Le projet SemanticHIFI : Manipulation par le contenu d'enregistrements musicaux

Hugues Vinet

IRCAM-CNRS - STMS
1, place Igor Stravinsky, F-75004 PARIS, FRANCE
vinet@ircam.fr
http://www.ircam.fr/recherche.html

Résumé

Le projet européen SemanticHIFI[1] vise la préfiguration des chaînes hi-fi de demain, proposant aux mélomanes des fonctions inédites de gestion et de manipulation par le contenu des enregistrements musicaux. Les limites des équipements existants sont liées à celles des formats de diffusion de la musique, qui, se présentant depuis plusieurs décennies sous la forme de signaux d'enregistrements stéréophoniques, n'autorisent que des modes de manipulation élémentaires. L'extension des supports d'informations musicales à des représentations plus riches, issues soit directement de processus de production renouvelés ou d'outils d'indexation personnalisés, rend possible la réalisation de fonctions innovantes : classification personnalisée, navigation par le contenu, spatialisation sonore, composition, partage sur les réseaux préservant les droits liés aux œuvres, etc. Ces fonctions sont le résultat d'activités de recherche menées dans le cadre du projet et se situant à la pointe de plusieurs disciplines : analyse et traitement des signaux audionumériques, ingénierie des connaissances musicales, interfaces homme-machine, architectures de réseaux distribuées. Le projet prévoit également une phase d'intégration, visant la réalisation de prototypes d'applications, permettant de valider l'ensemble de ces fonctions dans un environnement technique unifié et compatible avec les contraintes du marché de l'électronique grand public.
L'objet de cet article est de proposer une vue d'ensemble du projet. La première partie en introduit le contexte et les principaux objectifs. La seconde partie est consacrée à la problématique de description et d'extraction automatisée des contenus musicaux, qui traverse l'ensemble des réalisations du projet. Les troisième, quatrième et cinquième parties présentent les principales fonctions développées : navigation inter- et intra-documents, jeu, composition et partage sur les réseaux. La dernière partie en précise les modalités d'intégration techniques sous la forme d'applications prototypes.

[1] shf.ircam.fr

1 - Cadre et principaux objectifs du projet

Le développement, au cours des dernières décennies, des technologies musicales, en particulier dans le domaine du traitement du signal audionumérique, n'a eu que des répercussions limitées sur les produits destinés au grand public. Celles-ci se sont essentiellement traduites par une plus grande facilité d'accès aux enregistrements sonores, grâce à des modes de compression et de stockage de plus en plus performants, ainsi que par la généralisation des systèmes de *Home Cinema* tirant parti des formats audio multicanaux liés aux DVD (DTS 5.1). Ce dernier aspect représente une véritable rupture technologique, tirée par l'industrie de la vidéo, dans le domaine de la production sonore, dont le format cible en était resté à la stéréophonie depuis un demi-siècle. Dans le strict champ de la musique enregistrée, même si de nouveaux modèles économiques de distribution électronique commencent à voir le jour, ceux-ci apportent encore peu d'innovation en matière de manipulation musicale. Les raisons en sont liées à l'absence d'évolution des représentations sur lesquelles reposent les différents supports des informations musicales, qui se sont limitées pour le moment aux signaux des enregistrements sous forme stéréophonique, n'autorisant que des fonctions de manipulation élémentaires (marche/ arrêt/ morceau suivant, réglage du volume, etc.). Ces formats se situent aujourd'hui en deçà de l'état de l'art des technologies musicales, qui décrivent les différents aspects des phénomènes musicaux à travers plusieurs types de représentations numériques, organisés, par ordre d'abstraction croissant, selon les niveaux *physique*, *signal*, *symbolique* et *cognitif* [12].
L'objectif du projet SemanticHIFI est de dépasser ces limitations à travers l'élaboration d'une nouvelle génération de chaînes hi-fi, répondant orientations suivantes :
- application de recherches en analyse et traitement de signal audio à la réalisation de fonctions de manipulation interactive des matériaux musicaux, visant à promouvoir des modes d'*écoute active* à destination de mélomanes non spécialistes ;
- gestion de représentations plus riches des contenus musicaux pouvant être obtenus selon deux modes complémentaires, comme résultats d'une part d'un processus de production renouvelé, d'autre part d'outils d'indexation personnalisés opérant à partir d'enregistrements;

- dépassement des approches usuelles en *recherche d'informations musicales*, reposant sur des services en ligne accessibles par ordinateur, en se concentrant sur des dispositifs d'accès terminaux, d'utilisation simple. Cette démarche s'inscrit dans la continuité du projet européen CUIDADO [11],
- intégration des fonctions développés sous la forme de prototypes d'applications, comme résultat d'un compromis entre les innovations suscitées par l'avancée des recherches et de choix fonctionnels et techniques liés aux enjeux économiques du domaine de l'électronique grand public.

Le projet, réalisé dans le cadre du programme IST (*Information Society Technologies*) de la Commission européenne entre 2004 et 2006, associe environ 30 chercheurs et ingénieurs de 5 laboratoires de recherche[2] et de deux sociétés[3] parmi les plus en pointe dans leurs domaines respectifs.

2 Description et extraction des contenus musicaux

Le nom du projet doit davantage être considéré comme un acronyme promotionnel que d'un point de vue littéral, car la notion de sémantique s'applique difficilement au champ musical, même si celles de syntaxe et de vocabulaire permettent des rapprochements avec celui du langage. Dans la mesure où les éléments structurels du contenu musical sont figurés dans la partition, il pourrait paraître logique, dans le contexte des applications visées, de recourir à cette représentation comme support d'interaction musicale. Cette approche s'avère toutefois inappropriée car la notation est un outil d'écriture et de prescription, qui ne reflète généralement pas explicitement les structures formelles pertinentes du point de vue de la réception des œuvres et nécessite par ailleurs, pour pouvoir être appréhendée, plusieurs années de formation. De plus, la partition ne rend pas compte de certains aspects spécifiquement liés au son, dont la prégnance est particulièrement forte dans différents genres musicaux. Il est donc nécessaire d'élaborer des *descriptions* spécifiques des contenus musicaux, dont les exemples ci-après rendent compte des différentes approches existantes :

- *informations éditoriales* (titre, noms des musiciens, instrumentation,…),
- *catégories musicales* (dont genres), suivant éventuellement une organisation hiérarchique (taxinomies),
- *descripteurs numériques globaux*, automatiquement extraits à partir des signaux et décrivant statistiquement diverses propriétés d'un morceau : couleur orchestrale, tonalité, tempo, structures rythmiques, intensité subjective [1,10];
- *empreintes digitales*, données numériques compactes permettant d'identifier de manière univoque un morceau ;
- *informations structurelles* sur le contenu intrinsèque des morceaux (forme temporelle, voies de polyphonie, profil mélodique, etc.) ;
- texte des *paroles* des chansons ;
- analyses musicologiques sous forme d'*applications multimédias*, pouvant être exécutées sur la chaîne hi-fi.

Cette problématique de description des contenus musicaux représente l'un des principaux verrous scientifiques du projet se situant à la convergence de deux orientations de recherche complémentaires, d'une part l'élicitation et l'ingénierie de connaissances musicales adaptées aux fonctions visées (processus *top-down*), d'autre part l'état de l'art en matière d'analyse de signaux pour l'extraction automatisée d'informations pertinentes, dites de « bas niveau » à partir des données disponibles (processus *bottom-up*) et leur mise en correspondance à travers l'utilisation de techniques d'apprentissage automatique.

3 Navigation inter-documents

La gestion des morceaux, dont le nombre peut être de plusieurs dizaines de milliers, est assurée dans le système hi-fi par l'application *Music Browser* de Sony [8]. Elle comprend d'une part des fonctions de classement personnalisé, d'autre part de navigation inter-documents. Lors de l'insertion de nouveaux morceaux, les informations éditoriales correspondantes peuvent être téléchargées auprès de différents fournisseurs de métadonnées musicales en ligne. L'utilisateur a également la possibilité de définir ses propres classifications, caractérisées par la répartition des morceaux en plusieurs catégories dotées de noms arbitraires et par des exemples prototypiques de morceaux. Le système est doté d'une fonction de généralisation, qui apprend les caractéristiques acoustiques liées aux prototypes de chaque classe et peut ainsi classer l'ensemble des morceaux présents selon ces catégories personnalisées [13].

Les descripteurs numériques globaux introduits plus haut sont automatiquement calculés par analyse des enregistrements et interviennent dans plusieurs fonctions de navigation. Ils permettent notamment de spécifier une recherche de morceaux selon les valeurs prises par la quantification de leurs caractéristiques musicales, de manière transversale et complémentaire à toute classification. Une autre heuristique intéressante est celle de *recherche par similarité*, dans laquelle l'utilisateur recherche les morceaux se rapprochant le plus d'un exemple choisi en fonction d'une mesure de similarité combinant ces différents descripteurs de manière configurable, selon différentes pondérations.

Le système est également doté d'une fonction de *recherche par chantonnement*, trouvant les morceaux dont le profil mélodique, lorsqu'il existe, se rapproche le plus de celui chanté par l'utilisateur et capté par un microphone intégré dans la télécommande de la chaîne [5].

---

[2] IRCAM (Coordinateur, F), Sony Computer Science Laboratory (Sony-CSL,F), Fraunhofer IDMT (FhG, D), University Pompeu Fabra (UPF, E), Ben Gurion University (BGU, Is)

[3] Native Instruments (D) et Sony European Technology Center (D)

4 Navigation intra-documents et rendu spatial

Le dépassement des fonctions traditionnelles de lecture est conçu, à travers la représentation des structures internes des morceaux, sous la forme d'interfaces de navigation intra-documents, qui perfectionnent la chaîne hi-fi dans sa fonction d'*instrument d'écoute*. Les approches adoptées se fondent sur les deux dimensions musicales de temporalité et de superposition polyphonique et spatiale, ainsi que, de manière plus élaborée et complète mais non automatisée, sur des analyses multimédias des œuvres.

4.1 Navigation à travers la structure temporelle

Des travaux de recherche récents permettent la modélisation automatique d'un morceau, par analyse de signal, comme succession d'états stables du point de vue du contenu spectral, qui rendent compte de parties distinctes en matière d'instrumentation et de registre telles que l'introduction, le refrain, les couplet et les solos [9]. La représentation graphique de cette décomposition en plusieurs segments temporels appariés permet d'appréhender la structure globale de morceau et suscite une navigation interactive associant visualisation et écoute des extraits sélectionnés. Cette analyse est également mise à profit pour le calcul automatisé de *résumés sonores,* extraits sonores de durée brève (20 à 30 secondes), obtenus pas concaténation d'une instance de chaque partie et fournissant l'essentiel des changements intervenant dans le morceau en s'affranchissant des répétitions.

Une autre fonction de navigation temporelle proposée, déjà relativement banalisée en production sonore professionnelle, mais non moins spectaculaire, est celle d'étirement-contraction (*time stretching*) : il est possible de lire à vitesse lente ou rapide un morceau sans altération corrélative des hauteurs, permettant, selon les cas, une écoute approfondie, ou un balayage rapide de son contenu.

Enfin, dans le cas des chansons, il existe différents modes de synchronisation des paroles chantées avec les signaux des enregistrements, permettant d'afficher automatiquement, comme à l'opéra, leur texte pendant l'écoute.

4.2 Navigation polyphonique et spatialisation sonore

Même si les instruments ou groupes d'instruments sont souvent enregistrés séparément sous forme multipiste, cette information est perdue à l'étape du mixage dans le format stéréophonique. Dans la mesure où les supports de distribution de la musique sont amenés à évoluer, à la fois via les formats DVD et SACD actuels autorisant 6 à 8 canaux, mais également à plus long terme à travers la diffusion des morceaux par les réseaux, sous des formes numériques s'affranchissant des contraintes des supports physiques, l'une des directions d'investigation abordées par le projet concerne l'élaboration de fonctions de navigation à l'intérieur de la polyphonie, conférant à l'utilisateur la possibilité d'effectuer son propre mixage à partir d'interfaces interactives. Celles-ci sont réalisées en application directe de travaux de recherche menés sur la spatialisation des sons. Le système Spat de l'Ircam, utilisé en production musicale et sonore, assure la simulation et le rendu de l'effet produit par des sources sonores, placées à des positions données dans une salle virtuelle dont les caractéristiques sonores peuvent être configurées à partir de paramètres pertinents d'un point de vue perceptif [7]. Ainsi, l'interface développée, suivant la métaphore de l'orchestre, se présente sous la forme d'un espace bidimensionnel, dans lequel apparaissent les positions des différents instruments ou voies de polyphonie et de l'auditeur. Celui-ci a la possibilité de déplacer les instruments, de choisir sa position dans l'orchestre, de s'approcher d'un instrument donné pour ne plus entendre que lui, etc. Le système effectue en temps réel le rendu spatial de cette scène sonore selon différents types de dispositifs de restitution, notamment le mode binaural (par casque), autorisant le rendu le plus précis de l'espace sonore tridimensionnel.

Dans le cas où les signaux des différentes voies de polyphonie ne sont pas disponibles, le système comprend une fonction effectuant, sous certaines conditions, la séparation automatique d'une voie solo et de son accompagnement et permettant de les spatialiser avec ce dispositif, en modifiant notamment leurs niveaux relatifs [2].

4.3 Interfaces multimédias interactives

Le système hi-fi comprend un lecteur de documents multimédias, combinant interfaces graphiques et lecture d'extraits sonores synchronisés. Ce dispositif permet l'exécution d'analyses d'œuvres produites à ce format, selon l'approche développée par le projet « Ecoutes signées » [4], en dépassant les limites des algorithmes d'analyse automatique. La possibilité de diffusion de telles applications sur des bases techniques standardisées offre également aux compositeurs un support à l'élaboration de nouvelles formes musicales, conçues pour être appréhendées selon un mode d'interaction qui s'affranchisse des limites traditionnelles de la linéarité temporelle ou d'une polyphonie fixée.

5 Autres fonctions : jeu, composition, partage

Les modes d'écoute active décrits aux paragraphes précédents sont complétés par des fonctions assistant le mélomane dans la production personnalisée de séquences musicales, à partir de matériaux existants. Ainsi sont proposés différents instruments virtuels (trompette, basse, percussions), synthétisés en temps réel et pilotés par la voix, selon un contrôle intuitif reposant sur différentes métaphores d'interaction [6]. Des fonctions de composition, destinées à des utilisateurs plus avancés et reposant sur le logiciel Traktor de Native Instruments, permettent l'agencement (montage, mixage, interpolation) de morceaux existants, en prenant en charge automatiquement certains aspects de la production musicale comme la synchronisation en tempo des différents morceaux, la gestion des transitions ou le calcul automatisé de listes de lecture. La chaîne hi-fi est également dotée d'une fonction originale de partage *peer-to-peer*, favorisant les échanges entre internautes tout en garantissant le respect des

droits afférents aux œuvres manipulées : seules les données produites par les utilisateurs peuvent être partagées. Ces métadonnées partageables comprennent non seulement les informations de description des morceaux, mais également des opérations de jeu ou de composition effectuées à partir de ceux-ci, qui ne peuvent être reproduites par d'autres internautes que dans la mesure où ils disposent eux-mêmes de ces morceaux dans leur propre système.

8 Intégration technique

L'ensemble des fonctions décrites plus haut sont issues de travaux de recherche et développement menés par les différentes équipes participantes et font l'objet de modes de validation individuels. Le projet prévoit, dans une seconde étape, leur intégration technique dans deux applications prototypes : la chaîne hi-fi proprement dite, réalisée par Sony EuTEC et une application d'outil auteur développée par Native Instruments. Le prototype de chaîne hi-fi, destiné à des utilisateurs novices, se présente sous la forme d'un appareil indépendant, doté d'un écran tactile, d'un lecteur/ enregistreur CD et DVD, d'un disque dur de grande capacité, d'une connexion à Internet et d'une télécommande dotée de fonctions graphiques assurée par un assistant personnel. L'application d'outil auteur est développée sous la forme de logiciel pour ordinateur, communiquant avec la chaîne et intégrant des fonctions avancées d'indexation et de production des matériaux musicaux. L'objectif de ces réalisations est de démontrer la faisabilité et d'évaluer le caractère utilisable de ces différentes fonctions dans le cadre d'environnements unifiés, à la fois du point de vue de l'architecture technique et des interfaces homme-machine. La commercialisation de tout ou partie de ces fonctions n'est envisagée que dans une étape ultérieure, sous des formes qui seront déterminées en fonction des opportunités commerciales.

9 Conclusion

Cette vue d'ensemble du projet SemanticHIFI a présenté le concept original de système hi-fi dont il vise la réalisation et dont les fonctions se situent bien au-delà de celles des chaînes traditionnelles, mais aussi d'applications plus récentes de gestion de morceaux de musique. Les avancées technologiques issues des activités de recherche du projet se fondent sur l'extraction automatisée et la combinaison de représentations numériques étendues des informations musicales, autorisant des modes de manipulation inédits des enregistrements : classification personnalisée, navigation inter- et intra-documents, spatialisation sonore, jeu instrumental, composition, partage par les réseaux respectant les droits afférents aux œuvres. Ces fonctions suscitent des modes d'interaction plus riches avec les contenus musicaux, destinés à des utilisateurs néophytes, qui bouleversent les limites des fonctions techniques musicales traditionnelles (instruments de musique, outils de composition et de production, dispositifs d'écoute). Cette préfiguration vise également à démontrer l'opportunité d'une extension des formats de distribution musicale, dont la faisabilité technique est d'ores et déjà avérée à travers les réseaux numériques, à des données fournissant une description plus détaillée des contenus musicaux, voire à de nouvelles formes interactives.

Références

[1] J.J. Aucouturier, F. Pachet. "Improving Timbre Similarity: How high is the sky?" Journal of Negative Results in Speech and Audio Sciences, 1(1), (2004).

[2] A. Ben-Shalom, D. Dubnov. "Optimal Filtering of an Instrument Sound in a Mixed Recording Given Approximate Pitch Prior" Actes de l'International Computer Music Conference, Miami, USA, (2004).

[3] J. Bonada. "High Quality Voice Transformation Based on Modeling Radiated Voice Pulses in Frequency Domain", Actes de l'International Conference on Digital Audio Effects, Naples, (2004).

[4] N. Donin. "Towards organised listening: some aspects of the 'Signed Listening' project, IRCAM", Organised Sound 9(1), (2004).

[5] T. Heinz, A. Brückmann. "Using a Physiological Ear Model for Automatic Melody Transcription and Sound Source Recognition", Actes de la 114ème Convention de l' Audio Engineering Society, Amsterdam, (2003).

[6] S. Jorda. "Instruments and Players: Some thoughts on digital lutherie", Journal of New Music Research, 33(3), (2005).

[7] J.M. Jot. "Efficient Models for Distance and Reverberation Rendering in Computer Music and Virtual Audio Reality", Actes de l'International Computer Music Conference. San Francisco, USA, (1997).

[8] F. Pachet, J.J. Aucouturier, A. La Burthe, A. Beurive. "The Cuidado Music Browser : an end-to-end Electronic Music Distribution System", Multimedia Tools and Applications, Special Issue on the CBMI03 Conference, (2004).

[9] G. Peeters. "Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation : "Sequence" and "State" approach", Lecture Notes in Computer Science, Springer Verlag, Volume 2771, (2004).

[10] G. Peeters. "Time Variable Tempo Detection", Actes de l'International Computer Music Conference, (2005).

[11] H. Vinet, P. Herrera, F. Pachet. "The CUIDADO Project", Actes de l'International Conference on Music Information Retrieval (ISMIR), Paris, (2002).

[12] H. Vinet. "The Representation Levels of Music Information", Lecture Notes in Computer Science, Springer Verlag, Volume 2771, (2004).

[13] A. Zils, F. Pachet. "Automatic Extraction of Music Descriptors from Acoustic Signals using EDS", Actes de la 116ème Convention de l'Audio Engineering Society, Berlin, Germany, (2004).