Étude des descripteurs acoustiques pour l'alignement temporel audio-sur-partition musicale

Cyril Joder, Slim Essid, Gaël Richard

Institut Télécom - Télécom ParisTech - CNRS/LTCI
37 rue Dareau, 75014 Paris, France
{cyril.joder, slim.essid, gael.richard}@telecom-paristech.fr

Résumé – Dans cet article, nous comparons l'influence des descripteurs acoustiques utilisés dans les systèmes d'alignement temporel musique/partition, pour une tâche où la musique peut être polyphonique avec percussions. Différentes représentations de l'état de l'art sont employées dans un cadre formalisé avec deux stratégies d'alignement. Les résultats montrent que les descripteurs prenant en compte une dimension perceptive (échelle fréquentielle logarithmique) sont les plus pertinents; ils sont notamment plus robustes pour le cas polyphonique. De plus, la stratégie d'alignement à partir d'une resynthèse de la partition obtient des résultats globalement meilleurs que les modèles théoriques, même si sa complexité est supérieure.

Abstract – In this paper we review the acoustic features used for music-to-score alignment and study their influence on the performance in a challenging alignment task, where the audio data can be polyphonic, possibly containing percussion. Different state-of-the-art features are tested in a formalized framework, with two alignment strategies. Results show that the most efficient features are those which take into account a perceptual aspect (logarithmic frequency scale). The latter are found to be more robust, especially in the polyphonic case. Moreover, the alignment strategy which uses a synthesis of the musical score obtains better results than theoretical models, though it is computationnally more expensive.

1 Introduction

L'alignement temporel entre une interprétation musicale et sa partition consiste à associer, à chaque instant de l'exécution, la position correspondante dans la partition (ou réciproquement). Jusqu'à aujourd'hui, cette tâche a surtout été considérée d'un point de vue *temps réel*, pour le suivi d'un musicien en condition de concert [1, 2].

Or, un alignement temporel permet de nombreuses applications qui ne sont pas soumises aux contraintes du temps réel. Parmi celles-ci, on peut compter le codage très bas débit, l'aide à la séparation de sources, la correction/amélioration de la partition ou l'annotation automatique de bases de données. Pour ces raisons, nous nous intéressons au problème où l'interprétation est un enregistrement audio, dans le cas *hors ligne* (où tout l'enregistrement est connu).

Les systèmes d'alignement peuvent être décomposés en deux niveaux : le bas niveau, qui correspond à l'analyse « instantanée » de l'audio, et le haut niveau qui modélise la dimension temporelle de la partition. La plupart des travaux traitant du suivi de partition mettent l'accent sur le modèle temporel utilisé [3, 4, 1], et le choix des descripteurs de bas niveau n'est pas toujours clairement motivé. Il n'existe pas à notre connaissance d'étude comparant l'efficacité des descripteurs utilisés pour cette tâche. De plus, l'évaluation des systèmes d'alignement a été jusqu'à présent limitée presque exclusivement à de la musique classique monophonique ou faiblement polyphonique (voir la campagne d'évaluation MIREX 2006 [5], qui est la dernière en date sur cette tâche).

L'objectif de cet article est d'étudier spécifiquement l'influence de la paramétrisation de bas niveau pour deux stratégies d'alignement. Cette étude s'étend à de la musique populaire polyphonique multi-instrumentale avec présence éventuelle de percussions, ce qui représente le cas le plus général. Dans la partie 2 est formalisé le problème d'alignement temporel, puis les descripteurs utilisés dans notre étude sont exposés dans la partie suivante. Nous présentons les résultats des expériences dans la section 4 avant de conclure.

2 Formalisation du problème d'alignement temporel

La plupart des systèmes permettant de réaliser cette tâche peuvent être formalisés grâce à des modèles à états cachés. En effet, une hypothèse toujours utilisée est que le son perçu à un instant donné dépend uniquement de l'accord joué à cet instant. Cette hypothèse n'est en général pas rigoureusement vérifiée (par exemple dans des conditions d'enregistrement réverbérantes), mais elle permet de modéliser la partition comme une séquence d'états correspondant aux accords.

À chaque état A de la partition peut être associé la production d'une observation S (par exemple un spectre

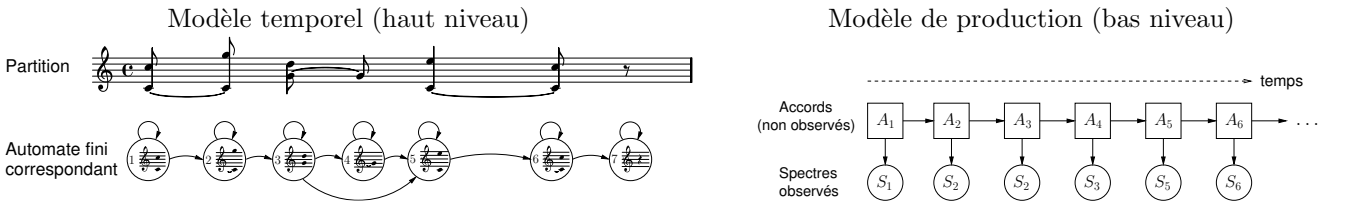

FIG. 1 – Représentation du modèle à états cachés utilisé pour l'alignement.

à court terme) selon une loi de probabilité supposée. Cette loi $p(S|A)$ est définie par le modèle *bas niveau*, qui permet ainsi d'estimer la vraisemblance d'une trame de son en connaissant la note jouée. La couche *haut niveau* correspond au modèle de transitions entre les états de l'automate ainsi formé. La figure 1 représente ces deux couches du modèle statistique. Il est possible d'utiliser plusieurs états par accord, pour modéliser une évolution temporelle plus fine.

Une interprétation de la partition est donc modélisée comme un *chemin* (c'est-à-dire une séquence d'états) dans l'automate. Connaissant la suite $S_1, \ldots, S_N$ des observations extraites de l'audio, la tâche d'alignement revient alors à trouver le chemin $A_1, \ldots, A_N$ optimal, selon un certain critère.

Le critère utilisé ici est celui du *Maximum de Vraisemblance*. Le chemin optimal $\hat{\mathbf{A}} = (A_1, \ldots, A_N)$, est défini comme

$$\hat{\mathbf{A}} = \underset{\mathbf{A} \in \mathcal{A}}{\operatorname{argmax}} \prod_{i=1}^{n} p(S_i|A_i),$$

où $\mathcal{A}$ est l'ensemble des chemins possibles. Ce critère est choisi pour donner le moins d'importance possible au modèle temporel, afin de mettre en évidence les différences causées par la couche bas niveau. Un algorithme de programmation dynamique permet de calculer simplement cet optimum.

3 Expériences

Les tests effectués ont pour but d'évaluer les performances relatives de différentes paramétrisations de bas niveau utilisées dans l'état de l'art et de les confronter à d'autres descripteurs qui n'ont pas été utilisées dans des systèmes d'alignement. Les descripteurs testés sont choisis pour capturer l'information de « hauteur » des notes, en vue de la comparer avec les indications de la partition.

3.1 Paramétrisations considérées

Transformée de Fourier Dans la littérature, les modélisations utilisées font presque toujours appel à la transformée de Fourier à court terme (TFCT) du signal sonore [2, 4, 1]. Cela s'explique notamment par la rapidité de calcul de cette transformée, qui est d'un grand intérêt dans le cas de contraintes temps-réel. La paramétrisation correspond donc au spectre de puissance à court terme.

Trois principales méthodes sont utilisées pour calculer la vraisemblance des accords à partir de cette représentation : le Peak Spectral Match (PSM) [2] qui considère la proportion de l'énergie du signal dans les bandes de fréquences correspondant aux notes attendues, ainsi que deux modèles explicites de puissance spectrale, utilisés par Raphael [4] (que nous désignons par MSR pour Modèle de Spectre de Raphael) et Cont [1] (MSC).

Pour calculer la vraisemblance d'un accord, ces deux méthodes font appel à un modèle g du spectre de puissance, correspondant à l'accord A. Ce modèle est en fait un mélange de Gaussiennes, dont chacune est centrée sur la fréquence d'un partiel d'une note de l'accord. Le MSR estime la vraisemblance par la formule :

$$p_{MSC}(S|A) = \prod_{\omega} g(\omega)^{S(\omega)}, \qquad (1)$$

où ω parcourt les fréquences. Comme l'indique Raphael, cette formule (à un facteur multiplicatif près) correspond au calcul de la vraisemblance si S est vue comme l'histogramme d'un tirage aléatoire d'après la distribution de probabilité g. Nous appelons cette méthode de calcul de la vraisemblance *modèle d'histogramme*.
Le MSC fait appel à une autre formule, utilisant une version normalisée $\bar{S}$ du spectre de puissance, telle que la somme de ses valeurs soit unitaire. Ce spectre peut alors être considéré comme une distribution de probabilité et la vraisemblance est calculée grâce à une mesure probabiliste : $p_{MSC}(S|A) = \exp\big(-D(\bar{S}||g)\big)$, où $D(\cdot||\cdot)$ est la divergence de Kullback-Leibler. L'exponentielle est utilisée pour obtenir une estimation de probabilité (dans l'intervalle $[0, 1]$) à partir d'une distance (dans l'intervalle $[0, \infty[$).

Énergie par bandes logarithmiques Dans le cas hors ligne, on peut faire appel à d'autres représentations que la TFCT. Müller *et al.* [3] considèrent l'énergie à la sortie d'un banc de filtres espacés logarithmiquement, correspondant à l'échelle des demi-tons de la gamme musicale. Nous calculons aussi une représentation similaire grâce à une transformée à Q constant (CQT).

La vraisemblance d'un accord est alors estimée par la proportion de l'énergie du signal dans les bandes correspondant aux notes de l'accord. Ces descripteurs sont nommés respectivement EBF (Énergie par Banc de Filtres) et ECQT (Énergie par CQT).

TAB. 1 – Récapitulatifs des modèles de descripteurs testés.

Acron.	Signification	Acron.	Signification
MSR	Modèle de Spectre de Raphael	HPCP	*Harmonic Pitch Class Profile*
MSC	Modèle de Spectre de Cont	CZ	Chroma de Zhu
PSM	*Peak Spectral Match*	CP	Chroma de Peeters
EBF	Énergie par Banc de Filtres	CBF	Chroma par Banc de Filtres
ECQT	Énergie par CQT		

Vecteurs de *Chroma* Afin d'être robuste à d'éventuelles erreurs d'octave dans la partition, nous nous intéressons à des représentations en *vecteurs de chroma*, qui intègrent l'énergie dans toutes les bandes correspondant à chacune des douze classes chromatiques de la gamme musicale (de *do* à *si*). Quatre manières de calculer ces descripteurs sont utilisées. La première (appelée CBF pour Chroma par Banc de Filtres) exploite le banc de filtres précédent. La suivante, d'après Peeters [6] (CP pour Chroma de Peeters) utilise une TFCT. Une autre de ces représentations, d'après Zhu[7] (CZ pour Chroma de Zhu) est calculée à partir d'une CQT. La dernière est la représentation HPCP (*Harmonic Pitch Class Profile*) de Gómez [8], qui utilise une interpolation du spectre de puissance.

La vraisemblance d'un accord est alors calculée en comparant le vecteur de chroma observé s avec un modèle théorique construit à partir des notes de cet accord. Chaque composante de ce modèle a pour valeur le nombre de notes correspondantes que comporte l'accord. Ainsi, à l'accord composé des notes do_3, mi_3, sol_3 et do_4 sera associé le modèle $g = (2, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0)$. Une composante constante est ajoutée à ces modèles pour modéliser un bruit. La mesure utilisée pour estimer la similarité de l'observation et du modèle est alors le *modèle d'histogramme* de l'équation (1) : $p_{\text{hist}}(s|g) = \prod_{i=1}^{12} g(i)^{s(i)}$.

Le tableau 1 récapitule les différents modèles utilisés dans cette étude.

3.2 Stratégies d'alignement

Deux types de systèmes d'alignement sont considérés. Dans le premier, la vraisemblance d'un accord est évaluée en comparant les valeurs des descripteurs à des modèles théoriques idéaux, comme présenté dans la section 3.1. Le chemin d'alignement est alors calculé comme le chemin de maximum de vraisemblance parmi ceux qui commencent avec le premier accord de la partition et finissent avec le dernier, sans « saut » d'accord.

Dans le second, les valeurs des descripteurs sont comparées à celles extraites d'un son synthétisé à partir de la partition. Cette deuxième stratégie est en fait l'alignement des deux sons par programmation dynamique (algorithme DTW pour Dynamic Time Warping). Cela correspond à une différente structure de modèle à états cachés, où le nombre d'états par accord est proportionnel à la durée théorique de l'accord.

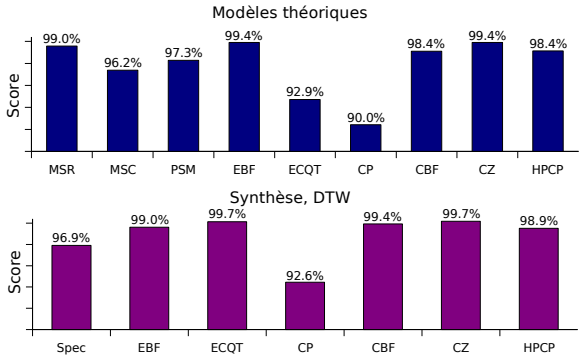

FIG. 2 – Résultats des différents systèmes testés sur la base de musique monophonique.

3.3 Bases de données

Deux ensembles de fichiers musicaux avec partitions au format MIDI sont utilisés. Le premier est la base de données constituée pour la campagne d'évaluation MIREX'06 [5], qui comprend 46 fichiers de musique classique monophonique, pour environ 45 min de son et 9000 évènements (attaques de notes). Le second est formé d'extraits de 19 chansons avec ou sans percussions de la base RWC-pop [9]. Cela représente environ 70 min et 16000 évènements. Pour tous ces morceaux, nous avons obtenu un alignement « parfait », afin d'évaluer les systèmes.

4 Résultats obtenus

Le protocole d'évaluation donne la proportion d'évènements bien détectés. Un évènement est considéré bien détecté s'il est détecté dans un intervalle de tolérance temporelle autour de la valeur de référence, intervalle que nous faisons varier dans notre étude. Les scores sont donnés ici pour un seuil fixé à 2s, égal à celui choisi pour MIREX'06.

4.1 Musique monophonique

La figure 2 compile les résultats obtenus par chaque représentation, pour leurs valeurs optimales de leurs paramètres. Notons que le score du meilleur système de MIREX'06 [5] est de 90,1%. Tous les systèmes testés (excepté le modèle théorique CP) obtiennent de meilleurs résultats. Cela est cohérent puisque nos systèmes ne sont pas soumis aux contraintes du temps-réel.

On remarque que pour une même représentation, les résultats de la deuxième stratégie d'alignement sont globalement meilleurs que ceux obtenus grâce à des modèles théoriques. Le score passe par exemple de 98,4% à 99,4% pour le descripteur CBF. Cela s'explique par le fait que la synthèse (prenant en compte les instruments) donne des modèles plus réalistes.

Les trois types de paramétrisations (spectre de puissance, énergie par bandes de fréquence et vecteurs de chro-

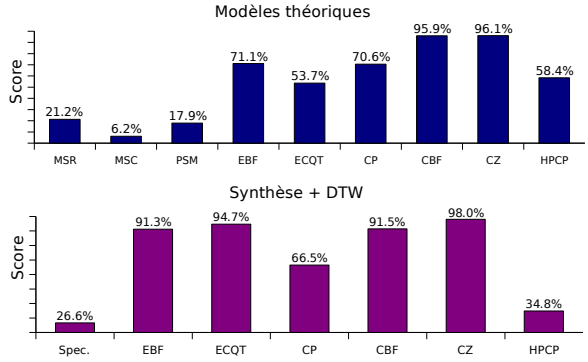

FIG. 3 – Résultats des différents systèmes testés sur la base de musique monophonique.

ma) permettent d'obtenir de bons résultats, supérieurs à 99%. On ne peut donc pas vraiment trancher en faveur d'un type de représentation particulier.

Le descripteur CP obtient des résultats significativement inférieurs aux autres (90,0% et 92,6% avec les deux méthodes). Cela s'explique par des problèmes de résolution fréquentielle de la TFCT en basse fréquence. Ces problèmes de résolutions sont moins sensibles avec les autres descripteurs utilisant cette même transformée. Les descripteurs spectraux (MSR, MSC et PSM) exploitent des informations venant d'harmoniques supérieures, et le descripteur HPCP utilise une interpolation du spectre de puissance pour localiser plus précisément les pics.

4.2 Passage au polyphonique

Les scores obtenus sur la base de musique pop sont représentés sur la figure 3. Toutes les méthodes obtiennent de moins bons résultats que dans le cas monophonique, et les écarts de scores sont beaucoup plus importants. Certains résultats très faibles peuvent être expliqués par le fait que les systèmes peuvent « se perdre » dans un morceau, et une grande partie de ce morceau est alors mal alignée.

Néanmoins, des tendances similaires aux expériences précédentes sont observées. La représentation la plus efficace reste CZ (98,0% avec synthèse et 96,1% avec modèle théorique). Dans chacun des trois types de descripteurs, les meilleurs systèmes sont les mêmes que dans le cas monophonique.

Par contre, les résultats relatifs de ces trois classes sont modifiés. Les représentations utilisant le spectre de puissance voient leurs résultats s'effondrer, jusqu'à 6,2% pour le modèle MSC. Ceci s'explique par les problèmes de résolution fréquentielle, qui ne peuvent plus être réglés en considérant les hautes fréquences du fait de la présence de nombreux partiels dans le cas polyphonique.

Les représentations d'énergie en sous-bandes ne permettent pas d'obtenir de meilleurs résultats que les représentations par vecteur de chroma. Cela indique que l'information d'octave n'est pas indispensable pour effectuer un alignement. Dans le cas des modèles théoriques, l'énergie en sous-bandes est même significativement moins efficace que les vecteurs de chroma. Ceci peut s'expliquer par la méthode de calcul de la vraisemblance, qui a tendance à favoriser les accords comportant un grand nombre de notes.

5 Conclusion

Nous avons mis en évidence l'importance de la paramétrisation de bas niveau utilisée dans un système d'alignement. Les descripteurs représentant le spectre sur une échelle logarithmique obtiennent des résultats supérieurs à ceux de l'état de l'art et sont plus robustes au passage à la musique polyphonique que ceux fondés sur une TFCT. L'information d'octave n'apparait pas indispensable à un bon alignement et les descripteurs de chroma obtiennent les meilleurs résultats.

Les perspectives ouvertes par ce travail comprennent la prise en compte d'erreurs dans la partition et l'ajout d'autres descripteurs, notamment représentant l'« impulsivité » pour localiser plus précisément les attaques de notes.

Références

[1] A. Cont, *Modeling musical anticipation : from the time of music to the music of time*, thèse de doctorat, Université Paris 6, 2008

[2] F. Soulez X. Rodet, D. Schwarz, *Improving Polyphonic and Poly-Instrumental Music to Score Alignment*, Proceedings of ISMIR, 2003

[3] M. Müller, F. Kurth, T. Röder, *Towards an Efficient Algorithm for Automatic Score-to-Audio Synchronization*, Proceedings of ISMIR, 2004

[4] C. Raphael, *Aligning music audio with symbolic scores using a hybrid graphical model*, Machine Learning Journal, 2006

[5] Music Information Retrieval Evaluation eXchange 2006, tâche de suivi de partition : `http://www.music-ir.org/mirex/2006/index.php/Score_Following_Proposal`

[6] G. Peeters, *Musical Key Estimation of Audio Signal Based on Hidden Markov Modeling of Chroma Vectors*, Proc. of DAFx, 2006

[7] Y. Zhu, M. Kankanhalli, *Precise pitch profile feature extraction from musical audio for key detection*, IEEE Transactions on Multimedia, 2006

[8] E. Gómez, *Tonal Description of Music Audio Signals*, thèse de doctorat, Université Pompeu Fabra, 2006

[9] M. Goto, H. Hashiguchi, T. Nishimura, R. Oka, *RWC Music Database : Popular, Classical, and Jazz Music Databases*, Proc. of ISMIR, 2002