HAL open science

Sample Orchestrator : gestion par le contenu d'échantillons sonores

Hugues Vinet, Gérard Assayag, Juan José Burred, Grégoire Carpentier, Nicolas Misdariis, Geoffroy Peeters, Axel Roebel, Norbert Schnell, Diemo Schwarz, Damien Tardieu

To cite this version:

Hugues Vinet, Gérard Assayag, Juan José Burred, Grégoire Carpentier, Nicolas Misdariis, et al.. Sample Orchestrator : gestion par le contenu d'échantillons sonores. Traitement du Signal, 2011, 2011 (3), pp.417-468. hal-01250794

HAL Id: hal-01250794

https://hal.science/hal-01250794v1

Submitted on 8 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sample Orchestrator : gestion par le contenu d'échantillons sonores

Hugues Vinet[1], Gérard Assayag[1], Juan José Burred[1], Grégoire Carpentier[1], Nicolas Misdariis[1], Geoffroy Peeters[1], Axel Roebel[1], Norbert Schnell[1], Diemo Schwarz[1], Damien Tardieu[1]

1. STMS IRCAM-CNRS-UPMC,
1 place Igor Stravinsky, F-75004 Paris
prenom.nom@ircam.fr

RÉSUMÉ. *Nous présentons les principaux travaux menés dans le projet Sample Orchestrator, destiné au développement de fonctions innovantes de manipulation d'échantillons sonores. Celles-ci se fondent sur des études consacrées à la description des sons, c'est-à-dire à la formalisation de structures de données pertinentes pour caractériser le contenu et l'organisation des sons. Ces travaux ont été appliqués à l'indexation automatique des sons, ainsi qu'à la réalisation d'applications inédites pour la création musicale - synthèse sonore interactive par corpus et aide informatisée à l'orchestration. Le projet a aussi comporté un important volet consacré au traitement de haute qualité des sons, à travers plusieurs perfectionnements du modèle de vocodeur de phase – traitement par modèle sinusoïdal dans le domaine spectral et calcul automatique des paramètres d'analyse.*

ABSTRACT. *The main advances of the R&D Sample Orchestrator project are presented, aiming at the development of innovative functions for the manipulation of sound samples. These features rely on studies on sound description, i.e. the formalization of relevant data structures for characterizing the sounds' content and organization. This work was applied to automatic sound indexing and to the development of new applications for musical creation - interactive corpus-based synthesis and computer-aided orchestration. The project also included an important part on high-quality sound processing, through several enhancements of the phase vocoder model – processing by sinusoidal model in the spectral domain and automatic computation of the analysis parameters.*

MOTS-CLÉS : *musique, son, échantillons sonores, synthèse et traitement sonore, indexation sonore, recherche d'informations musicales, cognition auditive, bases de données multimédia, apprentissage automatique, modèles de signaux, vocodeur de phase, transformée de Fourier à court terme, modèle sinusoïdal, synthèse concaténative, synthèse par corpus, composition assistée par ordinateur, orchestration, algorithmes génétiques, optimisation multicritères, interfaces homme-machine.*

KEYWORDS: *music, sound, sound samples, sound synthesis and processing, sound indexing, music information retrieval, auditory cognition, multimedia databases, machine learning, signal models, phase vocoder, short-term Fourier transform, sinusoidal model, concatenative synthesis, corpus-based synthesis, computer-aided composition, orchestration, genetic algorithms, multicriteria optimization, human-computer interaction.*

DOI:10.3166/TS.27.p1-n © 2010 Lavoisier, Paris

RÉSUMÉ ÉTENDU EN ANGLAIS

The Sample Orchestrator R&D project was aimed at the development of innovative functions for the content-based manipulation of sound samples. Its main objectives and advances are summarized hereinafter.

Sound description and automatic indexing

This objective concerns the development of new heuristics for managing sample databases. The sound description issue involves a top-down process for elaborating representations of the sound world relevant to the human cognition. Recent studies define three levels for sound similarity: semantic, causal and acoustic. Causal relates to the sound production in terms of sources and actions. For each homogeneous sound set, a specific timbre space, with dimensions directly correlated to acoustical attributes, can be obtained. A first study consisted in performing a meta-analysis of timbre spaces resulting from several studies related to different sound corpora. A second study was dedicated to an experience of categorization of a given corpus of environmental sounds targeted to the materials of the sources. The results show a good correlation between the experimental data and the actual categories. Three labeled corpora were elaborated from these experiences: Materials, Reduced materials and Onomatopoeia.

The second stage, automatic indexing, aimed at building a signal model of the classes in order to perform the automatic classification of new sounds from the analysis of their signals. First, an extensive set of audio descriptors, combining several models (spectral, harmonic, perceptual) was elaborated and implemented as the *IrcamDescriptor* library. For the learning phase of classification, an algorithm provides a selection of descriptors for a given set of classes. After applying an LDA, various supervised classification methods were evaluated. The best results were obtained with an SVM, with accuracies and f-measures greater than 70% for the 3 formerly produced corpora.

A last study concerned the categorization of sound morphologies. 6 classes emerged from an experience on intensity profiles and very good results were obtained for the automatic classification with a binary decision tree operating on selected descriptors.

High-quality sound processing

This objective consisted in extending the phase vocoder model for high-quality sound processing. The model, based on a STFT, requires processing the phases in the spectral domain in order to preserve their consistency both through time scale modifications and across FFT bins. Two main enhancements produced as part of the project are presented. First, the integration of an explicit non stationary sinusoidal model; sinusoidal components are manipulated in the spectral domain using pre-computed spectral data related to individual sinusoids for sampled values of frequency slopes. The model supports transposition or remixing of the sinusoidal and residual components in the spectral domain. Compared to previous approaches,

it improves the quality obtained for signals containing non-stationary sinusoids. The second advance concerns the automatic adaptive computation of the analysis window size, under the assumption of monophonic signals; the signal is segmented from a measure of harmonicity; a time-varying window size (for harmonic and non-harmonic segments) is computed from the fundamental frequency, and all parameters, including local hop sizes between successive segments and the global FFT size are adapted accordingly. Qualitative enhancements were evidenced for various classes of processing, although requiring an increase in computation cost. All extensions of the phase vocoder model were implemented in the *SuperVP* environment and licensed to various sound and music software publishers.

Corpus-based sound synthesis

Corpus-based methods were developed in the context of voice synthesis and the concept was applied to the interactive re-synthesis of any kinds of pre-recorded sounds in the *CataRT* system, also extending the notion of granular synthesis through a content-driven control. The integration of a new corpus in the system consists in automatically segmenting the sound files into small duration grains characterized by a vector of audio descriptors. Data reduction methods such as MDS, combined with sub-linear complexity search algorithms, were used for the 2D visualization of all grains in a descriptor space and for the interactive generation of sound sequences from a user-controlled trajectory. The system can be used for fast search into a sound database, and also features the generation of rhythmic loops from a chosen corpus in similarity to a target sound file. It is already broadly used in the context of contemporary music and multimedia production.

Computer-aided orchestration

Orchestration is the art of combining instrument sounds. This dimension of composition had been up to now subject to few developments in computer-aided systems, due noticeably to the unformalized status of the empirical knowledge on instrument sounds and auditory perception it relies on. The developed approach has been to simplify the problem to the search of solutions of combinations of individual instrument notes that perceptually match a sound target provided by the user. On the signal side, a statistical model of instrument sound descriptors was elaborated, together with rules of computation of the descriptor values of mixtures from those of their components. An independent statistical modeling of instruments and playing modes enables the extrapolation of missing data. The issue of combinatorial search was handled through a multicriteria approach, well adapted for managing all combinations of descriptor weightings in the computation of the sound similarity distance from preferences inferred from the user input. A genetic algorithm was used and based on a generic syntax of operators related to basic musical operations, enabling the application of global constraints to the search. The system was implemented as *Orchidee* in a client-server architecture with clients in several musical software applications. It has already been used by well-known composers and contributed to the creation of major contemporary music works.

1. Introduction

L'échantillonnage est l'une des méthodes de synthèse sonore les plus couramment utilisées dans les synthétiseurs musicaux. Elle consiste, pour simuler un instrument donné, à enregistrer, sous forme de signaux audionumériques, chaque note de manière isolée en fonction de toutes les valeurs pertinentes des paramètres de jeu (hauteur, nuance, mode de jeu). La synthèse restitue les parties d'attaque et de relâchement des notes et produit la partie entretenue (note tenue) au moyen de techniques telles que la lecture en boucle d'une section temporelle stable du signal enregistré. Après la généralisation de ce procédé pour les instruments électroniques, une seconde génération d'échantillonneurs logiciels a vu le jour, sous la forme d'applications de synthèse sonore pour ordinateur, avec la commercialisation de banques de sons relatives à divers instruments et effectifs instrumentaux. Au-delà des applications musicales, des collections d'échantillons sont produites dans différents domaines d'activités, notamment pour le bruitage en production audiovisuelle : les sons enregistrés, dits « environnementaux », concernent tous types de sources et situations (machines, animaux, sons ambiants, etc.).

La réalisation du projet Sample Orchestrator, soutenu par l'Agence Nationale de la Recherche dans le cadre de son programme « Audiovisuel et Multimédia 2006 », a été motivée par la conjonction de deux tendances : d'une part la généralisation des échantillonneurs logiciels combinée à la large diffusion commerciale de grandes banques d'échantillons; d'autre part les avancées récentes de la recherche en traitement de signal audio et ses applications au traitement sonore de haute qualité et à l'analyse automatisée de sons pour l'indexation et la recherche par contenu dans les grandes bases de données d'enregistrements.

Le projet associait la société Univers sons, l'un des principaux acteurs dans le monde pour l'édition de logiciels d'échantillonnage[1] et de vente de banques d'échantillons et quatre équipes de recherche de l'IRCAM, l'un des principaux centres de recherche dans le monde spécialisés dans les sciences et technologies de la musique et du son[2].

1.1. Objectifs du projet

Les objectifs du projet, dénommé dans la suite sous l'acronyme SOR, relèvent des applications du traitement de signal aux enjeux de production sonore et de création musicale, en articulant différents niveaux de représentations des informations musicales (Vinet, 2006 ; 2003). Ils s'attachent à la réalisation de fonctions inédites de manipulations d'échantillons sonores résumées dans les paragraphes suivants.

---

[1] http://www.ultimatesoundbank.com/uvi.html
[2] http://www.ircam.fr

Indexation automatique et recherche par le contenu

L'évolution des capacités de stockage disponibles, permettant l'accès simultané à un très grand nombre d'échantillons sonores sur un disque dur d'ordinateur[3], a nécessité une nouvelle approche dans la structuration et l'accès aux banques de sons, qui dépasse les modes d'organisation traditionnels fondés sur les paramètres de leur production (instrument, hauteur, nuance, mode de jeu). Leur référencement au moyen de techniques de bases de données permet de multiplier et combiner différents modes de classement, non plus limités aux caractéristiques de production, et proposer de nouvelles heuristiques de recherche et de navigation fondées sur leurs caractéristiques sonores intrinsèques. De plus, l'obtention de ces caractéristiques doit autant que possible être automatisée compte tenu du nombre de sons considérés. La méthode mise en œuvre pour atteindre ces objectifs consiste en l'articulation interdisciplinaire de trois champs de recherche : l'explicitation et la formalisation d'une structuration pertinente, du point de vue de la cognition sonore, de l'ensemble des sons manipulés (approche *top down*) ; la caractérisation de manière aussi exhaustive que possible des contenus sonores par un ensemble de descripteurs numériques, calculés par analyse des signaux des enregistrements (approche *bottom-up*) ; l'établissement de relations entre représentations de haut et de bas niveau issues des deux étapes précédentes, par méthodes d'analyse de données et apprentissage automatique.

Traitements sonores de haute qualité

Au-delà de la reproduction à l'identique des sons enregistrés, les fonctions des échantillonneurs ont été étendues à divers types de traitement sonores. Parmi ceux-ci, la transposition (changement de hauteur sans modification de durée) et la compression/expansion temporelle (changement de durée sans altération de la hauteur, ou *time stretching*), qui compensent de manière symétrique la variation conjointe des deux paramètres obtenue par changement du taux d'échantillonnage, sont d'usage fréquent. Le vocodeur de phase, dont le formalisme est présenté plus loin, est dans ses développements récents, l'un des modèles de signaux les mieux adaptés au traitement de larges classes de sons notamment par transposition et compression/expansion temporelle, ce avec une qualité sonore élevée, c'est-à-dire sans artefact perceptible pour des amplitudes de variation courantes des paramètres d'entrée. L'un des objectifs du projet était le perfectionnement du modèle et son application pour intégration dans une nouvelle génération d'échantillonneur logiciel.

Nouvelles applications de l'indexation d'échantillons

Le projet avait également pour ambition de susciter le développement de nouvelles applications pour la production sonore et la création musicale, proposant

[3] Un téraoctet contient environ 500.000 échantillons de 10 secondes enregistrés en stéréophonie échantillonnés sur 16 bits à 50 kHz

des modes inédits de manipulation des contenus sonores rendus possibles par l'avancée des recherches sur l'indexation automatique des échantillons.

1.2. Objet et structure du document

L'article présente une partie des recherches et développements menés dans le cadre du projet SOR. Sa structure suit celle des objectifs du projet introduits ci-dessus. Les parties 2 et 3 décrivent les recherches consacrées à la description et l'indexation d'échantillons de sons environnementaux et, pour le traitement sonore de haute qualité, à certains des perfectionnements opérés sur le modèle de vocodeur de phase : traitement par modèle sinusoïdal dans le domaine spectral, adaptation automatique des paramètres d'analyse. La partie 4 présente deux applications inédites pour la création musicale issues du projet : synthèse sonore par corpus et aide informatisée à l'orchestration.

2. Description et indexation d'échantillons sonores

L'un des objectifs du projet SOR est de permettre la description et l'indexation automatique des échantillons sonores, dans le but de fournir à chacun d'eux une ou plusieurs étiquettes permettant une recherche automatique au moyen d'un moteur de recherche.

Les premières recherches sur l'indexation automatique des échantillons sonores ont été initiées par la société californienne Muscle Fish (Wold et Blum, 1999) et par l'IRCAM et ses partenaires dans le cadre des projets Studio en ligne (Vinet *et al.*, 1999 ; Wöhrmann et Ballet, 1999), CUIDADO (Vinet *et al.*, 2002) et ECRINS (Mullon *et al.*, 2002 ; Peeters, 2003). Ces travaux s'étant particulièrement concentrés sur les échantillons instrumentaux, les objectifs spécifiques du projet SOR ont été consacrés, à des fins de généralisation, à l'indexation automatisée de sons environnementaux, qui présentent des caractéristiques sonores plus larges.

Dans cette section, nous aborderons donc successivement les problématiques de description et classification des sons environnementaux – qualifiés par nature de non musicaux – selon deux points de vue principaux : d'une part, l'expérimentation, afin de déterminer les structures perceptivement pertinentes qui sous-tendent la perception de cette nature de sons, et d'autre part, la modélisation, afin de proposer des outils de prédiction automatique adaptés à ces nouvelles représentations. Dans une seconde étape, la dimension temporelle, intrinsèque à tout phénomène sonore, sera plus précisément prise en compte en étudiant ces même questions du point de vue des profils morphologiques, c'est-à-dire de l'évolution dans le temps des propriétés des sons de l'environnement.

2.1. Description des échantillons sonores

2.1.1. Etat de l'art

Ce travail s'appuie en premier lieu sur les études de référence dans le domaine du timbre musical qui ont balisé une méthodologie expérimentale, d'analyse et d'interprétations des données aboutissant à la notion d'espace de timbres (Grey, 1977 ; Krumhansl, 1989 ; Krimphoff et McAdams, 1994 ; McAdams *et al.*, 1995). Cette approche fait l'hypothèse que, sous réserve d'homogénéité du corpus étudié, une structure continue sous-tend la perception de la similarité entre les sons aboutissant à la définition d'un espace dit « perceptif » grâce à des techniques d'analyses multi dimensionnelles (Winsberg et De Soete, 1993). On cherche ensuite à corréler chaque dimension de cet espace avec des indicateurs numériques – appelés descripteurs acoustiques – calculés à partir du signal sonore en intégrant, le cas échéant, des modèles perceptifs qui prennent en compte le fonctionnement de l'appareil auditif (on parle alors plus spécifiquement de descripteurs « psychoacoustiques ») ; le résultat constitue un espace acoustique associé à l'espace perceptif et caractérise au mieux le corpus considéré. Cette approche a depuis été transposée à plusieurs exemples de sons du quotidien, notamment dans le cadre d'habitacles de voiture (McAdams *et al.*, 1998), d'unités de climatisation (Susini *et al.*, 2004), de klaxons (Lemaitre *et al.*, 2007) ou de portières automobiles (Parizet *et al.*, 2006). Chacune de ces études a élaboré des espaces perceptifs et acoustiques relativement à son corpus ; ces données constituent le point de départ d'une partie du travail sur la description des sons de l'environnement effectué dans le projet.

Cela étant, lorsque l'hypothèse d'homogénéité du corpus sonore n'est plus valide, les mécanismes perceptifs mis en jeu relèvent davantage de processus catégoriels – donc en rapport avec une structure discrète –, comme l'a montré par exemple (Susini *et al.*, 2005) à partir de sons produits par différents moyens de transport. Dans ce cas, plusieurs niveaux de similarité peuvent rentrer en jeu comme l'a notamment montré (Gaver, 1993) ou, plus récemment, les travaux sur la classification des sons du quotidien menés dans le projet CLOSED[4] (Lemaitre *et al.*, 2010 ; Houix *et al.*, 2007). Ces études font apparaitre trois principaux niveaux de similarité : la similarité acoustique se basant sur des attributs perceptifs directement liés aux propriétés acoustiques, la similarité causale (ou évènementielle) se basant sur la cause physique du son (action ayant produit le son) et la similarité sémantique se basant sur une abstraction des propriétés associées à l'appartenance à une catégorie.

2.1.2. Problématique

Pour la description des échantillons sonores traités dans le projet SOR, la problématique s'articule donc autour de deux axes principaux :

---

[4] http://www.closed.ircam.fr

- d'un point de vue opérationnel, instruire et détailler les processus de catégorisation basés sur la similarité causale en développant une base de données sonores annotée selon l'identification des causes physiques du son (objet/source ou action) et validée expérimentalement, dans le but de fournir des données exploitables pour tester les méthodes de classification automatique.

- d'un point de vue exploratoire, étudier les relations entre les différents niveaux de similarité par une méta-analyse du sous-ensemble de sons issu de l'union des corpus de quatre études de référence mentionnées dans l'état de l'art (MacAdams *et al.*, 1998 ; Susini *et al.*, 2004 ; Lemaitre *et al.*, 2007 ; Parizet *et al.*, 2006) ; et en partant de l'hypothèse d'une organisation à deux niveaux : un niveau catégoriel représenté par des grandes classes d'évènements sonores correspondant à des types de source et des causes physique distincts ; un niveau continu où chacune des classes est décrite par un espace perceptif/acoustique possédant des dimensions propres mais aussi des dimensions partagées avec les autres classes. Ce deuxième axe a déjà fait l'objet d'une publication (Misdariis *et al.*, 2010) et ne sera donc pas traité dans la suite du document.

2.1.3. Expériences et résultats

Expériences.

La partie expérimentale consiste en une tâche d'identification de la cause physique du son, c'est-à-dire du type d'action physique qui l'a produit (indépendamment de l'objet, sa forme ou le matériau qui le constitue), selon deux niveaux de description (voir Figure 1) :

- Liquide, solide, gaz (causes mécaniques), électronique, électrique, machine (causes artificielles) ;

- Liquide / bulle, chute, écoulement, goutte, remous ; Solide / roulement, bris, choc, cliquetis, fermeture, tintement, frottement, grincement, écrasement, froissement ; Gaz / crépitement, explosion, sifflement, souffle.

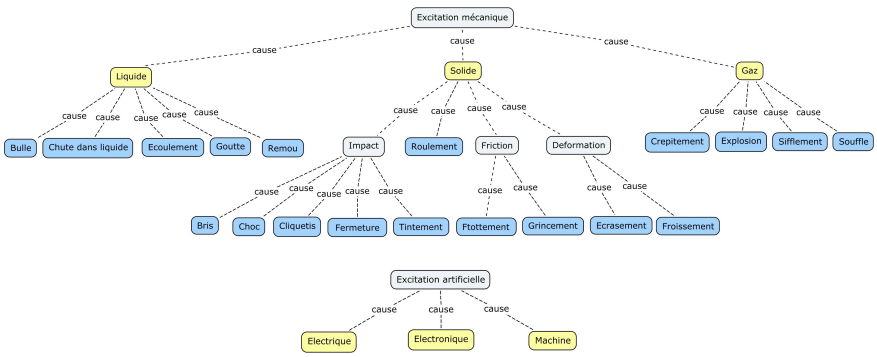

Figure 1. Corpus sonore représentatif - Taxonomie du niveau Evénementiel

Cette classification – donnée à priori aux participants à l'expérience – est notamment issue du travail de taxonomie effectué dans la cadre du projet CLOSED. Le corpus de l'expérience est constitué d'un ensemble de 213 sons extraits d'une base initiale constituée de manière arbitraire selon un critère de similarité causale, notamment à partir des données et notes d'enregistrements des sources (banques sonores commerciales : Sound Ideas[5], Hollywood Edge[6], Soundscan[7]). En outre, ce sous-ensemble respecte les distributions initiales selon le premier niveau de description et comprend le nombre suivant d'items par catégorie : 33– Liquide, 92– Solide, 49– Gaz, 10– Electrique, 12– Electronique et 17– Machine. Le test d'écoute est réalisé sur 20 personnes, selon un paradigme de catégorisation orientée au moyen d'une interface graphique (Smith, 1995) donnant un accès séquentiel aux deux niveaux de la taxonomie.

Résultats.

Les données de l'expérience ont été analysées en terme de matrice de confusion qui représente le taux d'identification de chaque classe en calculant, en moyenne sur les participants, le nombre de sons correctement assignés à la classe théorique à laquelle il appartient (classes arbitraires issues de la base initiale). Cette analyse est effectuée suivant les deux niveaux de la structure de classification (pour exemple, cf. Tableau 1 sur le premier niveau).

		Classes identifiées par les participants					
		Gaz	Liquide	Solide	Electrique	Electronique	Machine
Classes théoriques	Gaz (49)	67.9	0.8	15.8	1.5	2.8	11.3
	Liquide (33)	3.8	89.3	1.8	1.4	2.4	1.3
	Solide (92)	1.9	0.4	89.9	0.9	2.1	4.9
	Electrique (10)	1.1	0.0	0.0	61.6	25.3	12.1
	Electronique (12)	0.0	0.0	4.8	9.6	80.3	5.3
	Machine (17)	15.5	1.2	1.9	8.4	2.5	70.6

Tableau 1. Matrice de confusion (en %) – 1er niveau de classification

L'objectif fixé est donc atteint : générer un corpus de sons classifié et validé d'un point de vue perceptif afin de tester les procédures de classification automatique. Pour cela, une sélection des sons en fonction de la répartition des réponses est effectuée au premier niveau de description et aboutit – après un ré-équilibrage des classes nécessaire à la validation des modèles prédictifs – à une base de travail de 245 sons : 30– Liquide, 83– Solide, 38– Gaz, 35– Electrique, 31– Electronique, 28– Machine) et leurs sous-classes associées suivant la Figure 1.

[5] http://www.sound-ideas.com/
[6] http://www.hollywoodedge.com/
[7] http://www.univers-sons.com/sbsoundscan3.html

2.2. Indexation automatique des échantillons sonores

Le développement d'un module d'indexation automatique comporte deux grandes phases : l'apprentissage et l'estimation.

La première étape de l'apprentissage consiste à chercher la meilleure description du signal au vu des classes considérées. Cette description est obtenue en extrayant du signal audio des valeurs numériques mettant en évidence différentes propriétés des sons que nous appelons « descripteurs audio » (voir §2.2.1). Le choix des meilleurs descripteurs audio pour décrire un problème donné peut être fait par un sujet humain (lorsque le problème est facilement compréhensible, comme les profils dynamiques) ou déterminé de manière automatique par une machine (sélection automatique de descripteurs). La deuxième partie de l'apprentissage consiste à déterminer les valeurs types des descripteurs audio pour chacune des classes de la taxonomie considérée. Ceci constitue la phase d'apprentissage machine pour laquelle plusieurs types d'algorithmes sont utilisés.

La phase d'estimation consiste à appliquer le modèle de classes appris lors de l'apprentissage pour étiqueter un son inconnu. Pour cela, les mêmes descripteurs audio sont extraits du signal du son inconnu et les modèles de valeurs types de ces descripteurs sont utilisés pour assigner à ce son l'une des classes de la taxonomie considérée.

2.2.1. Les descripteurs audio : la bibliothèque IrcamDescriptor

Un descripteur audio (« audio feature », en anglais) désigne une caractéristique quantitative, sous forme numérique, extraite d'un signal audio. Chaque descripteur rend compte d'une caractéristique acoustique donnée (l'évolution de l'énergie, l'enveloppe spectrale, etc.) soit de manière isolée (le centre de gravité spectrale – CGS – est directement corrélé à la notion de brillance), soit combiné avec d'autres descripteurs (les Mel Frequency Cepstral Coefficients – MFCC – représentent ensemble l'enveloppe spectrale).

Ces dernières années, de nombreux descripteurs audio ont été proposés, issus généralement de communautés différentes (perception, parole, image, audio générique, musique) et pour des applications différentes (description, similarité, segmentation, classification) (Brown, 1999 ; Doval et Rodet, 1994 ; Krimphoff, 1994 ; Misdariis et Smith, 1998 ; Moore et Glasberg, 1997 ; MPEG7-02 ; Peeters et McAdams, 2000 ; Scheirer et Slaney, 1997 ; Serra et Bonada, 1998 ; Wold et Blum, 1999 ; Zwicker, 1990 ; Zwicker et Terhard, 1980). Ces descripteurs et leurs évolutions ont été implantés dans une bibliothèque logicielle nommée IrcamDescriptor (Peeters, 2004), qui calcule une centaine de descripteurs audio. Ils sont organisés selon leur mode d'extraction (voir Figure 2) : à partir de la forme d'onde (« Signal »), du spectre du signal à court terme (« FFT »), d'une représentation harmonique+bruit du signal à court terme (« Modèle sinusoïdal harmonique»), ou d'un modèle représentant le filtrage de l'oreille humaine (« Modèle perceptif »).

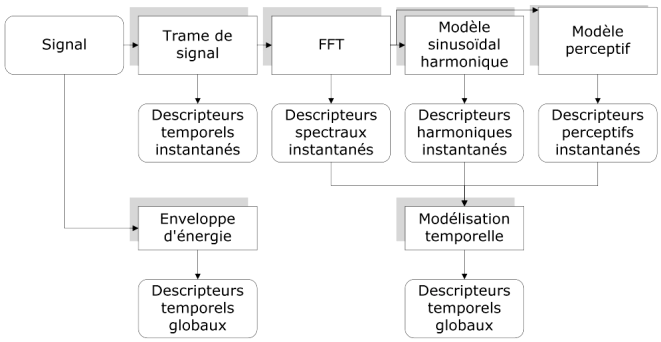

Figure 2. Organisation des descripteurs audio en fonction de leur mode d'extraction

Ces descripteurs audio sont extraits à partir d'une analyse du contenu local en temps du signal. Pour cela, l'analyse est effectuée sur une durée de 80 msec. avec un pas d'avancement de 20 msec. L'analyse est dite « à fenêtre glissante » et les descripteurs dits instantanés. Chaque son est donc représenté par une succession temporelle de descripteurs instantanés. Il est possible de résumer le comportement de ses descripteurs instantanés en utilisant des « modélisations temporelles ». On peut ainsi calculer la valeur moyenne, l'écart type ou la valeur minimale / maximale d'un descripteur spécifique sur une durée de 1 sec. en utilisant également une analyse « à fenêtre glissante ». Dans le cas extrême, la modélisation temporelle s'effectue en prenant en compte la globalité des valeurs du descripteur sur un signal ; on parle alors de descripteur « global ». D'autres types de descripteurs audio sont globaux par définition : ainsi la durée du signal, le temps d'attaque, le taux de croissance / décroissance de l'attaque n'ont de sens que calculés sur la totalité d'un signal. Les algorithmes de calcul des descripteurs de la bibliothèque IrcamDescriptor ont fait l'objet de plusieurs publications (Peeters, 2004) et ne seront donc pas davantage détaillés ici.

Au final, IrcamDescriptor calcule 54 concepts de descripteurs audio (soit 166 valeurs si l'on considère chacune des dimensions des descripteurs vectoriels). De plus, chacun des descripteurs instantanés peut être modélisé par différents types de modélisations temporelles. Du fait de ce grand nombre de possibilités, une normalisation des descripteurs audio a été proposée dans le cadre de la norme SDIF (Sound Description Interchange Format) (Burred *et al.*, 2008).

2.2.2. Classification automatique des échantillons sonores

L'objectif de l'apprentissage automatique est de créer un module prenant en entrée un fichier son et retournant la classe la plus vraisemblable du son au vue de son contenu. Pour chaque taxonomie, un modèle de classes est créé. Cette phase est appelée « apprentissage ». L'algorithme d'apprentissage analyse un grand nombre

d'exemples sonores appartenant à chacune des classes et cherche à trouver les meilleurs descripteurs audio à utiliser ainsi que les valeurs types de ces descripteurs propres à chaque classe.

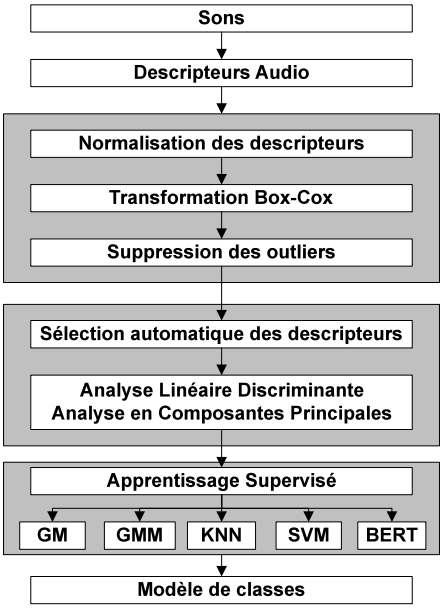

Figure 3. Séquence d'opérations relatives à la phase d'apprentissage

Le module d'apprentissage utilisé dans le projet SOR se base sur le classificateur (Peeters, 2003) issu du projet CUIDADO. L'enchaînement complet des opérations effectuées dans ce module est illustré par la Figure 3. Il est important de noter que chacune de ces opérations est optionnelle et comporte des paramètres. La mise au point d'un classificateur consiste précisément à trouver l'enchaînement idéal et les paramètres amenant aux meilleurs résultats de classification. La première étape de l'apprentissage consiste à extraire l'ensemble des descripteurs audio tels que décrits dans le paragraphe précédent pour l'ensemble des sons servant d'exemples aux classes à apprendre. Une série de pré-traitements de ces descripteurs est ensuite effectuée : normalisation (afin de ramener les plages de variation de chaque descripteur dans des domaines comparables), transformation box-cox (afin de maximiser la gaussiannité de la distribution de chaque descripteur), suppression des données *outliers* (afin de concentrer l'apprentissage sur les exemples types).

2.2.2.1. Réduction de la dimensionnalité

Le nombre de descripteurs audio considéré détermine la dimension de l'espace de représentation des sons. La réduction de la dimension est nécessaire à la fois pour éviter la "malédiction de la dimension" (*curse of dimensionality*) et pour permettre l'adaptabilité du système à différentes tâches de classification. Elle comprend les étapes suivantes : sélection des meilleurs descripteurs (manuel ou automatique), projection de l'espace de description.

La sélection automatique des descripteurs vise à déterminer le sous-ensemble de descripteurs le plus informatif au vue du problème donné et le plus compact possible. Nous utilisons l'algorithme IRMFSP (Inertia Ratio Maximization with Feature Space Projection) proposé par (Peeters, 2003). Celui-ci calcule pour chaque descripteur audio le rapport de l'inertie inter-classes sur l'inertie totale. Le descripteur produisant le plus grand ratio est jugé le plus discriminant entre classes et est donc sélectionné. L'espace de représentation est ensuite projeté sur ce descripteur afin de rendre tous les descripteurs orthogonaux au premier descripteur sélectionné. Ceci garantit que les descripteurs sélectionnés ensuite ne seront pas corrélés à ceux déjà sélectionnés. La procédure sélection/ projection est ensuite répétée pour la sélection des descripteurs suivants. Lors de l'utilisation d'un algorithme de sélection automatique de descripteurs, l'un des paramètres à optimiser est le nombre de descripteurs retenus.

L'analyse en composantes principales (ACP) et l'analyse linéaire discriminante (ALD) sont deux projections de l'espace des descripteurs en un sous-espace de dimension réduite. Etant donné la nature supervisée de notre problème, nous utiliserons l'ALD.

2.2.2.2. Apprentissage supervisé

Considérant les descripteurs sélectionnés (modifiés ou non), l'objectif de l'apprentissage supervisé est de trouver leurs valeurs types pour chacune des classes. Pour cela, il existe trois grands types de modèles d'apprentissage :

- Modèles basés sur les exemples (« instance-based »). Dans ce cas, les sons de la base d'apprentissage constituent des exemples des classes.

- Modèles génératifs. Dans ce cas, les sons de la base d'apprentissage servent à apprendre un modèle statistique. Deux modèles sont communément utilisés : le modèle gaussien unimodal multidimensionnel (GM) et le modèle de mélange de gaussiennes (GMM).

- Modèles discriminants. Dans ce cas, les sons de la base d'apprentissage servent à déterminer les valeurs types de descripteurs délimitant au mieux les frontières entre classes. Un premier exemple de ce type d'algorithme est l'arbre de décision binaire. Pour l'apprentissage de l'arbre, nous utilisons le critère de réduction entropique (« Binary Entropy Reduction Tree », BERT). Un deuxième exemple de modèle discriminant est la machine à vecteurs de support (SVM). Dans

la suite, nous utilisons un *soft-margin* SVM, aussi appelé C-SVM qui utilise un paramètre de coût *C* pour les pénalités de mauvaise classification et améliore la généralité. Nous choisissons un noyau RBF (*Radial Basis Function*), qui a l'avantage de n'être contrôlé que par un paramètre γ.

2.2.2.3. Evaluation

L'évaluation des performances d'un système de classification s'effectue en utilisant un ensemble de sons n'ayant pas servi au développement du système. Soit il s'agit d'un ensemble de données séparé (base dite de test), soit la base initiale est subdivisée en N sous-parties. Dans ce cas on en utilise N-1 pour l'apprentissage, et l'on mesure les performances sur la base restante. L'opération est alors répétée N fois en échangeant les sous-parties "apprentissage" et "test" parmi les N sous-parties. On appelle cette méthode validation à N-plis croisés (*N-fold cross-validation*). Plusieurs indices sont utilisés pour mesurer les performances d'un système, notamment l'exactitude (*accuracy*) et la f-mesure (*F-measure*) L'objectif du développement d'un système de classification consiste à trouver la succession de sous-modules et leurs paramètres permettant de maximiser ces indices.

2.2.3. Résultats de l'indexation automatique

Résultant des expériences perceptives de catégorisation décrites plus haut, trois types de taxonomies de classes sont proposés pour la description des sons environnementaux: une taxonomie de matériaux (causale), de matériaux réduite à un sous-ensemble de classes (causale réduite), d'onomatopées. Ces taxonomies sont résumées dans le Tableau 2. Un module de classification automatique a été développé pour chacune d'entre elles, de telle sorte qu'un échantillon sonore est décrit par trois étiquettes.

Taxonomie de Matériaux (Causale)	*Taxonomie réduite de Matériaux (Causale réduite)*	*Taxonomie d'Onomatopées*
Electrique *Feu* *Gaz* *Liquide* *Solide métal* *Solide papier* *Solide bois*	*Electrique* *Gaz* *Liquide* *Solide métal* *Solide papier*	*Buzzing (*bourdonnement) *Crackling* (crépitement, crissement) *Rushes (*ruées) *Scratching* (grattage, grincement) *Squeaking (*couinement, grincement)

Table 2. Taxonomies des sons environnementaux étudiées dans le projet SOR

L'apprentissage des modèles de classification a été effectué de manière entièrement automatique pour ces catégories. Les meilleurs résultats ont été obtenus à l'aide d'un classificateur de type SVM. Après optimisation des paramètres, les résultats suivant sont obtenus : 74.33% d'exactitude et 72.35% de f-measure pour la taxonomie causale, 83.72% / 80.78% pour la taxonomie causale réduite, 72.21% / 69.24% pour la taxonomie d'onomatopées. Nous indiquons ci-dessous (Tables 3, 4, 5) les matrices de confusion correspondantes.

Réel/Estimé	Gaz	Liquide	Solide Métal	Solide Papier	Solide Bois	Feu	Electrique
Gaz	**56,92**	1,54	12,31	1,54	21,54	1,54	4,61
Liquide	0	**59,38**	15,63	9,38	9,38	0	6,25
Solide Métal	1,49	2,24	**73,8**	14,18	6,72	0	1,49
Solide Papier	1,12	2,24	4,48	**85,82**	2,99	1,87	1,49
Solide Bois	2,07	2,07	4,15	2,59	**76,68**	3,11	9,33
Feu	0	0	0	3,08	4,62	**78,46**	13,85
Electrique	0	1,92	0	2,88	31,73	4,81	**58,65**

Table 3. Matrice de confusion pour la taxinomie causale, paramètres optimaux

Réel/Estimé	Gaz	Liquide	Solide Métal	Solide Papier	Electrique
Gaz	**55,38**	6,15	6,15	32,31	0
Liquide	1,48	**86,67**	6,67	5,19	0
Solide Métal	0	4,83	**87,36**	6,69	1,12
Solide Papier	1,04	4,15	3,63	**88,08**	3,11
Electrique	0	0	8,82	13,24	**77,94**

Table 4. Matrice de confusion pour la taxinomie causale réduite, paramètres optimaux

Réel/Estimé	Buzzing	Crackling	Squeaking	Scratching	Rushes
Buzzing (bourdonnement)	**75,81**	4,03	11,29	2,42	6,45
Crackling (crépitement, crissement)	3,77	**77,99**	6,29	1,26	10,69
Squeaking (couinement, grincement)	3,90	8,44	**77,27**	1,30	9,09
Scratching (grattage, grincement)	8,33	6,25	10,42	**52,08**	22,92
Rushes (ruées)	18,10	9,52	4,76	6,67	**60,95**

Table 5. Matrice de confusion pour la taxinomie d'onomatopées, paramètres optimaux

Etant donné l'aspect innovant de ces trois taxinomies, il n'est pas possible de comparer les résultats obtenus avec ceux de l'état de l'art du domaine. Le même algorithme d'indexation automatique appliqué à des taxinomies de genre et d'humeur musicales s'est classé parmi les meilleurs places lors des campagnes d'évaluation MIREX08, MIREX-09, MIREX-10[8].

2.3. Description morphologique des échantillons sonores.

La section précédente a présenté une démarche générale de caractérisation des propriétés d'un son au moyen de descripteurs acoustiques calculés sur le signal, et son application directe en terme de classification automatique. Cette première approche considère les descripteurs comme des indicateurs de type scalaire issus, soit d'une hypothèse de stationnarité du son (les paramètres ne varient pas de manière significative au cours du temps), soit d'une estimation statistique de la variation des descripteurs (si un paramètre varie au cours du temps, on considère sa moyenne et/ou des moments d'ordre supérieur de cette variation). En deuxième approche, et étant donné la nature implicitement temporelle des phénomènes sonores, il semble important de prendre en compte l'évolution de certains descripteurs au cours du temps : c'est la notion de profil morphologique, ou, plus exactement, de profil temporel morphologique du son. Dans la présente section,

[8] http://www.music-ir.org/mirex/wiki/MIREX_HOME

cette question sera abordée sous les deux mêmes angles de vue que précédemment – expérience et modélisation – et vise trois objectifs principaux : définir des classes de profils morphologiques perceptivement pertinents, concevoir un formalisme symbolique pour la description de ces profils et implémenter un modèle de calcul de descripteurs temporels associés à ces profils ; l'enjeu de la démarche étant de proposer une structure d'indexation et de classification prenant en compte cette dimension temporelle des propriétés sonores.

2.3.1. Etat de l'art

Dans un cadre général, Godoy (Godoy, 2006) et Van Nort (VanNort, 2009) ont soulevé la question des profils morphologiques et leur mode de représentation. Ils ont montré que l'objet sonore – défini à la manière de Schaeffer, c'est-à-dire en tant qu'« *unité sonore perçue dans sa matière, sa texture propre, ses qualités et ses dimensions perceptives propres* » (Schaeffer, 1966 ; Chion, 1994) – possède des composantes morphologiques intrinsèques et que ces propriétés dynamiques définissent un geste sonore associé qui participe à la caractérisation du son. Jensenius (Jensenius, 2007) a, quant à lui, abordé cette même question en adoptant une démarche expérimentale de tracé de profils : au cours d'une expérience, des participants devaient « *dessiner le mouvement associé au son* » tel qu'ils le percevaient, sans aucune consigne sur le type de dessin ou les critères à utiliser. L'une des observations résultantes est que la stratégie employée est souvent guidée par l'identification de la source physique du son ou de l'action l'ayant produit. Par ailleurs, dans le contexte de l'indexation de larges bases de données sonores, le projet ECRINS (Mullon *et al.*, 2002 ; Peeters, 2003) a également fourni une base à l'étude actuelle grâce à deux principaux résultats : d'une part, la définition d'un formalisme de description (différents types profils dynamiques et mélodiques, attaque, texture, etc.) et de représentations symboliques associées (icones, lettres, etc.) illustrées par une bibliothèque de sons prototypes (Peeters et Deruty, 2008); et, d'autre, part, le développement d'un paradigme expérimental à base de classification libre et de tracé de profils, dans le but de valider perceptivement les hypothèses de ce formalisme (Rioux, 2001).

2.3.2. Problématique

En partant de l'une des observations majeures de (Rioux, 2001) : « *malgré la consigne, des descripteurs typiquement causaux ont souvent été utilisés* », il apparaît nécessaire de prendre en compte le fait que la description morphologique d'un son fait difficilement abstraction de la source physique responsable de la production de ce son, surtout si le son en question est facilement identifiable – ce qui est globalement le cas dans notre étude. Par conséquent, pour éviter une trop grande dispersion des données dans l'étude de la morphologie des sons, il semble important de procéder à une phase de classification des sons en un nombre restreint de profils prototypiques, afin d'éviter d'avoir une description trop précise des évolutions temporelles (problème qui survient lorsque l'on demande le tracé de chaque son). Par ailleurs, il peut être également judicieux, dès cette première phase, de demander

aux participants de tracer les profils de leurs propres classes de sons, de manière à les obliger à se concentrer sur les principaux « traits » morphologiques communs aux sons de chacune des classes.

2.3.3. Expériences et résultats

Expériences.

A partir d'un corpus de sons réels de l'environnement (55 items issus de différentes base de données commerciales mentionnées au §2.1.3 et sélectionnés de sorte à échantillonner de manière exhaustive les différents types d'évènements sonores présentant un fort intérêt du point de vue de leur morphologie), la phase expérimentale de cette partie de l'étude se déroule donc en trois grandes phases :

1) une catégorisation libre des éléments du corpus par regroupement des sons ayant les mêmes profils; deux types de profils ont été initialement étudiés mais nous nous intéresserons par la suite uniquement au profil dynamique.

2) le tracé des profils correspondant aux classes identifiées ; l'intérêt de cette étape réside dans le fait que les participants, conscients de devoir tracer les profils de leurs classes, sont "forcés" à catégoriser selon des critères morphologiques.

3) sur la base de l'analyse des résultats de classification et l'extraction des classes 'moyennes', une tâche consistant à demander aux participants de tracer le profil associé à chacune de ces classes et de juger de la pertinence des classes proposées sur une échelle allant de 0 (« non pertinent ») à 10 (« très pertinent »).

Résultats.

Les données de la première phase de l'expérience sont traitées par une analyse de cluster (méthode 'average' (Legendre et Legendre, 1998)) qui permet d'extraire six classes de profils dynamiques pour les sons environnementaux : « stable », « train d'impulsions », « décroissant », « impulsif », « roulement » et « croissant ».

La troisième phase de l'expérience confirme, en premier lieu, la configuration de classification 'moyenne' obtenue à l'étape précédente : la moyenne des jugements de pertinence varie entre 6.5 et 8.5 (écarts-type entre 1 et 2.5). Elle permet également d'obtenir des éléments de symbolisme correspondant à chacune des classes identifiées, basés sur une analyse des tracés individuels (cf. Figures 4a, 4b)

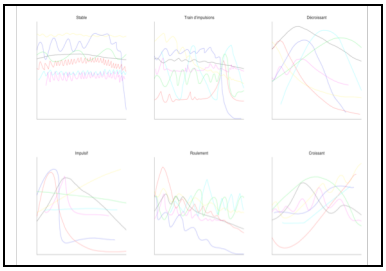

Figure 4a. Descripteurs morphologiques classes 'moyennes' - résultats individuels

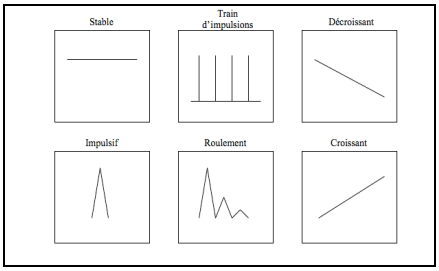

Figure 4b. Descripteurs morphologiques classes 'moyennes' - symboles

2.3.4. Résultats de l'indexation automatique

Le problème de classification morphologique est plus simple que celui qui concerne les tâches de classification statistique présentées plus haut (causale, causale réduite ou onomatopées). Comme les profils morphologiques considérés ont un lien direct avec le contenu du signal, les trois descripteurs sont sélectionnés manuellement, sans avoir besoin de recourir à une sélection ou une transformation à grande échelle des descripteurs ou bien une estimation des densités ou des paramètres de modèle. Ces trois descripteurs sont : la pente moyenne de montée de l'énergie durant l'attaque et l'extinction du son ainsi que sa durée effective perçue. La classification est directement implémentée sous la forme d'un arbre de décision en utilisant l'index de diversité de Gini comme critère de décision. La Figure 5 montre l'arbre de décision obtenu. La précision obtenue avec ce système est de 93.05%.

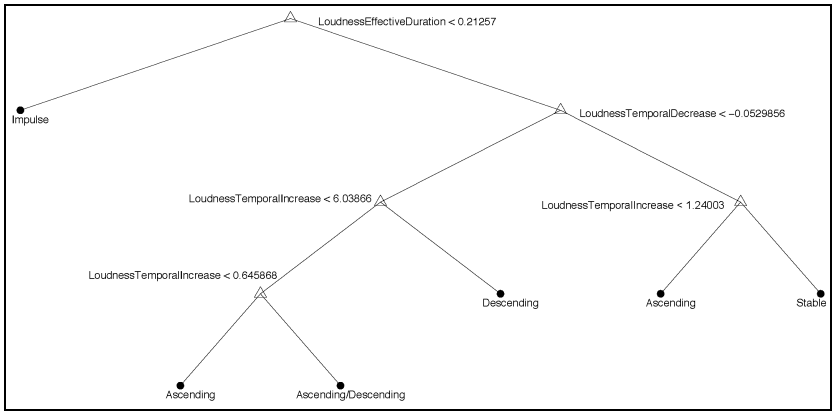

Figure 5. Arbre de décision binaire pour la classification en profil morphologique

2.4. Synthèse et perspectives

2.4.1. Synthèse

Pour la thématique de description et classification des échantillons sonores, la démarche adoptée dans le cadre du projet Sample Orchestrator a été double : d'une part, des expériences perceptives ont permis de déterminer les éléments de taxonomie pertinents (classes d'évènements, profils morphologiques) pour la perception des échantillons sonores considérés – essentiellement, des sons dit « de l'environnement » ; et d'autre part, des algorithmes d'indexation ont permis d'extraire les éléments et méthodes de traitement du signal et d'apprentissage les mieux adaptés pour implémenter des procédures de classification automatique de ces mêmes sons. La complémentarité de ces deux points de vue a permis, *in fine*, de

produire des classificateurs performants, fonctionnant de manière autonome à l'issue d'un apprentissage supervisé ; cet apprentissage ayant été effectué sur des structures (bases) proches de celles qui sous-tendent notre compréhension des sons de cette nature.

2.4.2. Perspectives

En termes de perspectives, la poursuite de ce travail peut s'envisager à plusieurs niveaux :

- sur le plan théorique, améliorer la définition de la structure de classification en affinant notamment la description de certaines grandes classes de sons (*e.g.*, Gaz ou Liquide) ;

- sur le plan expérimental, imaginer des solutions méthodologiques au problème d'incompatibilité entre la taille importante du corpus et la durée habituelle d'une expérience d'écoute. Une solution envisageable serait de développer des paradigmes expérimentaux adaptatifs, c'est-à-dire évoluant en fonction des décisions prises par le participant au cours du test ;

- sur le plan de la modélisation, adapter ou développer de nouveaux descripteurs prenant davantage en compte les spécificités des sons environnementaux qui, entre autres propriétés singulières, se caractérisent souvent par une absence de fréquence fondamentale.

3. Modèles de signaux pour le traitement sonore : extensions du vocodeur de phase

3.1. Modèle utilisé dans le vocodeur de phase

Le vocodeur de phase (Flanagan et Golden, 1966 ; Moorer, 1978) est une représentation temps-fréquence du signal basée sur un modèle sinusoïdal. Ce modèle représente le signal s(n) par une superposition de k sinusoïdes quasi stationnaires et une composante de bruit :

$$s(n) = \sum_k \alpha_k(n)\cos(\phi_k(n)) + \nu(n) \qquad [1]$$

L'utilisation de ce modèle nécessite une phase d'analyse des paramètres représentant $\alpha_k(n)$ et $\phi_k(n)$ (Amatriain *et al.*, 2002). Ces paramètres peuvent ensuite être modifiés et le signal correspondant peut être synthétisé. Une spécificité du vocodeur de phase est que les paramètres du modèle ne sont pas calculés de façon explicite. Le modèle reste toujours encodé dans la représentation temps fréquence de type transformée de Fourier à court terme (TFCT). Pour le cas des signaux numériques, la TFCT est constituée par une séquence des transformées de Fourier discrètes (TFD). Elle peut alors être écrite comme :

$$X(l,k) = \sum_{n=-\infty}^{\infty} w(n)x(n+lL)e^{-j2\pi kn/N} \qquad [2]$$

w(n) représente la fenêtre d'analyse de taille *M*, *L* le pas d'avancement entre les analyses, *x(n)* le signal d'entrée, *N* la taille de la TFD, et *X(l,k)* le coefficient du spectre de la TFCT à la position d'analyse $n_a=lL$ et avec fréquence $2\pi k/N$. Deux types de modifications du signal peuvent être effectuées dans cette représentation. Soit les amplitudes $A(l,k) = |X(l,k)|$ et phases $\Phi(l,k) = \arg(X(l,k))$ des spectres sont manipulées directement ; soit les positions de trames n_a sont modifiées afin par exemple d'obtenir une dilatation ou compression temporelle (modification de la durée du son sans changement des hauteurs). Dans le dernier cas, une adaptation des phases doit être appliquée à toutes les valeurs spectrales afin d'assurer une superposition cohérente des trames voisines lors de la synthèse. La TFCT du son est transformée à cet effet en une représentation en amplitude, phase et fréquence $\Omega(l,k)$ avec :

$$\Omega(l,k) = \frac{\left[\Phi(l,k) - \Phi(l-1,k) - L\frac{k2\pi}{N}\right]_{2\pi}}{L} + k\frac{2\pi}{N} \qquad [3]$$

La notation $[\Phi]_{2\pi}$ représente la valeur de la phase Φ dans l'intervalle $]-\pi,\pi]$. A partir de $\Omega(l,k)$, les phases peuvent donc être synthétisées de manière cohérente même si les positions des trames de synthèse sont modifiées. Pour une synthèse de haute qualité, la cohérence des phases doit aussi être assurée entre les bandes spectrales (Laroche et Dolson, 1999). Afin de pouvoir modifier les phases des données spectrales représentant des sinusoïdes, *M* doit être choisi suffisamment grand pour que les sinusoïdes, qui sont perceptivement importantes, soient résolues individuellement dans la représentation TFCT. En notant les trames de la TFCT après modification des coefficients sous la forme $Y(l,k)$, la synthèse peut être décrite par deux opérations. D'abord une transformation de Fourier inverse

$$\tilde{y}(n,l) = \sum_{k=0}^{N-1} Y(l,k)e^{j2\pi kn/N} \qquad [4]$$

qui donne les trames fenêtrées du signal modifié *y(n, l)* et ensuite une opération type « overlap-add » avec normalisation :

$$y(n) = \frac{\sum_l \tilde{y}(n-n_s(l),l)w(n-n_s(l))}{\sum_l w^2(n-n_s(l))} \qquad [5]$$

Les positions de synthèse $n_s(l)$ sont choisies en fonction de la modification de l'échelle temporelle souhaitée ; par exemple pour une dilatation constante par un

facteur 2, il faudrait $n_s(l)=2lL$. Pour une synthèse du signal, il est impératif que le dénominateur de [5] reste différent de zéro et que la distance entre deux trames synthétisées soit inférieure à M. Afin de minimiser les effets de bord dans les trames modifiées et de pouvoir synthétiser un signal modifié avec une très haute qualité, un pas d'avancement de $I<M/2$ est généralement sélectionné. L'application de la fenêtre d'analyse avant l'opération de « overlap-add » et la normalisation par la fenêtre au carré assurent que le signal synthétisé possède une TFCT avec une distance minimale au sens des moindre carrés par rapport à la TFCT modifiée $Y(l,k)$ (Griffin et Lim, 1984). Cela est important afin de minimiser la différence entre la TFCT du signal synthétisé et $Y(l,k)$.

Le fait que les paramètres du modèle sinusoïdal ne soient pas calculés de façon explicite est avantageux pour réduire le coût de calcul. Cependant l'utilisation des paramètres sinusoïdaux trouve un intérêt pour certaines opérations, notamment la séparation en composantes sinusoïdales et bruit ainsi que le déplacement des pics spectraux pour la transposition dans le domaine spectral (Laroche et Dolson, 1999a). L'un des axes de recherche dans le projet SOR portait notamment sur l'introduction de paramètres sinusoïdaux explicites dans le vocodeur afin de pouvoir améliorer le traitement des composantes sinusoïdales détectées dans le signal. Les objectifs étaient notamment la possibilité de pouvoir remixer les composantes sinusoïdales et le bruit en calculant le résiduel directement dans le domaine fréquentiel, et l'utilisation de ces paramètres sinusoïdaux pour la transposition dans le domaine fréquentiel. Un autre objectif était l'utilisation de la fréquence fondamentale pour le contrôle des paramètres d'analyse, notamment de la taille de la fenêtre.

3.1. Traitement par modèle sinusoïdal dans le domaine spectral

Afin de pouvoir séparer les composantes sinusoïdales du bruit dans le domaine fréquentiel, une représentation spectrale des sinusoïdes est nécessaire. Nous poursuivons ici l'idée proposée dans (Rodet et Depalle, 1992) consistant à calculer les tables des spectres des sinusoïdes qui peuvent ensuite être utilisées pour les opérations dans le spectre. La différence par rapport à cette méthode, dite FFT^{-1}, est que le cycle complet d'analyse et synthèse des sinusoïdes est intégré dans le vocodeur de phase.

Une extension de la méthode FFT^{-1} pour les sinusoïdes non stationnaires avec variation linéaire de la fréquence (chirp) a été proposée dans (Goodwin et Rodet, 1994). L'algorithme décrit ne considérait que la synthèse de sinusoïdes, et le problème de l'analyse des sinusoïdes non stationnaires ne se posait alors pas. Une application instantanée dans le vocodeur de phase, notamment pour un traitement en temps réel, nécessite par contre une estimation des paramètres sinusoïdaux directement à partir du spectre instantané. A cet effet, un algorithme récent permettant une estimation des paramètres sinusoïdaux pour les composantes sinusoïdales non stationnaires est utilisé (Abe et Smith, 2005). Cet algorithme suppose le modèle sinusoïdal suivant avec une amplitude exponentielle et phase polynomiale d'ordre 2:

$$s_k(n) = e^{a_k+\lambda_k n} e^{j(2\pi p_k n^2+\omega_k n+\varphi_k)} \qquad [6]$$

Les paramètres a_k, ω_k, et φ_k de ce modèle peuvent être facilement modifiés par multiplication et interpolation d'un pic pré-calculé. L'influence des paramètres λ et p sur le spectre de la sinusoïde n'est connue à ce jour que pour la fenêtre gaussienne. Pour la synthèse des sinusoïdes non stationnaires, un tableau bidimensionnel serait nécessaire. Afin de pouvoir initialiser le tableau en un temps raisonnable et de pouvoir limiter l'espace mémoire nécessaire, nous avons décidé de ne prendre en compte dans le tableau que le paramètre p et de fixer $\lambda=0$. Nous pourrons alors comparer l'impact de la prise en compte de la variation fréquentielle dans les traitements et ensuite décider si un tableau plus complet est nécessaire. Les spectres sinusoïdaux sont alors calculés pour une grille de valeurs p avec log amplitude $a=0$, phase $\varphi=0$ et fréquence $\psi=0.25 rad/s$ au centre de la fenêtre d'analyse. La TFD pour le calcul des spectres sinusoïdaux peut être suréchantillonnée afin d'améliorer l'interpolation. Seules les sinusoïdes à pentes positives ($p>0$) sont pré-calculées car les sinusoïdes avec $p<0$ peuvent être obtenues au moyen de la valeur conjuguée. Dans les expériences décrites ci-après, nous utilisons un tableau de 200 sinusoïdes couvrant les pentes $0<p<5/M^3$. Cette expression est motivée par la normalisation de la variation fréquentielle par la durée de la fenêtre d'analyse couvrant à peu près la bande d'un pic sinusoïdal stationnaire.

Le principe de l'algorithme proposé est le suivant. Pour chaque pic spectral, une classification en pics représentant des sinusoïdes ou bruit est appliquée (Roebel *et al.*, 2004 ; Zivanovic *et al.*, 2008 ; Wells et Murphy, 2010). Pour chaque pic sinusoïdal, les paramètres sinusoïdaux (a, φ, p, ω) sont estimés suivant l'algorithme décrit dans (Abe et Smith, 2005). Le spectre de la composante sinusoïdale du pic est alors extrait du tableau des pics sinusoïdaux en interpolant les spectres pré-calculés en fonction de la fréquence ω de la sinusoïde et en masquant le pic extrait du tableau de sorte qu'il respecte les limites fréquentielles du pic observé. Afin d'obtenir le modèle du pic de la composante sinusoïdale observé le pic normalisé construit à partir du tableau est ensuite multiplié par $ae^{i\varphi}$. Cette estimation de la composante sinusoïdale peut être utilisée pour séparer les composantes sinusoïdales de la composante bruitée du signal ou pour une amélioration de la cohérence entre les phases des bandes spectrales de la TFCT appartenant à des sinusoïdes non stationnaires par rapport à (Laroche et Dolson, 1999). Intéressons-nous ici à son application dans le cadre d'une transposition dans le domaine spectral. Dans ce cas, la sinusoïde transposée peut être générée directement dans le spectre. En utilisant les paramètres sinusoïdaux, la fréquence transposée peut être calculée à partir du facteur de transposition β. La méthode proposé dans (Laroche et Dolson, 1999a) déplace les pics en supposant que la sinusoïde correspondante est stationnaire. Ici, le pic représentant la composante sinusoïdale transposée est synthétisé à partir du tableau en sélectionnant une pente βp. Dans (Laroche et Dolson, 1999a), une modification de la phase instantanée par multiplication avec $Z = e^{j(\beta-1)\omega L}$ est proposée pour assurer la cohérence entre les phases des trames voisines. Si cette multiplication est appliquée de façon cumulative, prenant alors en compte les modifications appliquées pour les valeurs spectrales des trames précédentes, elle est correcte pour le cas des sinusoïdes

stationnaires. Pour les sinusoïdes non stationnaires, il faut tenir compte de la variation fréquentielle entre les trames. La représentation des fréquences $\Omega(l,k)$ peut être utilisée à cet effet, car elle reflète la variation de la fréquence d'une sinusoïde entre les trames. Les fréquences couvertes par le pic à modifier sont alors transformées comme suit :

$$\Omega'(l,k) = \Omega(l,k) + \beta\Omega(l,k_0) \quad \forall\ k \in \left[k_{\min}, k_{\max}\right], \qquad [7]$$

où k_0 est l'indice spectral le plus proche de la fréquence ω et k_{min} et k_{max} sont les indices spectraux limitant le pic traité. L'intérêt de [7] réside dans le fait qu'elle traite la représentation de la fréquence entre les trames de façon cohérente avec la modification de la pente pour le cas où la pente de la fréquence est constante. Le fait que la méthode proposée prend en compte les variations de la trajectoire de la fréquence des sinusoïdes peut être démontré expérimentalement.

Dans la Figure 6, nous comparons la qualité de la transposition d'un signal chirp d'amplitude constante et $p=3.2/M^3$. La fenêtre d'analyse est de type Hanning avec une taille M=90 ms (2000 échantillons) et N=2048. Pour comparer la qualité du signal transformé, nous affichons l'énergie du résiduel obtenue par soustraction de la sinusoïde estimée avec l'algorithme (Abe et Smith, 2005). Le résiduel augmentera avec la déviation du signal d'une sinusoïde, notamment avec les erreurs de gestion de la sinusoïde lors de la transformation. Le minimum du résiduel qui pourrait être mesuré dépend des erreurs d'analyse du modèle sinusoïdal. Ce niveau peut être déterminé par création du résiduel a partir d'une sinusoïde avec les paramètres cibles, elle est autour de -114dB (target). La courbe QI STAT montre le résultat de l'algorithme (Laroche et Dolson, 1999a). En remplaçant l'estimateur des paramètres sinusoïdaux par l'estimateur (Abe et Smith, 2005) et en utilisant [7] pour la modification des fréquences, une réduction du maximum du résiduel d'environ 10 dB est obtenue (courbe AS).
Pour cet algorithme, le résiduel est du même ordre de grandeur que celui obtenu par l'algorithme standard du vocodeur de phase travaillant dans le domaine temporel (courbe time) décrit dans (Laroche et Dolson, 1999a). En suréchantillonnant les spectres des sinusoïdes stockées dans le tableau par un facteur 4 et en adaptant la pente des sinusoïdes lors de la synthèse, l'énergie maximale du résiduel est encore diminuée de 10 dB (courbe AS+) et elle devient alors significativement plus basse que l'algorithme standard du vocodeur de phase.
Du point de vue des évaluations perceptives, il s'avère toutefois que la réduction d'erreurs n'est que très peu sensible pour les transpositions par rapport à (Laroche et Dolson, 1999a). Les avantages pour la perception sont plus importants notamment pour le cas du remixage quand les sinusoïdes sont fortement réduites. Par rapport à l'algorithme standard, la méthode offre également une réduction de coût de calcul, plus ou moins importante selon le taux de transposition.

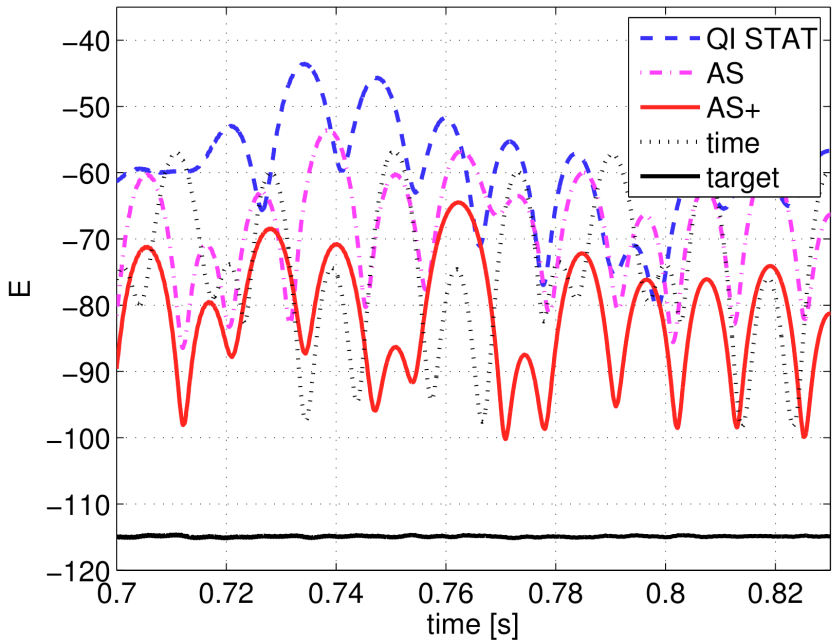

Figure 6. Energie du résiduel d'un chirp transposé avec plusieurs variantes de transposition dans le vocodeur de phase. Plus le résiduel est faible, meilleure est la qualité objective de la transposition.

3.2. Calcul automatique des paramètres d'analyse

Une partie centrale du deuxième champ de travail, concernant l'adaptation automatique de l'algorithme au signal, visait une simplification du choix des paramètres. Le paramètre fondamental dans ce contexte est la taille de la fenêtre d'analyse M qui détermine la résolution temps-fréquence de la représentation du signal. Le choix approprié nécessite une compréhension approfondie en traitement du signal qui ne peut être attendue des utilisateurs habituels de ces algorithmes (musiciens, ingénieurs du son). L'automatisation de ce choix permettrait d'une part d'améliorer l'ergonomie en évitant à l'utilisateur de chercher lui-même ces paramètres, mais aussi d'améliorer singulièrement la qualité des transformations dans le cas où la fréquence fondamentale change de façon significative. En effet, les artefacts d'une transformation sont souvent le résultat d'une transformation localement inadaptée. L'adaptation de la résolution est un problème difficile qui est actuellement beaucoup étudié (Wolfe *et al*., 2001 ; Jaillet et Torresani, 2007 ; Liuni *et al*., 2010). Cependant, ces travaux se limitent en général à l'adaptation de la représentation. Une solution générale couvrant notamment aussi les algorithmes de transformation ne pouvait pas être envisagée dans le cadre du projet. Pour le cas contraint des sons monophoniques et harmoniques, la résolution temps-fréquence

peut cependant être déterminée relativement facilement. Pour les segments voisés, la taille de la fenêtre peut être déterminée à partir de la fréquence fondamentale. Pour les segments non voisés, la stratégie de sélection de la taille de fenêtre consiste à interpoler les valeurs sélectionnées aux segments voisés voisins. L'objectif des travaux concernant la transformation des sons utilisant une résolution temps fréquence adaptative était d'adapter les algorithmes de l'état de l'art permettant une transformation de très haute qualité de sorte que la taille de la fenêtre d'analyse puisse être modifiée au cours du temps et d'évaluer la qualité des algorithmes modifiés dans des applications réelles.

Les travaux ont reposé sur un algorithme d'estimation de la fréquence fondamentale existant, la version monophonique de l'algorithme (Yeh *et al.*, 2010), mais tout autre algorithme aurait pu être sélectionné. Un algorithme d'estimation de l'harmonicité a été implanté en mesurant le rapport de la fonction d'autocovariance du signal pour un décalage du signal d'une période de la fréquence fondamentale estimé par la variance du signal. Les segments non voisés peuvent ensuite être déterminés par simple seuillage sur la mesure d'harmonicité. Afin de pouvoir minimiser l'impact des variations de la fréquence fondamentale sur la mesure de voisement, le signal est localement normalisé avant que la mesure de voisements soit calculée. Cette normalisation consiste à utiliser un nombre de périodes fixe autour du centre de la fenêtre à analyser en démodulant le signal de sorte que la variation de la fréquence fondamentale de la fenêtre soit neutralisée. L'adaptation des algorithmes de transformation a été effectuée sous l'hypothèse que la taille N de la TFD reste constante. Cela a permis de limiter le développement nécessaire à un niveau acceptable. L'ensemble des tailles de fenêtres peut alors être spécifié par l'utilisateur en donnant la fréquence fondamentale, le nombre de périodes à couvrir par la fenêtre et le nombre de tailles d'analyse différentes à utiliser. L'ensemble des fenêtres est créé de sorte que la grille des tailles soit uniforme en échelle logarithmique de durée. Le pas d'avancement I est variable entre fenêtres successives. Comme la qualité de transformation diminue avec la réduction du chevauchement entre les fenêtres, nous avons choisi de définir le pas d'avancement en fonction des tailles de fenêtres successives par $I(l) = \min(M(l-1),M(l))$.

Les algorithmes adaptés couvrent la détection des transitoires (Roebel, 2003), la classification des sinusoïdes décrite plus haut et le traitement avec préservation de la forme d'onde (Roebel, 2010) pour les signaux de parole. Un choix important était lié au pas d'avancement. L'évaluation informelle de la qualité des traitements pour tous les algorithmes de transformation qui ont été adaptés a montré des résultats très convaincants. Pour le traitement de la parole, nous observons une amélioration pour les signaux qui ont localement des fréquences fondamentales très basses, car la fenêtre peut être réduite pour les régions avec une fréquence fondamentale normale et les variations de fréquence peuvent alors être mieux suivies. Nous notons toutefois une augmentation considérable du temps de calcul du fait que le pas d'avancement doit toujours être adapté à la fenêtre la plus petite et la taille de la TFD à la taille de fenêtre le plus grande.

3.3. Synthèse et perspectives

Les perfectionnement du modèle de vocodeur de phase réalisés dans le cadre du projet SOR se traduisent selon les cas par une amélioration de la qualité des traitements, une réduction du coût de calcul à qualité équivalente, ainsi qu'une facilité accrue d'utilisation. L'ensemble des perfectionnements opérés sur le modèle du vocodeur de phase ont été intégrés dans le logiciel SuperVP de l'IRCAM et ont fait l'objet de plusieurs cessions de licences auprès d'éditeurs de logiciels spécialisés.

Les travaux décrits ont fourni des bases pour des recherches ultérieures. L'adaptation automatique de la résolution temps-fréquence de la TFCT utilisée dans le vocodeur de phase constitue un premier pas vers un algorithme de transformation fournissant des résultats de plus grande qualité sonore sans nécessiter d'expertise en traitement du signal de la part de l'utilisateur. Les travaux ont été poursuivis avec l'objectif de généraliser l'adaptation de la résolution temps fréquence sans notamment faire l'hypothèse que le signal soit monophonique (Liuni *et al*., 2010). L'utilisation et la gestion des sinusoïdes non stationnaires dans le vocodeur de phase nous a permis de renforcer les liens entre le modèle sinusoïdal et le vocodeur de phase.

4. Applications pour la création musicale

Cette section consacrée aux applications du projet SOR se concentre sur les plus originales d'entre elles, destinées à la création musicale – synthèse sonore interactive par corpus et aide à l'orchestration - ainsi qu'aux travaux de recherche dont elles sont issues. Ces applications sont principalement développées dans l'environnement Max/MSP[9], largement utilisé par l'IRCAM pour la réalisation d'œuvres musicales et multimédia interactives. D'autres prototypes applicatifs ont été réalisés pour la navigation dans les bases de données et la synthèse et le traitement sonores et ne sont pas présentés ici dans la mesure où ils découlent directement des travaux de recherche qui ont été exposés ou proposent des améliorations à des fonctions plus largement répandues en production musicale et sonore.

[9] Le logiciel Max/MSP permet la programmation visuelle et l'exécution temps réel d'algorithmes de traitement des informations musicales, sonores et vidéo. Il est commercialisé sous licence de l'IRCAM par la société californienne Cycling'74.
http://cycling74.com/whatismax/

4.1. Synthèse par corpus

4.1.1. Principes de base

La synthèse concaténative par corpus utilise une base de données de sons enregistrés, et fait appel à un algorithme de sélection d'unités qui choisit les segments de la base de données qui conviennent le mieux pour la séquence musicale que l'on souhaite synthétiser par concaténation (Schwarz, 2007). La sélection est fondée sur les caractéristiques de l'enregistrement, qui sont obtenues par analyse du signal et correspondent par exemple à la hauteur, à l'énergie ou au spectre. Les méthodes de synthèse musicale habituelles sont fondées sur un modèle du signal sonore, mais il est très difficile d'établir un modèle qui préserverait la totalité des détails et de la finesse du son. En revanche, la synthèse concaténative, qui utilise des enregistrements réels, préserve ces détails.

La mise en œuvre en temps réel de cette nouvelle approche de la synthèse sonore par le système CataRT (Schwarz *et al.*,2008) (cf. section 4.1.3.1) permet une exploration interactive d'une base sonore et une composition granulaire ciblée par des caractéristiques sonores précises, et permet aux compositeurs et musiciens d'atteindre de nouvelles sonorités, et aux designers sonore de rapidement explorer un corpus sonore constitué par un grand nombre d´enregistrements (cf. section 4.1.3.2).

La synthèse concaténative automatique par corpus s'inspire pour beaucoup de la synthèse de la parole à partir du texte *(text-to-speech)*, où depuis environ 15 ans, la plupart des systèmes de recherche et les applications commerciales utilisent un corpus de parole réelle et la sélection d'unités pour une synthèse concaténative hétérogène (avec taille d'unités variable). En effet, les résultats dans d'autres domaines, comme celui de la reconnaissance de la parole, confirment la supériorité générale de cette approche fondée sur les données. Le système logiciel Caterpillar (Schwarz, 2000; 2004) appliquait pour la première fois ces principes à la synthèse musicale. Le système CataRT développe ce travail en le transposant dans l'interaction en temps réel.

D'autres sources d'inspiration pour la synthèse concaténative sont la composition musicale par *collage* manuel, pratiquée depuis le milieu du 20ème siècle par les compositeurs de *musique concrète* à partir de Pierre Schaeffer (Schaeffer, 1966) et d'autres artistes (Oswald, 1993). La *synthèse granulaire* (Roads, 1988) permet également une exploration en temps réel d'un enregistrement sonore avec déplacement libre de la tête de lecture sur les *grains* (petites particules sonores) en fonction de leur position temporelle dans l'enregistrement. Le principe de la synthèse par corpus en temps réel pourrait donc être vu comme une extension de la synthèse granulaire basée sur le contenu, où le seul mode d'accès par position temporelle dans un son est augmenté par un accès ciblé à un ensemble de sons contrôlés par leurs caractéristiques sonores.

Les principes de la synthèse concaténative fondée sur les données apparaissent aussi dans d'autres applications et systèmes de synthèse musicale, présentés et

comparés dans (Schwarz, 2006)[10]. La plupart de systèmes existants ne fonctionnent pas en temps réel et n'offrent donc pas d'interaction directe avec le corpus. Pourtant, ces systèmes utilisent une approche spectrale pour la recherche des unités les plus semblables à la cible de synthèse, ce qui empêche la navigation ciblée dans l'espace sonore selon des caractéristiques du son. Il y a deux exemples d'applications commerciales : *Synful Orchestra*[11] (Lindemann, 2007) et *Vienna Symphonic Library Instruments*[12]. Ces deux applications proposent une extension du concept d'échantillonneur avec un système « intelligent » fondé sur les données, qui sélectionne le prochain extrait sonore à jouer, selon le contexte musical, à partir d'une grande base de sons (suffisamment grande pour rendre la sélection manuelle inefficace). Des travaux récents portent principalement sur des applications de la synthèse par corpus à la composition musicale (Einbond *et al.* 2009), le contrôle expressif de la synthèse par corpus (Maestre *et al.* 2009; Janer et de Boer, 2008; Tremblay et Schwarz, 2010; Stowell et Plumbley, 2010; Comajuncosas *et al.*, 2011), et les briques logicielles pour l'intégration dans des environnements de programmation temps réel multimédia (Brent, 2010; Schnell *et al.*, 2009).

Les premières applications de la synthèse visuelle par corpus, contrôlée par similarité acoustique, commencent à faire leur apparition avec les systèmes *Scrambled Hacks* (König, 2006), et les travaux de Casey et Grierson (Casey et Grierson, 2007) et Collins (Collins, 2007). Tous ces systèmes fonctionnent en temps réel et utilisent l'analyse spectrale d'un signal audio entrant, pour afficher un court segment vidéo choisi dans une base audiovisuelle, par similarité acoustique avec la bande son des vidéos.

4.1.1.1. Analyse de descripteurs

Dans la phase d'analyse, CataRT découpe un ou plusieurs enregistrements en petites unités appelées grains, et les caractérise par leurs contenus sonores. Ces paramètres sont donnés par la valeur moyenne de 17 descripteurs sonores tels que la fréquence fondamentale, le volume sonore, la brillance, le niveau de bruit, la répartition spectrale et d'autres caractéristiques du timbre. Ces descripteurs sont présentés dans la section 2.2.1. Les valeurs décrivant des informations de découpage sont également stockées comme descripteurs: position, durée et provenance du segment dans les fichiers son d'origine. Les descripteurs instantanés calculés à une périodicité de trames de signal d'environ 50 Hz sont condensés en valeurs scalaires par moyennage sur la durée de chaque segment. La caractérisation des segments est envisageable au moyen d'autres modélisations temporelles, telles que la moyenne, l'écart-type, la pente, ou une analyse de Fourier de l'évolution des descripteurs pour permettre de détecter un vibrato, par exemple.

---

[10] Cette veille technologique et scientifique est poursuivie et continuellement mise à jour sur le site http://imtr.ircam.fr/imtr/Corpus-Based_Sound_Synthesis_Survey
[11] http://www.synful.com/
[12] http://www.vsl.co.at/

4.1.1.2. Fonctions de distance

En général, les algorithmes de sélection d'unités utilisent soit la méthode classique de recherche du meilleur chemin à travers un réseaux d'états par l'algorithme de Viterbi, ou effectuent la sélection comme une résolution de contraintes. Les deux sont basées sur des fonctions de distance dont la distance cible exprime la similarité d'une unité cible avec des unités de la base, et la distance de concaténation la qualité de l'enchaînement entre deux unités de la base.

La similarité entre une unité de la base de données et la cible est déterminée par deux fonctions de distance entre leurs vecteurs de descripteurs respectives. D'abord, la distance cible C^t est une fonction de distance euclidienne pondérée qui exprime la similarité entre une unité-cible t et une unité de la base de données u_i,

$$C^t(u_i, t_\tau) = \sum_{k=1}^{K} w_k^t \, C_k^t(u_i, t_\tau)$$

basée sur les fonctions de distance individuelle C_k^t entre la valeur cible $t_\tau(k)$ et la valeur du descripteur de l'unité de la base de données $u_i(k)$, normalisée par l'écart type de ce descripteurs a travers le corpus σ_k

$$C_k^t(u_i, t_\tau) = \left(\frac{t_\tau(k) - u_i(k)}{\sigma_k} \right)^2$$

Soit l'unité avec la moindre distance C^t est sélectionnée, soit elle est choisie au hasard dans l'ensemble des unités vérifiant $C_t < r^2$, si un rayon de sélection r à été spécifié, soit sinon l'un des k plus proches voisins de la cible. La recherche des plus proches unités peut être effectuée en temps logarithmique par une structure de données en arbre de recherche, tel que le *kD-Tree* (Schwarz *et al.*, 2009).

L'unité sélectionnée est ensuite éventuellement transformée par transposition, changement de gain, filtrage, fenêtrage, puis copiée dans le flux de la sortie audio, soit au moment de la sélection, soit quand l'unité précédente a fini de jouer.

4.1.2. Visualisation

La recherche et le jeu musical de sons à partir d'une grande base de données sonores est semblable à l'exploration de données en manipulant une représentation graphique. Cette tâche est bien décrite et beaucoup de recherches ont été effectuées dans le domaine de la visualisation d'information (Shneiderman et Plaisant, 2005). La navigation interactive dans des espaces de données multidimensionnels nécessite soit une exploration des dimensions une à une, soit une réduction à deux ou trois des multiples dimensions pour permettre leur affichage. Des méthodes de réduction de dimensionnalité tel que l'analyse multidimensionnelle *(MDS ou multidimensional scaling)* avec l'intégration interactive de poids par dimension, l'analyse en composantes principales *(ACP*), le *mapping* linéaire (*matmap*) (Bevilacqua *et al.*, 2005), et, si des labels de classes sont disponibles, l'analyse linéaire discriminante (ALD), peuvent aider à rendre accessibles les espaces de grande dimensionnalité.

Ainsi, l'article (Schwarz et Schnell, 2009) décrit l'algorithme *Hybrid MDS* qui est une version optimisée itérative de MDS avec des itérations de complexité sub-linéaire basées sur la recherche des plus proches voisins (KPPV) (Morrison et Chalmers, 2003). Cet algorithme se base sur une modélisation masses–ressorts. L'approche itérative du placement des points vers la solution optimale peut être visualisé, ce qui permet à l'utilisateur d'interrompre l'algorithme dès que la solution intermédiaire est suffisante pour la tâche de navigation requise.

4.1.3. Applications

4.1.3.1. Synthèse interactive par navigation

Le parcours interactif de l'espace des descripteurs à l'aide de périphériques de contrôle, tels que des pointeurs 2D (souris) ou 3D, les surfaces tactiles, les potentiomètres linéaires, ou les systèmes de captation du mouvement, permet un jeu musical ou le design sonore par navigation dans l'espace des caractéristiques audio. Une trajectoire dans cet espace ou dans un sous-espace génère une évolution sonore maitrisée, mais en gardant la texture et les nuances sonores sous-jacentes.

Cette application est réalisée par le système modulaire CataRT, implanté dans le logiciel temps réel Max/MSP avec les extensions FTM (Schnell *et al.*, 2005), Gabor (Schnell et Schwarz, 2005) et MnM (Bevilacqua *et al.*, 2005). CataRT est utilisé dans des contextes musicaux de composition, de performance, et d'installations variés. L'interaction repose sur une interface simple consistant en l'affichage d'une projection 2D de l'espace de descripteurs (Figure 7), et une navigation avec la souris, où les grains sont sélectionnés et joués par proximité géométrique.

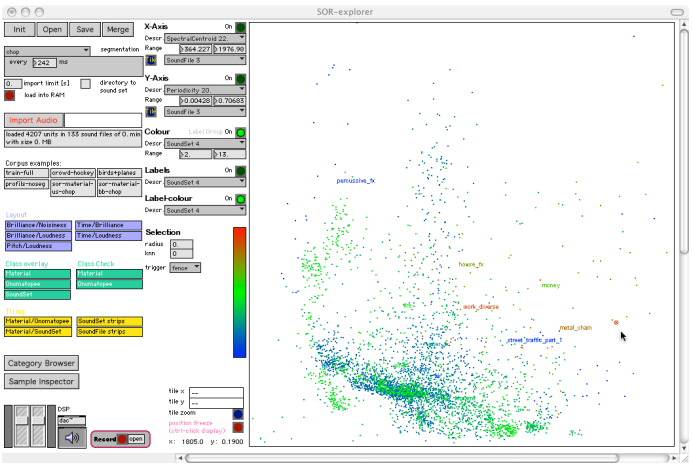

Figure 7. Interface de navigation dans un corpus par descripteurs sonores.

4.1.3.2. Recherche interactive de sons

CataRT est un modèle d'application de navigation dans des espaces sonores, et un cadre de test pour des nouveaux paradigmes de visualisation et recherche dans des grandes bases de sons. En effet, un prototype à été adapté à la recherche de sons dans des grandes collections sonores (effets sonores) pour le *sound design* au cinéma, dans les jeux, et la création multimédia. Le nombre et la taille des bases de sons augmente continuellement, ce qui déplace le problème de l'exploitation efficace de ces bases de la question de savoir si un son spécifique se trouve dans la base, à celle de comment le trouver. Ici, notre approche de navigation interactive dans un espace de descripteurs avec retour audio immédiat accélère grandement la recherche, en comparaison des interfaces usuelles basées sur des menus hiérarchiques, formulaires de recherche, et listes de résultats, à écouter l'un après l'autre. L'article (Schwarz et Schnell, 2009) décrit comment les algorithmes scalables de recherche rapide des *k* plus proches voisins, de réduction de dimensionnalité, et le système de visualisation avancé sont appliqués à ce problème.

4.1.3.3. Modélisation de la continuation de boucles rythmiques

Une application de la synthèse par segments à la génération de boucles rythmiques peut profiter pleinement de la description riche du contenu : par la modélisation des continuations typiques inhérentes à un corpus de boucles de batterie, nous pouvons régénérer des variations de boucles avec un contrôle interactif, tandis que le style de base du corpus reste intact. Cette modélisation du comportement temporel d'un corpus de sons et de ses transitions est fondée sur une distance dite de continuation à partir du corpus. Cette distance exprime les comportements temporels inhérents au corpus et peut être appliquée pour guider la synthèse par navigation dans un espace de descripteurs (Schwarz, Cadars, Schnell, 2008).

Elle est définie en prenant en compte le contexte des unités du corpus, en le comparant avec le contexte des unités cibles, ce qui prédit la meilleure continuation de ce contexte. Pour ceci, nous définissons la fonction de distance de continuation C^c sur un contexte de taille r basée sur la distance cible C^t comme suit :

$$C^c(u_i, t_\tau) = \sum_{j=0}^{r} w_j^c \, C^t(u_{i-j}, t_{\tau-j})$$

Le prototype réalisé sur cette base permet de synthétiser des séquences sonores avec la même structure rythmique que celle du son analysé à partir d'éléments issus d'un autre corpus sonore.

4.1.4. Synthèse et perspectives

Le projet SOR a permis l'application de techniques de synthèse par corpus à tous types d'échantillons sonores et le développement d'interfaces homme-machine fournissant des modes de visualisation et de manipulation interactive de leurs

espaces de descripteurs sonores. L'application CataRT qui en résulte a été utilisée pour la création de nombreuses œuvres de musique contemporaine.

Ces travaux ouvrent plusieurs directions de recherche nouvelles. D'une part l'intégration du vocodeur de phase pour la synthèse par corpus. L'enjeu est ici d'améliorer la qualité des transformations sonores intervenant dans la synthèse, ainsi que des transitions entre segments, pouvant être opérées directement par interpolation dans le domaine spectral, et non plus seulement par fondu enchaîné calculé dans le domaine temporel comme c'est le cas avec le modèle de synthèse actuel. Les deux modèles de synthèse partageront en partie une description sonore commune et assureront ensemble le contrôle de haut niveau d'un rendu sonore de très haute qualité.

D'autre part le développement de stratégies d'analyse et de contrôle pour la synthèse de textures sonores. Cet aspect fait l'objet du projet Topophonie[13] (Schwarz et Schnell, 2010; Schwarz, 2011), visant la réalisation de dispositifs *audiographiques*, c'est-à-dire produisant la synthèse d'éléments visuels et sonores synchronisés à partir de contrôles communs.

4.2. Aide à l'orchestration

Cette section aborde un problème nouveau en informatique musicale, et plus particulièrement en composition assistée par ordinateur, celui de l'assistance logicielle à l'orchestration. En regard des autres disciplines de l'écriture musicale (harmonie, rythme, contrepoint), l'orchestration occupe une position particulière : en tant qu'art de choisir et de mélanger les timbres individuels des instruments de musique pour produire des « couleurs » particulières, elle touche comme nulle autre à la « réalité sonore » orchestrale. Par exemple, le même accord joué par des cordes, des vents ou par un piano engendre des timbres radicalement différents, bien qu'il s'agisse dans chaque cas des mêmes notes. Là réside sans doute la difficulté de cette discipline : ces effets ne sont pas déductibles des seules données qu'elle manipule, hauteurs, dynamiques, modes de jeu, sourdines… Ces symboles abstraits par lesquels la musique se conçoit, s'écrit et se pense, ne livrent rien du son en tant qu'objet perçu, indépendant de la chaîne causale qui le met en œuvre. Par sa nature même, l'orchestration reste un savoir (et un savoir-faire) empirique et difficilement formalisable. Il n'est donc pas étonnant qu'elle soit restée jusqu'à récemment la grande absente de la composition musicale assistée par ordinateur (Maresz, 2006).

Parmi les équipes de recherche impliquées dans le projet SOR, deux se sont tout particulièrement employées à pallier ce manque. L'originalité de leurs travaux a été d'aborder simultanément les dimensions « signal» et « symbolique » de l'orchestration. Un premier groupe de chercheurs s'est ainsi concentré sur la construction d'un modèle de signal pour décrire des mélanges instrumentaux et évaluer des distances perceptives entre mélanges. En parallèle, une autre équipe a

[13] http://www.topophonie.fr

travaillé sur la définition d'un cadre formel pour l'orchestration assistée par ordinateur, la rapprochant en l'occurrence d'un problème d'optimisation multicritères sous contraintes. Ces deux axes de travail ont dans un second temps conduit à la conception et le développement du logiciel d'orchestration *Orchidée*, aujourd'hui utilisé par de nombreux compositeurs.

4.2.1. Modèle de description et de combinaison des échantillons

4.2.1.1. Introduction

Le système d'orchestration proposé fonctionne de la manière suivante : étant donné un timbre cible, défini sous la forme d'un son pré-enregistré et un ensemble d'échantillons de sons instrumentaux, le système cherche une combinaison de sons instrumentaux s'approchant le plus possible du timbre cible. Trois problèmes liés à la description des échantillons instrumentaux et des combinaisons ont été traités dans le cadre du projet. Tout d'abord nous avons défini un ensemble de descripteurs des sons instrumentaux adapté au problème de l'orchestration. Ensuite nous avons proposé un modèle probabiliste génératif des descripteurs sachant l'instrument, le mode de jeu, la hauteur et la nuance (liée à l'intensité sonore). Ce modèle, entrainé sur une base d'échantillons, permet de rendre compte de la variabilité intrinsèque au son instrumental. Cette variabilité peut être due à la facture de l'instrument, au jeu de l'interprète ou aux conditions d'enregistrement des échantillons. En outre ce modèle permet de généraliser les connaissances extraites de la base d'échantillons et ainsi d'extrapoler la description de hauteurs, de nuances ou de modes de jeu absents de l'ensemble d'échantillons. Enfin nous avons déterminé comment ces descripteurs et ces modèles pouvaient être combiné pour obtenir le modèle de description de combinaison de sons. Ces modalités de combinaisons tiennent compte de la variabilité intrinsèque des sons instrumentaux et de l'effet de salle.

4.2.1.2. Description des sons pour l'orchestration

Le choix des descripteurs pour l'orchestration est soumis à plusieurs contraintes. Tout d'abord, ils doivent permettre une bonne description des sons instrumentaux et une distinction entre les différents modes de jeu. Ensuite, ils doivent pouvoir être extraits de sons monodiques et polyphoniques puisque la cible à atteindre peut être polyphonique. Les descripteurs doivent aussi décrire des caractéristiques de suffisamment haut niveau pour permettre une interaction avec l'utilisateur, c'est-à-dire des caractéristiques proches de dimensions perceptives. Enfin, il doit être possible de définir une procédure "d'addition" des descripteurs. Nous avons proposé un ensemble de descripteurs de formes spectrales (enveloppe spectrale…), et de formes temporelles (temps d'attaque, modulation d'amplitude…) permettant de décrire un large ensemble d'instruments et de modes de jeu. L'ensemble complet ainsi que les procédures d'extraction sont détaillés dans (Tardieu *et al.*, 2011) et (Peeters, 2004).

4.2.1.3. Modèle d'instrument

Pour tenir compte de la variabilité intrinsèque des sons instrumentaux et pour généraliser les connaissances contenues dans l'ensemble d'échantillons, nous avons proposé un modèle probabiliste génératif des descripteurs du timbre. Au lieu de représenter un son instrumental par ses descripteurs, nous le représentons par une distribution des descripteurs. Le modèle est constitué de la distribution des descripteurs sachant l'instrument i, le mode de jeu t, la hauteur p et la nuance d de la note jouée.

$$f((\mathbf{d_j}) \mid i,p,d,t) = \prod_j f(\mathbf{d_j} \mid i,p,d,t)$$

Où l'on suppose l'indépendance des descripteurs conditionnellement à i,p,d et t.

	Violon	Violoncelle	Flûte	
Vibrato	100	100	0	Modèle de vibrato 200
Non vibrato	30	30	100	Modèle de non vibrato 160
	Modèle de violon 130	Modèle de violoncelle 130	Modèle de flûte 100	

Figure 8 : Exemple de regroupement des modes de jeu pour l'apprentissage

Au lieu d'apprendre directement les paramètres de cette distribution, ce qui demanderait de disposer d'échantillons en grand nombre pour chaque mode de jeu, chaque nuance et chaque hauteur, nous proposons une réécriture du modèle permettant un apprentissage avec moins de données et l'extrapolation de sons absents de la base. Cette modification consiste à supposer que les différents modes

de jeu ont des effets similaires sur les instruments. Par exemple, l'effet tremolo se caractérise par une modulation périodique d'énergie et un rapport harmonique sur bruit faible quel que soit l'instrument à cordes frottées utilisé. Ainsi il est possible d'apprendre d'une part les caractéristiques liées à l'instrument et d'autres part les caractéristiques liées au mode de jeu et finalement d'agréger les descriptions. Cette approche présente plusieurs avantages. Premièrement, en regroupant les différents échantillons d'un même mode de jeu pour l'apprentissage, les modèles appris sont plus généraux et plus robustes, grâce à un ensemble d'apprentissage plus grand. Prenons l'exemple du vibrato et des cordes frottées : l'instrumentiste qui a enregistré les sons vibrato d'alto a, pour des raisons d'habitude ou par hasard, joué le vibrato avec une certaine fréquence et un certain ambitus. Un modèle simple prendrait ces valeurs pour représentatives de l'alto, ce qui n'est évidemment pas le cas, et pourrait éliminer l'alto dans une orchestration sous prétexte qu'il ne peut pas jouer, par exemple, de vibrato plus rapide. Grouper tous les vibratos permet d'apprendre un modèle plus général de vibrato en supposant que les différents instruments ont les mêmes possibilités. Deuxièmement, les modèles de modes de jeu absents de la base d'apprentissage peuvent être déduits par combinaison. La Figure 8 montre un exemple dans lequel, en appliquant le modèle de vibrato appris sur les cordes nous déduisons le modèle d'une flûte vibrato. La figure représente le nombre d'échantillons disponibles pour chaque type de son. En bas et à droite se trouve le nombre d'échantillons obtenus si on regroupe par instrument et par présence ou absence de vibrato. Dans ce cas, il est possible d'apprendre des modèles génériques de violon, violoncelle et flûte d'une part, vibrato et non vibrato d'autre part et enfin de combiner les modèles obtenus. Nous pouvons donc extrapoler un modèle approximatif de flûte vibrato sans en avoir d'échantillon. Une fois de plus, cela suppose que les modes de jeu ont les mêmes effets sur tous les instruments, ce qui dans le présent exemple n'est pas tout à fait juste puisque le vibrato de flûte présente beaucoup moins de modulation de fréquence que le vibrato de corde, donc l'application à une flûte du modèle de vibrato appris sur des cordes constitue une approximation. Nous supposons cette approximation valide dans notre cadre et choisissons de favoriser une plus grande généralité des modèles à une trop grande précision.

Cette hypothèse ce traduit par la reformulation suivante du modèle :

$$f(\mathbf{d_j} \mid i,p,d,(u_k)_{k\in[1,K]}) = \frac{1}{Z} f_{ji} \prod_{k=1}^{K} f_{jk}$$

Où f_{ji} représente la distribution du descripteur j pour l'instrument i, f_{jk} représente la distribution du descripteur d_j pour un mode de jeu donné, c'est-à-dire pour une valeur donnée de u_k. La variable t représentant le mode de jeu à été remplacée par un ensemble de variables représentant les modes de jeu élémentaires (vibrato, trémolo…) appris séparément. Z est une constante de normalisation. Les distributions sont approximées par des lois normales à matrice de covariance pleine.

4.2.1.4. Addition des descripteurs et des modèles

A partir des modèles d'un ensemble de sons, nous avons besoin de procédures permettant d'obtenir le modèle d'une combinaison de ces sons. Il s'agit d'abord de savoir comment les descripteurs se combinent puis de déterminer comment les distributions de ces descripteurs se combinent. Lorsque deux ou plusieurs sons sont émis simultanément, ils interagissent entre eux. Dans le contexte de l'aide à l'orchestration, nous devons modéliser cette interaction pour obtenir des opérateurs simples nous permettant de combiner des sons, ce qui signifie d'être capable, à partir de la connaissance des propriétés des sons émis, de déterminer les propriétés de la combinaison de ces sons. Cette interaction dépend de plusieurs phénomènes liés aux propriétés des émetteurs (les instruments), du transmetteur (la salle) et du récepteur (l'auditeur). Sa compréhension fait donc intervenir des notions d'acoustique et de perception. La perception détermine comment est traitée l'information par l'auditeur tandis que l'acoustique permet de comprendre comment le signal émis par les instruments est transformé par la salle. Il faut aussi prendre en compte les variations du signal émis par les instruments, conséquences des propriétés acoustiques des instruments mais aussi des gestes des instrumentistes. Les résultats obtenus nous ont permis de proposer une approximation de l'interaction des sons instrumentaux dans un orchestre.

Concernant le spectre, nous supposons une addition en puissance. Pour les autres descripteurs, comme les modulations d'amplitude, des mesures ont été réalisées sur des sommes de sons pour déduire des formules d'addition empiriques. Cependant pour certains types de sons ou certain descripteurs, déterminer ces formules est plus complexe et relève plus de la perception que de l'acoustique. Ainsi, il est difficile de prédire comment sera perçue une combinaison impliquant des sons percussifs. Pour comprendre la perception de telles combinaisons deux expériences ont été réalisées en collaboration avec l'université de McGill (Tardieu et McAdams, 2011). Elles nous ont permis de comprendre quels facteurs influençaient la fusion entre un son percussif et un son tenu, de déterminer les corrélats acoustiques de la perception de ce type de combinaison et enfin de trouver des formules d'addition pour certains descripteurs.

4.2.2. Optimisation multicritères sous contraintes

Peut-on concevoir un outil informatique tirant parti de nos connaissances récentes en psychoacoustique et traitement du signal et capable de calculer la meilleure combinaison instrumentale qui réponde à une idée de timbre donnée ? Telle fut, peut ou prou, la question posée à la science par les compositeurs. Elle en soulève en réalité plusieurs. La caractérisation objective et quantitative des propriétés perceptives de combinaisons instrumentales introduite au paragraphe précédent constitue une première étape. Nous nous attaquons ici aux enjeux combinatoires posés par la recherche d'une solution optimale.

Dans le cadre de notre recherche, les orchestrations (i.e. les combinaisons instrumentales) sont représentées par des superpositions de signaux sonores. Ces

derniers proviennent d'échantillons instrumentaux en nombre suffisant pour couvrir l'étendue des possibilités orchestrales, aussi bien en termes de techniques de jeu que de particularités timbrales. Le problème devient alors celui de trouver, parmi une vaste collection d'éléments, un sous-ensemble de cardinalité bornée vérifiant un critère d'optimalité que nous définissons comme une distance « perceptive » minimale avec un « timbre cible ». Un tel timbre, tout comme celui de chaque échantillon et de toute combinaison d'échantillons, est caractérisé, nous l'avons vu, par un ensemble de valeurs de descripteurs perceptifs, voire par un jeu de lois de probabilités pour chacun de ces descripteurs. Nous faisons alors l'hypothèse que pour chaque descripteur perceptif f il existe pour tout couple de timbres (a,b) une mesure $D_f(a,b) = D_f(f_a, f_b)$ reflétant la dissimilarité entre a et b le long de la dimension perceptive associée à f (Carpentier *et al.*, 2010a). Par exemple, si f est le centroïde spectral, descripteur reconnu comme fortement corrélé au percept de brillance, alors $D_f(a,b) = 0$ si et seulement si les deux timbres a et b ont la même brillance. En allant plus loin, si $\{f \in F\}$ est un jeu de descripteurs perceptifs caractérisant pleinement le timbre, alors deux timbres a et b sont indiscernables perceptivement si et seulement si $\forall f \in F, D_f(a,b) = 0$. Etant donné alors un timbre cible t caractérisé par un jeu de descripteurs $\{f_t, f \in F\}$, notre problème d'orchestration revient à trouver la configuration x vérifiant : $\forall f \in F, x = \arg\min D_f(f_x, f_t)$.

En règle générale, il n'existe pas de solution minimisant conjointement toutes les distances perceptives. Aussi nous faut-il soit nous résoudre à une transformation du problème, par exemple $x = \arg\min \sum_f \lambda_f D_f(f_x, f_t)$, soit opter pour une approche multicritère. C'est cette dernière option que nous avons retenue, car il n'est pas possible de connaître au préalable l'importance relative de chaque dimension perceptive dans un jugement de similarité. Si le temps et le niveau de bruit de l'attaque sont des indices perceptifs pertinents pour différencier des sons très brefs, l'oreille se concentrera en revanche sur la brillance ou encore l'étendue spectrale pour des sons entretenus. L'expérience montre par ailleurs que d'un auditeur à l'autre, et d'une situation d'écoute à l'autre, les indices perceptifs mobilisés par l'écoute peuvent être différents, même pour des « classes » de sons similaires.

Par opposition aux problèmes « classiques » d'optimisation, les méthodes multicritères ont l'avantage de retourner non pas une unique et meilleure solution, mais un ensemble de solutions réalisant différents compromis entre les critères. Dans notre cas, elles correspondent à des propositions d'orchestration satisfaisant différentes préférences d'écoute, qui varient selon le contexte et la subjectivité de l'utilisateur. L'intérêt majeur de ce type d'approche est de permettre une recherche itérative au cours de laquelle l'utilisateur émet des choix, retient certaines configurations, en écarte d'autres. Il fournit ainsi au moteur de calcul des informations sur des préférences personnelles qu'il n'est bien souvent ni en mesure de formuler, ni de quantifier, et oriente ainsi la recherche vers des solutions les plus à même de satisfaire ses exigences esthétiques.

En termes techniques, il a été montré dans (Carpentier, 2008) qu'ainsi formulée, la recherche d'orchestrations peut s'apparenter à un problème de sac-à-dos multiobjectif, attestant ainsi de son caractère NP-difficile. A cette complexité combinatoire se rajoute, en vertu de l'approche psychoacoustique retenue et des modèles perceptifs qui en découlent, de fortes non-linéarités et non-monotonies dans l'expression des fonctions objectifs, interdisant tout recours à des méthodes complètes d'optimisation. Nous nous sommes donc tournés vers les métaheuristiques (Talbi, 2009), et plus précisément vers les algorithmes génétiques (Goldberg, 1989), principalement en vertu du couplage entre génotype et phénotype, particulièrement adapté à notre problème. Une représentation spécifique permet d'encoder directement les variables de l'écriture musicale dans le génotype des individus (chacun étant vu comme une proposition d'orchestration particulière), et de proposer des opérateurs génétiques s'inspirant directement du travail des compositeurs : ajout ou retrait d'un instrument, substitution par un autre de timbre semblable, croisement par groupes, etc. Le phénotype, quant à lui, correspond aux propriétés perceptives de chaque individu, qui s'expriment dans un espace métrisé, (donc dans lequel on mesure des distances). En pratique, l'algorithme retenu s'inspire du MOGLS (Jaszkiewicz, 2001) : l'évaluation des individus se fait à l'aide de normes de Tchebycheff pondérées, desquelles on peut explicitement déduire un jeu de préférences subjectives quantifiées, et utiliser cette information dans les étapes ultérieures du calcul (Carpentier *et al.*, 2010b).

Enfin, il est possible, à l'aide d'une syntaxe d'opérateurs génériques, de définir sur le génome un ensemble de contraintes globales garantissant une recherche cohérente avec un contexte musical donné (Carpentier, 2011). Un algorithme de recherche locale spécifique a été développé, étendant le principe de la recherche adaptative (Truchet et Codognet, 2004) aux contraintes globales, et s'interfaçant avec l'algorithme génétique principal grâce à une modification du concept de dominance (Deb, 2000). Les problèmes d'optimisation et de contraintes sont donc abordés de façon conjointe.

4.2.1. Applications

Les recherches exposées dans les deux paragraphes précédents ont conduit au développement du logiciel d'aide à l'orchestration *Orchidée* (Carpentier et Bresson, 2010). Son but : permettre aux compositeurs de découvrir en un temps raisonnable des combinaisons instrumentales dont le timbre s'apparente à un « son cible » pré-enregistré. *Orchidée* s'inscrit ainsi dans la catégorie des outils d'aide à l'écriture musicale. Il produit en sortie une partition qui peut être simulée par des échantillonneurs.

Afin de permettre une interaction maximale avec les outils déjà existants de composition assistée par ordinateur (Miranda, 2001). *Orchidée* n'a pas été conçu comme une solution intégrée, mais comme une suite d'outils modulaire, selon un modèle conceptuel proche des architectures trois tiers, et décrit sur la Figure 9

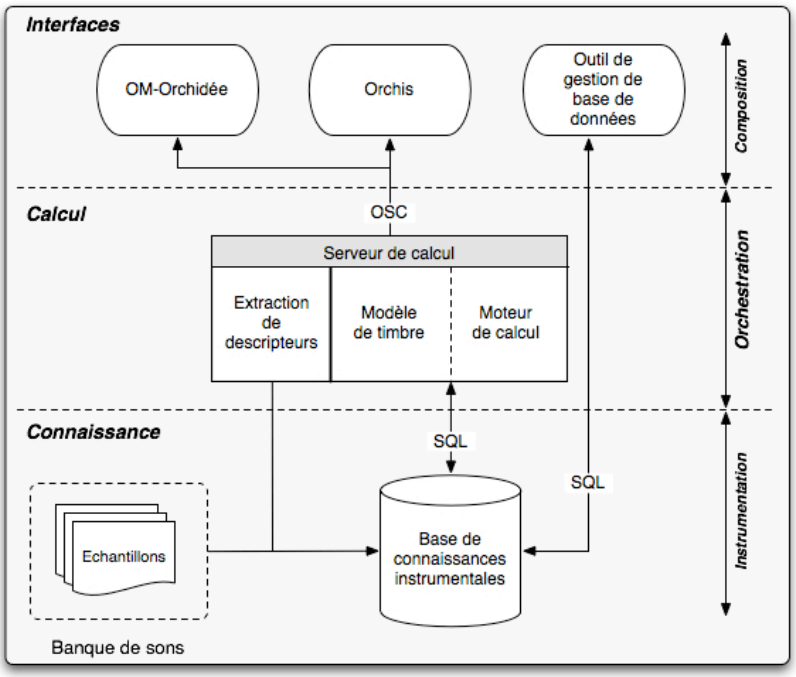

Figure 9. Architecture de la suite d'outils Orchidée. Parmi les interfaces de contrôle, OM-Orchidée est une bibliothèque du logiciel OpenMusic, Orchis est en ensemble d'outils développés dans Max/MSP.

Le cœur d'Orchidée est un « serveur » de calcul, application locale qui s'exécute en processus de fond, à l'écoute des messages que lui envoient ses différents « clients ». Il est composé d'un module d'extraction de descripteurs perceptifs à partir de signaux audio, d'un modèle de timbre permettant d'estimer les propriétés timbrales des mélanges instrumentaux, et enfin d'un moteur de calcul dans lequel est implanté l'algorithme de recherche multicritère sous contraintes décrit au paragraphe précédent. Ce « serveur » communique d'une part – via un ensemble de requêtes SQL – avec une base de données relationnelle contenant une représentation des possibilités instrumentales, d'autre part – via un ensemble de messages OSC (Wright, 2005) – avec des interfaces visuelles que les utilisateurs ont tout loisir de développer dans leurs environnements favoris (comme *OpenMusic*, ou *PWGL* par exemple). Il est intéressant de remarquer qu'une telle architecture, déjà éprouvée dans bien des domaines d'application, trouve ici également une justification musicale. Ainsi le module « connaissance » (en bas de la Figure 9) peut-il être considéré comme un traité numérique d'instrumentation, puisqu'il réunit à la fois les caractéristiques de jeu des instruments ainsi que leurs propriétés timbrales. Cette connaissance est exploitée par le serveur d'orchestration, lui-même « commandé »

par des actions entreprises au niveau compositionnel. En d'autres termes, l'architecture d'*Orchidée* est à l'image d'une hiérarchie musicale liant les trois disciplines de l'instrumentation, l'orchestration et de la composition.

Le logiciel *Orchidée* est livré avec *Orchis*, une application cliente développée dans Max/MSP permettant aux compositeurs, de manière rapide et intuitive, d'exécuter des requêtes vers une base de données instrumentales, d'analyser un timbre cible, de naviguer efficacement dans un ensemble d'orchestrations s'en approchant, de contraindre l'espace de recherche à l'aide de filtres sur les variables symboliques de l'écriture musicale. Il permet également de simuler les résultats avec un échantillonneur et de les exporter sous forme de partitions.

Orchidée a déjà été utilisé par de nombreux compositeurs internationalement reconnus, tels Gérard Buquet, Yan Maresz ou Jonathan Harvey (Nouno *et al.*, 2009), ainsi que des étudiants en composition, à l'IRCAM, à la Haute Ecole de Musique de Genève et à l'Université de Columbia à New York. Le succès rencontré par cet outil nouveau tient d'après nous à deux raisons. D'une part, il comble un manque dans la panoplie des outils d'informatique musicale. D'autre part, les compositeurs ont été associés à la recherche tout au long de ce projet, au cours duquel la pertinence musicale des avancées scientifiques et technologiques n'a cessé d'être questionnée, afin d'aboutir à un outil répondant au mieux au besoins de ses utilisateurs finaux.

4.2.2. Synthèse et perspectives

La question de l'aide à l'orchestration qui a fait l'objet de deux thèses de doctorat et du développement d'un logiciel utilisé en grandeur réelle par les compositeurs a considérablement progressé dans le cadre du projet SOR Si les résultats obtenus sont satisfaisants, c'est cependant au prix de deux simplifications majeures.

Tout d'abord, l'orchestration conçue comme réalisation d'une cible audio fournie à titre d'exemple n'est qu'une manière d'envisager le problème. Quelquefois la cible timbrale est imaginaire, voire abstraite, c'est-à-dire conditionnée par un ensemble de paramètres musicaux symboliques liés à l'harmonie, la position dans le mètre, la conduite horizontale des voix, qui surdétermine en un lieu de la partition, en fonction des instruments disponibles, les choix timbraux. A vrai dire, l'orchestration imitative est l'exception dans la tradition musicale, même elle est plus présente dans la musique du XXème siècle (que l'on pense par exemple aux œuvres de Steve Reich telles The Cave ou City life dans lesquelles l'orchestre parle ou imite les sons de la ville).

L'autre simplification majeure consiste à considérer la cible comme statique, en moyennant les descripteurs sur une plage d'analyse. L'effet immédiat de cette simplification est que les orchestrations résultantes sont elles même statiques, consistant en des sortes d'accords plats d'instruments jouant individuellement des notes stables. Ce mode de résolution conduit à un style d'écriture par « colonnes

verticales » qui existe certes, mais, une fois de plus, n'épuise pas la richesse et la variété de l'écriture musicale.

Le problème de l'orchestration abstraite a été traité partiellement dans le cours du projet SOR en remplaçant l'idée de cible audio fixée par celle d'un outil graphique interactif de construction de la cible à travers ses composantes symboliques (notes, durées) et spectrales (partiels, formants), processus contrôlé pas à pas par de la synthèse. Le point principal ici est le cycle d'interaction dans un cadre constructiviste permettant l'utilisation des éléments atomiques de la musique, de se figurer sommairement par la synthèse le timbre visé, de lancer des calculs d'orchestrations intermédiaires, de les évaluer, et de reprendre les étapes d'édition jusqu'à convergence vers une solution satisfaisante. Notons que du point de vue du moteur d'orchestration rien ne change, puisque les cibles « abstraites » intermédiaires lui sont soumises sous forme de cibles audio grâce à la synthèse.

La question de la dynamique temporelle des cibles et des orchestrations est très complexe et a fait l'objet, à la suite de SOR, du démarrage d'une nouvelle thèse sur le principe du remplacement des descripteurs par des séries temporelles décrivant leurs évolutions. Les volumes de données à traiter ont alors naturellement tendance à exploser. Un des premiers problèmes traités est l'optimisation du codage et de l'indexation des ces séries dans les bases de données pour permettre un traitement des requêtes efficace. Un deuxième défi, en lien avec les problématiques du §2.3, est ensuite de détecter, décrire et de reproduire correctement les spectro-morphologies avec leur dynamique temporelle, ce qui a des implications non seulement sur la sélection des échantillons ou modèles instrumentaux mais aussi sur leur ordonnancement temporel (Esling *et al.*, 2010).

5. Conclusion

Nous avons décrit certaines avancées des recherches issues du projet Sample Orchestrator qui ont porté sur la description et l'indexation automatiques d'échantillons environnementaux, les perfectionnements du vocodeur de phase pour le traitement sonore de haute qualité et de nouvelles applications des descripteurs de sons. La réalisation de ces travaux a fait intervenir la coordination interdisciplinaire de plusieurs approches scientifiques, associant traitement de signal audio (analyse, traitement, synthèse), psychologie cognitive (structures de description et mesures de similarité) et informatique (apprentissage automatique, langages et algorithmique, interfaces homme-machine).

Les résultats du projet débouchent sous la forme de fonctions de gestion par le contenu de grandes bases de données sonores et de traitement de haute qualité des sons, qui trouvent de larges applications en production sonore et musicale et ont fait l'objet de plusieurs cessions de licences industrielles auprès d'éditeurs de logiciels.

De plus, le projet s'est traduit par le développement d'applications radicalement nouvelles, destinées à la création musicale et proposant des modes inédits de synthèse et de manipulation des contenus sonores. Ces réalisations – synthèse sonore interactive par corpus et aide à l'orchestration - ont été rendues possibles par l'arrivée à maturation de techniques de description et de gestion de grandes bases de données de sons, dont elles dépassent les fonctions de recherche pour lesquelles elles ont été initialement conçues en les appliquant à la production de sons et de structures musicales. Les travaux sur la description des sons trouvent ainsi aujourd'hui un large intérêt dans la création contemporaine, dans la mesure où ils instancient, sous la forme d'outils opérationnels, une extension du vocabulaire musical traditionnel fondé sur un ensemble limité de paramètres - hauteurs, intensités, et les structures temporelles de leurs occurrences – à de nouvelles dimensions descriptives du sonore.

Projet Sample Orchestrator, programme ANR Audiovisuel et Multimédia Edition 2006, durée : 30 mois.

Partenaires : IRCAM (coordinateur), société Univers sons.

6. Bibliographie

Abe M., Smith J. O. (2005). AM/FM rate estimation for time-varying sinusoidal modelling, *Proc. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'05)*, vol. III, pp. 201-204.

Amatriain X., J. Bonada, A. Loscos et X. Serra (2002) Spectral Processing, *Digital Audio Effects*, chapter 10, pp. 373-438, John Wiley & Sons.

Aucouturier, J.-J. et F. Pachet (2006). Jamming With Plunderphonics: Interactive Concatenative Synthesis Of Music. *Journal of New Music Research,* 35(1), 35–50. Special Issue on Audio Mosaicing.

Bevilacqua, F., R. Muller, et N. Schnell (2005). MnM: a Max/MSP mapping toolbox. *Proc. Conference for New Interfaces for Musical Expression (NIME'05)*, Vancouver, pp. 85–88.

Brent, W. (2010). A Timbre Analysis and Classification Toolkit for Pure Data. Proc. *International Computer Music Conference (ICMC'10)*, New York City, NY.

Brown J. (1999). Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *Journal of the Acoustical Society of America*, 105(3): 1933-1941.

Burred J. J., Cella C. E., et al. (2008). Using the SDIF Sound Description Interface Format for Audio Features. *Proc. International Conference on Music Information Retrieval (ISMIR'08)*. Philadelphia, USA.

Carpentier G. (2008). *Approche computationnelle de l'orchestration musicale - Optimisation multicritère sous contraintes de combinaisons instrumentales dans de grandes banques de sons*. Thèse de doctorat, UPMC Paris-6.

Carpentier G. (2011). Global Constraints in Orchestration. In Truchet & Assayag (Eds.), *Constraint Programming in Music*, Wiley.

Carpentier G. et Bresson J. (2010). Interacting with Symbolic, Sound and Feature Spaces in Orchidée, a Computer-Aided Orchestration Environment. *Computer Music Journal*, 34(1) : 10-27, MIT Press.

Carpentier G., Assayag G., Saint-James E. (2010b). Solving the Musical Orchestration Problem using Multiobjective Constrained Optimization with a Genetic Local Search Approach. *Journal of Heuristics*, 16(5) : 681-714, Springer.

Carpentier G., Tardieu D., Harvey J., Assayag G., Saint-James E. (2010a). Predicting Timbre Features of Instrument Sound Combinations: Application to Automatic Orchestration. *Journal of New Music Research,* 39(1) : 47-61, Taylor & Francis.

Casey, M. et M. Grierson (2007). Soundspotter and remix-tv: Fast approximate matching for audio and video performance. *Proc. International Computer Music Conference (ICMC'07)*, Copenhagen, Denmark.

Chion M. (1994). *Guide des objets sonores*, Editions Buchet Chatel, Paris.

Collins, N. (2007). Audiovisual Concatenative Synthesis. *Proc. International Computer Music Conference* (ICMC'07), pp. 389–392.

Comajuncosas, J. M., A. Barrachina, et J. O. Connell (2011). Nuvolet : 3D Gesture-driven Collaborative Audio Mosaicing. *Proc. of New Interfaces for Musical Expression (NIME'11)*, pp. 252–255.

Deb K. (2000). An Efficient Constraint Hand ling Method for Genetic Algorithms. *Computer Methods in Applied Mechanics and Engineering*, 186 : 311-338.

Doval B., Rodet X. (1993). Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. *Proc. of International conference on audio, speech and signal processing (ICASSP'93)*, Minneapolis.

Einbond, A., D. Schwarz, et J. Bresson (2009). Corpus-based transcription as an approach to the compositional control of timbre. *Proc. of the International Computer Music Conference (ICMC'09)*, Montreal, QC, Canada.

Esling P., Carpentier G., Agon C. (2010) Dynamic Musical Orchestration using Genetic Algorithms and a Spectro-Temporal Description of Musical Instruments, *Lecture Notes in Computer Science,* vol. 6025, EvoApplications Part II, Springer.

Flanagan J. L., R.M. Golden (1966). *Phase vocoder*, Bell Syst. Tech. J., vol. 45, pp. 1493-1509.

Gaver W. (1993). "How do we hear in the world ? explorations in ecological acoustics". *Ecological Psychology*, 5(4) :285–313.

Gaver W. (1993). What in the world do we hear ? an ecological approach to auditory event perception. *Ecological Psychology*, 5 :1–29.

Godoy R.I. (2006). "Gestural-Sonorous Objects : embodied extensions of Schaeffer's conceptual apparatus". *Organised Sound*, 11 (2), 149-157.

Goldberg D.E. (1989). *Genetic Algorithms in Search, optimization and Machine Learning*. Addison-Wesley.

Goodwin M. , Rodet X. (1994). Efficient Fourier synthesis of nonstationary sinusoids, *Proc. International Computer Music Conference (ICMC'94)*, pp. 333-334.

Grey J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America,* vol. 61, no. 5, pp. 1270–1277.

Griffin, D., Lim J. (1984). Signal Estimation from Modified Short-Time Fourier Transform, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236-243.

Houix O., Lemaitre G., Misdariis N., Susini P. (2007). *Everyday sound classification. Part 2 : Experimental classification of everyday sounds*. Deliverable 4.1 - part I, projet CLOSED.

Jaillet F., Torresani B (2007). Time–frequency jigsaw puzzle: adaptive multiwindow and multilayered Gabor expansions., *Int. J. Wavelets Multiresolut. Inf. Process.*, vol. 2, pp. 293–316.

Janer, J. et M. de Boer (2008). Extending voice-driven synthesis to audio mosaicing. *Proc. of the International Conference on Sound and Music Computing (SMC'08)*.

Jaszkiewicz, A. (2001). Genetic Local Search for Multiple Objective Combinatorial Optimization. *Foundations of Computing and Design Sciences*, vol. 26(1) : 99-120.

Jensenius A. R. (2007) *Action - Sound : Developing Methods and Tools to Study Music-Related Body Movement*. Ph.D. thesis, Department of Musicology, University of Oslo.

König, S. (2006). scrambled? hackz! web page. http://popmodernism.org/scrambledhackz/.

Krimphoff J., McAdams S., et al. (1994). Caractérisation du timbre des sons complexes. II: Analyses acoustiques et quantification psychophysique. *Journal de physique*. 4: 625-628.

Krumhansl C. L. (1989). Why is musical timbre so hard to understand ? *Structure and Perception of Electroacoustic Sound and Music*, S. Nielzen and O. Olsson, Eds., Elsevier, Amsterdam, The Netherlands, (Excerpta Medica 846), pp. 43–53,.

Laroche J., Dolson M. (1999). Improved Phase Vocoder Time-Scale Modification of Audio. *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323-332.

Laroche J., Dolson M. (1999a), New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing and other exotic audio modifications, *Journal of the Audio Engineering Society*, vol. 47, no. 11, pp. 928-936.

Legendre P., Legendre L. (1998) Numerical Ecology, *Development in environmental modelling*, Elsevier, second English edition.

Lemaitre G., Houix O., Misdariis N., Susini P. (2010). Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental psychology*, Vol. 16, No. 1, 16–32

Lemaitre G., Susini P., Winsberg S., McAdams S., Letinturier B. (2007). The sound quality of car horns: a psychoacoustical study of timbre. *Acta Acustica United with Acustica*, vol. 93, no. 3, pp. 457–468.

Lindemann, E. (2007). Music synthesis with reconstructive phrase modeling. *IEEE Signal Processing Magazine*, 24(2), 80–91.

Liuni M., Roebel A., Romito M., Rodet X. (2010). A reduced multiple Gabor frame for local time adaptation of the spectrogram, *Proc. International Conference on Digital Audio Effects (DAFx'10)*, pp. 338 – 343.

Maestre, E., R. Ramirez, S. Kersten, et X. Serra (2009). Expressive Concatenative Synthesis by Reusing Samples from Real Performance Recordings. *Computer Music Journal*, 33(4), 23–42.

Maresz Y. (2006). Pour un traité d'orchestration contemporain. *L'Etincelle, journal de la création à l'IRCAM*, IRCAM, Paris.

McAdams S., Susini P., Misdariis N., Winsberg S. (1998). Multidimensional characterisation of perceptual and preference judgements of vehicle and environmental noises. *Proc. Euronoise Conference*, Munich, Germany.

McAdams S., Winsberg S., Donnadieu S., de Soete G., Krimphoff J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychological Research*, vol. 58, no. 3, pp. 177–192.

Miranda E. (2001). *Composing Music with Computers*. Focal Press.

Misdariis N., Minard A., Susini P., Lemaitre G., McAdams S., Parizet E., (2010) Environmental sound perception : meta-description and modeling based on independent primary studies. *EURASIP Journal on Audio, Speech, and Music Processing*, Volume 2010. Article ID 362013.

Misdariis N., Smith B., et al. (1998). Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres. *Proc. 135th Meet. Ac. Soc. of America / 16th Int. Cong. on Acoustics,* Seattle, Washington, USA.

Moore B., Glasberg B. R., et al. (1997). A Model for the Prediction of Thresholds Loudness and Partial Loudness. *Journal of the Audio Engineering Society*, 45: 224-240.

Moorer J. A., (1978). The use of the phase vocoder in computer music applications, *Journal of the Audio Engineering Society*, vol. 26, no. 1/2, pp. 42-45.

Morrison, A. et M. Chalmers (2003). Improving hybrid MDS with pivot-based searching. *IEEE Symposium on Information Visualization*, Los Alamitos, CA, USA, pp. 11.

MPEG-7 (2002). Information Technology - Multimedia Content Description Interface - Part 4: Audio. ISO/IEC JTC 1/SC 29. ISO/IEC FDIS 15938-4:2002.

Mullon, P., Geslin, Y., & Jacob, M. (2002). Ecrins: an audio-content description environment for sound samples. *Proc. International Computer Music Conference (ICMC'02)*, Göteborg.

Nouno G., Cont A., Carpentier G. Harvey J. (2009). Making an Orchestra Speak. *Proc. Sound and Music Computing Conference (SMC'09)*, Porto, Portugal, 277-282.

Oswald, J. (1993). Plexure. CD. http://plunderphonics.com/xhtml/xdiscography.html#plexure.

Parizet E., Guyader E., Nosulenko V. (2006). Analysis of car door closing sound quality, *Applied Acoustics*, vol. 69, no. 1, pp. 12– 22.

Peeters G., McAdams S., et al. (2000). Instrument sound description in the context of MPEG-7. *Proc. International Computer Music Conference (ICMC'00)*. Berlin, Germany.

Peeters G. (2003). Automatic Classification of Large Musical Instrument Databases Using Hierachical Classifiers with Inertia Ratio Maximization, *Proc. of the Audio Engineering Society*, AES 115th Convention. New-York.

Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the Cuidado project*. Technical Report version 1.0, IRCAM – Centre Pompidou, Paris, France.

Peeters G., Deruty E. (2008). Automatic morphological description of sounds. *Proc. Acoustics'08*, Paris.

Rioux V. (2001). *Projet ECRINS / validation expérimentale phase I: descripteurs morphologiques*. Rapport interne IRCAM.

Roads, C. (1988). Introduction to granular synthesis. *Computer Music Journal.* 12(2), 11–13.

Rodet X., Depalle P., (1992). A new additive synthesis method using inverse Fourier transform and spectral envelopes, *Proc. International Computer Music Conference (ICMC'92)*, pp. 410-412.

Roebel A., (2003). A new approach to transient processing in the phase vocoder », *Proc. International Conference on Digital Audio Effects (DAFx'03)*, pp. 344-349.

Roebel A., (2010). A Shape-Invariant Phase Vocoder for Speech Transformation, *Proc. International Conference on Digital Audio Effects (DAFx'10)*.

Roebel A., Zivanovic M., Rodet X., (2004). Signal decomposition by means of classification of spectral peaks, *Proc. International Computer Music Conference (ICMC'04)*, pp. 446-449.

Schaeffer, P. (1966). *Traité des objets musicaux*. Editions du Seuil. Paris, France.

Scheirer E. and Slaney M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. *Proc. International Conference on Audio, Speech and Signal Processing (ICASSP'97)*, Munich, Germany.

Schnell, N. et D. Schwarz (2005). Gabor, Multi-Representation Real-Time Analysis/Synthesis. *Proc. International Conference on Digital Audio Effects (DAFx'05)*, Madrid, Spain.

Schnell, N., A. Röbel, D. Schwarz, G. Peeters, et R. Borghesi (2009). MuBu & friends – assembling tools for content based real-time interactive audio processing in Max/MSP. *Proc. of the International Computer Music Conference (ICMC'09)*, Montréal, Canada.

Schnell, N., R. Borghesi, D. Schwarz, F. Bevilacqua, et R. Muller (2005). FTM-Complex Data Structures for Max. *Proc. International Computer Music Conference (ICMC'05)*, Barcelona, Spain.

Schwarz, D. (2000). A System for Data-Driven Concatenative Sound Synthesis. *Proc. International Conference on Digital Audio Effects (DAFx'00)*, Verona, Italy, pp. 97–102.

Schwarz, D. (2004). *Data-Driven Concatenative Sound Synthesis*. Thèse de doctorat, Université Paris 6 – Pierre et Marie Curie, Paris.

Schwarz, D. (2006). Concatenative sound synthesis: The early years. *Journal of New Music Research*, 35(1), 3–22. Special Issue on Audio Mosaicing.

Schwarz, D. (2007). Corpus-based concatenative synthesis. *IEEE Signal Processing Magazine*, 24(2), 92–104. Special Section: Signal Processing for Sound Synthesis.

Schwarz, D. (2011). State of the art in sound texture synthesis. *Proc. International Conference on Digital Audio Effects (DAFx'11)*, Paris, France.

Schwarz, D. et N. Schnell (2009). Sound search by content-based navigation in large databases. *Proc. of the International Conference on Sound and Music Computing (SMC'09)*, Porto, Portugal.

Schwarz, D. et N. Schnell (2010). Descriptor-based sound texture sampling. *Proc. of the International Conference on Sound and Music Computing (SMC'10)*, Barcelona, Spain, pp. 510–515.

Schwarz, D., N. Schnell, et S. Gulluni (2009). Scalability in content-based navigation of sound databases. *Proc. International Computer Music Conference (ICMC'09)*, Montréal, QC, Canada.

Schwarz, D., R. Cahen, et S. Britton (2008). Principles and applications of interactive corpus-based concatenative synthesis. *Proc. Journées d'Informatique Musicale (JIM'08)*, GMEA, Albi, France.

Schwarz, D., S. Cadars, et N. Schnell (2008). What next? continuation in real-time corpus-based concatenative synthesis. *Proc. International Computer Music Conference (ICMC'08)*, Belfast, Northern Ireland.

Serra X. and Bonada J. (1998). Sound transformations based on SMS High Level Attributes. *Proc. International Conference on Digital Audio Effects (DAFx'98)*, Barcelona (Spain).

Shneiderman, B. et C. Plaisant (2005). *Designing the User Interface*, Chapter Information visualization, Boston, USA: Pearson, pp. 580-603..

Smith B. (1995). PsiExp: an environment for psychoacoustic experimentation using the IRCAM musical workstation. *Proc. Society for Music Perception and Cognition conference*, University of Berkeley.

Stowell, D. et M. Plumbley (2010). Timbre remapping through a regression-tree technique. In *Proc. of the International Conference on Sound and Music Computing (SMC'10)*.

Susini P., McAdams S., Misdariis N., Lemaitre G., Winsberg S. (2005). Timbre des sons environnementaux. *CIM*, Montréal.

Susini P., McAdams S., Winsberg S., Perry I., Vieillard S., Rodet X. (2004). Characterizing the sound quality of airconditioning noise. *Applied Acoustics*, vol. 65, no. 8, pp. 763–790.

Talbi E-G. (2009). *Metaheuristics: From Design to Implementation*. Wiley.

Tardieu D., McAdams S. (2011). Perception of dyads of percussive and sustained instruments. En attente de parution, *Music Perception*.

Tardieu D., Peeters G., Rodet X. (2011). Instrument Model for Computer Aided Orchestration. Soumis à *IEEE Transactions on Audio Speech and Language Processing*.

Tremblay, P. A. et D. Schwarz (2010). Surfing the waves : Live audio mosaicing of an electric bass performance as a corpus browsing interface. *Proc. Internationa Conference for New Interfaces for Musical Expression (NIME'10)*, Sydney, Australia, pp. 447–450.

Truchet C., Codognet P. (2004). Musical Constraint Satisfaction Problems Solved with Adaptive Search. *Soft Computing*, 8-9 : 633-640.

Van Nort D. (2009). Instrumental Listening : sonic gesture as design principle. *Organised Sound*, 14 (2), 177-187.

Vinet, H., Ballet, G., Puig, V.,(1999) *Rapport final du projet Studio en ligne*, rapport interne IRCAM, http://articles.ircam.fr/textes/Vinet98b/

Vinet, H., Herrera, P. et Pachet, F., (2002). The CUIDADO Project. *Proc. International Conference on Music Information Retrieval (ISMIR'02)*, IRCAM, Paris, 2002.

Vinet, H. (2003), The Representation Levels of Musical Information, *Lecture Notes in Computer Science*, 2771, Springer Verlag

Vinet, H. (2006), Applications musicales du traitement de signal : synthèse et prospective, *Traitement du signal*, vol. 5-6, N°3, GRETSI.

Wells J.J., Murphy D. T., (2010). A comparative evaluation of techniques for single-frame discrimination of nonstationary sinusoids, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 498-508.

Winsberg S. and De Soete G. (1993). A latent class approach to fitting the weighted Euclidean model, clascal. *Psychometrika*, vol. 58, no. 2, pp. 315–330.

Wöhrmann, R., & Ballet, G. (1999). Design and Architecture of Distributed Sound Processing and Database Systems for Web-Based Computer Music Applications. *Computer Music Journal*, 23(3), MIT Press, 73-84.

Wold E., Blum T., et al. (1999). Classification, search and retrieval of audio. *CRC Handbook of Multimedia Computing*. B. Furth. Boca Raton, FLA, CRC Press: 207-226.

Wolfe P., Godsill S. J., Dörfler M., (2001) Multi-gabor dictionaries for audio time-frequency analysis. *Proc. IEEE WASPAA*, pp. 43–46.

Wright M. (2005). Open Sound Control: An Enabling Technology for Musical Networking. *Organised Sound*, 10(3) : 193-200.

Yeh C., Roebel A., Rodet X., (2010). Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1116-1126.

Zivanovic M., Roebel A., Rodet X., (2008). Adaptive Threshold Determination for Spectral Peak Classification, *Computer Music Journal*, vol. 32, no. 2, pp. 57-67.

Zwicker E. (1990). *Psychoacoustics*. Springer-Verlag, Berlin.

Zwicker, E. and Terhardt E. (1980). Analytical expression for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, 68, 1523-1525.