Applications musicales du traitement de signal : synthèse et prospective

Musical Applications of Signal Processing: Synthesis and Prospect

Hugues Vinet [1]
[1]STMS, Ircam-CNRS, 1, place Igor Stravinsky 75004 Paris
vinet@ircam.fr

Manuscrit reçu le 15 novembre 2005

Résumé et mots clés

L'objet de cet article est de proposer une synthèse des applications musicales du traitement de signal, des problématiques de recherche qui leur sont liées et des directions prospectives qui se dégagent sur la base de travaux récents dans ce domaine. Après l'exposé de notions préliminaires, relatives au système technique musical et à l'analyse des différentes représentations numériques des informations musicales, cette synthèse se concentre sur trois types de fonctions principales : la synthèse et le traitement des sons musicaux, la spatialisation sonore et les technologies d'indexation et d'accès.

Musique, Signaux audionumériques, Acoustique, Modélisation, Spatialisation sonore, Indexation musicale, Recherche d'informations musicales, Ingénierie des connaissances.

Abstract and key words

This article aims at providing a synthesis of the musical applications of digital signal processing, of related research issues, and of future directions that emerge from recent works in that field. After introducing preliminary notions related to the music technical system and to the analysis of different digital representations of music information, it focuses on three main function types: audio synthesis and processing, sound spatialization and audio indexing and access technologies.

Music, Digital Audio, Acoustics, Modeling, Sound Spatialization, Music Indexing, Music Information Retrieval, Knowledge Engineering.

Remerciements

Les développements qui suivent ont été notamment inspirés par les recherches, auxquelles ils font largement référence, dirigées à l'Ircam par Xavier Rodet (Analyse/ synthèse des sons), Olivier Warusfel (Spatialisation sonore) et René Caussé (Modélisation physique).

1. Introduction

Les applications musicales du traitement de signal ont longtemps été le domaine réservé de centres spécialisés, notamment en lien avec la création contemporaine (synthèse, transformation des sons) et les télécommunications (codage). Elles ont connu au cours des 20 dernières années deux importantes phases successives d'extension, la première liée à la généralisation des techniques de production et de diffusion audionumérique, la seconde, en cours, à la diffusion d'enregistrements par les réseaux et aux méthodes d'indexation et de recherche dans les bases de données musicales. L'objet de cet article est de proposer une synthèse de l'ensemble des méthodes et problématiques existantes et de présenter les directions prospectives qui se dégagent sur la base de recherches en cours, notamment en lien avec les travaux de l'Ircam [1]. Après l'exposé de notions préliminaires, il se concentre sur trois types de fonctions principales : la synthèse et le traitement sonores, la spatialisation et les technologies d'indexation et d'accès.

2. Notions préliminaires

2.1. Le système technique musical

Toute activité musicale peut se décomposer en trois phases de production, transmission et réception des artefacts sonores (Figure 1). La transmission est effectuée directement par le milieu acoustique dans le cas du concert, et dans les autres cas par l'intermédiaire de supports codant les informations musicales et nécessitant à la réception une étape d'accès (sélection et obtention des éléments) et de reproduction sonore. Les supports ainsi produits peuvent être réinvestis pour de nouvelles étapes de production, liées par exemple à la diffusion des contenus selon de nouveaux formats (*re-mastering*), moyennant éventuellement certaines étapes de traitement (restauration d'archives sonores par exemple).
La phase de production réside dans l'utilisation d'instruments et appareils issus d'une activité de lutherie ou production d'outils,

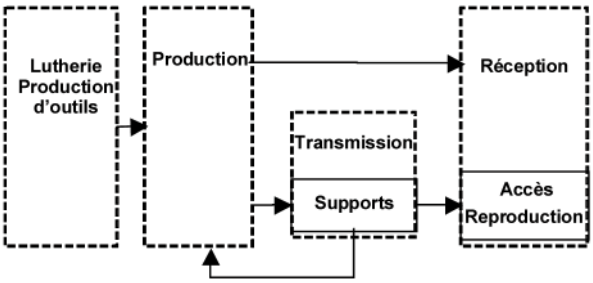

Figure 1. Phases du système technique musical.

qui est ici mentionnée car les environnements informatiques de production musicale offrent désormais à l'utilisateur musicien, comme nous le verrons plus loin, de telles possibilités à travers des fonctions de programmation. Elle se décompose elle-même en trois phases distinctes (Figure 2), relatives respectivement à la production des matériaux musicaux isolés (signaux issus de sons instrumentaux ou synthétiques, éventuellement traités, éléments de partitions, etc.), à leur composition, à la fois en temps (montage) et en superposition (mixage) en un contenu composite, puis au codage de ce produit musical au format de diffusion (mastering CD et DVD par exemple).

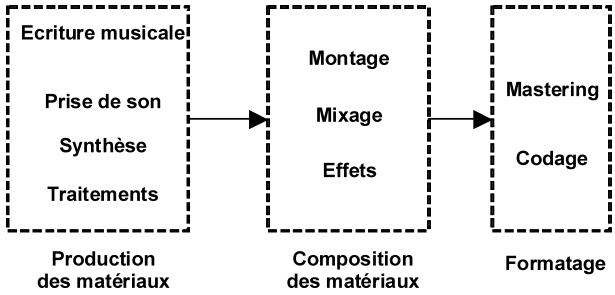

Figure 2. Phases de production musicale.

Notre étude des différentes fonctions techniques musicales comme applications du traitement de signal se concentrera dans la suite sur les étapes de lutherie, de production de matériaux et de dispositifs d'accès et de reproduction, et ne traitera pas des aspects de codage des informations à travers les différents formats de compression et de diffusion.

2.2. Représentations numériques des informations musicales

Même si les signaux audionumériques constituent l'une des principales formes de diffusion des informations musicales, il convient de les situer du point de vue des différentes représentations existantes des phénomènes musicaux et dans leurs interrelations avec celles-ci. J'ai proposé récemment une typologie des représentations numériques de la musique telles qu'elles apparaissent dans les différentes applications, en dégageant quatre principaux types et en montrant qu'ils s'organisent, dans un sens d'abstraction croissant, selon les niveaux « Physique », « Signal », « Symbolique » et « Cognitif » [37] (cf. Figure 3).
Le niveau symbolique décrit les contenus musicaux sous forme de structures à un niveau de discrétisation relevant de la théorie musicale : hauteurs (gamme tempérée), intensités (nuances), rythmes (structures temporelles définies comme multiples et sous-multiples d'une pulsation de base). Le niveau physique rend compte de modélisations acoustiques des phénomènes sonores, en particulier pour la synthèse et le rendu spatial. Le niveau cognitif code des informations de plus haut niveau que le niveau symbolique et inclut tous types de connaissances et des-

1. Institut de recherche et communication acoustique/musique.

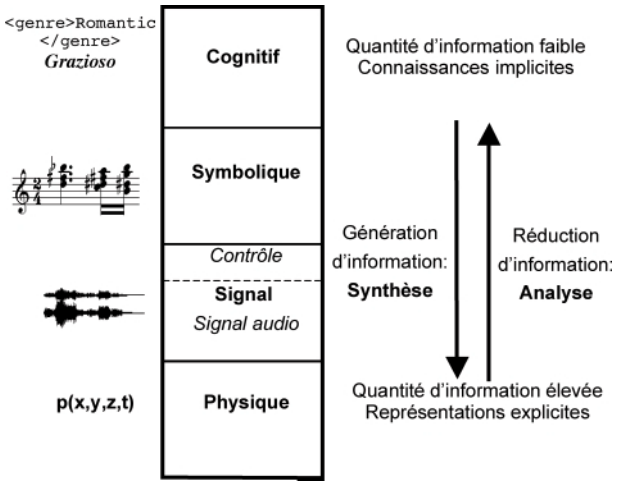

Figure 3. Les niveaux de représentations musicales.

cripteurs, notamment qualitatifs, des contenus musicaux. Il intervient notamment sous la forme de métadonnées dans les applications de bases de données musicales dotées de fonctions de recherche par contenu. Quant aux représentations de type signal, elles concernent tout autant les signaux audionumériques sous différentes formes (signaux multicanaux, compressés, etc.) que les signaux de contrôle, issus par exemple de captations gestuelles, selon des variations plus lentes. Je reviens ici sur mes définitions initiales dans lesquelles les signaux de contrôle étaient liés au niveau symbolique pour des raisons d'homogénéité quantitative, car les méthodes de traitement de ces informations s'avèrent davantage relever de techniques de signal à des taux d'échantillonnage faibles (typiquement 30 Hz à 1 kHz) que de méthodes symboliques.

Il est aisé de montrer que ces niveaux sont ordonnés par quantité d'information décroissante, les informations contenues dans les représentations de plus haut niveau étant associées à des connaissances implicites (culturelles par exemple) extérieures aux données de représentation. Les processus d'analyse et de synthèse constituent les opérateurs génériques de conversion entre niveaux, et prennent de formes spécifiques en ce qui concerne les représentations musicales. Par exemple, la réduction d'information opérée par l'analyse concerne notamment, dans le passage du physique au signal, une réduction spatiale, équivalent de la *prise de son*, consistant à passer d'une pression acoustique fonction de l'espace et du temps à des signaux d'une variable temporelle. Le passage du signal au symbolique concerne une double discrétisation, à la fois des occurrences temporelles et des valeurs prises par les grandeurs concernées (par exemple, passage du continuum des fréquences et énergies à des échelles discrètes de hauteurs et de nuances). Le passage du signal et/ou symbolique au niveau cognitif concerne une discrétisation de plus haut niveau, s'attachant à l'obtention de catégories et descriptions globales à l'ensemble du morceau.

3. Synthèse et traitement sonores

Les fonctions de synthèse et traitement sonore, relevant des phases « Production des matériaux » et « Effets » de la Figure 2, constituent une large classe d'applications musicales du traitement de signal. L'étude de leurs caractéristiques passe par celle des modèles de base utilisés (§ 3.1), mais aussi, d'un point de vue plus macroscopique, des architectures à travers lesquelles ces modèles sont mis en œuvre (§ 3.2). En effet, des méthodes récentes s'orientent vers des fonctions de synthèse et traitement *par le contenu*, d'une part à travers la différenciation des traitements opérés en fonction de l'analyse des signaux d'entrée, d'autre part à travers leur contrôle de haut niveau. De plus, ces questions d'architecture interviennent également dans la réalisation de *langages musicaux*, offrant à l'utilisateur musicien des fonctions de programmation d'algorithmes de synthèse et traitement relevant d'une « lutherie électronique ».

3.1. Modèles de synthèse et de traitement

3.1.1. Modèles de synthèse

Les premières synthèses audionumériques ont été réalisées par l'équipe de Max Mathews aux Bell Labs dans les années 50 et reposaient sur la combinaison d'oscillateurs périodiques, produisant des sons électroniques très typés. Dans les années 80, la mise au point de techniques de synthèse par modulation de fréquence à l'Université de Stanford, permettant la production de structures spectrales riches et variées avec des ressources de calcul limitées, a permis l'essor des synthétiseurs de la famille DX7 de Yamaha [8]. Les techniques utilisées dans les synthétiseurs se sont par ailleurs concentrées sur l'échantillonnage de sons enregistrés dotés de fonctions simples de post-traitement (variation de taux d'échantillonnage pour la transposition, enveloppes en amplitude, etc.), avec comme métaphore de contrôle celle du clavier à travers la norme MIDI[2]. Dans un contexte plus expérimental, la mise au point par Xavier Rodet à l'Ircam des FOF (fonctions d'ondes formantiques), modèle temporel de synthèse source-filtre adapté aux formants vocaux, a rendu possibles les premières synthèses de voix chantée.

Les méthodes de *synthèse par modélisation physique*, fondées sur la modélisation acoustique des sources sonores, se sont développées plus récemment. Bien que nécessitant des ressources de calcul plus importantes, elles présentent de nombreux avantages par rapport aux modèles de signaux : production de sons plus riches et proches des sons réels, paramètres physiques plus signifiants, reproduction de comportements non-linéaires caractéristiques de certains instruments (transitoires, bifurcations et régimes chaotiques). Il existe plusieurs approches adaptées à la modélisation de différentes classes d'instruments :

2. www.midi.org.

- les *guides d'ondes* [34], technique peu coûteuse en calcul mais limitée à certaines structures algorithmiques (monodimensionnelles essentiellement), même si l'application à des structures bidimensionnelles a été formalisée [35]. Cette méthode a connu une première application commerciale avec le synthétiseur VL1 de Yamaha. Des travaux récents liés à la modélisation de résonateurs d'instruments à vent fournissent des formalismes pour la résolution et la simulation de la propagation des ondes dans des tubes évasés en prenant en compte les pertes viscothermiques [14] ;
- la *synthèse modale*, formalisme utilisé dans le logiciel Modalys de l'Ircam [1,16], permettant la modélisation d'un grand nombre de classes d'objets de base mono- et bidimensionnels (tubes, plaques, cordes, membranes,…) et leur assemblage, par l'intermédiaire d'interactions non-linéaires, pour « construire » des instruments virtuels sans limite de complexité. Au-delà de ces possibilités de « lutherie virtuelle » permises par son formalisme unificateur, l'intérêt de la synthèse modale est aussi que les données modales d'un objet peuvent soit résulter d'une résolution analytique dans le cas de structures simples, soit de mesures d'analyse modale pour des corps réels plus complexes. Des travaux récents ont également permis l'obtention des modes de structures tridimensionnelles à partir de méthodes par éléments finis [2] ;
- les systèmes masses-ressorts-amortissements, permettant de constituer des réseaux de corps oscillants. Cette approche a montré son intérêt dans des applications expérimentales associant synthèse d'image, de son et contrôle gestuel, ces différentes modalités étant reliées à travers le même modèle [6].

Au-delà de problèmes de modélisation, l'une des principales difficultés inhérentes aux modèles physiques est celle de leur contrôle : certains modèles se présentent comme des systèmes dynamiques non-linéaires et leur contrôle pour obtenir les sons voulus nécessite l'adjonction d'une interface supplémentaire, modélisant l'expertise de l'interprète, qui effectue la variation conjointe des différents paramètres à partir de données d'entrée plus musicales, telles que hauteur à produire, mode de jeu choisi. Plusieurs approches peuvent être menées à cet effet à partir de signaux produits par des instruments réels : problèmes inverses [13], apprentissage automatique. Cette problématique, encore relativement peu abordée dans les applications existantes, a récemment débouché, dans le cadre du projet RIAM Windset, sur un ensemble de modules commerciaux de synthèse d'instruments à vent dotés de contrôles musicaux [3].

3.1.2. Modèles pour le traitement musical

Il existe une grande variété de fonctions de traitement sonore dans les applications existantes, allant du filtrage au débruitage en passant par différents types d'effets. Nous nous concentrerons ici sur les traitements les plus spécifiquement musicaux : transposition, changement de la durée sans changement de hauteur (*time-stretching*), morphing et hybridation entre sons, etc. Plusieurs critères doivent être pris en compte pour l'évaluation des différents modèles : les classes de sons reproduits, l'espace de variations possibles, la présence ou non d'artefacts à l'écoute, la fidélité de simulation de fonctions de traitements par rapport à leurs référents acoustiques, le coût de calcul. À titre d'illustration des principaux modèles utilisés, on peut citer :
- le *vocodeur de phase*, consistant en une répartition du contenu spectral par bandes de fréquences, pouvant être traitées indépendamment. Cette approche, adaptée à de larges classes de sons, est aujourd'hui généralement modélisée par une analyse/synthèse par TFCT [4] et permet des traitements tels que filtrage, time-stretching et transposition avec une bonne qualité sonore. Les avancées récentes sur ce modèle concernent notamment la reconstitution des phases pour le time stretching [12], ainsi que la prise en compte des transitoires, un rendu plus naturel du time stretching étant obtenu par un traitement limité aux portions quasi-stationnaires des sons [26]. Le logiciel Audiosculpt de l'Ircam (Figure 4) [5], reposant sur le noyau d'analyse/ synthèse SuperVP, propose ainsi des fonctions d'édition temps-fréquence des sons à partir d'une analyse de type TFCT, ainsi que la visualisation de différentes analyses superposées au sonagramme : fréquence fondamentale, analyses de partiels et de formants, détection de transitoires, etc.

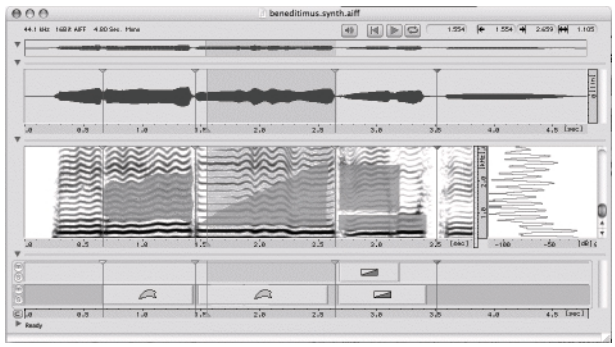

Figure 4. Logiciel Audiosculpt de l'Ircam.

- le *modèle additif ou sinusoïdal*, consistant à décomposer un son en une somme de sinusoïdes de fréquences et d'amplitudes à variations lentes et d'un signal résiduel de bruit [24]. Ce modèle, plus compact en données d'analyse que le précédent, convient à des classes de sons plus limitées (spectre discret). Des méthodes efficaces de synthèse par TFR inverse existent et les problèmes d'analyse varient selon les types de sons : calcul de fréquence fondamentale pour les sons harmoniques, méthodes de détection de pics et de suivi de partiels pour les sons non-harmoniques [27]. Au-delà de fonctions de transposition et de time-stretching, ou de variation du contenu spectral par modification des amplitudes des partiels, une fonction musicale intéressante de ce modèle est celle de « morphing sonore », par interpolation des fréquences et amplitudes des partiels appariés entre le son de départ et celui de destination.

3. www.arturia.com/en/brass.php.org.

4. Transformée de Fourier à court terme.

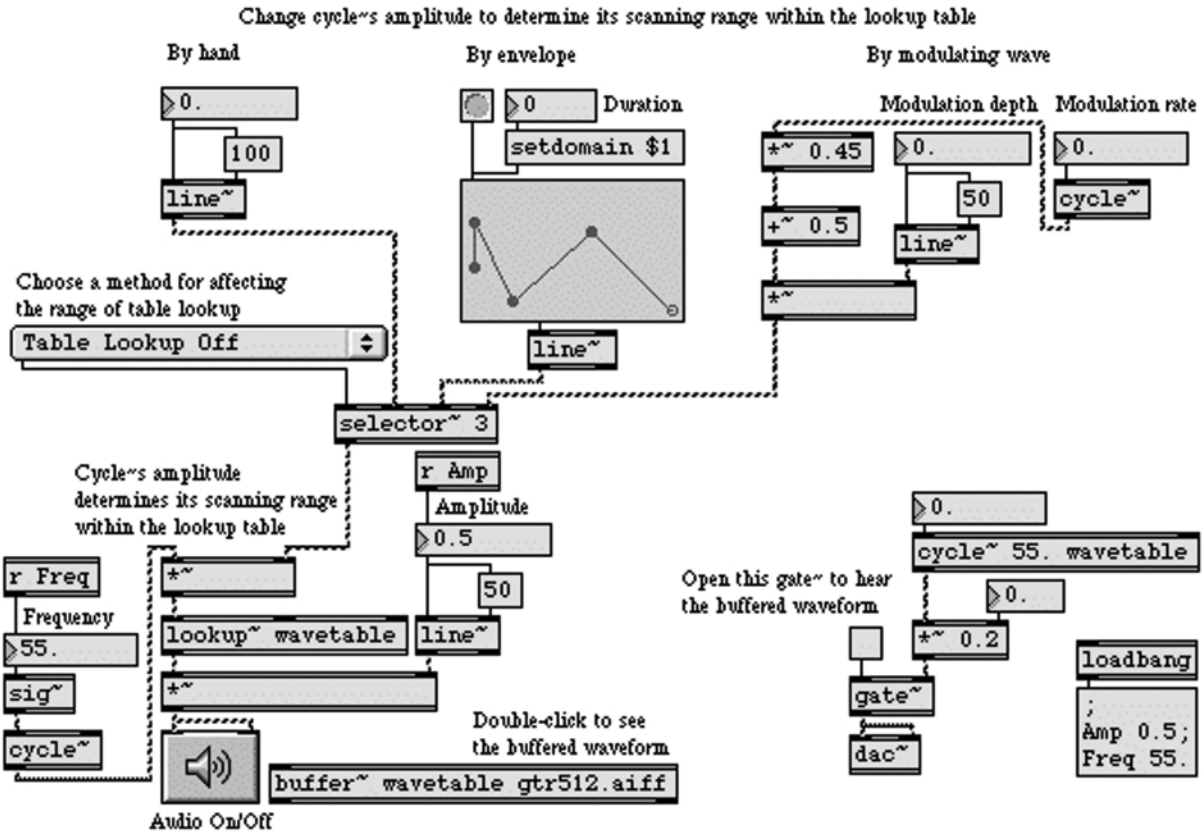

Figure 5. Exemple de patch du logiciel Max.

- les méthodes de type PSOLA (*pitch synchronous overlap add*) sont également utilisées, en particulier associées avec le modèle additif [20], et sont notamment intéressantes dans le cas des sons de spectre harmonique du fait de leur faible coût de calcul pour la synthèse par des méthodes superposition-addition temporelles,
- même si leurs applications musicales restent encore à notre connaissance expérimentales, les méthodes de décomposition sur des familles d'atomes de type *Matching Pursuit* présentent un intérêt pour l'analyse des signaux audio, de même que, pour la synthèse, les méthodes de concaténation d'unités appliquées aux signaux musicaux [30, 31, 33].

Le format SDIF (*Sound Description Interchange Format*), est devenu un standard de fichiers dans la communauté d'informatique musicale pour représenter sous forme binaire tous types de données d'analyse issues des modèles présentés ci-dessus, ainsi que de nombreuses autres (enveloppes spectrales, diphones, etc.) [32].

3.2. Architectures pour la synthèse et le traitement

3.2.1. Langages musicaux

Si les applications industrielles du traitement de signal font appel à des circuits spécialisés pour la réalisation de fonctions spécifiques, les modes d'implantation des algorithmes dans le domaine de l'informatique musicale suivent généralement une autre logique, consistant à exploiter les ressources de calcul existantes pour offrir une gamme aussi large que possible de fonctions de synthèse et de traitement. Dès les débuts de la synthèse audionumérique, la conception de fonctions de calcul élémentaires a été associée à la réalisation de *langages musicaux*, langages informatiques spécialisés pour la musique, dotés notamment de modèles du temps et permettant de programmer des algorithmes de calcul complexes à partir de ces opérateurs élémentaires. De nombreux langages musicaux ont ainsi été développés dans la communauté d'informatique musicale, notamment MusicV et plus récemment Csound[5], dont les concepts reposent sur l'articulation de deux métaphores musicales, la notion d'« instrument » lié à l'algorithme de synthèse ou de traitement, et celle de « partition », définissant la variation des paramètres d'entrée de l'instrument au cours du temps.

Le lien entre langages musicaux et traitement temps réel a été rendu possible par la diffusion du logiciel Max/MSP[6], qui, à raison de plusieurs milliers de licences vendues par an, est devenu un environnement de référence pour toute la communauté d'informatique musicale. Max est un langage de programmation visuelle permettant de réaliser un algorithme de traitement en

5. www. csounds.com.
6. http://www.cycling74.com/products/maxmsp.html.

temps réel sous forme de « patch » à partir de fonctions de base, ou « objets » (cf. Figure 5). Une interface de programmation permet à des tierces parties de développer leurs propres objets, et l'un des atouts de Max est qu'il est accompagné de plusieurs milliers d'objets et patchs assurant un ensemble très large de fonctions (synthèse et traitement audio, vidéo, interfaces à des dispositifs d'entrée, à des protocoles réseau, etc.).
Les travaux récents liés à Max visent notamment à doter celui-ci de types de données complexes, permettant, comme dans le cas de la bibliothèque « Gabor » [28], de réaliser des familles d'objets d'analyse/ synthèse de signal en temps réel s'échangeant des structures de données d'analyse indexées sur le temps (listes de partiels, enveloppes spectrales, etc.) et les resynchronisant automatiquement pour la synthèse à travers un mécanisme superposition-addition généralisé. Cette approche repose sur des travaux récents dans cet environnement visant à concilier efficacité pour le traitement en temps réel et allocation dynamique des données [29].

3.2.2. Structures de traitement et de contrôle : vers un traitement par le contenu

L'objet de cette section est de reprendre les notions abordées au §3.1 dans une perspective plus large en étudiant les différentes structures de traitement du point de vue de la prise en compte qu'elles opèrent des contenus musicaux, et en particulier des modes de contrôle qu'elles autorisent. Cette analyse se fondera sur les notions de niveaux de représentation des informations musicales définies au §2.2.

Le cas de la Figure 6-a, le plus courant, est celui de traitements n'opérant aucune différenciation en fonction des signaux d'entrée. La Figure 6-b illustre la production de certains paramètres de traitement par analyse du signal d'entrée, par exemple le calcul d'une enveloppe spectrale avant transposition de signaux produits par une structure de type source-filtre. La figure 6-c rend compte d'une classe assez large de modèles effectuant une décomposition des signaux d'entrée et opérant des traitements différenciés selon leurs contenus (par exemple time-stretching sur parties stationnaires uniquement et parties transitoires inchangées). La figure 6-d concerne en particulier les structures de traitement abordées au §3.1.2 sous la forme de modèles paramétriques d'analyse/synthèse, les traitements étant effectués par transformations dans l'espace des paramètres. Un cas particulier est celui où les traitements sont contrôlés par des paramètres de niveau symbolique, dont la signification musicale est directement pertinente pour l'utilisateur. Certaines applications commerciales intègrent de telles approches. Ainsi, le logiciel Melodyne permet l'édition musicale de sons monodiques après une pré-analyse dans le domaine symbolique, à travers des fonctions musicales telles que transposition ou calage sur la grille temporelle d'un rythme donné [7]. Le logiciel GarageBand [8] d'Apple permet à des utilisateurs non-musiciens de constituer facilement des séquences musicales à partir de boucles instrumentales préenregistrées, le calage des hauteurs et des instants associés à chaque note étant effectué automatiquement en fonc-

7. www.celemony.com/cms/index.php?id=melodyne.
8. www.apple.com/ilife/garageband.

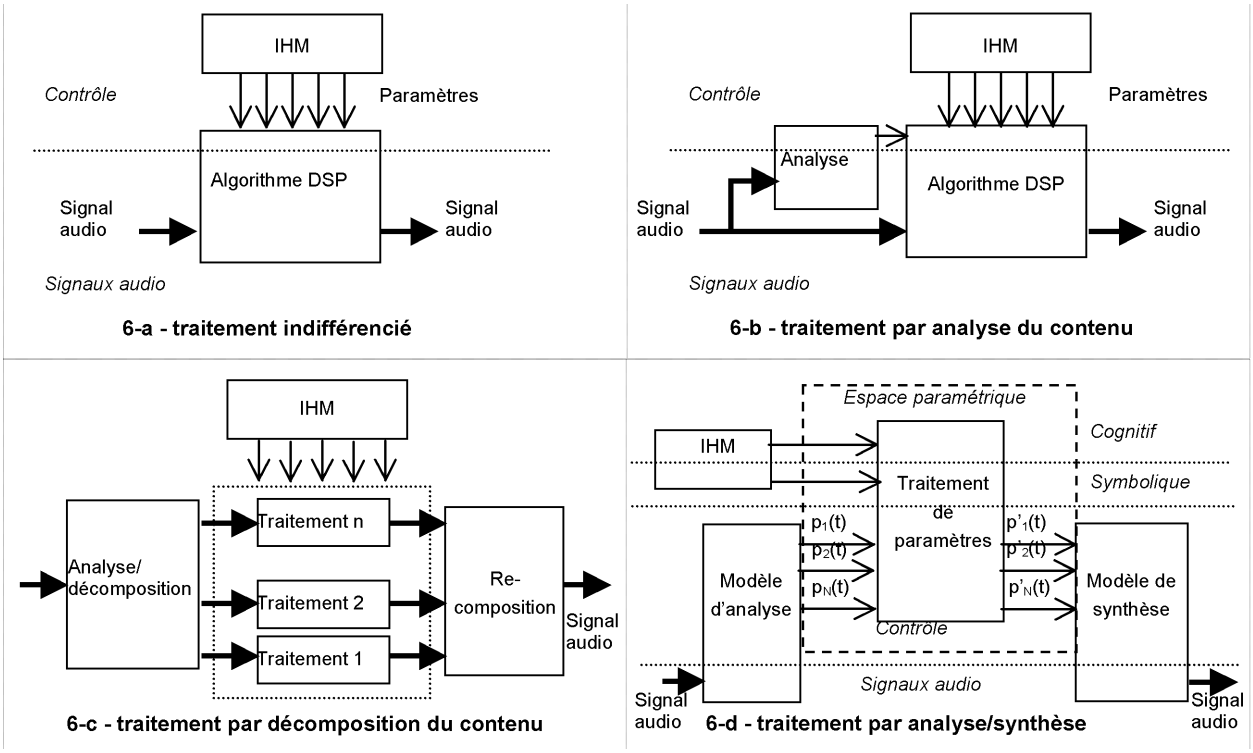

Figure 6. Structures de traitement et de contrôle.

tion du contexte musical où est placé le nouveau son (harmonie, tempo) à partir de formats (AppleLoops) combinant représentations de types signaux et symboliques. La figure 6-d dans sa généralité concerne une acception plus large de la notion de traitement par le contenu, celui-ci étant spécifié non seulement selon ses paramètres symboliques, mais aussi à partir de descripteurs relevant du niveau cognitif, s'attachant à spécifier par exemple des caractéristiques qualitatives du résultat sonore attendu : son « plus métallique », « moins sourd », etc. Cette problématique est liée à celle de la description des contenus sonores abordée au § 5.1. Une approche complémentaire consiste à élaborer des *modèles de contrôle* pour l'analyse et la synthèse, décrivant les évolutions des paramètres de contrôle selon le contexte (symbolique et cognitif) de plus haut niveau. Cette visée, encore prospective, relève notamment de recherches sur la modélisation de l'interprétation, comme fonction équivalant à instancier des signaux de contrôle à variation lente (gestes), pilotant un modèle de synthèse (instrument), à partir d'une prescription de type symbolique + cognitif (partition). Le cas le plus général du traitement par le contenu est celui qui combine les structures des figures 6-c et 6-d, les traitements spécifiques à chaque partie du son pouvant faire l'objet d'un modèle d'analyse/synthèse adapté à chacune d'elle.

3.2.3. Extensions de la notion d'instrument

La notion traditionnelle d'instrument de musique se trouve bouleversée par l'évolution des techniques de synthèse en temps réel. Une première extension de cette notion d'instruments conduit à les définir comme des systèmes reliant interfaces gestuelles et dispositifs de synthèse sonore en temps réel, selon un mode d'interaction immédiate. Une différence notable entre les instruments électroniques et acoustiques tient en effet dans le découplage que les premiers permettent entre l'énergie du geste et celle du son, autorisant tous types de mises en correspondance (*mapping*) entre paramètres issus du geste et paramètres de synthèse [7]. Au-delà de fonctions de correspondance statique reliant p paramètres gestuels à n paramètres de synthèse, des recherches récentes visent également à extraire, de signaux gestuels multidimensionnels issus d'ensembles de capteurs, des formes temporelles de plus haut niveau (modes de jeu, geste dansé) pour piloter des processus musicaux. Cette approche relève de problèmes de modélisation et de traitement de signaux de contrôle abordés au paragraphe précédent.

4. Spatialisation sonore

Cette acception désigne toutes techniques s'attachant au caractère spatial des sons à travers des fonctions de représentation/ manipulation, de captation et de reproduction/ simulation/ synthèse des scènes sonores. Je me concentrerai ici sur les modèles de spatialisation et sur les dispositifs de reproduction et de synthèse spatiale, en laissant de côté les problématiques de captation, qui reposent à certains égards sur des méthodes symétriques à celles liées à la reproduction (transducteurs électroacoustiques).

4.1. Modèles pour la spatialisation

Sans parler à ce stade de modèles, les principales formes de codage des informations sonores spatiales dans les applications usuelles reposent sur des systèmes à plusieurs canaux de signaux audionumériques (stéréo, 5.1), captés, transmis et reproduits séparément. Ces procédés, de mise en œuvre relativement simple, ne permettent cependant la reproduction et simulation de scènes sonores spatiales que de manière approximative, en particulier du fait du manque de prise en compte des conditions de restitution (réponse de la salle, directivité des haut-parleurs, etc.). Le format Ambisonic [19], qui propose la modélisation de champs acoustiques sous la forme d'une décomposition en harmoniques sphériques, offre un mode de codage de scènes tridimensionnelles pouvant être obtenues à partir de microphones multi-capteurs. Les modèles permettant de décrire les informations spatiales selon un point de vue plus abstrait s'attachent à caractériser d'une part des sources sonores isolées (position, rayonnement), d'autre part la réponse de l'espace dans lequel elles sont placées. Il existe deux classes principales de modèles de spatialisation, toutes deux normalisées dans MPEG4 [9] : l'approche géométrique et l'approche perceptive. Reposant sur un modèle géométrique de la salle (discrétisation des parois selon des portions de surface de caractéristiques acoustiques homogènes), les modèles géométriques permettent de calculer avec précision la réponse impulsionnelle caractérisant la réponse de la salle sur un couple (source, récepteur). Ils conviennent davantage à des simulations statiques, le cas de sources en déplacement nécessitant de recalculer la réponse impulsionnelle à chaque position ou d'avoir pré-calculé les réponses sur tous les couples de positions (émetteur, récepteur). Le modèle perceptif, issu d'études expérimentales sur la qualité acoustique des salles [18], caractérise celles-ci selon un ensemble de paramètres perceptifs liés à la source, à la salle et à leurs interactions. À partir d'un modèle simplifié de réponse impulsionnelle en zones spectro-temporelles, il est possible de produire automatiquement une telle fonction à partir des paramètres perceptifs et donc de simuler la qualité acoustique correspondante [17]. L'avantage de cette approche, qui combine simulation de la localisation des sources et de l'effet de salle, est que la spatialisation peut être spécifiée selon des paramètres de haut niveau, indépendamment du dispositif de reproduction utilisé en s'adaptant automatiquement à celui-ci : cette approche relève des méthodes de traitement par le contenu décrites plus haut.

9. ISO/IEC 14496-1: 2000. MPEG-4 Systems standard, 2nd Edition.

4.2. Reproduction et synthèse spatiales

Les techniques les plus courantes pour simuler un espace sonore et notamment des sources de position donnée, éventuellement mobiles, reposent sur des systèmes à base de plusieurs haut-parleurs, par pondération des amplitudes des signaux envoyés à chaque source. Cette méthode est notamment à l'œuvre dans les potentiomètres panoramiques des consoles. Cependant, elle est approximative, car les indices d'amplitude ne sont pas les seuls pertinents dans la perception auditive de la localisation, et parce que le rendu de tels effets est fortement dépendant de la position des auditeurs dans l'espace de restitution. La synthèse binaurale (par casque) résulte quant à elle d'études systématiques sur la perception spatiale des sons [4] et se fonde sur des ensembles de réponses impulsionnelles mesurées aux niveaux des deux tympans pour chaque direction d'incidence de son : les HRTF (*Head Related Transfer Functions*). Dans le rendu par casque, la convolution du signal monophonique à spatialiser par une paire de HRTF donnée provoque l'impression d'un son venant de la direction d'incidence correspondante. Cependant, chaque individu dispose d'un jeu de HRTF qui lui sont propres et la conception de modèles minimisant les différences interindividuelles constitue une problématique actuelle de recherche, de même que le calcul automatique des HRTF à partir des caractéristiques morphologiques de la tête. La cohérence de perception de la localisation des sources peut être renforcée lorsque le système de restitution par casque est couplé, comme dans le cas du projet européen LISTEN [10] à un système de suivi qui permet de compenser en temps réel la variation de la position de la tête pour synthétiser des sources dont la localisation reste stable, même si l'auditeur a « revêtu » des HRTF qui ne sont pas les siennes.

Un autre système de restitution spatiale prometteur est l'holophonie (analogie acoustique de l'holographie), en particulier selon la méthode de Wavefield synthesis (WFS). Celle-ci, initiée par l'Université de Delft, se fonde sur le principe de Huygens pour simuler le rayonnement acoustique de sources par l'intermédiaire de transducteurs, jouant le rôle de sources secondaires, entourant l'auditoire, en particulier sous la forme de panneaux multi-actuateurs [3,9]. L'intérêt de cette méthode, qui fait l'objet de recherches actuelles et est vouée à supplanter les systèmes à haut-parleurs pour certaines applications, réside notamment dans le rendu homogène qu'elle produit dans tout l'espace de restitution et dans les possibilités nouvelles de simulation qu'elle ouvre, notamment de sources virtuelles à l'infini (ondes planes) voire même situées dans l'espace de diffusion.

Une approche en quelque sorte duale consiste à réaliser des sources multi-haut-parleurs à rayonnement contrôlé par filtrage des signaux envoyés à chaque haut-parleur. Il est notamment possible d'approximer une figure de directivité donnée (celle d'un instrument particulier par exemple) avec une configuration donnée de haut-parleurs. Un tel système a fait l'objet d'expérimentations dans le contexte de la création musicale, mettant en scène ce dispositif excitant la salle de concert de manières différentes selon les signaux envoyés et le contexte musical : diffusion polyphonique de signaux selon des directivités différentes, variation de la directivité au cours du déroulement du son [40]. Les principales applications existantes de ces différentes techniques de spatialisation sonore se trouvent dans des contextes expérimentaux ou spécialisés (musique contemporaine, simulation, télécommunications et autres applications de réalité virtuelle), à l'exception notable de l'industrie du jeu, qui met déjà en œuvre ces procédés sous différentes formes. Leur exploitation à une plus large échelle est dépendante de deux facteurs principaux : l'évolution des supports de distribution des enregistrements sonores, et la complexité de mise en œuvre de dispositifs de reproduction avancés. La généralisation récente de systèmes grand public de Home Cinema, exploitant le format DTS 5.1 présent dans les DVD, a déjà constitué une première révolution dans le monde de la production sonore, qui en était resté au format stéréophonique depuis plusieurs décennies. La distribution électronique des contenus musicaux, objet du paragraphe suivant, ouvre des possibilités beaucoup plus larges en matière de spatialisation sonore, en s'affranchissant du goulot d'étranglement lié au support matériel, et en autorisant notamment la diffusion des informations spatiales sous différentes formes (signaux multiples liés aux différentes sources, codages de l'information spatiale liée à chaque source, caractéristiques géométriques ou perceptives de l'espace virtuel dans lequel elles sont placées).

5. Indexation audio et technologies d'accès

La généralisation conjuguée de l'Internet et de codecs de compression de signaux audio provoque un bouleversement du système de diffusion de musiques enregistrées. Les problèmes de protection des contenus conjugués aux nouvelles possibilités ouvertes par la diffusion des enregistrements par les réseaux (accès potentiel à des millions de morceaux, systèmes de recommandation personnalisée) ont suscité une forte demande qui a contribué à la structuration, depuis 2000, d'une importante communauté pluridisciplinaire dans le domaine de la recherche d'informations musicales [11]. Au-delà de technologies liées à la protection des droits, largement débattues et qui ne seront pas abordées ici, la problématique centrale liée à ce domaine est celle de la caractérisation des contenus musicaux, à travers la mise en œuvre de métadonnées de description, selon la terminologie de la norme MPEG-7 [12], adaptées à chaque classe d'applications. Après une introduction à cette problématique en lien

10. http://listen.imk.fraunhofer.de.

11. MIR ou Music information retrieval. Voir : www.ismir.net.

12. www.chiariglione.org/MPEG/standart/mpeg-7.htm.

à celle de l'extraction automatisée des informations, les avancées récentes dans ce domaine seront illustrées à partir d'exemples issus des projets européens CUIDADO [13] et SemanticHIFI [14] qui se consacrent respectivement à concevoir des bases de données audio reposant sur l'utilisation systématique de descriptions automatiquement extraites des signaux [38], et à préfigurer les chaînes Hi-fi de demain fondées sur des fonctions de manipulation par le contenu des informations musicales [39].

5.1. Description et extraction des contenus musicaux

Il n'existe pas de manière unique de décrire les contenus musicaux, chaque application particulière nécessitant la mise en œuvre de structures de données adaptées. Une grille d'analyse utile des différents types de descriptions consiste à différencier d'une part celles qui sont d'ordre subjectif (et donc dépendantes des contextes d'utilisation) et celles d'ordre objectif, ces dernières se divisant en celles qui peuvent être automatiquement calculées à partir des informations musicales (signaux, symboliques) et celles qui nécessitent une intervention humaine. Ainsi, les informations éditoriales (artistes, titre) sont de type objectif-manuel, et les codages d' « empreintes digitales » (identifiants compacts caractérisant un enregistrement), la durée ou l'énergie moyenne du signal sont de type automatique. La constitution de descripteurs subjectifs adaptés, qui relèvent du niveau cognitif dans la terminologie du §2.2, nécessite la mise en œuvre de méthodes d'ingénierie des connaissances musicales (processus *top-down*) et leur obtention automatisée passe par leur mise en relation, généralement par apprentissage, avec des paramètres de bas niveau automatiquement extraits des signaux (processus *bottom-up*). Ainsi, les réalisations opérationnelles dans ce domaine se fondent généralement sur des compromis entre structuration des connaissances relevant de la cognition musicale et état de l'art en matière d'analyse automatisée. Il reste encore beaucoup à faire en analyse de signal pour en arriver à des descriptions qui restent relativement triviales du point de vue cognitif, si l'on considère par exemple l'état de l'art en analyse de fréquence fondamentale, qui progresse encore dans le cas de sons monodiques [10] et produit de premiers résultats probants dans des cas polyphoniques [41].

5.2. Navigation inter-documents

Ce type de fonctions, qui vise la *navigation par le contenu musical* dans les bases de données d'enregistrements, fait appel à la constitution de descripteurs caractérisant globalement le contenu d'un morceau donné selon différents critères. Divers descripteurs pertinents peuvent être extraits des signaux : tempo, intensité, tonalité, « timbre » orchestral, présence de voix, etc. Complétant les métadonnées éditoriales usuelles (titre, artistes), ces descripteurs permettent une recherche de morceau par le contenu musical (morceaux de tempo lent de tel compositeur par exemple). L'utilisation de catégories de type subjectif, telles que les taxonomies de genres musicaux, sont également utiles en tant que mode de classification et, de manière indirecte, de personnalisation de l'offre en fonction des goûts de l'utilisateur. Des expérimentations récentes visent également à apprendre les caractéristiques sonores de catégories définies par les utilisateurs eux-mêmes (« rock », « calme », « ambiance »,…) à partir d'ensembles de morceaux d'exemples [15,42]. Une autre heuristique de navigation, transversale à cette notion de catégories, est celle de recherche par l'exemple, sur la base de mesures de similarités entre sons. À partir d'un son de départ, le système calcule une mesure de similarité avec tous les sons de la base et rend une liste d'items classés par similarité décroissante. De telles mesures peuvent combiner des critères très variés à partir de descripteurs extraits des signaux, selon des poids relatifs préétablis ou configurables par l'utilisateur. Leur intérêt réside notamment dans le caractère souvent imprévu des appariements obtenus, tout en restant cohérent du point de vue des critères retenus pour la similarité.

5.3. Navigation intra-documents

Ce type de fonction, expérimenté dans le cadre du projet SemanticHIFI, vise à fournir des interfaces de représentation et de navigation à l'intérieur du contenu musical d'un morceau, qui dépasse les interfaces traditionnelles des chaînes hi-fi (lecture, arrêt, volume, balance, etc.). À la différence de celles utilisées pour la navigation inter-documents, les descriptions musicales adaptées à cet effet s'attachent à décrire les structures musicales intrinsèques au morceau. Les approches existantes distinguent d'une part l'analyse de la structure temporelle des morceaux, soit en alignant les enregistrements audio à des représentations symboliques de type MIDI [25] et en autorisant ainsi la navigation dans le fichier son à partir d'une interface proche de la partition, soit en analysant directement une structure temporelle de haut niveau comme succession et appariement d'état stables du point de vue de critères d'évolution spectrale (introduction, refrain, couplets) [21]. D'autre part, des expérimentations sont menées sur la navigation à l'intérieur de la polyphonie d'un morceau à partir d'interfaces permettant de définir les positions relatives des différentes sources instrumentales ou polyphoniques et de l'auditeur et de produire le rendu spatial correspondant à la situation où l'auditeur se trouverait au milieu de l'orchestre et se rapprocherait de tel ou tel instrument [11]. Une telle fonction, mettant en œuvre les techniques de spatialisation présentées au §4, impose de disposer de méthodes fournissant les différents canaux polyphoniques, soit à partir des formats de diffusion stéréophoniques ou 5.1 par séparation

13. Content-based Unified Interfaces and Descriptors for digital Audio Databases available Online.

14. shf.ircam.fr.

automatique des sources [36], ou directement à partir des enregistrements sous forme multipiste. Ce dernier cas montre l'intérêt d'une extension des formats actuels de distribution musicale à des contenus plus riches (métadonnées, canaux multiples), dont la réalisation technique serait d'ores et déjà envisageable à travers l'ensemble de la chaîne numérique de production, de diffusion et d'accès des enregistrements. Un autre enjeu, d'ordre artistique, de ces interfaces de navigation inter-documents est qu'elles ouvrent un nouvel espace à la production musicale, en permettant la diffusion d'œuvres interactives. Ces potentialités, déjà viables techniquement, n'attendent, pour s'actualiser, que l'émergence d'usages et de marchés nouveaux.

6. Conclusion

Ce tour d'horizon des applications musicales du traitement de signal et des problématiques de recherche qui leur sont liées met en évidence les tendances globales suivantes :

- la généralisation des techniques de traitement de signal, précédemment concentrées sur les outils de production et méthodes de codage, à l'ensemble de la chaîne production/ diffusion/ interfaces d'accès/ reproduction, ces deux derniers maillons étant actuellement l'objet d'importants développements ;
- l'évolution des fonctions de traitement des signaux audionumériques vers des techniques de manipulation par le contenu, en lien avec d'autres types de représentations musicales (niveaux physique, symbolique, cognitif). Les conversions entre ces différents niveaux se fondent sur des méthodes d'analyse et de synthèse des signaux, à travers notamment l'utilisation de techniques d'apprentissage et d'analyse statistique ;
- une fois les modèles de signaux audionumériques stabilisés, les perspectives ouvertes par la modélisation du contrôle, en tant que signaux à variation lente, résultant notamment d'une analyse de l'interprétation, soit à partir des enregistrements ou de données issues de dispositifs de captation gestuelle ;
- le potentiel d'évolution des modes de production, de diffusion et d'accès de la musique vers des formats étendus (représentations symboliques, métadonnées, enregistrements multipistes) et des fonctions de manipulation configurables et personnalisables (programmation, lutherie virtuelle, interfaces d'accès) ;
- corrélativement au point précédent, l'effacement potentiel des frontières entre les différents dispositifs techniques musicaux traditionnels (instruments de musique, logiciels de production sonore, techniques de lutherie, chaînes hi-fi et autres dispositifs d'écoute) et l'émergence de fonctions musicales inédites, susceptibles de renouveler radicalement les pratiques et usages à travers de nouveaux modes d'interaction avec les artefacts musicaux.

Références

[1] ADRIEN J.M., *The Missing Link: Modal Synthesis* in Representations of Musical Signals, ss. la dir. de G. De Poli, A. Picalli et C. Roads, MIT Press, 1991.

[2] BENSOAM J., Représentation intégrale appliquée à la synthèse sonore par modélisation physique, Thèse de doctorat, Université du Maine (Le Mans, Académie de Nantes), 2003.

[3] BERKHOUT A. J., *A Holographic Approach to Acoustic Control*. Journal of the Audio Engineering Society, Vol. 36, N°12, 1988.

[4] BLAUERT J., Spatial Earing, The psychophysics of Human Sound Localization, MIT Press, 1997.

[5] BOGAARDS N., RÖBEL A., RODET X., *Sound Analysis and Processing with AudioSculpt 2*. Proc. Int. Computer Music Conf. (ICMC'04), 2004.

[6] CADOZ C., LUCIANI A., FLORENS J.-L., *CORDIS-ANIMA: A Modeling and Simulation System for Sound and Image Synthesis- The General Formalism*, Computer Music Journal, MIT Press, Vol. 17, N°1 Spring 1993.

[7] CADOZ, C., *Continuum énergétique du geste au son, simulation multisensorielle interactive d'objets physiques*, in Interfaces homme-machine et création musicale, ss. la dir. de H. Vinet et F. Delalande, Hermes Science, Paris, 2002.

[8] CHOWNING J., *The synthesis of complex audio spectra by means of frequency modulation*. Journal of the Audio Engineering Society, 21:526-534, 1973.

[9] CORTEEL E., Adaptations de la Wave Field Synthesis aux conditions réelles. Thèse de doctorat, Université Paris 6, 2004.

[10] DE CHEVEIGNÉ A., KAWAHARA H., *YIN, a fundamental frequency estimator for speech and music*. Journal of the Acoustical Society of America, Vol. 111, 2002.

[11] DELERUE O., Spatialisation du son et programmation par contraintes : le système MusicSpace. Thèse de doctorat, Université Paris 6, 2004.

[12] DOLSON M., LAROCHE J., *Improved phase vocoder time-scale modification of audio,* IEEE Transactions on Speech and Audio Processing, Vol. 7, N°3, 1999.

[13] HÉLIE T., Modélisation physique d'instruments de musique en systèmes dynamiques et inversion. Thèse de doctorat, Université Paris 11, 2002.

[14] HÉLIE T., MATIGNON D., *Diffusive Representations for Analyzing and Simulating Flared Acoustic Pipes with Visco-thermal Losses*, Mathematical Models and Methods in Applied Sciences, 16 (2006), pp. 503-536.

[15] HERRERA P., PEETERS G., DUBNOV S., *Automatic Classification of Musical Sounds*. Journal of New Musical Research, 2003.

[16] IOVINO F., CAUSSÉ, R., DUDAS R., *Recent work around Modalys and Modal Synthesis*. Proc. Int. Computer Music Conf. (ICMC'97), 1997.

[17] JOT J.M., *Efficient Models for Distance and Reverberation Rendering in Computer Music and Virtual Audio Reality*. Proc. Int. Computer Music Conf. (ICMC'97), 1997.

[18] JULLIEN J.-P. 1995., *Structured Model for the Representation and the Control of Room Acoustical Quality*. Proc. Int. Congress on Acoustics (ICA'95), 1995.

[19] MALHAM D.G., MYATT A., *3-D Sound Spatialization using Ambisonic Techniques*. Computer Music Journal, Vol. 19 N°4, MIT Press, 1995.

[20] PEETERS G., RODET X., *Non-stationary Analysis/Synthesis using Spectrum Peak Shape Distortion, Phase and Reassignment*, Proc. Int. Conf. on Signal Processing Applications and Technology (ICSPAT'99), 1999.

[21] PEETERS G., *Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation: « Sequence » and « State » approach*, Lecture Notes in Computer Science, Vol. 2771, Springer Verlag, 2003.

[22] RASAMIMANANA N., FLÉTY E., BEVILACQUA F., Gesture Analysis of Violin Bow Strokes, GW 2005, Lecture Notes in Artificial Intelligence Vol. 3881, ss. la dir. de S. Gibet, N. Courty, and J.-F. Kamp, Springer Verlag, 2006.

[23] RODET X., *Time-domain formant-wave-function synthesis*. Computer Music Journal, Vol. 8, N°3, MIT Press, 1984.

[24] RODET X., *Sinusoidal + Residual Models for Musical Sound Signals Analysis/Synthesis*. Applied Signal Processing, Vol. 4, N°3, 1998.

[25] RODET X., ESCRIBE J., DURIGON S., *Improving score to audio alignment: Percussion alignment and Precise Onset Estimation*, Proc. Int. Computer Music Conf. (ICMC'04), 2004.

[26] RÖBEL A., *A new approach to transient processing in the phase vocoder*. Proc. Int. Conf. on Digital Audio Effects (DAFx'03), 2003.

[27] RÖBEL A., *Adaptive additive modeling with continuous parameter trajectories*. IEEE Transactions on Audio, Speech and Signal Processing (à paraître).

[28] SCHNELL N., SCHWARZ D., *Gabor, Multi-Representation Real-Time Analysis/Synthesis*. Proc. Int. Conf. on Digital Audio Effects (DAFx'05), 2005.

[29] SCHNELL N., BORGHESI R., SCHWARZ D., BEVILACQUA F., MÜLLER R., *FTM-Complex Data Structures for Max*, Proc. Int. Computer Music Conf. (ICMC'05), 2005.

[30] SCHWARZ D., Data-Driven Concatenative Sound Synthesis. Thèse de doctorat, Université Paris 6, 2004.

[31] SCHWARZ D., *Current Research in Concatenative Sound Synthesis*, Proc. Int. Computer Music Conf. (ICMC'05), 2005.

[32] SCHWARZ D., WRIGHT M., *Extensions and Applications of the SDIF Sound Description Interchange Format*. Proc. Int. Computer Music Conf. (ICMC'00), 2000.

[33] SIMON I., BASU S., SALESIN D, AGRAWALA M., *Audio Analogies : Creating New Music from an Existing Performance by Concatenative Synthesis*, Proc. Int. Computer Music Conf. (ICMC'05), 2005.

[34] SMITH J., *Physical Modeling using Digital Waveguides*, Computer Music Journal, Vol. 16, N°4, MIT Press, 1992.

[35] VAN DUYNE S.A., SMITH J.O., *Physical modeling with the 2-D digital waveguide mesh*, Proc. Int. Computer Music Conf. (ICMC'93), 1993.

[36] VINCENT E., Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux. Thèse de doctorat, Université Paris 6, 2004.

[37] VINET H., *The Representation Levels of Music Information*. Lecture Notes in Computer Science, Vol. 2771, Springer Verlag, 2003.

[38] VINET H., HERRERA P., PACHET F., *The CUIDADO Project*. Proc. Int. Conf. on Music Information Retrieval (ISMIR'02), Ircam, Paris, 2002.

[39] VINET H., *The SemanticHIFI Project : Content-based Manipulation of Digital Audio Recordings*, Proc. European Workshop on the Integration of knowledge, semantic, and digital Media Technology (EWIMT'05), 2005.

[40] WARUSFEL O., MISDARIIS N., *Directivity synthesis with 3D array of loudspeakers : application for stage performance.*, Proc. Int. Conf. on Digital Audio Effects (DAFx'01), 2001.

[41] YEH C., RÖBEL A., RODET X., *Multiple Fundamental Frequency Estimation of Polyphonic Music Signals*, IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 225-228 (Vol. III), 2005.

[42] ZILS A., PACHET F., *Automatic Extraction of Music Descriptors from Acoustic Signals using EDS*. Proc. 116th AES Convention, 2004.

Hugues **Vinet**

Hugues Vinet est, depuis 1994, directeur scientifique de l'Ircam, dont il dirige le département Recherche et développement et l'UMR STMS (Sciences et Technologies de la Musique et du Son). Après des études scientifiques et musicales, il a précédemment travaillé comme ingénieur en chef à l'Institut national de l'audiovisuel (INA) en tant que responsable des recherches du Groupe de recherches musicales (GRM). Il a entre autres assuré la coordination des projets européens CUIDADO et SemanticHIFI.