste au-dessus de l'aléatoire avec la divergence de Kullback-Leibler symétrisée.

Pour mieux comprendre, on a tracé sur la figure 15 les quatre descripteurs dans le cas de la classe $\mathcal{A}$ [gauche] et de la classe $\mathcal{D}$ [droite]. On voit sur la figure de gauche que les trois premiers descripteurs ont bien leurs pics en commun. Par contre, dans le cas du désaccord, on voit que le tempo de la fonction de similarité à court-terme correspond à $\frac{1}{4}$ du tempo représenté par la fonction d'énergie, et celui de la balance spectrale à $\frac{1}{2}$ de celui de l'énergie.

Le problème de ce modèle vient sûrement des fonctions modélisant le partage d'information entre les indices acoustiques, qui ne mesurent pas exactement ce que l'on veut. La corrélation de Pearson, par exemple, prend en compte toute l'énergie commune entre deux indices, même celle hors des pics principaux. On peut se retrouver donc avec des corrélations élevées alors qu'aucun pic n'est commun entre les indices.

Ces deux fonctions modélisent donc mal cette notion d'accord entre indices acoustiques, ce qui nous a poussé au dernier modèle, qui estime d'abord le tempo sur chaque indice, puis estime les

classes à partir de ces quatre tempi.

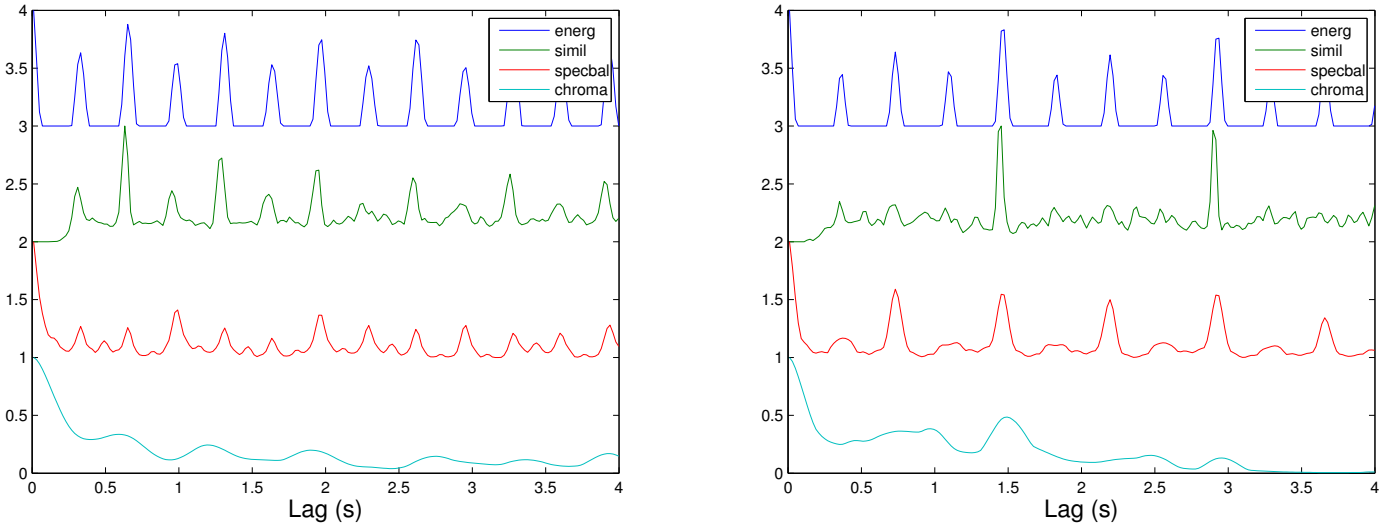


FIGURE 15 – [Gauche] De haut en bas, les indices d'énergie, de similarité, de balance spectrale et d'harmonicité, pour un titre de la classe $\mathcal{A}$. [Droite] La même chose pour un titre de la classe $\mathcal{D}$.

4.6 Analyse du modèle Tempo-GMM

Ce modèle donne 70.10 % de rappel moyen. Ce résultat semble assez prometteur, et nous allons l'analyser un peu plus en détails.

Dans le tableau 6, nous présentons les résultats détaillés pour le modèle Tempo-GMM. Ils montrent que les deux classes sont estimées à peu près au même niveau, même si la classe $\mathcal{A}$ est un peu mieux reconnue que la classe $\mathcal{D}$.

	$\hat{\mathcal{A}}$	$\hat{\mathcal{D}}$
$\mathcal{A}$	**73.7 %**	26.3 %
$\mathcal{D}$	33.5 %	**66.5 %**

TABLE 6 – Matrice de confusion pour le modèle Tempo-GMM. [Gauche] en nombre de titres. [Droite] en pourcentage.

Pour mieux observer les classes, la figure 16 montre la relation entre les tempi estimés $t_1 = \hat{t}_{ener}$, $t_2 = \hat{t}_{sim}$ et $t_3 = \hat{t}_{specbal}$. Les signes '+' rouges représentent les données de la classe $\mathcal{A}$ et les 'x' bleus de la classe $\mathcal{D}$. Les éléments de la classe $\mathcal{A}$ sont majoritairement situés sur la diagonale (ce qui signifie que les tempi estimés par les différents indices sont les mêmes), et les éléments de la classe

$\mathcal{D}$ sont majoritairement situés hors de la diagonale. Ces figures valident donc notre hypothèse que si l'information est partagée entre les indices, les utilisateurs seront d'accord sur la perception du tempo.

Il est à noter que, comme dans [3, 18], l'harmonicité est un indice qui fonctionne mal. On peut le constater dans le tableau 7 qui résume les performances des régressions GMM pour l'estimation de tempo individuelle sur chaque indice. C'est pourquoi nous n'avons pas tracé les relations entre $t_4 = \hat{t}_{harmo}$ et les autres t_i, $i = 1..3$. Les résultats du tableau 7 ont été obtenus en faisant une validation croisée à 5 plis lors de la phase d'estimation du tempo sur chaque indice.

Indice acoustique	Tempo trouvé (%)
ener	85.9 %
sim	70.3 %
specbal	64.3 %
harmo	38.2 %

TABLE 7 – Performances des régressions GMM pour l'estimation de tempo individuelle sur chaque indice.

Pour compléter l'analyse, nous avons décidé de tester une deuxième méthode de classification. Au lieu d'utiliser une classification GMM à partir des 4 tempi, nous utilisons une classification par SVM (Machines à Vecteurs Support). Les résultats sont présentés dans le tableau 8. Ils ont été obtenus en cherchant les meilleurs paramètres sur une grille ($c = 1.59, \gamma = 0.001$). Le rappel moyen est meilleur de 3 % comparé à la classification par GMM. Par contre, les rappels des deux classes sont très déséquilibrés, ce qui nous fait préférer les résultats par GMM.

	Rappel $\mathcal{A}$	Rappel $\mathcal{D}$	Rappel moyen
Modèle Tempo-SVM	87.35 %	44.35 %	**74.85 %**

TABLE 8 – Résultats obtenus pour le Modèle Tempo-SVM.

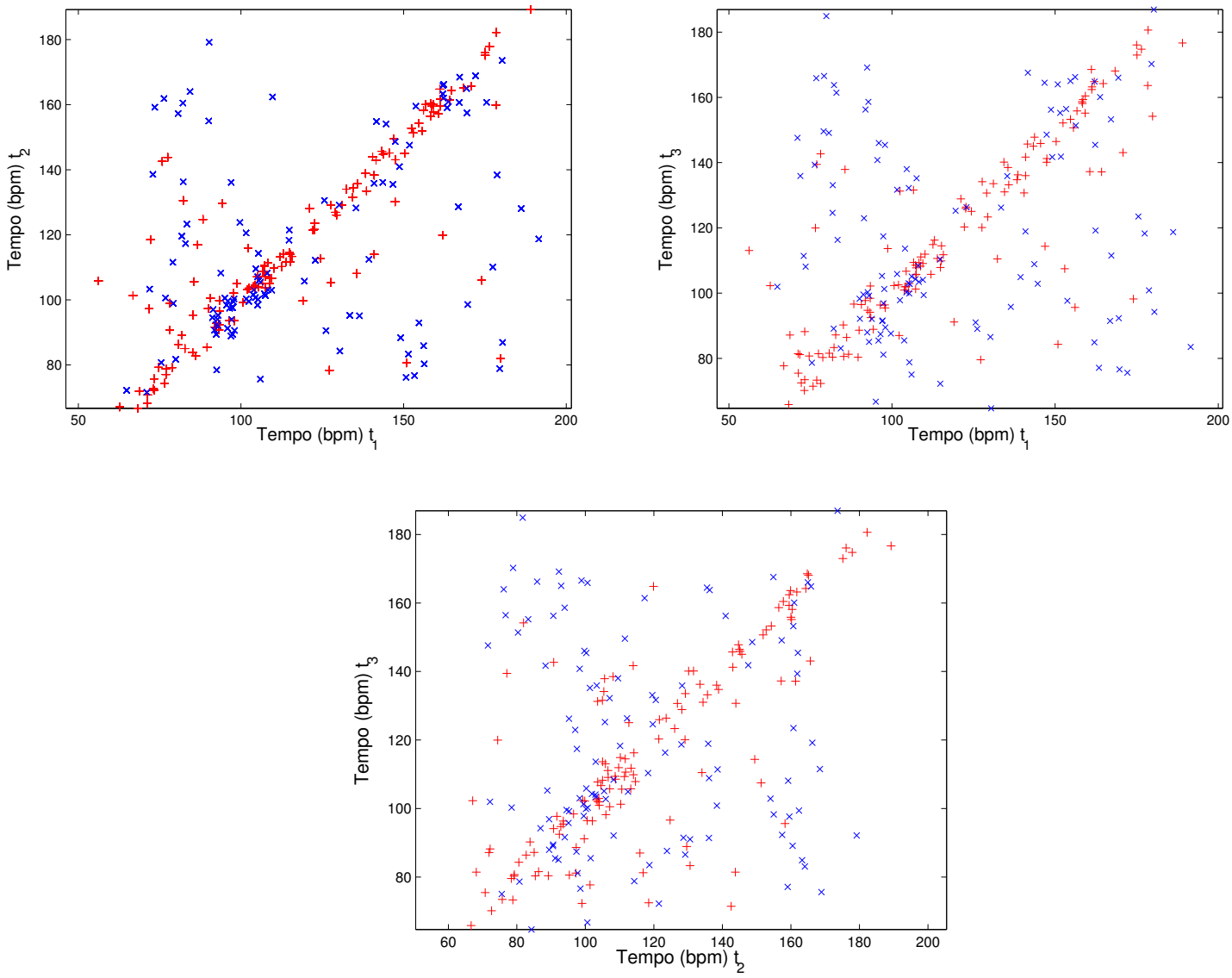


FIGURE 16 – Chaque figure montre la relation entre les tempi estimé [de gauche à droite] $t_1 = \hat{t}_{ener}/t_2 = \hat{t}_{sim}$, $t_1 = \hat{t}_{ener}/t_3 = \hat{t}_{specbal}$ et $t_2 = \hat{t}_{sim}$ / $t_3 = \hat{t}_{specbal}$. Les signes '+' rouges représentent les données de la classe $\mathcal{A}$, les 'x' bleues de la classe $\mathcal{D}$.

Conclusion et perspectives

Durant ce stage, nous avons étudié la prédiction de l'$\mathcal{A}$ccord ou du $\mathcal{D}$ésaccord de la perception du tempo entre utilisateurs. Nous avons proposé plusieurs modèles de prédiction du tempo, basés sur 4 indices acoustiques. Ces indices sont la variation d'énergie, la similarité à court-terme, la balance spectrale et l'harmonicité. Ils présentent tous un aspect perceptif.

Les 4 modèles que nous avons proposés cherchent à modéliser la relation entre ces 4 indices et deux classes : $\mathcal{A}$ (signifiant accord entre les utilisateurs) ou $\mathcal{D}$ (signifiant désaccord entre les utilisateurs).

Le premier modèle est basé sur l'hypothèse de McKinney et Moelants [15] qu'il existe un tempo préférentiel, autour de 120 bpm, par lequel les utilisateurs sont naturellement attirés. Notre corpus (basé sur celui de Last.Fm [11]) ne montre pas une telle résonance à 120 bpm, mais plutôt deux tempi préférentiels vers 90 et 180 bpm. Le modèle d'$\mathcal{A}/\mathcal{D}$ dérivant de cette hypothèse ne donne qu'un résultat de 57% de rappel moyen, c'est-à-dire juste au-dessus d'une classification aléatoire.

Le second modèle classifie directement à partir des 4 indices acoustiques, et donne les mêmes résultats qu'un classifieur aléatoire.

Les troisième et quatrième modèles sont basés sur l'hypothèse que si l'information de tempo est partagée entre les indices acoustiques, alors la perception du tempo sera elle aussi partagée entre les utilisateurs. Pour modéliser ce partage d'informations, le troisième modèle utilise la corrélation de Pearson et la divergence de Kullback-Leibler symétrisée. Un modèle GMM prédit ensuite les classes $\mathcal{A}/\mathcal{D}$. Ce modèle atteint aussi une séparation juste au-dessus de l'aléatoire pour la divergence de Kullback-Leibler symétrisée.

Le dernier modèle estime d'abord le tempo pour chacun des indices acoustiques pris séparément, puis modélise les classes par GMM à partir des quatre tempi obtenus. Ce modèle obtient un résultat de 70% de rappel moyen. On a aussi montré que les tempi estimés, pour des titres appartenant à $\mathcal{A}$, sont corrélés entre eux. C'est le contraire pour les tempi des titres appartenant à $\mathcal{D}$. Ces résultats valident donc notre hypothèse que la cohérence des indices acoustiques facilite l'accord inter-utilisateurs sur le tempo.

Nous avons donc un modèle capable de prédire si la perception du tempo entre deux utilisateurs u et u', pour un extrait audio a, va être partagée ($f(a,u) = f(a,u')$) ou non ($f(a,u) \neq f(a,u')$). L'objectif suivant est de prédire le tempo en prenant en compte l'utilisateur ($f(a,u) = \hat{t}_u \simeq t_u$). Cependant, le corpus que nous possédons, bien qu'il semblait prometteur pour ce type d'expérience, n'est pas assez fourni. En effet, l'étape de fiabilisation du corpus a drastiquement réduit le nombre de titres fiables, et il ne reste plus que 15 annotateurs dans notre corpus, ayant annoté plus de 10 extraits. Pour continuer dans cette voie, une idée serait donc de re-faire une expérience perceptive en laboratoire, pour s'affranchir des limitations qui nous sont imposées par le corpus de LastFM.

Références

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

[2] Ching-Wei Chen, Markus Cremer, Kyogu Lee, Peter DiMaria, and Ho-Hsiang Wu. Improving perceived tempo estimation by statistical modeling of higher-level musical descriptors. In *Audio Engineering Society Convention 126*, 2009.

[3] Joachim Floncon-Cholet. Estimation du tempo perceptif et réduction des erreurs d'octave du tempo.

[4] 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.

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

[6] Aggelos Gkiokas, Vassilis Katsouros, and George Carayannis. Reducing tempo octave errors by periodicity vector coding and svm learning. *Proc. of ISMIR, Porto, Portugal*, 2012.

[7] J Hockman and Ichiro Fujinaga. Fast vs slow : Learning tempo octaves from user data. *Proc. of ISMIR, Ultrecht, The Nederlands*, 2010.

[8] Andre Holzapfel, Matthew EP Davies, José R Zapata, João Lobato Oliveira, and Fabien Gouyon. Selective sampling for beat tracking evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(9) :2539–2548, 2012.

[9] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 113–116. IEEE, 2002.

[10] Jean Laroche. Efficient tempo and beat tracking in audio recordings. *Journal of the Audio Engineering Society*, 51(4) :226–233, 2003.

[11] Mark Levy. Improving perceptual tempo estimation with crowd-sourced annotations. *Proc. of ISMIR,(Miami, USA)*, 2011.

[12] Cory McKay, Ichiro Fujinaga, and Philippe Depalle. jaudio : A feature extraction library. In *Proceedings of the International Conference on Music Information Retrieval*, pages 600–3, 2005.

[13] Martin F McKinney and Dirk Moelants. Deviations from the resonance theory of tempo induction. In *Proc. Conference on Interdisciplinary Musicology*, 2004.

[14] Martin F McKinney and Dirk Moelants. Ambiguity in tempo perception : What draws listeners to different metrical levels ? *Music Perception*, 24(2) :155–166, 2006.

[15] Dirk Moelants and M McKinney. 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*, pages 558–562, 2004.

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

[17] Geoffroy Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. *Proc. ISMIR, Vienna, Austria*, 2007.

[18] Geoffroy Peeters and Joachim Flocon-Cholet. Perceptual tempo estimation using gmm-regression. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 45–50. ACM, 2012.

[19] Geoffroy Peeters and Helene Papadopoulos. Simultaneous beat and downbeat-tracking using a probabilistic framework : theory and large-scale evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(6) :1754–1769, 2011.

[20] Klaus Seyerlehner, Gerhard Widmer, and Dominik Schnitzer. From rhythm patterns to perceived tempo. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 519–524, 2007.

[21] Leon van Noorden and Dirk Moelants. Resonance in the perception of musical pulse. *Journal of New Music Research*, 28(1) :43–66, 1999.

[22] Linxing Xiao, Aibo Tian, Wen Li, and Jie Zhou. Using a statistic model to capture the association between timbre and perceived tempo. In *Proceedings of the International Conference on Music Information Retrieval*, pages 659–662, 2008.

[23] José R Zapata, André Holzapfel, Matthew EP Davies, Joao L Oliveira, and Fabien Gouyon. Assigning a confidence threshold on automatic beat annotation in large datasets. In *Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR 2012), Porto*, 2012.