Mesure rapide de similarités musicales

Perception du rythme

Luigi.lancieri, Lucille.Tanquerel

[1] France Telecom R&D

42 rue des coutures 1400 Caen

Résumé

Cet article décrit une technique de caractérisation rapide de documents sonores basée sur une mesure statistique de la variation du signal. Nous avons montré qu'un échantillonnage très limité des morceaux était suffisant pour obtenir une performance de la caractéristique raisonnable tout en étant 300 fois plus rapide à calculer qu'un échantillonnage complet. Nous avons réalisé une première validation de notre approche en mettant en évidence une corrélation de 0,7 entre la perception humaine du rythme et le rendu de notre caractéristique ainsi qu'une erreur de reconnaissance inférieure à 5%.

Mots clefs

Similarité musicale, rythme, variation.

1 Introduction

De nombreuses sources font état des besoins et du fort potentiel commercial lié à la gestion automatisée de documents sonores. Par exemple l'IFPI (International Federation of the Phonographic Industry) a annoncé que le chiffre d'affaires global des services de vente de musique numérique en ligne a été multiplié par 10 en 2004 par rapport à 2003. Les analystes sont très confiants et annoncent qu'en 2005-2006 ce type de services devrait générer un chiffre d'affaires de l'ordre 330 millions de dollars [12].

La description des caractéristiques sonores d'un document est un élément clé pour réaliser des traitements automatiques impliquant des données audio. Ce type de mesure peut être utile non seulement pour caractériser les données mais aussi pour décrire les goûts musicaux des usagers sur la base de leurs activités d'écoute. Ces techniques deviennent critiques compte tenu de la quantité croissante de documents sonores, que ce soit sur le Web ou dans les bases de données musicales des fournisseurs de contenus. De nombreux travaux ont été réalisés dans ce domaine mais les techniques de traitement restent lourdes à mettre en œuvre et manquent de standards.

L'objectif de ce document est de décrire une méthode permettant de caractériser de manière compacte et rapide le rythme associé à un fichier sonore par l'extraction de caractéristiques physiques réparties sur le fichier (analyse spectrale du signal). L'innovation de notre proposition porte sur l'organisation de l'extraction des échantillons et sur le mode d'analyse pour fournir très rapidement une signature représentative de la nature rythmique du contenu musical.

L'organisation de l'extraction définit la manière dont les échantillons sont prélevés. Il paraît possible, par exemple, de déterminer le spectre sur tout le fichier musical ou seulement sur la première minute. Notre proposition vise à réaliser un échantillonnage statistique séquentiel minimal réparti sur le fichier sonore selon une loi de probabilité particulière. Le principe de cette proposition est basé sur le postulat que la collecte d'une faible quantité d'échantillons de petite durée suffit pour avoir une information résumant de manière efficace le rythme perçu. L'état de l'art montre, par exemple, qu'un individu est capable de reconnaître un genre musical dans 70 % des cas après avoir écouté seulement 3 secondes d'une bande son [1]. Notre méthode de validation repose d'une part sur la comparaison de la signature rythmique avec la perception humaine et d'autre par sur une mesure d'erreur de reconnaissance objective. Dans ce dernier cas, nous montrons que la signature rythmique permet de comparer les morceaux entre eux et d'identifier fidèlement les morceaux identiques même si ceux ci ne sont pas complets.

Dans la suite de ce document, après avoir détaillé les différents éléments de notre approche, nous proposons un état de l'art de travaux comparables ainsi qu'une présentation de quelques résultats.

2 Description générale

La figure suivante montre les bases du processus d'obtention de la signature à partir de l'analyse d'échantillons prélevés dans un fichier sonore. L'idée est de capturer l'image du balancement du spectre sonore tel que l'on peut le percevoir en observant le barre-graph d'un lecteur audio. Les échantillons à analyser sont collectés par triplets (E0, E1,..) de spécimen contigus de durée k. Dans cette première étude, chaque triplet est

collecté de manière aléatoire mais en respectant un ordre chronologique. C'est-à-dire que si on décide de prélever 10 triplets, la seule contrainte sera que le premier précède le second qui devra précéder le troisième, etc. L'espace de temps entre chaque triplet pourra être quelconque.

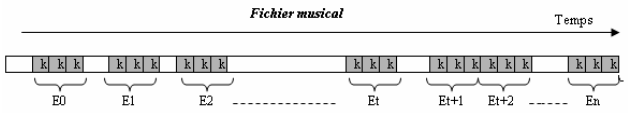

Figure 1 - *Collecte des échantillons par triplets dans un fichier sonore*

Sur chaque échantillon k de chaque triplet est calculée la répartition de fréquences au sens de Fourrier [2] puis, le coefficient directeur *p* de la droite de régression liant le niveau *(y)* à chaque classe de fréquence *(x)* du spectre. Cette droite de régression s'exprime de la manière suivante : $y = px + b$.

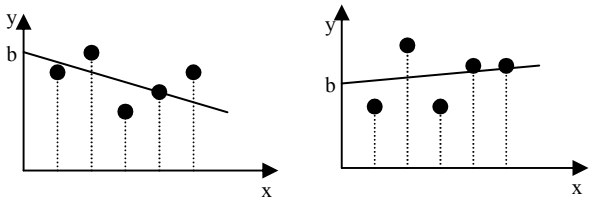

Figure 2 - *La pente du spectre de 2 éléments d'un triplet.*

L'analyse du comportement de *p* (pente de la droite de la figure 2) va contribuer à évaluer le comportement rythmique en mesurant le balancement du spectre sur une période, et en valeur moyenne sur les différents échantillons. Par référence à la mécanique, ce balancement, sa vitesse et son accélération sont évalués de la manière qui suit.

La première étape consiste à identifier le nombre de triplets ainsi que leur position dans le signal. Sur une fraction du fichier sonore, on extrait le premier triplet sur lequel on calcule les 3 spectres puis les coefficients directeurs des droites de régression. On obtient ainsi 3 valeurs de pente $(p1_1, p1_2, p1_3)$. La vitesse du balancement est obtenue en calculant l'écart entre 2 pentes consécutives. On obtient 2 valeurs de vitesses $(v1_1, v1_2)$, pour chaque triplet. L'accélération a_1, unique par triplet, est évaluée sur l'écart des vitesses. On recalcule ces données sur le triplet suivant et ainsi de suite jusqu'à la fin du fichier. A la fin de l'opération on dispose d'un ensemble de valeurs de coefficients $(p1_1, p1_2, p1_3, p2_1, p2_2, p2_3, ..., pn_1, pn_2, pn_3)$, de vitesses $(v1_1, v1_2, v2_1, v2_1, ..., vn_1, vn_2)$ et d'accélérations *(a1, a2, ..., an)* pour n triplets représentatifs du morceau de musique. Le comportement du balancement (position, vitesse et accélération) est obtenu par une combinaison des valeurs moyennes et de l'écart type de toutes ces données (pi, vi, a_i).

Une difficulté importante que nous n'abordons que partiellement ici est de définir la proportion idéale de ces caractéristiques (pi, vi, ai). Pour commencer, nous n'utiliserons que la vitesse comme image du rythme. Dans d'autres travaux, en cours, nous évaluons l'influence des autres grandeurs (pente et accélération) pour optimiser la représentativité ou pour définir d'autres caractéristiques que le rythme. La signature du fichier musical est donc constituée par une valeur numérique combinant la moyenne et l'écart type de la vitesse. Cette valeur sera utilisée dans des métriques de comparaison avec l'évaluation humaine.

3 Etat de l'art

Le procédé décrit dans ce document se distingue de l'art antérieur par une meilleure capacité descriptive rapportée aux ressources de calcul et de stockage nécessaires. La capacité descriptive est liée à l'évaluation de la rythmique par l'analyse de structure de balancement. Ces éléments n'ont pas besoin d'être obtenus sur tout le fichier sonore, un échantillonnage statistique limité suffit. La signature ne nécessite a priori que le stockage d'une quantité très limitée de données numériques (une seule ici). D'autre part, la signature sera quasiment indépendante du format ou de la qualité sonore du morceau, même si ce dernier est incomplet.

Les techniques existantes pour la caractérisation de fichiers musicaux et les recherches de similarités (MIR – Music Information Retrieval) sont très variées. Il existe trois principales approches : celles basées sur le traitement du signal, le filtrage collaboratif, et la fouille de données. Les approches basées sur le traitement du signal consistent à analyser directement le contenu du morceau (signal et spectre) et peuvent être appliquées à n'importe quel fichier audio. En général, ces caractéristiques sont modélisées par des systèmes d'apprentissage, et des comparaisons sont effectuées pour la recherche de similarités [3, 4]. Par exemple, dans ses travaux, Georges Tzenakis [3] extrait une liste de caractéristiques obtenues à partir de l'enveloppe du signal et des données spectrales, notamment le centroïd (mesure de la luminance spectrale), le rollof (mesure de la forme du spectre), le ZeroCrossings (nombre de fois où la courbe du signal passe par le zéro) et parfois même les MFCC (Mel-frequency spectral coefficients) [5], caractéristiques couramment utilisées dans la reconnaissance vocale. Ces caractéristiques sont calculées dans des fenêtres d'analyse successives de taille fixe et seulement sur les 30 premières secondes du morceau. Un autre exemple de technologie en matière d'empreintes acoustiques est la TRM (This Recognizes Music) [11]. Cette technologie a été mise au point par la société américaine Relatable. Concrètement, ce système permet la reconnaissance de morceaux de musique par analogie acoustique exploitant une empreinte de type "code barre audio" qui génère une signature unique. Dès

que l'empreinte numérique a été créée, elle est envoyée au serveur TRM, qui compare l'empreinte à celle d'une chanson existante dans la base de données d'un client. La dernière version commerciale du serveur TRM peut gérer plus de 5000 empreintes par seconde, ou jusqu'à plusieurs milliards de requêtes par jour.

Avec le développement du Web, d'autres techniques basées sur des données publiques ont émergé [6, 7]. Elles utilisent l'analyse de texte et des techniques de filtrage afin de combiner des données provenant de divers individus pour déterminer des similarités basées sur des informations subjectives. Les techniques de filtrage collaboratif sont basées sur la comparaison de profils utilisateurs et représentent la technique principale utilisée aujourd'hui dans les systèmes de recommandations (Amazon, AllMusicGuide, etc.). L'avantage du filtrage collaboratif est que c'est une technique relativement simple à implémenter. Le principal inconvénient est le fait qu'elle requiert un très grand nombre d'utilisateurs d'un système donné pour être significative. Les méta-données culturelles sont des informations décrivant l'opinion publique et les tendances culturelles provenant de divers textes non structurés associés aux contenus et produits par le public. L'utilisation de ces informations pour juger de la similarité entre artistes musicaux a l'avantage d'exploiter des données complémentaires largement distribuées.

En dehors de l'aspect proprement musical, certaines techniques (dont la nôtre) peuvent être utilisées dans un contexte de DRM (Digital Right Management). Un exemple d'avantage est l'identification de fichiers musicaux tronqués ou piratés qui ne pourrait pas forcement être pris en charge par des techniques de DRM plus traditionnelles comme le watermarking. Comparée aux systèmes traditionnels, une DRM basée sur la similarité acoustique a de nombreux avantages (facile à mettre en œuvre, mieux tolérée par l'usager final, ..) même si la fiabilité peut être plus limitée.

4 Mesure de performances

Pour évaluer la pertinence de notre méthode, nous utilisons 2 ensembles de morceaux de musiques différents. L'un contrôlé, l'autre composé aléatoirement. Ces 2 sélections vont, dans un premier temps, être confrontées à l'opinion de 10 évaluateurs humains dont nous comparerons la perception à celle de notre système. Dans un second temps, le premier ensemble sera utilisé pour évaluer le taux d'erreur de reconnaissance et la robustesse de la signature (reproductibilité de la reconnaissance).

Le premier ensemble « calibré » comporte 26 morceaux de musique que nous avons choisis a priori de manière à couvrir une large plage de spectre rythmique. A titre d'exemple de morceaux rythmés citons « la Sonate n 9 pour piano » de Wolfgang Amadeus Mozart ou « The easy winner » de Scott Joplin. Pareillement pour les morceaux peu rythmés citons « l'Allemande » de la Suite pour violoncelle de Jean-Sébastien Bach, ou « Pièce pour haut-bois et harpes » de Gabriel Fauré. Nous avons volontairement limité le spectre des genres à la musique classique et au jazz de manière à ne pas introduire trop de paramètres dans l'étude. Pour augmenter la représentativité de cet ensemble nous avons réalisé 50 mesures de signature pour chacun des 26 fichiers (1300 signatures au total) sachant que chacune de ces 50 signatures peut être différente, en particulier pour les taux de couverture faibles. Il est en effet important de vérifier que toutes les signatures d'un même morceau sont cohérentes. Le second ensemble n'est pas calibré et correspond à 50 morceaux de musique choisis aléatoirement parmi 700 figurant au programme actuel de quelques radios généralistes comme SKY FM.

En plus de la comparaison de ces 2 ensembles avec la perception humaine, nous déterminons la capacité intrinsèque de discrimination de la signature par le biais d'une matrice de confusion. Cette technique permet, en comparant 2 à 2 les signatures, de calculer les erreurs de reconnaissance (similitudes reconnues à tort) et de non reconnaissance (similitudes réelles non reconnues). Cette technique sera appliquée aux 1300 signatures du premier ensemble.

Avant d'évaluer les performances de notre méthode, nous étudions la sensibilité de la signature aux différents paramètres qui y sont liés. Pour mémoire, ces paramètres sont le taux de couverture du morceau (T : de 5 à 75 %) et la taille d'un élément du triplet de base (K : de 1024 à 16384 octets), (voir figure 1).

4.1 Capacité descriptive de la signature

L'écart type est une mesure intéressante de la stabilité et de la performance des résultats d'un processus. En effet un écart type faible implique qu'au travers des nombreux tests, les résultats sont très proches (reproductibilité). Dans notre cas, il se trouve que l'écart type de la vitesse est l'élément de base de la signature. Pour évaluer l'influence de cette composante aux différents paramètres (K, T), nous l'agrégeons pour tous les morceaux. C'est donc l'influence de K et T sur cet agrégat que nous évaluons. L'agrégat est produit de la manière suivante. Pour chaque morceau de musique nous calculons l'écart-type EC correspondant à chaque morceau (i.e écart type de la vitesse). Comme nous calculons 50 signatures pour chaque morceau et pour un taux de couverture donné, nous obtenons par exemple pour une couverture de 5% :

Morceau 1 (T=5%) : EC1-1,…. EC1-50
Morceau i (T=5%) : ECi-1……ECi-50
Morceau 26 (T=5%) : EC26-1….EC26-50

Nous calculons ensuite pour chaque morceau la moyenne Mi des (ECi-1,.., ECi-50) et enfin ME, l'agrégat correspondant à la l'écart-type des Mi. Les courbes qui suivent montrent comment ME est influencé par la taille de l'échantillon élémentaire (K) ainsi que du taux de couverture du morceau (T). Naturellement, cette influence est moyennée mais elle permet de se faire une opinion globale.

Nous interprétons ces courbes de la manière suivante. Une discrimination importante entre les différents morceaux implique une valeur de ME élevée. A la limite, une valeur de ME nulle, indique que chaque morceau produit une caractéristique identique aux autres ce qui implique un pouvoir de discrimination nul.

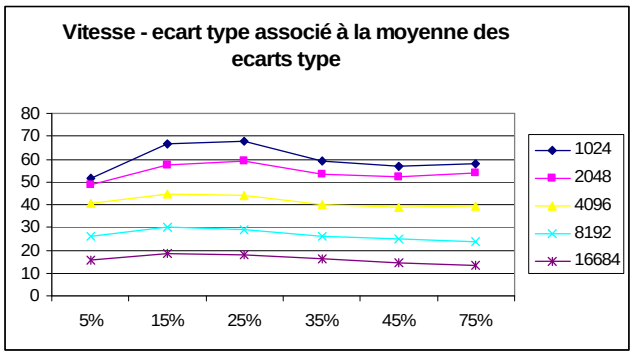

Figure 2 : Evolution de l'agrégat ME (ord) en fonction du taux de couverture (abs) et de la taille de l'échantillon élémentaire (paramètre).

On constate de manière assez prévisible qu'une taille d'échantillon élémentaire importante engendre une stabilité de la métrique. Ceci est vérifié par le fait que ME est au minimum et varie peu en fonction du taux de couverture pour une taille de 16684 octets. Ce résultat s'explique par l'effet d'intégration produit par l'évaluation du spectre sur des échantillons larges. A la limite, une taille d'échantillon élémentaire très grande recouvrant, par exemple, la première moitié d'un morceau de musique aurait une très forte probabilité de produire une métrique quasi identique (écart-type quasi nul) comparée à celle de la seconde moitié. Une taille d'échantillon trop importante est donc à proscrire si l'on souhaite une métrique représentative du contenu. De la même manière, un taux de couverture trop faible ou trop élevé finit par être pénalisant.

Ce qui est aussi intéressant dans ce graphique c'est la mise en évidence d'une relation non linéaire entre le taux de couverture et le niveau de discrimination de la métrique. Ceci implique qu'au-delà d'un certain niveau de taux de couverture, le gain en performance de discrimination s'affaiblit. C'est ce que l'on peut observer par une valeur de ME quasi identique entre 15 et 25 %, puis décroissante ensuite. Ainsi non seulement l'utilisation d'un taux de couverture important est pénalisant en terme de temps de calcul mais en plus il diminue la performance de la métrique. Le raisonnement est de même nature pour des valeurs faibles. L'idéal semble être un taux compris entre 10 et 20 %. En réduisant ce taux, on affaiblit les performances de la discrimination mais de manière très limité comparé au gain en temps de calcul. En effet, en prenant 3 fois moins d'échantillons (passage de 15% à 5% du taux de couverture) on ne réduit la « performance » de la catégorisation que de l'ordre de 20 %. (passage de 68% à 53 % de ME pour une taille d'échantillon de 1024). Ainsi, puisque notre objectif est d'obtenir une mesure rapide, nous utiliserons dans les tests de performances qui suivent un taux de couverture volontairement très faible compris entre 1 et 5 %.

4.2 Matrice de confusion

La matrice de confusion porte sur le premier ensemble et a pour objectif de comparer les échantillons deux à deux afin de tester la capacité de reconnaissance de la mesure de similarité. Pour une caractéristique et une métrique données, il devrait être possible de reconnaître les morceaux identiques et ceux qui sont différents. Le pourcentage de réussite permet d'apprécier la fiabilité du couple caractéristique-métrique. Nous évaluons cette performance pour l'échantillon des 26 morceaux représentés par les 1300 signatures. En plus de tester la fonction discriminante, ceci permet d'évaluer la capacité de l'algorithme à reconnaitre les mêmes morceaux échantillonnés différemment. Ceci est particulièrement intéressant dans le cas des taux de couverture très faible où la probabilité d'échantillonner les mêmes parties de chaque morceau est faible.

La matrice de confusion peut être déterminée pour un taux de recouvrement et une largeur d'échantillon donnés. La similitude entre 2 morceaux est déterminée par l'écart entre la signature de chaque morceau. La décision de similarité est prise en fonction d'un seuil en-deçà duquel les 2 morceaux sont considérés comme identiques. Le choix de ce seuil est naturellement fondamental, nous évaluons donc son influence. La courbe qui suit représente en pourcentage l'évolution de l'erreur d'association (courbe du haut) et de dissociation (en bas) en fonction de ce seuil. L'erreur d'association survient lorsque l'on considère que 2 morceaux sont identiques alors qu'ils sont différents. L'erreur de dissociation survient lorsque l'on considère que 2 morceaux sont différents alors qu'ils sont identiques. Ces deux visions inverses de la notion d'erreur de reconnaissance sont mesurées en fonction du seuil avec un taux de couverture de 5 % et une durée d'échantillon de 1024.

Dans la figure qui suit On observe clairement que les 2 types d'erreurs évoluent de manière inverse avec l'accroissement du seuil. En effet, il est logique de

constater qu'un seuil plus grand donne plus de chance de ne pas omettre de bons morceaux mais augmente aussi les chances de laisser passer de mauvaises associations. En fonction des souhaits on peut donc minimiser les erreurs d'association en utilisant un seuil minimal ou minimiser les erreurs de dissociation en maximisant le seuil.

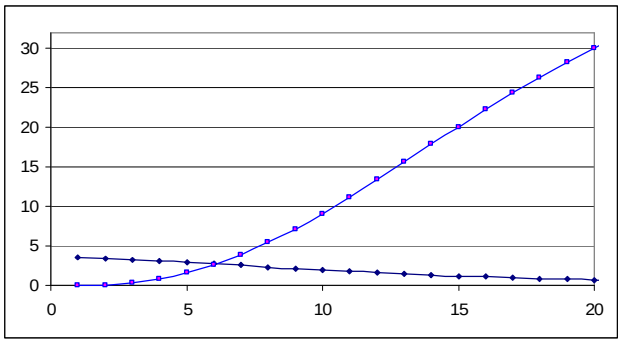

Figure 3 : *Pourcentage d'erreurs d'association (haut) et de dissociation (bas) en fonction du seuil de similarité.*

Un bon compromis semble être obtenu avec un seuil de 6. Pour cette valeur, on obtient une équivalence des 2 types d'erreurs autour de 3 %. Ces résultats sont encourageants car ils mettent en évidence un bon niveau de discrimination compte tenu du temps de calcul.

4.3 Comparaison avec la perception humaine

Pour mieux évaluer la pertinence de notre approche, nous avons soumis nos deux échantillons à un groupe de 10 individus auxquels nous avons demandé s'ils considéraient que les morceaux étaient rythmés ou non. Chaque individu a été interrogé de manière isolée sans contacts avec les autres. Nous avons ensuite calculé la moyenne des 10 avis afin d'obtenir pour chaque morceau une valeur comprise entre 0 et 1. La courbe qui suit exprime la relation entre le rythme perçu par les usagers et la valeur de la signature (normalisée) pour une couverture de 1%. Chaque point sur ce premier graphique représente un des 26 fichiers du premier ensemble.

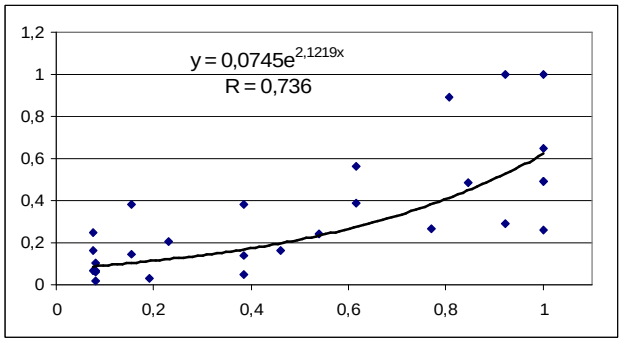

Figure 4-*: Relation entre le rythme perçu par les testeurs (abs) et la valeur de la signature (ord) pour chacun des 26 morceaux (premier ensemble).*

De la même manière le second graphique concerne le second ensemble de fichiers.

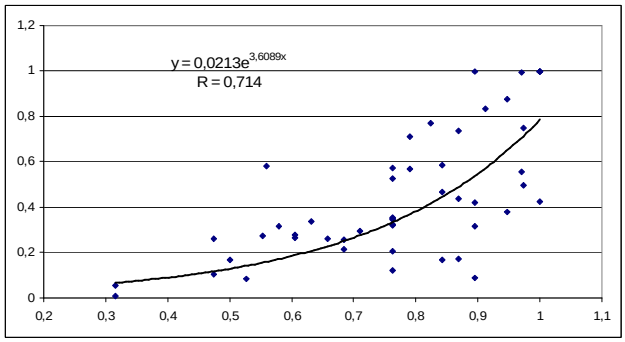

Figure 5-*: Relation entre le rythme perçu par les testeurs (abs) et la valeur de la signature (ord) pour chacun des 50 morceaux (second ensemble).*

Ces courbes permettent de faire plusieurs remarques. Tout d'abord on observe que les 26 échantillons (figure 4) sont répartis de manière assez homogène dans le spectre plus ou moins rythmé. Ceci est confirmé par la valeur moyenne du rythme perçu sur les 26 morceaux se situant à 0,49. On observe aussi que la signature basée sur la moyenne et l'écart-type de la vitesse permet une évaluation raisonnable du rythme avec, dans le cas d'une régression exponentielle, un coefficient de corrélation égal à 0,7. Ceci est confirmé par la seconde courbe avec des niveaux de performances comparables.

Compte tenu de la très faible quantité de signal prélevé, ces coefficients de corrélation doivent être considérés avec précautions. En effet, il est fort possible que 2 signatures de 1 % successives sur le même fichier correspondent en fait à 2 parties du signal complètement différentes. La matrice de confusion montre que malgré les différences, d'une signature sur l'autre, la cohérence est bien présente avec un bon niveau de discrimination. Les résultats de la comparaison avec l'avis des testeurs semblent plus nuancés, même s'ils restent très convenables compte tenu des temps de calcul (taux de couverture 1%). Sur ce point, il faut aussi considérer le caractère aléatoire et ambigu de l'évaluation humaine. Ceci dit, la moyenne des opinions réalisées sur les 10 usagers est de nature à limiter ce facteur.

4.4 Temps de calcul

Il est bien connu que les traitements multimédias sont lourds en temps de calcul. Une de nos motivations en abordant cette étude était d'ailleurs de limiter cette contrainte tout en conservant des performances raisonnables en termes de caractérisation des contenus.

Dans notre cas tous les traitements ont été réalisés sur un PC P4 datant de 2003. A titre d'exemple, le temps unitaire de traitement pour obtenir la moyenne et l'écart type pour

la vitesse et l'accélération avec un taux de couverture de 1% et une taille d'échantillon élémentaire de 2048 octets est de 0,12 secondes (0,08 sec si comme dans nos essais seule la vitesse est nécessaire). Ce temps passe à 36 secondes pour un taux de couverture à 75 % en conservant les autres paramètres identiques.

Sans compter le temps d'évaluation humaine, les mesures de l'influence de tous les paramètres (taux de couverture de 1 à 75 %, taille d'échantillons entre 1024 octets et 131172 octets, calcul des matrices de confusions, etc.) ont nécessité 700 h de traitement (équivalent à un mois de calcul continu). Cette durée importante est une des raisons qui nous ont poussés à limiter le nombre de morceaux de musique distincts évalués dans cette étude.

5 Conclusion

La caractérisation des fichiers musicaux représente un enjeu important dans la mesure où elle permet d'envisager l'indexation et la gestion automatisée et performante des contenus multimédias. Cette automatisation peut être appliquée de plusieurs manières impliquant le document sonore lui-même ou l'usager dans une perspective de modélisation de la perception musicale.

Dans ce contexte, nous avons développé et breveté une technique de caractérisation rapide basée sur la prise en compte de la variation du signal. Nous avons montré qu'un échantillonnage limité de séquences interne était suffisant pour obtenir une performance raisonnable de la caractéristique tout en étant plus de 300 fois plus rapide à calculer qu'un échantillonnage complet. Nous avons abordé la méthodologie de validation suivant deux angles différents : la matrice de confusion et la comparaison avec la perception humaine. Chacune de ces méthodes permet de conclure que la technique offre une représentation cohérente des fichiers sonores.

L'évaluation de notre algorithme en fonction des différentes variables d'influence comme le taux de couverture ou la taille des échantillons internes a nécessité une période de traitement longue. Cette contrainte et la volonté de prendre en compte l'évaluation humaine explique le nombre limité d'échantillons musicaux pris en compte dans cette expérience. Dans les phases ultérieures de nos travaux nous envisageons de valider ces résultats sur la base d'une plus grande quantité de fichiers, mais en limitant l'étendue des variables aux valeurs identifiées comme pertinentes (e.g taux de couverture 1 à 5 %).

Par ailleurs, il nous semble possible d'optimiser la représentativité de la signature en combinant de manière plus pertinente les diverses composantes extraites de notre approche.

Références

[1] Perrot, D., and R. O. Gjerdingen. Scanning the dial: An exploration of factors in identification of musical style. Research notes. Department of Music, Northwestern University, Illinois, USA. 1999

[2] Oppenheim, A. and Schafer, R. Discrete-Time Signal Processing. Prentice Hall. Edgewood Cliffs, NJ. 1989.

[3] George Tzanetakis, George Essl, Perry Cook. Automatic musical genre classification of audio signals. ISMIR, 2001

[4] Cory McKay, Ichiro Fujinaga. Automatic genre classification using large high-level musical. ISMIR, 2004

[5] Hunt, M., Lenning, M., and Mermelstein, P. Experiments in syllabe-based recognition of continuous speech. In Proceedings of International Conference on Acoustics, Speech and Signal Processing, 1996, 880-883

[6] Mark Zadel et Ichiro Fujinaga. Web services for music information retrieval ISMIR 2004

[7] François Pachet, Gert Westermann, Damien Laigre. Musical Data Mining for Electronic Music Distribution ISMIR, 2004

[8] http://www.servicedoc.info/article.php3?id_article=174

[9] http://www.xrml.org

[10] http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm

[11] http://rm.relatable.com/

[11] http://www.clubic.com/actualite-18188-le-marche-de-la-musique-en-ligne-multiplie-par-10-.htmle