Compression et Représentation de Signaux Audiovisuels (CORESA'03), 16-17 janvier 2003, Lyon, France, pp. 45-48.

Dimensions perceptives pour un espace musical

Y. Fernandez[1] M. Crucianu[1]

[1] LI (Laboratoire d'Informatique)

EA 2101, Université de Tours
64, avenue Jean Portalis, 37200 Tours – France

yann.fernandez@caramail.com, michel.crucianu@inria.fr

Résumé

Notre objectif est de mettre au point de nouveaux outils, plus expressifs, pour l'exploration de collections de pièces de musique. Dans ce but nous proposons d'utiliser des dimensions perceptives interprétables, obtenues à partir du signal sonore et représentées visuellement. Le positionnement d'une nouvelle pièce dans l'espace musical est réalisé par l'extraction de caractéristiques du signal, suivie par l'application d'une transformation apprise à partir de réponses fournies par des utilisateurs.

Mots clefs

Analyse du signal musical, perception musicale, similarité perceptive, dimensions de similarité.

1 Introduction

Les méthodes d'organisation et d'exploration des collections de pièces de musique n'ont pas évolué au même rythme que les catalogues des distributeurs ou les moyens d'accès en ligne aux contenus musicaux. Le genre musical reste encore largement exploité pour regrouper des pièces de musique. La mise au point de méthodes automatiques d'étiquetage par genre [1] permet de pallier le caractère relativement arbitraire de cette notion. Néanmoins, le genre musical reste, au mieux, une mesure de similarité *globale* entre pièces de musique.

Les moyens d'exploration disponibles à ce jour sont essentiellement fondés sur l'interrogation, à travers des mots-clefs, de bases de données textuelles concernant les pièces de musique (genre, titre, interprète, etc.). Lors de l'exploration de la collection l'utilisateur n'a pas accès à des informations concernant le contenu musical réel d'une pièce ou la similarité musicale entre des pièces.

Les corrélations constatées entre les actes d'achat des clients ont permis à certains distributeurs de proposer une autre mesure globale de similarité, qui dépend dans une large mesure d'éléments étrangers au contenu musical (publicité, effets de mode, proximité chronologique des publications, etc.).

Les utilisateurs potentiels se retrouvent ainsi devant deux difficultés majeures : la pauvreté des critères d'exploration et le caractère rudimentaire de l'interface. Nous remarquerons que la relation est forte entre les critères retenus et l'interface employée.

2 Définition d'un espace musical

Notre objectif est d'arriver à un meilleur équilibre entre le caractère synthétique (permettant d'éviter le recours systématique à l'écoute) et l'expressivité des représentations des pièces de musique. La représentation d'une pièce doit à la fois rester synthétique par rapport à un extrait sonore et traduire sous une forme directement perceptible les caractéristiques sonores de cette pièce. Ensemble, ces deux qualités assurent la lisibilité de l'organisation de la collection et l'évaluation fiable de chaque pièce.

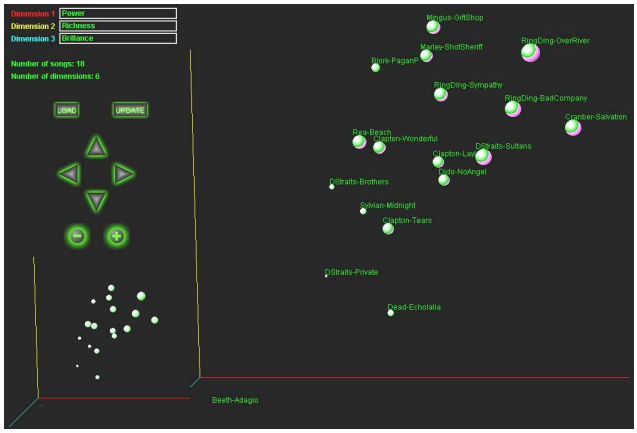

Figure 1 - *Interface de navigation dans l'espace musical*

L'impression que laisse l'écoute d'une pièce dépend de phénomènes perceptifs complexes et subjectifs [2]. Pour arriver à une représentation à la fois synthétique et expressive, il est donc naturel de chercher à caractériser la pièce suivant des *dimensions perceptives interprétables* (richesse, vigueur, caractère mélodieux, etc.). Ces dimensions permettent d'organiser la collection de pièces de musique, qui sera représentée par un ensemble de points dans un espace multidimensionnel (figure 1). Au lieu d'être exprimée par une valeur globale, la similarité entre deux pièces de musique se sépare en composantes

correspondant aux différentes dimensions perceptives. L'utilisateur peut ainsi avoir un aperçu de ce qui distingue deux pièces de musique sans avoir à les écouter, effectuer une recherche en se déplaçant suivant une ou plusieurs dimensions interprétables, etc.

Plusieurs méthodes sont envisageables pour positionner des pièces musicales dans un espace défini par de telles dimensions perceptives. Une première méthode, directe, consiste à recueillir et à faire la synthèse des avis d'utilisateurs volontaires ; cette méthode présente un « temps de réponse » assez long et peut seulement être appliquée à un nombre relativement restreint de pièces qui suscitent un intérêt suffisant de la part du public.

Afin de pouvoir traiter tout enregistrement musical disponible, de façon automatique, il est nécessaire d'introduire une représentation intermédiaire, obtenue par l'analyse du signal sonore. La méthode que nous avons retenue, indirecte mais automatique, comporte une première étape d'extraction d'indices à partir du signal musical, suivie par une étape qui associe à ces indices des valeurs pour les dimensions perceptives (figure 2). Cette méthode indirecte exploite aussi – dans sa deuxième étape – des avis d'utilisateurs volontaires, mais uniquement sur un sous-ensemble restreint de la collection. En tirant profit des capacités de généralisation de techniques d'apprentissage automatique, la relation entre indices extraits du signal et dimensions perceptives est étendue au-delà de cet ensemble.

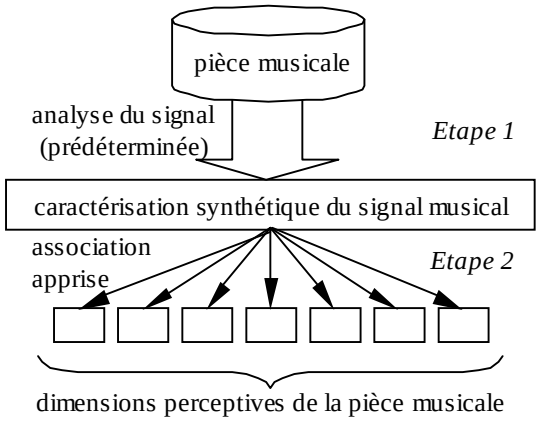

Figure 2 - *De la musique aux dimensions perceptives*

3 Extraction de caractéristiques

La première étape du traitement d'une pièce de musique est l'analyse du signal. Contrairement à d'autres travaux qui cherchent à décrire de façon très fine des sons de courte durée, souvent pour fournir une aide à la composition, nous nous intéressons à la caractérisation d'une pièce de musique entière à partir du signal sonore. Les paramètres extraits doivent permettre de caractériser le signal musical dans sa globalité, de façon aussi riche et fidèle que possible.

Nous procédons à une analyse de l'évolution dans le temps de l'amplitude et du spectre de fréquences pour aboutir à un ensemble de séries chronologiques décrivant la pièce de musique. Les séries extraites du signal concernent :

- L'amplitude et l'enveloppe temporelle du signal. Par amplitude nous entendons la racine carrée de la moyenne des carrés des valeurs du signal dans une fenêtre de courte durée. L'enveloppe temporelle est obtenue par l'application d'un filtre passe-bas sur la série des amplitudes.
- Les temps d'attaque, de maintien et de descente, mesurés à partir de l'enveloppe temporelle. Nous retenons uniquement les maxima significatifs, pour lesquels la différence entre le maximum et le minimum qui le précède est supérieure à 0,9 de l'amplitude moyenne du signal.
- La hauteur du son (*pitch*). À travers le calcul des auto-corrélations pour des délais correspondant à des fréquences audibles, nous cherchons à identifier la fréquence fondamentale dominante pour chaque position de la fenêtre temporelle. Afin d'améliorer la fidélité de l'identification de la hauteur du son, les harmoniques de cette fréquence sont recherchées aussi et une interpolation de corrélogramme est mise en œuvre.
- Le centroïde spectral et la largeur de bande [3]. Le centroïde spectral est le centre de gravité du spectre du signal dans la fenêtre temporelle considérée. Dans notre cas, par largeur de bande du signal dans la fenêtre nous entendons la moyenne pondérée des différences en valeur absolue entre les composantes du spectre et le centroïde. Dans l'état actuel du développement, nous n'exploitons pas des caractéristiques comme le flux spectral ou l'irrégularité spectrale [3].
- Le nombre d'harmoniques et le nombre de ruptures harmoniques. Les pics significatifs du spectre du signal sont extraits dans chaque fenêtre temporelle. Les nombres d'harmoniques présentes et de gaps sont enregistrés. La continuité des harmoniques lors du déplacement de la fenêtre temporelle n'est pas directement évaluée dans la version actuelle.

Enfin, une extraction d'indications de rythme dans différentes bandes de fréquence est réalisée. Le spectre du signal est découpé en bandes de largeur 1 Bark et, dans chaque bande, l'amplitude du pic principal (dans un intervalle qui peut correspondre à un rythme) de l'auto-corrélogramme est enregistrée. Afin d'éviter un nombre trop important de descripteurs, seule l'information concernant la première bande du spectre (basses fréquences) est actuellement exploitée. Une méthode de représentation synthétique mais plus complète de l'information rythmique est en cours de mise au point.

L'enchaînement des composantes de l'analyse du signal musical est indiqué dans la figure 3. L'analyse s'applique actuellement à des fichiers au format *wave* et

chaque canal est traité séparément. Cette séparation permet souvent une extraction plus fiable des paramètres.

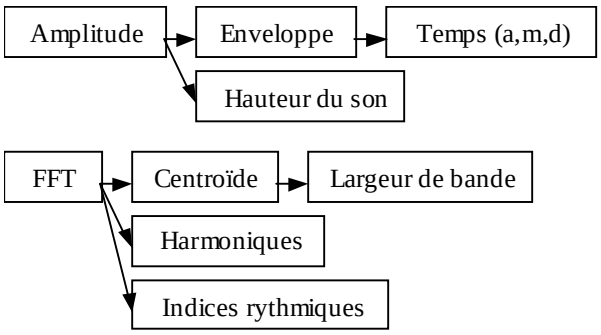

Figure 3 - *Processus d'analyse du signal musical*

Pour aborder l'étape d'identification des dimensions perceptives, l'information contenue dans les séries extraites est résumée dans des valeurs correspondant aux moyennes, variances, corrélogrammes et histogrammes synthétiques de ces séries.

Les techniques d'extraction de caractéristiques ont été appliquées sur un ensemble de 102 pièces de musique variées (pop, rock, jazz, classique, etc.).

4 Détermination des dimensions

Nous avons sollicité des volontaires pour noter une partie des pièces musicales mentionnées selon les dimensions perceptives suivantes : aptitude à être dansée, présence de basses, brillance, effet émotionnel, caractère mélodieux, qualité des arrangements, puissance, qualité du chant, qualité de l'interprétation, qualité sonore, richesse, tempo, vigueur. Ces dimensions ont été retenues après une étude de la littérature, de forums et de sites Internet spécialisés. Nous pouvons remarquer que la subjectivité attendue n'est pas la même pour ces différentes dimensions.

Parmi les conséquences importantes de la recherche de dimensions perceptives (donc subjectives) et de l'utilisation de volontaires pour remplir les questionnaires nous pouvons mentionner :

- La perception d'une même pièce musicale n'est pas la même d'un utilisateur à un autre. De plus, cette perception peut éventuellement varier dans le temps et suivant le contexte pour un utilisateur donné. L'interprétation d'une note à donner (par exemple 10/20) change aussi d'un utilisateur à un autre.
- Les différents volontaires ne notent pas nécessairement les mêmes pièces musicales. Cette difficulté peut être évitée dans la phase initiale d'une expérimentation, mais notre souhait de profiter des notes que certains utilisateurs fourniraient lors de l'exploitation ultérieure du système nous impose de ne pas contourner cette difficulté.

Dans la suite, nous faisons l'hypothèse simplificatrice suivante : la perception d'un individu ne change pas de façon significative dans le temps ou avec le contexte. D'abord, les interactions d'un utilisateur avec le système devraient se dérouler sur une période de temps relativement courte. Ensuite, nous sommes intéressés par le positionnement *relatif* des pièces musicales suivant les dimensions perceptives, qui devrait être plus stable que leur positionnement absolu.

Les questionnaires remplis par les volontaires lors de la phase initiale de notre expérimentation sont pré-traités : les pièces qui ont été notées par tous sont identifiées et, pour chaque dimension perceptive, les notes données par chaque volontaire sont normalisées à partir des notes pour des pièces communes (même moyenne et même variance pour tous les volontaires). L'opération de normalisation ne peut pas être effectuée de façon aussi simple dans l'étape ultérieure d'exploitation du système, car le pourcentage de pièces notées par tous serait probablement très faible. La normalisation devra alors être faite sur des sous-ensembles d'utilisateurs.

Les notes ainsi obtenues doivent nous permettre de trouver une correspondance entre les caractéristiques extraites du signal et des dimensions perceptives. Certaines relations sont examinées dans la littérature et nous les avons observées à partir de nos résultats, par exemple la relation entre l'amplitude moyenne et la puissance ou celle entre la valeur moyenne du centroïde spectral et la brillance.

Nous avons néanmoins constaté, sur les pièces musicales traitées, qu'aucune dimension perceptive ne pouvait être expliquée à partir d'une seule caractéristique du signal. Il est donc nécessaire d'explorer des modèles plus complexes, multivariés, pour ces dimensions. Trois questions importantes doivent trouver réponse :

- Comment sélectionner les variables explicatives utiles pour chaque dimension perceptive ?
- Comment choisir la complexité des modèles ?
- Comment tenir compte de la variabilité des notes (nombre, écart-type) données par différents volontaires à une même pièce musicale ?

En raison du faible nombre de pièces pour lesquelles nous avons obtenu des notes, nous avons décidé d'imposer un *a priori* linéaire fort sur tous les modèles. Sous cette hypothèse, la régression linéaire multiple nous permet de sélectionner les variables explicatives significatives. Une présélection a aussi été réalisée, pour quelques dimensions perceptives, à partir de publications existantes.

Si des modèles linéaires peuvent éventuellement suffire en première approximation, nous avons préféré faire appel aux réseaux de neurones de type perceptron multi-couches (PMC), avec une couche cachée, capables de modéliser des relations non linéaires. Dans ce cas, l'*a priori* linéaire est donné par l'utilisation, dans l'expression de l'erreur, d'un terme de régularisation spécifique, appelé aussi « oubli » (*weight decay*) dans [4]. Si on note par $\mathbf{w}$ le vecteur de paramètres du modèle, par $\mathbf{w}_h$ les poids des connexions entre les entrées et les neurones cachés, par N le nombre de pièces musicales de l'ensemble

d'apprentissage, par c_i la pondération de la pièce i, par $\bar{y}_i$ la moyenne des réponses pour cette pièce et par y_i la valeur donnée par le réseau en sortie, l'expression de l'erreur du réseau de neurones est

$$E(\mathbf{w}) = \sum_{i=1}^{N} c_i (\bar{y}_i - y_i)^2 + \mu \|\mathbf{w}_h\|^2 .$$

En pénalisant la norme de $\mathbf{w}_h$ suivant la valeur de $\mu > 0$, cette méthode encourage le fonctionnement des neurones cachés dans la région quasi-linéaire de leur fonction de transfert (dans notre cas, la tangente hyperbolique) et permet d'obtenir l'*a priori* linéaire.

Nous avons introduit la pondération c_i afin de tenir compte de la variabilité des notes données par différents volontaires à une même pièce musicale. Si n_i est le nombre de réponses concernant la pièce i et σ_i l'écart-type des réponses pour cette pièce, $c_i = n_i / \sigma_i$. Pour arriver à cette relation, nous considérons que les réponses fournies par les différents volontaires pour une même pièce constituent autant d'observations indépendantes et que l'objectif est d'obtenir la moyenne de la distribution associée (modélisée par une loi normale). La régularisation forte ne nous permettant pas d'atteindre l'estimation de la moyenne pour toutes les pièces, nous essayerons de nous en rapprocher. La mesure de distance (l'erreur) utilisée doit tenir compte de l'imprécision variable de l'estimation de la moyenne pour chaque pièce. La forme indiquée pour la constante c_i est obtenue en considérant une métrique de Mahalanobis.

Cette approche permet de sur-pondérer les pièces musicales pour lesquelles il y a beaucoup de notes et un bon accord entre ces notes. L'approche peut être appliquée telle quelle dans l'étape ultérieure d'exploitation du système et permet de profiter de chaque nouvelle note donnée par un volontaire à une pièce musicale.

Des modèles de type PMC ont été développés pour les dimensions suivantes : aptitude à être dansée, brillance, puissance, richesse, vigueur. L'ensemble de test représente 40% de l'ensemble des pièces musicales étudiées. Les modèles les plus fiables (la différence est négligeable entre l'erreur d'apprentissage et l'erreur sur l'ensemble de test) ont été obtenus pour la puissance et la richesse. Suivent, par ordre croissant de la différence, la brillance, l'aptitude à être dansée et la vigueur. Les valeurs de l'erreur sur les ensembles de test s'échelonnent entre 6% et 13%, ce qui reste faible si on tient compte de la relative dispersion des réponses données par les utilisateurs volontaires.

Pour tenir compte de la subjectivité indissociable de la perception musicale, les relations obtenues entre indices extraits du signal et dimensions perceptives ne doivent servir que de base de départ pour un utilisateur spécifique. Ces relations seront personnalisées par la suite grâce à l'interaction de l'utilisateur avec la représentation de la collection de pièces de musique, à travers l'outil d'exploration (figure 1). L'interaction peut être explicite, par le déplacement de points représentant des pièces, ou implicite, par le choix de dimensions et d'angles de vue. Sur les données ainsi modifiées une nouvelle correspondance, personnalisée, doit être obtenue par apprentissage automatique.

5 Conclusion et travaux actuels

Dans le but de mettre au point des outils d'exploration plus expressifs de collections de pièces de musique, nous avons proposé l'utilisation de dimensions perceptives interprétables, obtenues à partir du signal sonore et représentées visuellement. L'extraction de caractéristiques du signal, suivie par l'application d'une transformation apprise, permet le positionnement d'une nouvelle pièce de musique dans l'espace défini par ces dimensions.

Les résultats obtenus peuvent être améliorés par une augmentation de la base d'apprentissage, par l'extraction plus fiable de caractéristiques à partir du signal musical et, éventuellement, par l'enrichissement contrôlé de l'ensemble de caractéristiques extraites.

L'outil d'exploration actuel doit principalement permettre d'évaluer la faisabilité de notre approche. Bien que la navigation 3D soit familière à un nombre croissant d'utilisateurs potentiels, l'interface peut certainement être améliorée pour une utilisation courante. Aussi, différentes pistes peuvent être explorées pour l'adaptation aux terminaux légers, comme l'exploitation de synesthésies musique-couleur (synopsie) ou musique-texture.

Références

[1] R.B. Dannenberg, B. Thom, D. Watson. A machine learning approach to musical style recognition. Dans *Proceedings of the 1997 International Computer Music Conference*, pages 344-347, Septembre 1997.

[2] S. McAdams, C. Drake. Auditory perception and cognition. Dans H. Pashler, S. Yantis (éditeurs) *Stevens' Handbook of Experimental Psychology*, Vol. 1: Sensation and Perception, pages 397-452, New York: Wiley, 2002.

[3] J. Krimphoff, S. McAdams, S. Winsberg. Caractérisation du timbre des sons complexes, II : analyses acoustiques et quantification psychophysique. *Journal de Physique*, 4(C5) : 625-628.

[4] A.S. Weigend, D.E. Rumelhart, B.A. Huberman. Generalisation by weight-elimination applied to currency exchange rate prediction. Dans *Proceedings of IJCNN'91*, pages 837-841, Seattle, USA, 1991.