Étude d'algorithmes pour la détection de la tonalité de fichiers musicaux et implémentation en Clojure

Antoine Passemiers

Université Libre de Bruxelles
apassemi@ulb.ac.be

Résumé

Le projet consiste en la discussion de différents algorithmes relatifs à la détection automatisée de tonalité de fichiers musicaux, le prototypage de ceux-ci en Python, ainsi qu'une réflexion sur l'adaptabilité de ces derniers avec le paradigme fonctionnel. Le choix de l'algorithme a concevoir selon l'approche fonctionnelle sera basé sur des critères de rapidité d'exécution et de précision de la détection. L'algorithme final sera alors implémenté en Clojure.

1. Introduction

La tonalité d'une oeuvre musicale se caractérise par l'ensemble des sons formant une même gamme diatonique. A la différence de la gamme, où les sons se succèdent de façon contigüe, la tonalité (ou ton) regroupe des sons qui peuvent être disjoints et/ou superposés (Danhauser, 1929). En conséquence, nous nous intéressons à l'analyse de mélodies polyphoniques, où plusieurs notes peuvent être jouées en même temps.

En particulier, nous allons nous pencher sur deux catégories d'algorithmes : ceux basés sur des modèles cognitifs, et ceux incluant des notions d'apprentissage automatique. Les premiers tentent d'intégrer au mieux les connaissances de la théorie musicale et reposent sur la façon dont les personnes reconnaissent les différentes tonalités, alors que les seconds utilisent l'inférence statistique pour déterminer celles-ci.

L'approche cognitive utilisée pour la détection de la tonalité repose en partie sur la solution proposée par Ibrahim Sha'ath lors de la conception du logiciel KeyFinder (Sha' ath, 2011). La précision de la détection est évaluée à l'aide d'une base de données, constituée de 250 fichiers musicaux au format wav, dont les tonalités sont connues et inscrites dans un fichier csv. Ces fichiers font partie de ceux utilisés par Ibrahim Sha'ath dans le cadre de sa recherche.

Pour ce qui est de la partie apprentissage automatique, les méthodes présentées seront principalement en lien avec les modèles de Markov cachés. Leur évaluation se fera en divisant la base de données en un jeu de données d'apprentissage et un jeu de données de validation (respectivement 60 % et 40 % du jeu de données d'origine). Ce mini-mémoire se voulant concis et centré sur les objectifs décrits, le lecteur est supposé déjà disposer de connaissances suffisantes en apprentissage automatique, en théorie musicale et en traitement logiciel de signaux.

TODO : (Pauws, 2004) (Takeuchi, 1994) -> Partie 2.1.1.

2. Considérations théoriques

2.1. Pré-traitement

Le signal audio est premièrement extrait du fichier wav, puis la moyenne entre les deux canaux est effectuée si le fichier a été enregistré en stéréo. En effet il n'est pas nécessaire de prendre en compte le panoramique puisque celui-ci n'a que peu d'influence sur la mélodie dans le domaine spectral. Étant donné que les notes jouées sont uniquement caractérisées par leur fréquence fondamentale, il n'est pas nécessaire de considérer l'entièreté du spectre du fichier musical. De fait, la fréquence d'échantillonnage est abaissée à un dixième de la fréquence standard (4410 Hz), mais ce sous-échantillonnage est susceptible de provoquer des phénomènes d'aliasing. La solution de Sh'ath implique de gérer les problèmes d'aliasing sonore par l'application d'un filtre passe-bas. La taille de la fenêtre temporelle est un hyper-paramètre fixé durant l'évaluation de l'algorithme. (Sha' ath, 2011)

2.2. Estimation spectrale

Il existe deux catégories de techniques d'estimation de densité spectrale : les méthodes paramétriques et les méthodes non-paramétriques.

2.2.1. Constant-Q Transform (CQT) Avant de procéder à l'estimation de la CQT, une fenêtre de Blackman est appliquée sur les données observées afin

d'éviter les distortions spectrales dues à l'étroitesse de la fenêtre (et au principe d'incertitude d'Heisenberg). Le spectre est alors approximé à l'aide de la transformée de Fourier rapide (FFT). L'intuition derrière la Constant-Q Transform (CQT) est de penser que les coefficients de la FFT dont les fréquences ne correspondent pas à des notes de musique prennent plus de poids que les autres coefficients. Ceci doit être réajusté en appliquant des fenêtres spectrales centrées sur les notes de musiques. Pour chaque fenêtre, les coefficients résultants de cette opération sont alors sommés pour ne former qu'un seul coefficient spectral. L'ensemble des coefficients globaux constitue alors la CQT.

2.2.2. Décomposition harmonique de Pisarenko (PHD) (Mujahid F. Al-Azzo, 2014)

2.2.3. Estimation spectrale par moindres carrés Il est possible d'estimer le spectre par régression, au travers d'algorithmes n'exigeant pas de grandes quantités de calcul. Parmi les algorithmes les plus importants résident celui de Vaníek et celui de Lomb-Scargle. Leur particularité est de pouvoir traiter des données qui ne sont pas reçues à intervalles réguliers (la majorité des observations se font la nuit). En outre, contrairement à la transformée de Fourier rapide qui possède une résolution aussi précise que la fenêtre d'observation est grande, ces algorithmes calculent un périodogramme dont la taille est fixée par le nombre de fréquences qui nous intéressent. Au plus le nombre de fréquences analysées est élevé, au plus le temps d'exécution sera conséquent.

$$\hat{\theta}_k = (A_k^T A_k)^{-1} A_k^T y \tag{1}$$

Selon la méthode de Vanícek, le signal est supposé être centré sur zéro et représenter une combinaison linéaire de sinusoïdes et d'un bruit blanc. Les coefficients de cette combinaison linéaire sont réunis dans un vecteur θ_k, de telle manière que le signal équivaut à $y \approx A\theta_k$. La matrice A est telle que chaque ligne de celle-ci constitue une sinusoïde de fréquence d'intérêt. Les coefficients sont alors finalement donnés par l'équation 1. (Petr Stoica, 2009)

Le défaut de la méthode est de ne pas considérer les phases des fréquences concernées. En effet chaque sinusoïde contenue dans la matrice A est supposée posséder une phase nulle. La méthode de Lomb-Scargle permet de résoudre ce problème en pré-calculant les phases des sinusoïdes utilisées et en tenant compte au mieux de celles-ci lors du calcul du spectre. Le déphasage τ correspondant à la fréquence f est donné par l'équation 2. (Lomb, 1975)

$$\tan 2\pi f\tau = \frac{\sum_{j=1} \sin 2\pi f t_j}{\sum_{j=1} \cos 2\pi f t_j} \tag{2}$$

Tout comme dans la méthode de Vanícek, la régression consiste en la recherche de coefficients qui expliquent au mieux le signal sous la forme d'une pondération de sinusoïdes. Le coefficient associé à la fréquence f est donné par l'équation 3 :

$$\Delta R(f) = \frac{(YC)^2}{CC} + \frac{(YS)^2}{SS} \tag{3}$$

Les valeurs CC et SS peuvent être pré-calculées, car les fréquences d'intérêt ne changent pas d'un fichier à l'autre.

$$CC = \sum_{j=1} \cos^2 2\pi f(t_j - \tau) \tag{4}$$

$$SS = \sum_{j=1} \sin^2 2\pi f(t_j - \tau) \tag{5}$$

Les valeurs YC et YS cherchent à mesurer respectivement les niveau d'orthogonalité du signal y avec une sinusoïde déphasée de $\pi/2$ et avec une sinusoïde de phase nulle. Ces sinusoïdes peuvent également être pré-calculées. Les formules permettant de calculer YC et YS sont :

$$YC = \sum_{j=1} y_j \cos 2\pi f(t_j - \tau) \tag{6}$$

$$YS = \sum_{j=1} y_j \sin 2\pi f(t_j - \tau) \tag{7}$$

Ainsi, les seules opérations restantes lors de l'exécution de l'algorithme sont les produits scalaires entre les sinusoïdes et le signal délimité par la fenêtre glissante, ce qui réduit drastiquement la quantité de calculs requis pour l'estimation du spectre.

2.2.4. Autres méthodes TODO : Cepstre, spectre de puissance, ...

2.3. Prédiction de la tonalité

Une fois le spectre estimé, celui-ci est compacté dans un vecteur chromatique composé de douze coefficients. Dans le cadre de l'évaluation de l'algorithme, il a été trouvé que 6 octaves suffisent à approximer le spectre :

l'ajout d'une septième octave n'améliore pas la précision des prédictions. De fait, le spectre dont la résolution est de 72 coefficients (12 x 6 octaves) est redimensionné en une matrice $M_{i,j}$ de dimensions (6 x 12). Enfin, le vecteur chromatique C_i est obtenu en formalisant la méthode élaborée par Sha'ath :

$$C_i = (1-p) * \max_j M_{i,j} + p * \sum_j M_{i,j} \qquad (8)$$

où p est un hyper-paramètre déterminé durant l'évaluation/validation de l'algorithme. La dernière étape consiste alors à identifier localement la tonalité sur base unique de ce vecteur chromatique. Différentes méthodes ont été discutées, la première se base sur la méthode de Sh'ath alors que les suivantes reposent sur des algorithmes d'apprentissage automatique.

2.3.1. Modèle cognitif Cette solution repose sur les expériences menées par Carol L. Krumhansl et Lola L. Cuddy sur la perception des tonalités chez l'Homme. Des gammes incomplètes ont été jouées, suivies de notes additionnelles que les sujets devaient évaluer sur leur capacité à compéter ces gammes. Les scores ont été moyennés pour donner les deux graphiques suivants :

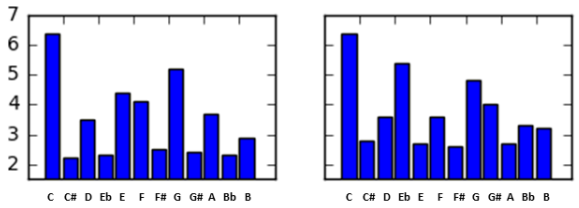

Figure 1: Profils de tonalités dérivés de l'expérience de Krumhansl, et représentant respectivement la gamme do majeur et la gamme do mineur.

Bien que ces deux profils de tonalités fournissent des résultats satisfaisants lorsque le spectre est estimé à l'aide de la CQT, il ne sont pas adaptés lorsque la méthode de Lomb-Scargle est utilisée, car il a été observé que les coefficients renvoyés par cette dernières sont globalement plus equidistribués. Cette distribution doit se refléter dans les profils de tonalité, c'est pourquoi l'algorithme a été évalué plusieurs fois jusqu'à obtenir les profils fournissant les meilleurs résultats, et donnés ci-dessous :

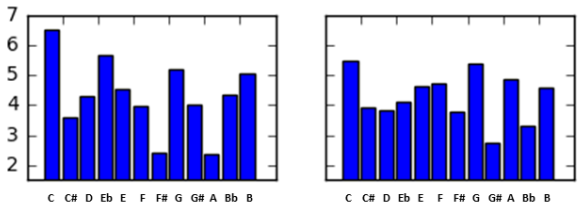

Figure 2: Meilleurs profils de tonalités, déterminés empiriquement

2.3.2. Modèles de Markov Cachés (HMM) Les modèles de Markov cachés sont des machines à états discrets cherchant à représenter des séries multivariées par leurs distributions, ainsi que par les probabilités de transition entre les états cachés de la machine. De plus, chaque état caché de cette dernière possède ses propres probabilités d'émission. Contrairement à des modèles d'apprentissage automatique plus populaires tels que les réseaux de neurones ou les machines à vecteurs de support, les HMM sont capables de traiter des séquences de longueur non fixée. Cette caractéristique est appréciable dans le cadre de l'analyse de morceaux de musique, qui ont des durées de nature très variables.

TODO : (Peeters, 2006) (Ramage, 2007)

2.2.3. Modèles de Markov Cachés de type entrée-sortie (IO-HMM) TODO : (Bengio and Frasconi, 1996)

2.4. Évaluation

Dans le cadre de ce travail, seules les douze gammes majeures et leurs douze gammes mineures relatives correspondantes sont considérées. Deux mesures différentes sont utilisées afin d'évaluer la précision de l'algorithme : la précision (le ratio du nombre de prédictions correctes sur le nombre de fichiers analysés), ainsi que l'indice du MIREX. Ce dernier est égal à une combinaison du nombre de bonnes prédictions (avec une pondération de 1,0), du nombre de prédictions décalées de 5 demi-tons (pondération de 0,5), du nombre de prédictions décalées de 4 demi-tons (pondération de 0,5), du nombre de prédictions de gamme relative (pondération de 0,3), et du nombre de prédictions de gamme parallèle (pondération de 0,2).

2.5. Résultats

Méthode	Précision	MIREX
CQT + profiles	30,7%	-
Lomb-Scargle + profiles	-	-
CQT + HMM	–	–
FFT + IO-HMM	–	–
CQT + IO-HMM	–	–
Lomb-Scargle + IO-HMM	–	–

Table 1: Évaluation des méthodes présentées selon la précision et l'indice du MIREX

3. Implémentation en Clojure

TODO : (Kumar, 2015)

References

Bengio, Y. and Frasconi, P. (1996). Input-output hmm's for sequence processing. *IEEE transactions on neural networks, vol. 7, no. 5.*

Danhauser, A. (1929). *Théorie de la Musique.*

Kumar, S. (2015). *Clojure High Performance Programming.*

Lomb, N. R. (1975). Least-squares frequency analysis of unequally spaced data.

Mujahid F. Al-Azzo, K. I. A.-S. (2014). High resolution techniques for direction of arrival estimation of ultrasonic waves. *American Journal of Signal Processing*, pages 49–59.

Pauws, S. (2004). Musical key extraction from audio. *Proceedings of the 9th International Conference on Digital Audio Effects, DAFx-06, Montreal, Canada*, pages 96–99.

Peeters, J. (2006). Musical key estimation of audio signal based on hidden markov modeling of chroma vectors. *Proceedings of the 9th International Conference on Digital Audio Effects, DAFx-06, Montreal, Canada.*

Petr Stoica, Jian Li, H. H. (2009). Spectral analysis of nonuniformly sampled data : A new approach versus the periodogram. *IEEE Transactions on signal processing, vol 57, no. 3*, pages 843–857.

Ramage, D. (2007). Hidden markov models fundamentals. *CS229 Section Notes.*

Sha' ath, I. (2011). Estimation of key in digital music recordings. pages 1–64.

Takeuchi, A. H. (1994). Maximum key-profile correlation (mkc) as a measure of tonal structure in music. *Perception & Psychophysics*, pages 335–346.